Mathematical Principles of Theoretical Physics

（理论物理的数学原理）

Tian Ma Shouhong Wang（马 天 汪守宏）

SCIENCE PRESS
Beijing

To our families

Mathematical Principles of Theoretical Physics

(理论物理的数学原理)

Tian Ma　Shouhong Wang　(马天　汪守宏)

Science Press
Beijing

Responsible Editors: Jingke Li and Yanchao Zhao

Published by Science Press
16 Donghuangchenggen North Street
Beijing 100717, P. R. China

Printed in Beijing

ISBN 978-7-03-045289-4

Chapter 1

General Introduction

1.1 Challenges of Physics and Guiding Principle

Challenges of theoretical physics

Physics is an important part of science of Nature, and is one of the oldest science disciplines. It intersects with many other disciplines of science such as mathematics and chemistry.

Great progresses have been made in physics since the second half of the 19th century. The Maxwell equation, the Einstein special and general relativity and quantum mechanics have become cornerstones of modern physics. Nowadays physics faces new challenges. A partial list of most important and challenging ones is given as follows.

1. What is dark matter?
2. What is dark energy?

Dark matter and dark energy are two great mysteries in physics. Their gravitational effects are observed and are not accounted for in the Einstein gravitational field equations, and in the Newtonian gravitational laws.

3. Is there a Big-Bang? What is the origin of our Universe? Is our Universe static? What is the geometric shape of our Universe?

These are certainly most fundamental questions about our Universe. The current dominant thinking is that the Universe was originated from the Big-Bang. However, there are many unsolved mysteries associated with the Big-Bang theory, such as the horizon problem, the cosmic microwave radiation problem, and the flatness problem.

4. What is the main characteristic of a black hole?

Black holes are fascinating objects in our Universe. However, there are a lot of confusions about black holes, even its very definition.

5. Quark Confinement: Why has there never been observed free quarks?

There are 12 fundamental subatomic particles, including six leptons and six quarks. This is a mystery for not being able to observe free quarks and gluons.

6. Baryon asymmetry: Where are there more particles than anti-particles?

Each particle has its own antiparticle. It is clear that there are far more particles in this Universe than anti-particles. What is the reason? This is another mystery, which is also related to the formation and origin of our Universe.

7. Are there weak and strong interaction/force formulas?

We know that the Newton and the Coulomb formulas are basic force formulas for gravitational force and for electromagnetic force. One longstanding problem is to derive similar force formulas for the weak and the strong interactions, which are responsible for holding subatomic particles together and for various decays.

8. What is the strong interaction potential of nucleus? Can we derive the Yukawa potential from first principles?

9. Why do leptons not participate in the strong interaction?

10. What is the mechanism of subatomic decays and scattering?

11. Can the four fundamental interactions be unified, as Einstein hoped?

Objectives and guiding principles

The objectives of this book are

1) to derive experimentally verifiable laws of Nature based on a few fundamental mathematical principles, and

2) to provide new insights and solutions to some outstanding challenging problems of theoretical physics, including those mentioned above.

The main focus of this book is on the symbiotic interplay between theoretical physics and advanced mathematics. Throughout the entire history of science, the searching for mathematical representations of the laws of Nature is built upon the believe that the Nature speaks the language of Mathematics. The Newton's universal law of gravitation and laws of mechanics are clearly among the most important discoveries of the mankind based on the interplay between mathematics and natural sciences. This viewpoint is vividly revealed in Newton's introduction to the third and final volumes of his great Principia Mathematica: "*I now demonstrate the frame of the system of the world.*"

It was, however, to the credit of Albert Einstein who envisioned that the laws of Nature are dictated by a few fundamental mathematical principles. Inspired by the Albert Einstein's vision, our general view of Nature is synthesized in two guiding principles, Principles 2.1 & 2.2, which can be recapitulated as follows:

> *Nature speaks the language of Mathematics: the laws of Nature 1) are represented by mathematical equations, 2) are dictated by a few fundamental principles, and 3) always take the simplest and aesthetic forms.*

1.2 Law of Gravity, Dark Matter and Dark Energy

Gravity is one of the four fundamental interactions/forces of Nature, and is certainly the first interaction/force that people studied over centuries, dating back to Aristotle (4th century BC), to Galileo (late 16th century and early 17th century), to Johannes Kepler (mid 17th century), to Isaac Newton (late 17th century), and to Albert Einstein (1915).

Newtonian gravity

Newton's universal law of gravity states that the gravitational force between two massive objects with m and M is given by

$$F = -\frac{GmM}{r^2}, \tag{1.2.1}$$

which is an empirical law.

Einstein's General theory of relativity

One of the greatest discovery in the history of science is Albert Einstein's general theory of relativity (Einstein, 1915, 1916). He derives the law of gravity, his gravitational field equations by postulating two revolutionary fundamental principles: the principle of equivalence (PE) and the principle of general relativity (PGR):

1) *PE says that the space-time is a 4-dimensional Riemannian manifold* $\{\mathscr{M}, g_{\mu\nu}\}$ *with the Riemannian metric* $\{g_{\mu\nu}\}$ *representing the gravitational potential*;
2) *PGR says that the law of gravity is covariant under general coordinate transformations of both the inertial and non-inertial reference frames*;
3) *PGR, together with simplicity principle of law of Nature, uniquely dictates the Lagrangian action, also called the Einstein-Hilbert functional*:

$$L_{EH}(\{g_{\mu\nu}\}) = \int_M \left(R + \frac{8\pi G}{c^4} S\right) \sqrt{-g} dx; \tag{1.2.2}$$

4) *The Einstein gravitational field equations are then derived using the least action principle, also called the principle of Lagrangian dynamics* (*PLD*):

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu}. \tag{1.2.3}$$

This is the most profound theory of science. The PGR is a symmetry principle, and the law of gravity, represented as a set of differential equations (1.2.3), is dictated by this profound and simple looking symmetry principle. The connection to the Newtonian gravitational law (1.2.1) is achieved through the following Schwarzschild solution of the Einstein field equations in the exterior of a ball of spherically symmetric matter field with mass M:

$$ds^2 = -\left(1 - \frac{2MG}{c^2 r}\right) c^2 dt^2 + \left(1 - \frac{2MG}{c^2 r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2. \tag{1.2.4}$$

Here the temporal component of the metric and the gravitational force of the ball exerted on an object of mass mass m are

$$g_{00} = -\left(1 - \frac{2MG}{c^2 r}\right) = -\left(1 + \frac{2}{c^2}\psi\right), \qquad F = -m\nabla\psi = -\frac{GMm}{r^2}. \tag{1.2.5}$$

New law of gravity (Ma and Wang, 2014e)

Gravity is the dominant interaction governing the motion and structure of the large scale astronomical objects and the Universe. The Einstein law of gravity has been a tremendous success when it received many experimental and observational supports, mainly in a scale of the solar system.

Dark matter and dark energy are two great mysteries in the scale of galaxies and beyond (Riess et al., 1989; Perlmutter et al., 1999; Zwicky, 1937; Rubin and Ford, 1970). The gravitational effects are observed and are not accounted for in the Einstein gravitational field equations. Consequently, seeking for solutions of these two great mysteries requires a more fundamental level of examination for the law of gravity, and has been the main inspiration for numerous attempts to alter the Einstein gravitational field equations. Most of these attempts, if not all, focus on altering the Einstein-Hilbert action with fine tunings, and therefore are phenomenological. These attempts can be summarized into two groups: (a) $f(R)$ theories, and (b) scalar field theories, which are all based on artificially modifying the Einstein-Hilbert action.

Our key observation is that due to the presence of dark matter and dark energy, the energy-momentum tensor $T_{\mu\nu}$ of visible baryonic matter in the Universe may not conserved: $\nabla^\mu T_{\mu\nu} \neq 0$. which is a contradiction to the Einstein field equations (1.2.3), since the left-hand side of the (1.2.3) is conserved: $\nabla^\mu \left[R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R\right] = 0$. The direct consequence of this observation is to take the variation of the Einstein-Hilbert action under energy momentum conservation constraint:

$$(\delta L_{EH}(g_{\mu\nu}), X) = \frac{d}{d\lambda}\Big|_{\lambda=0} L_{EH}(g_{\mu\nu} + \lambda X_{\mu\nu}) = 0 \quad \forall X = \{X_{\mu\nu}\} \text{ with } \nabla^\mu X_{\mu\nu} = 0. \tag{1.2.6}$$

The div-free condition, $\nabla^\mu X_{\mu\nu} = 0$, imposed on the variation element represents energy-momentum conservation. We call this variational principle, the principle of interaction dynamics (PID), which will be discussed hereafter again.

Using PID (1.2.6), [1] we then derive the new gravitational field equations (Ma and Wang, 2014e):

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu} - \nabla_\mu \nabla_\nu \phi, \tag{1.2.7}$$

$$\nabla^\mu \left[\frac{8\pi G}{c^4} T_{\mu\nu} + \nabla_\mu \nabla_\nu \phi \right] = 0. \tag{1.2.8}$$

In summary, we have derived new law of gravity based solely on first principles, the PE, the PGR, and the constraint variational principle (1.2.6):

Law of gravity (Ma and Wang, 2014e)

1) *(Einstein's PE). The space-time is a 4D Riemannian manifold $\{\mathscr{M}, g_{\mu\nu}\}$, with the metric $\{g_{\mu\nu}\}$ being the gravitational potential*;
2) *The Einstein PGR dictates the Einstein-Hilbert action (1.2.2)*;
3) *The gravitational field equations (1.2.7) and (1.2.8) are derived using PID, and determine gravitational potential $\{g_{\mu\nu}\}$ and its dual vector field $\Phi_\mu = \nabla_\mu \phi$*;
4) *Gravity can display both attractive and repulsive effect, caused by the duality between the attracting gravitational field $\{g_{\mu\nu}\}$ and the repulsive dual vector field $\{\Phi_\mu\}$, together with their nonlinear interactions governed by the field equations (1.2.7) and (1.2.8).*

We remark that it is the duality and both attractive and repulsive behavior of gravity that maintain the stability of the large scale structure of our Universe.

Modified Newtonian formula from first principles (Ma and Wang, 2014e)

Consider a central matter field with total mass M and with spherical symmetry. We derive an approximate gravitational force formula:

$$F = mMG\left[-\frac{1}{r^2} - \frac{k_0}{r} + k_1 r\right], \qquad k_0 = 4\times 10^{-18}\text{km}^{-1}, \qquad k_1 = 10^{-57}\text{km}^{-3}. \tag{1.2.9}$$

[1] The new field equations (1.2.7) can also be equivalently derived using the orthogonal decomposition theorem, reminiscent to the Helmholtz decomposition, so that

$$T_{\mu\nu} = \widetilde{T}_{\mu\nu} - \frac{c^4}{8\pi G} \nabla_\mu \Phi_\nu.$$

See Chapter 4 and (Ma and Wang, 2014e) for details.

Here the first term represents the Newton gravitation, the attracting second term stands for dark matter and the repelling third term is the dark energy. We note that the dark matter property of the gravity and the approximate gravitational interaction formula are consistent with the MOND theory proposed by (Milgrom, 1983); see also (Milgrom, 2014) and the references therein. In particular, our modified new formula is derived from first principles.

Dark matter and dark energy: a property of gravity (Ma and Wang, 2014e)

We have shown that it is the duality between the *attracting* gravitational field $\{g_{\mu\nu}\}$ and the *repulsive* dual field $\{\Phi_\mu = \nabla_\mu \phi\}$ in (1.2.7), and their nonlinear interaction that gives rise to gravity, and in particular the gravitational effect of dark energy and dark matter.

Also, we show in (Hernandez, Ma and Wang, 2015) that consider the gravitational field outside of a ball of centrally symmetric matter field. There exist precisely two physical parameters dictating the two-dimensional stable manifold of asymptotically flat space-time geometry, such that, as the distance to the center of the ball of the matter field increases, gravity behaves as Newtonian gravity, then additional attraction due to the curvature of space (dark matter effect), and repulsive (dark energy effect). This also clearly demonstrates that both dark matter and dark energy are just a property of gravity.

Of course, one can consider dark matter and dark energy as the energy carried by the gravitons and the dual gravitons, to addressed in the unified field theory later in this book.

1.3 First Principles of Four Fundamental Interactions

The four fundamental interactions of Nature are the gravitational interaction, the electromagnetism, the weak and the strong interactions. Seeking laws of the four fundamental interactions is the most important human endeavor. In this section, we demonstrate that laws for the four fundamental interactions are determined by the following principles:

1) *the principle of general relativity, the principle of gauge invariance, and the principle of Lorentz invariance, together with the simplicity principle of laws of Nature, dictates the Lagrangian actions of the four interactions, and*
2) *the principle of interaction dynamics and the principle of representation invariance determines the field equations.*

Symmetry principles

We have shown that for gravity, the basic symmetry principle is the Einstein PGR, which dictates the Einstein-Hilbert action, and induces the gravitational field equations (1.2.7) using PID.

Quantum mechanics provides a mathematical description about the Nature in the molecule, the atomic and subatomic levels. Modern theory and experimental evidence have suggested that the electromagnetic, the weak and the strong interactions obey the gauge symmetry. In fact, these symmetries and the Lorentz symmetry, together with the simplicity of laws of Nature dictate the Lagrangian actions for the electromagnetic, the weak and the strong interactions:

Symmetry Dictates Actions of fundamental interactions:

(a) *The principle of general relativity dictates the action for gravity, the Einstein-Hilbert action.*

(b) *The principle of Lorentz invariance and the principle of gauge invariance, together with the simplicity principle of laws of Nature, dictate the Lagrangian actions for the electromagnetic, the weak and the strong interactions.*

This represents clearly the intrinsic beauty of Nature.

The abelian $U(1)$ gauge theory describes quantum electrodynamics (QED). The non-abelian $SU(N)$ gauge theory was originated from the early work of (Weyl, 1919; Klein, 1938; Yang and Mills, 1954). Physically, gauge invariance refers to the conservation of certain quantum property of the underlying interaction. Such quantum property of the N particles with wave functions cannot be distinguished for the interaction, and consequently, the energy contribution of these N particles associated with the interaction is invariant under the general $SU(N)$ phase (gauge) transformations:

$$(\widetilde{\Psi},\ \widetilde{G}^a_\mu\tau_a) = \left(\Omega\Psi,\ G^a_\mu\Omega\tau_a\Omega^{-1} + \frac{i}{g}(\partial_\mu\Omega)\Omega^{-1}\right), \quad \forall\Omega = e^{i\theta^k(x)\tau_k} \in SU(N). \quad (1.3.1)$$

Here the N wave functions: $\Psi = (\psi_1, \cdots, \psi_N)^{\mathrm{T}}$ represent the N particle fields, the N^2-1 gauge fields G^a_μ represent the interacting potentials between these N particles, and $\{\tau_a \mid a = 1,\cdots,N^2-1\}$ is the set of representation generators of $SU(N)$. The gauge symmetry is then stated as follows:

Principle of Gauge Invariance. *The electromagnetic, the weak, and the strong interactions obey gauge invariance. Namely, the Dirac equations involved in the three interactions are gauge covariant and the actions of the interaction fields are gauge invariant under the gauge transformations (1.3.1).*

The field equations involving the gauge fields G^a_μ are determined by the corresponding Yang-Mills action, is uniquely determined by both the gauge invariance and the Lorentz invariance, together with simplicity of laws of nature:

$$L_{YM}\left(\Psi,\{G^a_\mu\}\right) = \int_{\mathscr{M}} \left[-\frac{1}{4}\mathscr{G}_{ab}g^{\mu\alpha}g^{\nu\beta}G^a_{\mu\nu}G^b_{\alpha\beta} + \overline{\Psi}\left(i\gamma^\mu D_\mu - \frac{mc}{\hbar}\right)\Psi\right] dx, \qquad (1.3.2)$$

which is invariant under both the Lorentz and gauge transformations (1.2.3). Here $\overline{\Psi} = \Psi^\dagger\gamma^0$, $\Psi^\dagger = (\Psi^*)^{\mathrm{T}}$ is the transpose conjugate of Ψ, $\mathscr{G}_{ab} = \frac{1}{2}\mathrm{tr}(\tau_a\tau_b^\dagger)$, λ_{ab}^c are the structure constants of $\{\tau_a \mid a = 1,\cdots,N^2-1\}$, D_μ is the covariant derivative and $G_{\mu\nu}^a$ stands for the curvature tensor associated with D_μ:

$$\begin{aligned} &D_\mu = \partial_\mu + igG_\mu^k\tau_k,\\ &G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g\lambda_{bc}^a G_\mu^b G_\nu^c. \end{aligned} \tag{1.3.3}$$

Principle of Interaction Dynamics (PID) (Ma and Wang, 2014e, 2015a)

With the Lagrangian action at our disposal, the physical law of the underlying system is then represented as the Euler-Lagrangian equations of the action, using the principle of Lagrangian dynamics (PLD). For example, all laws of classical mechanics can be derived using PLD. However, in Section 1.1, we have demonstrated that the law of gravity obeys the principle of interaction dynamics (PID), which takes the variation of the Einstein-Hilbert action under energy-momentum conservation constraint (1.2.6).

We now state the general form of PID, and then illustrate the validity of PID for all fundamental interactions.

PID was discovered in (Ma and Wang, 2014e, 2015a), and requires that for the four fundamental interactions, the variation be taken under the energy-momentum conservation constraints:

PID (Ma and Wang, 2014e, 2015a)

1) *For the four fundamental interactions, the Lagrangian actions are given by*

$$L(g,A,\psi) = \int_{\mathscr{M}} \mathscr{L}(g_{\mu\nu},A,\psi)\sqrt{-g}dx, \tag{1.3.4}$$

where $g = \{g_{\mu\nu}\}$ is the Riemannian metric representing the gravitational potential, A is a set of vector fields representing the gauge potentials, and ψ are the wave functions of particles.

2) *The actions (1.3.4) satisfy the invariance of general relativity, Lorentz invariance, gauge invariance and the gauge representation invariance.*
3) *The states (g,A,ψ) are the extremum points of (1.3.4) with the div_A-free constraint.*

We now illustrate that PID is also valid for both the weak and strong interactions, replacing the classical PLD.

First, PID applied to the Yang-Mills action (1.3.2) takes the following form:

$$\mathscr{G}_{ab}\left[\partial^{\nu}G^{b}_{\nu\mu}-g\lambda^{b}_{cd}g^{\alpha\beta}G^{c}_{\alpha\mu}G^{d}_{\beta}\right]-g\overline{\Psi}\gamma_{\mu}\tau_{a}\Psi=(\partial_{\mu}+\alpha_{b}G^{b}_{\mu})\phi_{a}, \tag{1.3.5}$$

$$(i\gamma^{\mu}D_{\mu}-m)\Psi=0, \tag{1.3.6}$$

where the Dirac equations (1.3.6) are the gauge covariant, and are the dynamic equations for fermions participating the interaction. The right hand side of (1.3.5) is due to PID, with the operator

$$\nabla_A=\partial_{\mu}+\alpha_{b}G^{b}_{\mu}$$

taken in such a way that $\alpha_b G^b_\mu$ is PRI-covariant under the representation transformations (1.3.8) below.

One important consequence of the PID $SU(N)$ theory is the natural introduction of the scalar fields ϕ_a, reminiscent of the Higgs field in the standard model in particle physics.

Second, for the Yang-Mills action for an $SU(N)$ gauge theory, the resulting Yang-Mills equations are the Euler-Lagrange equations of the Yang-Mills action:

$$\mathscr{G}_{ab}\left[\partial^{\nu}G^{b}_{\nu\mu}-g\lambda^{b}_{cd}g^{\alpha\beta}G^{c}_{\alpha\mu}G^{d}_{\beta}\right]-g\overline{\Psi}\gamma_{\mu}\tau_{a}\Psi=0, \tag{1.3.7}$$

supplemented with the Dirac equations (1.3.6).

Historically, the most important difficulty encountered by the gauge theory and the field equations (1.3.7) is that the gauge vectorial field particles $\{G^a_\mu\}$ for the weak interaction are massless, in disagreement with experimental observations.

A great deal of efforts have been made toward to the modification of the gauge theory by introducing proper mass generation mechanism. A historically important breakthrough is the discovery of the mechanism of spontaneous symmetry breaking in subatomic physics. Although the phenomenon was discovered in superconductivity by Ginzburg-Landau in 1951, the mechanism of spontaneous symmetry breaking in particle physics was first proposed by Y. Nambu in 1960; see (Nambu, 1960; Nambu and Jona-Lasinio, 1961a,b). The Higgs mechanism, introduced in (Higgs, 1964; Englert and Brour, 1964; Guralnik, Hagen and Kibble, 1964), is an ad hoc method based on the Nambu-Jona-Lasinio spontaneous symmetry breaking, leading to the mass generation of the vector bosons for the weak interaction.

In all these efforts associated with Higgs fields, the modification of the Yang-Mills action is artificial as in the case for modifying the Einstein-Hilbert action for gravity.

As we indicated before, the Yang-Mills action is uniquely determined by the gauge symmetry, together with the simplicity principle of laws of Nature. All artificial modification of the action will only lead to certain approximations of the underlying physical laws.

Third, the PID $SU(N)$ gauge field equations (1.3.5) and (1.3.6) provide a first principle based mechanism for the mass generation and spontaneous gauge symmetry-breaking: The $\alpha_b G^b_\mu$ on the right-hand side of (1.3.5) breaks the $SU(N)$ gauge symmetry, and the mass generation follows the Nambu-Jona Lasinio idea.

Fourth, one of the most challenging problems for strong interaction is the quark confine ment-no free quarks have been observed. One hopes to solve this mystery with the quantum chromodynamics (QCD) based on the classical $SU(3)$ gauge theory. Unfortunately, as we shall see later that the classical $SU(3)$ Yang-Mills equations produces only repulsive force, and it is the dual fields in the PID gauge field equations (1.3.5) that give rise to the needed attraction for the binding quarks together forming hadrons.

Hence experimental evidence of quark confinement, as well as many other properties derived from the PID strong interaction model, clearly demonstrates the validity of PID for strong interactions.

Fifth, from the mathematical point of view, the Einstein field equations are in general non well-posed, as illustrated by a simple example in Section 4.2.2. In addition, for the classical Yang-Mills equations, the gauge-fixing problem will also pose issues on the well-posedness of the Yang-Mills field equations; see Section 4.3.5. The issue is caused by the fact that there are more equations than the number of unknowns in the system. PID induced model brings in additional unknowns to the equations, and resolves this problem.

Principle of representation invariance (PRI)

Recently we have observed that there is a freedom to choose the set of generators for representing elements in $SU(N)$. In other words, basic logic dictates that the $SU(N)$ gauge theory should be invariant under the following representation transformations of the generator bases:

$$\widetilde{\tau}_a = x_a^b \tau_b, \tag{1.3.8}$$

where $X = (x_a^b)$ are non-degenerate (N^2-1)-th order matrices. Then we can define naturally $SU(N)$ tensors under the transformations (1.3.8). It is clear then that θ^a, G_μ^a, and the structure constants λ_{ab}^c are all $SU(N)$-tensors. In addition, $\mathscr{G}_{ab} = \frac{1}{2}\text{Tr}(\tau_a \tau_b^\dagger)$ is a symmetric positive definite 2nd-order covariant $SU(N)$-tensor, which can be regarded as a Riemannian metric on $SU(N)$. Consequently we have arrived at the following principle of representation invariance, first discovered by the authors (Ma and Wang, 2014h):

PRI (Ma and Wang, 2014h) *For the $SU(N)$ gauge theory, under the representation transformations (1.3.8),*

1) *the Yang-Mills action (1.3.2) of the gauge fields is invariant, and*
2) *the gauge field equations (1.3.5) and (1.3.6) are covariant.*

It is clear that PRI is a basic logic requirement for an $SU(N)$ gauge theory, and has profound physical implications.

First, as indicated in (Ma and Wang, 2013a, 2014h) and in Chapter 4, the field model based on PID appears to be the only model which obeys PRI. In fact, based on PRI, for the gauge interacting fields A_μ and $\{W_\mu^a\}_{a=1}^3$, corresponding to two different gauge groups $U(1)$ for electromagnetism and $SU(2)$ for the weak interaction, the following combination

$$\alpha A_\mu + \beta W_\mu^3 \tag{1.3.9}$$

is prohibited. The reason is that A_μ is an $U(1)$-tensor with tensor, and W_μ^3 is simply the third component of an $SU(2)$-tensor. The above combination violates PRI. This point of view clearly shows that the classical electroweak theory violates PRI, so does the standard model. The difficulty comes from the artificial way of introducing the Higgs field. The PID based approach for introducing Higgs fields by the authors appears to be the only model obeying PRI.

Another important consequence of PRI is that for the term $\alpha_b G_\mu^b$ in the right-hand side of the PID gauge field equations (1.3.5), both $\{\alpha_b \mid b = 1, \cdots, N^2-1\}$ and $\{G_\mu^b \mid b = 1, \cdots, N^2-1\}$ are $SU(N)$ tensors under the representation transformations (1.3.8).

The coefficients α_b represent the portions distributed to the gauge potentials by the charge, represented by the coupling constant g. Then it is clear that

In the field equations (1.3.5) and (1.3.6) of the $SU(N)$ gauge theory for an fundamental interaction,

(a) *the coupling constant g represents the interaction charge, playing the same role as the electric charge e in the $U(1)$ abelian gauge theory for quantum electrodynamics (QED)*;

(b) *the potential*

$$G_\mu \stackrel{\text{def}}{=} \alpha_b G_\mu^b \tag{1.3.10}$$

represents the total interacting potential, where the $SU(N)$ covector α_b represents the portions of each interacting potential G_μ^b contributed to the total interacting potential; *and*

(c) *the temporal component G_0 and the spatial components $\vec{G} = (G_1, G_2, G_3)$ represent, respectively the interaction potential and interaction magnetic potential. The force and the magnetic force generated by the interaction are given by*:

$$F = -g\nabla G_0, \qquad F_{\text{m}} = \frac{g}{c}\vec{v} \times \text{curl}\, \vec{G},$$

where ∇ and curl are the spatial gradient and curl operators.

Geometric interaction mechanism

A simple yet the most challenging problem throughout the history of physics is the mechanism or nature of a force.

One great vision of Albert Einstein is his principle of equivalence, which, in the mathematical terms, says that the space-time is a 4-dimensional (4D) Riemannian manifold $\{\mathscr{M}, g_{\mu\nu}\}$ with the metric $g_{\mu\nu}$ representing the gravitational potential. In other words, gravity is manifested as the curved effect of the space-time manifold $\{\mathscr{M}, g_{\mu\nu}\}$. In essence, gravity is manifested by the gravitational fields $\{g_{\mu\nu}, \nabla_\mu \phi\}$, determined by the gravitational field equations (1.2.7) and (1.2.8) together with the matter distribution $\{T_{\mu\nu}\}$.

The gauge theory provides a field theory for describing the electromagnetic, the weak and the strong interactions. The geometry of the $SU(N)$ gauge theory is determined by

1) the complex bundle [1] $\mathscr{M} \otimes_p (\mathbb{C}^4)^N$ for the wave functions $\Psi = (\psi_1, \cdots, \psi_N)^{\mathrm{T}}$, representing the N particles,

2) the gauge interacting fields $\{G^a_\mu \mid a = 1, \cdots, N^2 - 1\}$, and their dual fields $\{\phi^a \mid a = 1, \cdots, N^2 - 1\}$, and

3) the gauge field equations (1.3.5) coupled with the Dirac equations (1.3.6).

In other words, the geometry of the complex bundle $\mathscr{M} \otimes_p (\mathbb{C}^4)^N$, dictated by the gauge field equations (1.3.5) together with the matter equations (1.3.6), manifests the underlying interaction.

Consequently, it is natural for us to postulate the Geometric Interaction Mechanism 4.1 for all four fundamental interactions:

Geometric Interaction Mechanism(Ma and Wang, 2014d)

1) *(Einstein, 1915) The gravitational force is the curved effect of the time-space*; *and*
2) *the electromagnetic, weak, strong interactions are the twisted effects of the underlying complex vector bundles* $\mathscr{M} \otimes_p \mathbb{C}^n$.

We note that Yukawa's viewpoint, entirely different from Einstein's, is that the other three fundamental forces—the electromagnetism, the weak and the strong interactions–take place through exchanging intermediate bosons such as photons for the electromagnetic interaction, the $\mathrm{W}^\pm$ and Z intermediate vector bosons for the weak interaction, and the gluons for the strong interaction.

It is worth mentioning that the Yukawa Mechanism is oriented toward to computing the transition probability for particle decays and scattering, and the above Geometric Interaction Mechanism is oriented toward to establishing fundamental laws, such as interaction potentials, of the four interactions.

[1] Throughout this book, we use the notation $\otimes_p$ to denote "gluing a vector space to each point of a manifold" to form a vector bundle. For example,

$$\mathscr{M} \otimes_p \mathbb{C}^n = \bigcup_{p \in \mathscr{M}} \{p\} \times \mathbb{C}^n$$

is a vector bundle with base manifold M and fiber complex vector space $\mathbb{C}^n$.

1.4 Symmetry and Symmetry-Breaking

As we have discussed so far, symmetry plays a fundamental role in understanding Nature. In mathematical terms, each symmetry, associated with particular physical laws, consists of three main ingredients: 1) the underlying space, 2) the symmetry group, and 3) tensors, describing the objects which possess the symmetry.

For example, the Lorentz symmetry is made up of 1) the 4D Minkowski space-time $\mathcal{M}^4$ with the Minkowski metric, 2) the Lorentz group LG, and the Lorentz tensors. For example, the electromagnetic potential A_μ is a Lorentz tensor, and the Maxwell equations are Lorentz invariant.

One important point to make is that different physical systems enjoy different symmetry. For example, gravitational interaction enjoys the symmetry of general relativity, which, amazingly, dictates the Lagrangian action for the law of gravity. Also, the other three fundamental interactions obey the gauge and the Lorentz symmetries.

In searching for laws of Nature, one inevitably encounters a system consisting of a number of subsystems, each of which enjoys its own symmetry principle with its own symmetry group. To derive the basic law of the system, one approach is to seek for a large symmetry group, which contains all the symmetry groups of the subsystems. Then one uses the large symmetry group to derive the ultimate law for the system.

However, often times, the basic logic would dictate that the approach of seeking large symmetry group is not allowed. For example, as demonstrated earlier, PRI specifically disallow the mixing the $U(1)$ and $SU(2)$ gauge interacting potentials in the classical electroweak theory.

In fact, this demonstrates an inevitably needed departure from the Einstein vision of unification of the four interactions using large gauge groups.

Our view is that the unification of the four fundamental interactions, as well as the modeling of multi-level physical systems, is achieved through a symmetry-breaking mechanism, together with PID and PRI. Namely, we postulated in (Ma and Wang, 2014a) and in Section 2.1.7 the following Principle of Symmetry-Breaking 2.14:

1) *The three sets of symmetries — the general relativistic invariance, the Lorentz and gauge invariances, and the Galileo invariance — are mutually independent and dictate in part the physical laws in different levels of Nature*; *and*
2) *for a system coupling different levels of physical laws, part of these symmetries must be broken.*

Here we mention three examples.

First, for the unification of the four fundamental interaction, the PRI demonstrates that the unification through seeking large symmetry is not feasible, and the gauge symmetry-breaking is inevitably needed for the unification. The PID-induced gauge symmetry-breaking, by the authors (Ma and Wang, 2015a, 2014h,d), offers a symmetry-breaking mechanism based only on the first principle; see also Chapter4 for details.

Second, for a multi-particle and multi-level system, its action is dictated by a set of $SU(N_1), \cdots, SU(N_m)$ gauge symmetries, and the governing field equations will break some of these gauge symmetries; see Chapter 6 .

Third, in astrophysical fluid dynamics, one difficulty we encounter is that the Newtonian Second Law for fluid motion and the diffusion law for heat conduction are not compatible with the principle of general relativity. Also, there are no basic principles and rules for combining relativistic systems and the Galilean systems together to form a consistent system. The distinction between relativistic and Galilean systems gives rise to an obstacle for establishing a consistent model of astrophysical dynamics. This difficulty can be circumvented by using the above mentioned symmetry-breaking principle, where we have to chose the coordinate system

$$x^\mu = (x^0, x), \quad x^0 = ct \text{ and } x = (x^1, x^2, x^3),$$

such that the metric is in the form:

$$ds^2 = -\left(1 + \frac{2}{c^2}\psi(x,t)\right)c^2 dt^2 + g_{ij}(x,t)dx^i dx^j.$$

Here g_{ij} $(1 \leqslant i, j \leqslant 3)$ are the spatial metric, and ψ represents the gravitational potential. The resulting system breaks the symmetry of general coordinate transformations, and we call such symmetry-breaking as relativistic-symmetry breaking.

1.5 Unified Field Theory Based on PID and PRI

One of the greatest problems in physics is to unify all four fundamental interactions. Albert Einstein was the first person who made serious attempts to this problem.

Most attempts so far have focused on unification through large symmetry, following Einstein's vision. However, as indicated above, one of the most profound implication of PRI is that such a unification with a large symmetry group would violate PRI, which is a basic logic requirement. In fact, the unification should be based on coupling different interactions using the principle of symmetry-breaking (PSB) instead of seeking for a large symmetry group.

The basic principles for the four fundamental interactions addressed in the previous sections have demonstrated that the three first principles, PID, PRI and PSB, offer an entirely different route for the unification, which is one of the main aims of this book:

1) *the general relativity and the gauge symmetries dictate the Lagrangian*;

2) *the coupling of the four interactions is achieved through PID, PRI and PSB in the unified field equations, which obey the PGR and PRI, but break spontaneously the gauge symmetry*; *and*
3) *the unified field model can be easily decoupled to study individual interaction, when the other interactions are negligible.*

Hereafter we address briefly the main ingredients of the unified field theory.

Lagrangian action

Following the simplicity principle of laws of Nature as stated in Principle 2.2, the three basic symmetries—the Einstein general relativity, the Lorentz invariance and the gauge invariance—uniquely determine the interaction fields and their Lagrangian actions for the four interactions:

- The Lagrangian action for gravity is the Einstein-Hilbert functional given by (1.2.2);
- The field describing electromagnetic interaction is the $U(1)$ gauge field $\{A_\mu\}$, representing the electromagnetic potential, and the Lagrangian action density is

$$\mathscr{L}_{EM} = -\frac{1}{4}A_{\mu\nu}A^{\mu\nu} + \overline{\psi}^e(i\gamma^\mu D_\mu - m_e)\psi^e, \tag{1.5.1}$$

  in which the first term stands for the scalar curvature of the vector bundle $\mathscr{M} \otimes_p \mathbb{C}^4$. The covariant derivative and the field strength are given by

$$D_\mu\psi^e = (\partial_\mu + ieA_\mu)\psi^e, \qquad A_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu.$$

- For the weak interaction, the $SU(2)$ gauge fields $\{W_\mu^a \,|a = 1,2,3\}$ are the interacting fields, and the $SU(2)$ Lagrangian action density $\mathscr{L}_W$ for the weak interaction is the standard Yang-Mills action density as given by (1.3.2).
- The $SU(3)$ gauge action density $\mathscr{L}_S$ for the strong interaction is also in the standard Yang-Mills form given by (1.3.2), and the strong interaction fields are the $SU(3)$ gauge fields $\{S_\mu^k \mid 1 \leqslant k \leqslant 8\}$, representing the 8 gluon fields.

It is clear that the action coupling the four fundamental interactions is the natural combination of the Einstein-Hilbert functional, the standard $U(1), SU(2), SU(3)$ gauge actions for the electromagnetic, weak and strong interactions:

$$L\Big(\{g_{\mu\nu}\}, A_\mu, \{W_\mu^a\}, \{S_\mu^k\}\Big) = \int_{\mathscr{M}} [\mathscr{L}_{EH} + \mathscr{L}_{EM} + \mathscr{L}_W + \mathscr{L}_S]\sqrt{-g}dx, \tag{1.5.2}$$

which obeys all the symmetric principles, including principle of general relativity, the Lorentz invariance, the $U(1)\times SU(2)\times SU(3)$ gauge invariance and PRI.

PID unified field equations

With PID, the PRI covariant unified field equations are then given by:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \frac{8\pi G}{c^4}T_{\mu\nu} = \left[\nabla_\mu + \alpha^0 A_\mu + \alpha_b^1 W_\mu^b + \alpha_k^2 S_\mu^k\right]\phi_\nu^G, \tag{1.5.3}$$

$$\partial^\mu(\partial_\mu A_\nu - \partial_\nu A_\mu) - eJ_\nu = \left[\nabla_\nu + \beta^0 A_\nu + \beta_b^1 W_\nu^b + \beta_k^2 S_\nu^k\right]\phi^e, \tag{1.5.4}$$

$$\begin{aligned}\mathscr{G}_{ab}^w &\left[\partial^\mu W_{\mu\nu}^b - g_w \lambda_{cd}^b g^{\alpha\beta} W_{\alpha\nu}^c W_\beta^d\right] - g_w J_{\nu a} \\ &= \left[\nabla_\nu + \gamma^0 A_\nu + \gamma_b^1 W_\nu^b + \gamma_k^2 S_\nu^k - \frac{1}{4}m_w^2 x_\nu\right]\phi_a^w,\end{aligned} \tag{1.5.5}$$

$$\begin{aligned}\mathscr{G}_{kj}^s &\left[\partial^\mu S_{\mu\nu}^j - g_s \Lambda_{cd}^j g^{\alpha\beta} S_{\alpha\nu}^c S_\beta^d\right] - g_s Q_{\nu k} \\ &= \left[\nabla_\nu + \delta^0 A_\nu + \delta_b^1 W_\nu^b + \delta_k^2 S_\nu^k - \frac{1}{4}m_s^2 x_\nu\right]\phi_k^s,\end{aligned} \tag{1.5.6}$$

$$(i\gamma^\mu D_\mu - m)\Psi = 0, \tag{1.5.7}$$

where $\Psi = (\psi^e, \psi^w, \psi^s)^{\mathrm{T}}$ stands for the wave functions for all fermions, participating respectively the electromagnetic, the weak and the strong interactions, and the current densities are defined by

$$J_\nu = \overline{\Psi}^e \gamma_\nu \psi^e, \qquad J_{\nu a} = \overline{\Psi}^w \gamma_\nu \sigma_a \psi^w, \qquad Q_{\nu k} = \overline{\Psi}^s \gamma_\nu \tau_k \psi^s. \tag{1.5.8}$$

Equations (1.5.7) are the gauge covariant Dirac equations for all fermions participating the four fundamental interactions. The left-hand sides of the field equations (1.5.3)-(1.5.6) are the same as the classical Einstein equations and the standard $U(1)\times SU(2)\times SU(3)$ gauge field equations. The right-hand sides of are due to PID, and couple the four fundamental interactions.

Duality

In the field equations (1.5.3)-(1.5.6), there exists a natural duality between the interaction fields $(g_{\mu\nu}, A_\mu, W_\mu^a, S_\mu^k)$ and their corresponding dual fields $(\phi_\mu^G, \phi^e, \phi_a^w, \phi_k^s)$. This duality can be viewed as the duality between mediators and the duality between interacting forces, summarized as follows:

(a) **Duality of mediators.** *Each interaction mediator possesses a dual field particle, called the dual mediator, and if the mediator has spin-k, then its dual mediator has spin-*$(k-1)$.

(b) **Duality of interacting forces.** *Each interaction generates both attracting and repelling forces. Moreover, for each pair of dual fields, the even-spin field generates an attracting force, and the odd-spin field generates a repelling force.*

A few remarks are now in order.

First, from the (phenomenological) weakton theory of elementary particles to be introduced later, that each mediator and its corresponding dual mediator are the same type of particles with the same constituents, but with different spins.

Second, from the duality of the mediators induced from the first principle PID, we conjecture that in addition to the neutral Higgs found by LHC in 2012, there must be two charged massive spin-0 Higgs particles.

Third, the duality of interacting forces demonstrates that all four forces can be both attractive and repulsive. For example, as the distance of two massive objects increases, the gravitational force displays attraction as the Newtonian gravity, more attraction than the Newtonian gravity in the galaxy scale, and repulsion in a much larger scale.

Fourth, as we shall see later, the 8 gluon fields produces repulsive force, and it is their dual fields that produce the needed attraction for quark confinement. This gives a strong observational evidence for the existence of the dual gluons and for the valid and necessity of PID.

Decoupling

The unified field model can be easily decoupled to study each individual interaction when other interactions are negligible. In other words, PID is certainly applicable to each individual interaction. For gravity, for example, PID offers to a new gravitational field model, leading to a unified model for dark energy and dark matter as described earlier in this introduction.

1.6 Theory of Strong Interactions

The strong interaction is responsible for the nuclear force that binds protons and neutrons (nucleons) together to form the nucleus of an atom, and is the force that holds quarks together to form protons, neutrons, and other hadron particles. The current theory of strong interaction is the quantum chromodynamics (QCD), based on an $SU(3)$ non-Abelian gauge theory.

Two most important observed basic properties of the strong interaction are the asymptotic freedom and the quark confinement. The theoretical understanding of these properties are still lacking. There have been many attempts such as the lattice QCD, which is developed in part for the purpose of understanding the quark confinement.

Modern theory of QCD leaves a number of key problems open for a long time, including

1) (Quark confinement) Why is there no observed quark?

2) Why is there asymptotic freedom?

3) What is the strong interaction potential?

4) Can one derive the Yukawa potential from first principles?

5) Why is the strong force short-ranged?

These are longstanding problems, and the new field theory for strong interaction based on PID and PRI completely solves these open problems. We now present some basic ingredients of this new development, and we refer the interested readers to (Ma and Wang, 2014c) and Section 4.5 for details.

PID *SU*(3) gauge field equations

The strong interaction field equations decoupled from the unified field model are

$$\mathscr{G}^s_{kj}\left[\partial^\mu S^j_{\mu\nu} - g_s\Lambda^j_{cd}g^{\alpha\beta}S^c_{\alpha\nu}S^d_\beta\right] - g_sQ_{\nu k} = \left[\partial_\nu + \delta^2_k S^k_\nu - \frac{1}{4}m^2_s x_\nu\right]\phi^s_k, \tag{1.6.1}$$

$$\left[i\gamma^\mu\left(\partial_\mu + ig_sS^k_\mu\tau_k\right) - m_q\right]\psi = 0. \tag{1.6.2}$$

1. *Gluons and dual scalar gluons.* For the strong interaction, the field equations induce a duality between the eight spin-1 massless gluons, described by the $SU(3)$ gauge fields $\{S^k_\mu \mid k = 1,\cdots,8\}$, and the eight spin-0 dual gluons, described by the dual fields $\{\phi^s_k \mid a = 1,\cdots,8\}$. As we shall see from the explianation of quark confinements gluons and their dual gluons are confined in hadrons.

2. *Attracting and repulsive behavior of strong force.* As indicated before and from the strong interaction potentials derived hereafter, strong interaction can display both attracting and repulsive behavior. The repulsive behavior is due mainly to the spin-1 gluon fields, and the attraction is caused by the spin-0 dual gluon fields.

Layered strong interaction potentials

Different from gravity and electromagnetic force, strong interaction is of short-ranged with different strengths in different levels. For example, in the quark level, strong interaction confines quarks inside hadrons, in the nucleon level, strong interaction bounds nucleons inside atoms, and in the atom and molecule level, strong interaction almost diminishes. This layered phenomena can be well-explained using the unified field theory based on PID and PRI.

One key ingredient is total interaction as the consequence of PRI addressed above. For the strong interaction, in view of (1.3.10), the total strong interaction potential is defined by $\delta^2_k S^k_\nu$ on the right hand side of (1.6.1):

$$S_\mu \overset{\text{def}}{=} \delta^2_k S^k_\mu = (S_0, S_1, S_2, S_3), \qquad \Phi \overset{\text{def}}{=} S_0. \tag{1.6.3}$$

The temporal component S_0, denoted by Φ, gives rise to the strong force between two elementary particles carrying strong charges:

$$F = -g_s \nabla \Phi. \tag{1.6.4}$$

From the field equations (1.6.1) and (1.6.2), we derive in Section 4.5.2 that for a particle with N strong charges and radius ρ, its strong interaction potential are given by (4.5.39) and recast here convenience:

$$\begin{aligned} &\Phi = g_s(\rho)\left[\frac{1}{r} - \frac{A}{\rho}(1+kr)e^{-kr}\right], \\ &g_s(\rho) = N\left(\frac{\rho_w}{\rho}\right)^3 g_s, \end{aligned} \tag{1.6.5}$$

where g_s is the strong charge of w^*-weakton, the ρ_w is the radius of the w^*-weakton, A is a dimensionless constant depending on the particle, and $1/k$ is the radius of strong attraction of particles. Phenomenologically, we can take

$$\frac{1}{k} = \begin{cases} 10^{-18}\ \text{cm} & \text{for } w^* - \text{ weaktons}, \\ 10^{-16}\ \text{cm} & \text{for quarks}, \\ 10^{-13}\ \text{cm} & \text{for neutreons}, \\ 10^{-7}\ \text{cm} & \text{for atom/molecule}, \end{cases} \tag{1.6.6}$$

and the resulting layered formulas of strong interaction potentials are called the w^*-weakton potential Φ_0, the quark potential Φ_q, the nucleon/hadron potential Φ_n and the atom/molecule potential Φ_a.

Quark confinement

Quark confinement is a challenging problem in physics. Quark model was confirmed by many experiments. However, no any single quark is found ever. This fact suggests that the quarks were permanently bound inside a hadron, which is called the quark confinement. Up to now, no other theories can successfully describe the quark confinement phenomena. The direct reason is that all current theories for interactions fail to provide a successful strong interaction potential to explain the various level strong interactions.

With the strong interaction potential (1.6.5), we can calculate the strong interaction bound energy E for two same type of particles:

$$E = g_s^2(\rho)\left[\frac{1}{r} - \frac{A}{\rho}(1+kr)e^{-kr}\right]. \tag{1.6.7}$$

The quark confinement can be well explained from the viewpoint of the strong quark bound energy E_q and the nucleon bound energy E_n. In fact, by (1.6.7) we can show that

$$E_q = 10^{20} E_n \sim 10^{18}\text{GeV}. \tag{1.6.8}$$

This clearly shows that the quarks is confined in hadrons, and no free quarks can be found.

Asymptotic freedom and short-range nature of the strong interaction

Asymptotic freedom was discovered and described by (Gross and Wilczek, 1973; Politzer, 1973). Using the strong interaction potential (1.6.5), we can clearly demonstrate the asymptotic freedom property and the short-ranged nature of the strong interaction.

1.7 Theory of Weak Interactions

The new weak interaction theory based on PID and PRI was first discovered by (Ma and Wang, 2013a). As addressed earlier, the weak interaction obeys the $SU(2)$ gauge symmetry, which dictates the standard $SU(2)$ Yang-Mills action. By PID and PRI, the field equations of the weak interaction are given by:

$$\mathscr{G}^w_{ab}\left[\partial^\mu W^b_{\mu\nu} - g_w\lambda^b_{cd}g^{\alpha\beta}W^c_{\alpha\nu}W^d_\beta\right] - g_w\overline{\Psi}^w\gamma^\nu\sigma_a\Psi^w = \left[\partial_\mu + \gamma^1_b W^b_\mu - \frac{1}{4}m_w^2 x_\mu\right]\phi^w_a, \quad (1.7.1)$$

$$(i\gamma^\mu D_\mu - m_l)\Psi^w = 0. \quad (1.7.2)$$

1. *Higgs fields from first principles.* The right-hand side of (1.7.1) is due to PID, leading naturally to the introduction of three scalar dual fields. The left-hand side of (1.7.1) represents the intermediate vector bosons $W^\pm$ and Z, and the dual fields represent two charged Higgs $H^\pm$ (to be discovered) and the neutral Higgs H^0, with the later being discovered by LHC in 2012.

It is worth mentioning that the right-hand side of (1.7.1), involving the Higgs fields, can not be generated by directly adding certain terms in the Lagrangian action, as in the case for the new gravitational field equations derived in (Ma and Wang, 2014e).

2. *Duality.* We establish a natural duality between weak gauge fields $\{W^a_\mu\}$, representing the $W^\pm$ and Z intermediate vector bosons, and three bosonic scalar fields ϕ^w_a, representing both two charged and one neutral Higgs particles $H^\pm$, H^0.

3. *Spontaneous gauge symmetry-breaking.* PID induces naturally spontaneous symmetry breaking mechanism. By construction, it is clear that the Lagrangian action $\mathscr{L}_W$ obeys the $SU(2)$ gauge symmetry, the PRI and the Lorentz invariance. Both the Lorentz invariance and PRI are universal principles, and, consequently, the field equations (1.7.1) and (1.7.2) are covariant under these symmetries. The gauge symmetry is spontaneously breaking in the field equations (1.7.1), due to the presence of the terms in the right-hand side, derived by PID. This term generates the mass for the vector bosons.

4. *Weak charge and weak potential.* By PRI, the $SU(2)$ gauge coupling constant g_w plays the role of weak charge, responsible for the weak interaction.

Also, PRI induces an important $SU(2)$ constant vector $\{\gamma_b^1\}$. The components of this vector represent the portions distributed to the gauge potentials W_μ^a by the weak charge g_w. Hence the (total) weak interaction potential is given by the following PRI representation invariant

$$W_\mu = \gamma_a^1 W_\mu^a = (W_0, W_1, W_2, W_3), \tag{1.7.3}$$

and the weak charge potential and weak force are as

$$\begin{aligned} &\Phi_w = W_0 && \text{the time component of } W_\mu, \\ &F_w = -g_w(\rho)\nabla\Phi_w, \end{aligned} \tag{1.7.4}$$

where $g_w(\rho)$ is the weak charge of a particle with radius ρ.

5. *Layered formulas for the weak interaction potential.* The weak interaction is also layered, and we derive from the field equations the following

$$\begin{aligned} &\Phi_w = g_w(\rho)e^{-kr}\left[\frac{1}{r} - \frac{B}{\rho}(1+2kr)e^{-kr}\right], \\ &g_w(\rho) = N\left(\frac{\rho_w}{\rho}\right)^3 g_w, \end{aligned} \tag{1.7.5}$$

where Φ_w is the weak force potential of a particle with radius ρ and carrying N weak charges g_w, taken as the unit of weak charge g_s for each weakton (Ma and Wang, 2015b), ρ_w is the weakton radius, B is a parameter depending on the particles, and $1/k = 10^{-16}$ cm represents the force-range of weak interactions.

6. The layered weak interaction potential formula (1.7.5) shows clearly that the weak interaction is short-ranged. Also, it is clear that the weak interaction is repulsive, asymptotically free, and attractive when the distance of two particles increases.

1.8 New Theory of Black Holes

The concept of black holes is essentially developed following the Karl Schwarzschild's derivation of the Schwarzschild solution for the Einstein gravitational field equations. In the exterior of spherically symmetric ball of mass, the solution is given by

$$ds^2 = -\left(1 - \frac{R_s}{r}\right)c^2dt^2 + \left(1 - \frac{R_s}{r}\right)^{-1}dr^2 + r^2d\theta^2 + r^2\sin^2\theta d\varphi^2, \tag{1.8.1}$$

where

$$R_s = \frac{2MG}{c^2} \tag{1.8.2}$$

is called the Schwarzschild radius, at which the metric displays a singularity. There have been many confusions about black holes throughout the history of general relativity and black holes.

In this section, we present the black hole theorem proved in Section 7.3, and clarify some of the confusions in the literature. This black hole theorem leads to important insights to many problems in astrophysics and cosmology, which will be addressed in details in Chapter 7.

Blackhole Theorem (Ma and Wang, 2014a) *Assume the validity of the Einstein theory of general relativity, then the following assertions hold true*:

1) *black holes are closed*: *matters can neither enter nor leave their interiors,*
2) *black holes are innate*: *they are neither born to explosion of cosmic objects, nor born to gravitational collapsing, and*
3) *black holes are filled and incompressible, and if the matter field is non-homogeneously distributed in a black hole, then there must be sub-blackholes in the interior of the black hole.*

This theorem leads to drastically different views on the structure and formation of our Universe, as well as the mechanism of supernovae explosion and the active galactic nucleus (AGN) jets. Here we only make a few remarks, and refer interested readers to Section 7.3 for the detailed proof.

1. *An intuitive observation.* One important part of the theorem is that all black holes are closed: matters can neither enter nor leave their interiors. Classical view was that nothing can get out of blackholes, but matters can fall into blackholes. We show that nothing can get inside the blackhole either.

To understand this result better, let's consider the implication of the classical theory that matters can fall inside a blackhole. Take as an example the supermassive black hole at the center of our galaxy, the Milky Way. By the classical theory, this blackhole would continuously gobble matters nearby, such as the cosmic microwave background (CMB). As the Schwarzschild radius of the black hole $r = R_s$ is proportional to the mass, then the radius R_s would increase in cubic rate, as the mass M is proportional to the volume. Then it would be easy to see that the black hole will consume the entire Milky Way, and eventually the entire Universe. However, observational evidence demonstrates otherwise, and supports our result in the blackhole theorem.

2. *Singularity at the Schwarzschild radius is physical.* One important ingredient is that the singularity of the space-time metric at the Schwarzschild radius R_s is essential, and cannot be removed by any differentiable coordinate transformations. Classical transformations such as those by Eddington and Kruskal are **non-differentiable**, and are not valid for removing the singularity at the Schwarzschild radius. In other words, the singularity displayed in both the Schwarzschild metric and the Tolman-Oppenheimer-Volkoff (TOV)

metric at $r = R_s$

$$ds^2 = -e^u c^2 dt^2 + \left(1 - \frac{r^2}{R_s^2}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2, \tag{1.8.3}$$

is a true physical singularity, and defines the black hole boundary.

3. *Geometric realization of a black hole.* As described in Section 4.1 in (Ma and Wang, 2014a) and in Section 7.3.1, the geometrical realization of a black hole, dictated by the Schwarzschild and TOV metrics, clearly manifests that the real world in the black hole is a hemisphere with radius R_s embedded in $\mathbb{R}^4$, and at the singularity $r = R_s$, the tangent space of the black hole is perpendicular to the coordinate space $\mathbb{R}^3$.

This geometric realization clearly demonstrates that the disk in the realization space $\mathbb{R}^3$ is equivalent to the real world in the black hole. If the outside observer observes that nothing gets inside the black hole, then nothing will get inside the black hole in the reality as well.

1.9 The Universe

Geometry and origin of our Universe

Modern cosmology adopts the view that our Universe is formed through the Big-Bang or Big-Bounce; see among others (Harrison, 2000; Kutner, 2003; Popławski, 2012; Roos, 2003). Based on our new insights on black holes, we have reached very different conclusions on the structure and formation of our Universe.

It is clear that the large scale structure of our Universe is essentially dictated by the law of gravity, which is based on Einstein's two principles: the principle of general relativity and the principle of equivalence, as addressed in Section 1.1. Also, strong cosmological observational evidence suggests that the large scale Universe obey the cosmological principle that the Universe is homogeneous and isotropic.

Basic assumptions

(a) *the Einstein theory of general relativity, and*

(b) *the cosmological principle.*

With Assumption (a) above, we have our blackhole theorem, Theorem 7.15, at our disposal. Then we can draw a number of important conclusions on the structure of our Universe.

1. Let E and M be the total energy and mass of the Universe:

$$E = \text{kinetic} + \text{electromagnetic} + \text{thermal} + \Psi, \qquad M = E/c^2, \tag{1.9.1}$$

where Ψ is the energy of all interaction fields. The total mass M dictates the Schwarzschild radius R_s.

If our Universe were born to the Big-Bang, assuming at the initial stage, all energy is concentrated in a ball with radius $R_0 < R_s$, by the theory of black holes, then the energy contained in B_{R_0} must generate a black hole in $\mathbb{R}^3$ with fixed radius R_s as defined by (1.8.2).

If we assume that at certain stage, the Universe were contained in ball of a radius R with $R_0 < R < R_s$, then we can prove that the Universe must contain a sub-black hole with radius r given by

$$r = \sqrt{\frac{R}{R_s}}R.$$

Based on this property, the expansion of the Universe, with increasing R to R_s, will give rise to an infinite sequence of black holes with one embedded to another. Apparently, this scenario is clearly against the observations of our Universe, and demonstrates that our Universe cannot be originated from a Big-Bang.

2. By the cosmological principle, given the energy density $\rho_0 > 0$ of the Universe, based on the Schwarzschild radius, the Universe will always be bounded in black hole, which is an open ball of radius:

$$R_s = \sqrt{\frac{3c^2}{8\pi G\rho_0}}.$$

This immediately shows that there is no unbounded universe. Consequently, since a black hole is unable to expand and shrink, we arrive immediately from the above analysis that our Universe must be static, and not expanding.

Notice that the isotropy requirement in the cosmological principle excludes the globular open universe scenario. Consequently, we have shown that our Universe must be a closed 3D sphere S^3.

In summary, we have proved two theorems, Theorems 7.27-7.28, on the geometry and structure of our Universe, which have been discovered and proved in (Ma and Wang, 2014a, Theorems 6.2 & 6.3):

Theorem on Structure of our Universe *Assume the Einstein theory of general relativity, and the principle of cosmological principle, then the following assertions hold true*:

1) *our Universe is not originated from a Big-Bang, and is static*;
2) *the topological structure of our Universe is the 3D sphere S^3 such that to each observer, the corresponding equator with the observer at the center of the hemisphere can be viewed as the black hole horizon*;
3) *the total mass M_{total} in the Universe includes both the cosmic observable mass M and the non-observable mass, regarded as dark matter, due to the space curvature energy*; *and*

4) *a negative pressure is present in our Universe to balance the gravitational attracting force, and is due to the gravitational repelling force, also called dark energy.*

It is clear that this theorem changes drastically our view on the geometry and the origin of the Universe. Inevitably, a number of important questions need to be addressed for this scenario of our Universe. Hereafter we examine a few most important problems.

Redshift problem

The natural and important question that one has to answer is the consistency with astronomical observations, including the cosmic edge, the flatness, the horizon, the redshift, and the cosmic microwave background (CMB) problems. These problems can now be easily understood based on the structure of the Universe and the blackhole theorem we derived. Hereafter we focus only on the redshift and the CMB problems.

The most fundamental problem is the redshift problem. Observations clearly show that light coming from a remote galaxy is redshifted, and the farther away the galaxy is, the larger the redshift. In modern astronomy and cosmology, it is customary to characterize the redshift by a dimensionless quantity z in the formula

$$1+z=\frac{\lambda_{\text{observ}}}{\lambda_{\text{emit}}}, \tag{1.9.2}$$

where λ_{observ} and λ_{emit} represent the observed and emitting wavelenths.

There are three sources of redshifts: the Doppler effect, the cosmological redshift, and the gravitational redshift. If the Universe is not considered as a black hole, then the gravitational redshift and the cosmological redshift are both too small to be significant. Hence, modern astronomers have to think that the large port of the redshift is due to the Doppler effect.

However, due to black hole properties of our Universe, the black hole and cosmological redshifts cannot be ignored. Due to the horizon of the sphere, for an arbitrary point in the spherical Universe, its opposite hemisphere relative to the point is regarded as a black hole. Hence, g_{00} can be approximatively taken as the Schwarzschild solution for distant objects as follows

$$-g_{00}(r)=\alpha(r)\left(1-\frac{R_s}{\widetilde{r}}\right), \qquad \alpha(0)=2, \qquad \alpha(R_s)=1, \qquad \alpha'(r)<0,$$

where $\widetilde{r}=2R_s-r$ for $0\leqslant r<R_s$ is the distance from the light source to the opposite radial point, and r is the distance from the light source to the point. Then we derive the following redshift formula, which is consistent with the observed redshifts:

$$1+z=\frac{1}{\sqrt{\alpha(r)\left(1-\frac{R_s}{\widetilde{r}}\right)}}=\frac{\sqrt{2R_s-r}}{\sqrt{\alpha(r)(R_s-r)}} \qquad \text{for } 0<r<R_s. \tag{1.9.3}$$

CMB problem

In 1965, two physicists A. Penzias and R. Wilson discovered the low-temperature cosmic microwave background (CMB) radiation, which fills the Universe, and it has been regarded as the smoking gun for the Big-Bang theory. However, based on the unique scenario of our Universe we derived, it is the most natural thing that there exists a CMB, because the Universe has always been there as a black-body, and CMB is a result of blackbody equilibrium radiation.

Dark matter and dark energy

Conclusion 4) in the above results for our Universe shows that the observable cosmic mass M, and the total mass M_{total} which includes both M and the non-observable mass caused by space curvature energy, enjoy the following relation:

$$M_{\text{total}} = 3\pi M/2 \tag{1.9.4}$$

The difference $M_{\text{total}} - M$ can be regarded as the dark matter. Astronomical observations have shown that the measurable mass M is about one fifth of total mass M_{total}.

Also, the static Universe has to possess a negative pressure to balance the gravitational attracting force. The negative pressure is actually the effect of the gravitational repelling force, attributed to dark energy.

Equivalently, the above interpretation of dark matter and dark energy is consistent with the theory based on the new PID gravitational field equations discovered by (Ma and Wang, 2014e) and addressed in detail in Chapter 4. It is clear now that gravity can display both attractive and repulsive effect, caused by the duality between the *attracting* gravitational field $\{g_{\mu\nu}\}$ and the *repulsive* dual vector field $\{\Phi_\mu\}$, together with their nonlinear interactions governed by the field equations. Consequently, dark energy and dark matter phenomena are simply a property of gravity. The detailed account of this relation is addressed in Section 7.6, based in part on (Hernandez, Ma and Wang, 2015).

PID-cosmological model

One of the main motivations for the introduction of the Big-Bang theory and for the expanding universe is that the Friedamann solution of the Einstein gravitational field equations demonstrated that the Einstein theory must produce a variable size universe; see Conclusions of Friedmann Cosmology 7.23. Also, it is classical that bringing the cosmological constant Λ into the Einstein field equations will lead to an unstable static Universe.

We have demonstrated that the right cosmological model should be derived from the new gravitational field equations (1.2.7), taking into consideration the presence of dark matter and dark energy. In this case, based on the cosmological principle, the the metric of a homogeneous spherical universe is of the form

$$ds^2 = -c^2dt^2 + R^2\left[\frac{dr^2}{1-r^2} + r^2(d\theta^2 + \sin^2\theta d\varphi^2)\right], \tag{1.9.5}$$

where $R = R(t)$ is the cosmic radius. We deduce then from (1.2.7) the following PID-cosmological model, with $\varphi = \phi''$:

$$\begin{aligned} R'' &= -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2} + \frac{\varphi}{8\pi G}\right)R, \\ (R')^2 &= \frac{1}{3}(8\pi G\rho + \varphi)R^2 - c^2, \\ \varphi' + \frac{3R'}{R}\varphi &= -\frac{24\pi G}{c^2}\frac{R'}{R}p, \end{aligned} \tag{1.9.6}$$

supplemented with the equation of state:

$$p = f(\rho, \varphi). \tag{1.9.7}$$

Note that only two equations in (1.9.6) are independent.

Also, the model describing the static Universe is the equation of state (1.9.7) together with the stationary equations of (1.9.6), which are equivalent to the form

$$\begin{aligned} \varphi &= -8\pi G\left(\rho + \frac{3p}{c^2}\right), \\ p &= -\frac{c^4}{8\pi G R^2}. \end{aligned} \tag{1.9.8}$$

1.10 Supernovae Explosion and AGN Jets

Supernovae explosion and the active galactic nucleus (AGN) jets are among the most important astronomical phenomena, which are lack of reasonable explanation.

Relativistic, magnetic and thermal effects are main ingredients in astrophysical fluid dynamics, and are responsible for many astronomical phenomena. The thermal effect is described by the Rayleigh number Re:

$$\mathrm{Re} = \frac{mGr_0r_1\beta}{\kappa\nu}\frac{T_0 - T_1}{r_1 - r_0}, \tag{1.10.1}$$

where T_0 and T_1 are the temperatures at the bottom and top of an annular shell region $r_0 < r < r_1$; see e.g. (7.1.74) for other notations used here.

Based on our theory of black holes, including in particular the incompressibility and closedness of black holes, the relativistic effect is described by the δ-factor:

$$\delta = \frac{2mG}{c^2r_0}. \tag{1.10.2}$$

where m is the mass of the core $0 < r < r_0$.

Consider e.g. an active galactic nucleus (AGN), which occupies a spherical annular shell region $R_s < r < R_1$, where R_s is the Schwarzschild radius of the black hole core of the galaxy. Then $r_0 = R_s$ and the δ-factor is $\delta = 1$. The relativistic effect is then reflected in the radial force

$$F_r = \frac{\nu}{2\alpha}\frac{\partial}{\partial r}\left(\frac{1}{\alpha}\frac{d\alpha}{dr}P_r\right), \qquad \alpha = \left(1 - \frac{R_s}{r}\right)^{-1},$$

which gives rise to a huge explosive force near $r = R_s$:

$$\frac{\nu}{1 - R_s/r}\frac{R_s^2}{r^4}P_r. \tag{1.10.3}$$

The relativistic effect is also reflected in the electromagnetic energy:

$$\frac{\nu_0}{2\alpha}\frac{\partial}{\partial r}\left(\frac{1}{\alpha}\frac{d\alpha}{dr}H_r\right),$$

which consists of a huge explosive electromagnetic energy near $r = R_s$:

$$\frac{\nu_0}{1 - R_s/r}\frac{R_s^2}{r^4}H_r. \tag{1.10.4}$$

The basic mechanism of the formation of AGN jets is that the radial temperature gradient causes vertical Bénard convection cells. Each Bénard convection cell has a vertical exit, where the circulating gas is pushed by the radial force, and then erupts leading to a jet. Each Bénard convection cell is also an entrance, where the external gas is attracted into the nucleus, is cycloaccelerated by the radial force as well, goes down to the interior boundary $r = R_s$, and then is pushed toward to the exit. Thus the circulation cells form jets in their exits and accretions in their entrances.

This mechanism can also be applied to supernovae explosion. When a very massive red giant completely consumes its central supply of nuclear fuels, its core collapses. Its radius r_0 begins to decrease, and consequently the δ-factor increases. The huge mass m and the rapidly reduced radius r_0 make the δ-factor approaching one. The thermal convection gives rise to an outward radial circulation momentum flux P_r. Then the radial force as in (1.10.3) will lead to the supernova explosion. Also, $P_r = 0$ at $r = r_0$, where r_0 is the radius of blackhole core of supernovae. Consequently, the supernova's huge explosion preserves a smaller ball, yielding a neutron star.

1.11 Multi-Particle Systems and Unification

The field theory for multi-particle system was discovered by (Ma and Wang, 2014d). Classical quantum dynamic equations describe single particle systems. The existing model for a multi-particle system is non-relativistic and is based on prescribing the interaction between particles using such potentials as the Coulomb potential.

As far as we know, there is still no good model for a multi-particle system, which takes also into consideration the dynamic interactions between particles. The main obstacle for establishing a field theory for an interacting multi-particle system is the lack of basic principles to describe the dynamic interactions of the particles.

Basic postulates for interacting multi-particle systems

To seek the needed principles, we proceed with three observations.

1. One natural outcome of the field theory of four interactions developed recently by the authors and addressed in the previous sections is that the coupling constants for the $U(1)\times SU(2)\times SU(3)$ gauge theory play the role of the three charges e, g_w and g_s for electromagnetism, the weak and the strong interaction. These charges generate interacting fields among the interacting particles.

Now we consider an N-particle system with each particle carrying an interaction charge g. Let this be a fermionic system, and the Dirac spinors be given by

$$\Psi=(\psi_1,\cdots,\psi_N)^{\mathrm{T}},$$

which obeys the Dirac equations:

$$i\gamma^\mu D_\mu\Psi+M\Psi=0, \tag{1.11.1}$$

where M is the mass matrix, and

$$D_\mu\Psi=\partial_\mu\Psi+igG\Psi, \tag{1.11.2}$$

where $G=(G^{ij}_\mu)$ is an Hermitian matrix, representing the interacting potentials between the N-particles generated by the interaction charge g.

Now let

$$\{\tau_0,\ \tau_1,\ \cdots,\ \tau_K \mid K=N^2-1\}$$

be a basis of the set of all Hermitian matrices, where $\tau_0=I$ is the identity, and τ_a $(1\leqslant a\leqslant N^2-1)$ are the traceless Hermitian matrices. Then the Hermitian matrix $G=(G^{ij}_\mu)$ and the differential operator D_μ in (1.11.1) can be expressed as

$$\begin{aligned}&G=G^0_\mu I+G^a_\mu\tau_a,\\&D_\mu=\partial_\mu+igG^0_\mu+igG^a_\mu\tau_a.\end{aligned}$$

Consequently the Dirac equations (1.11.1) are rewritten as

$$i\gamma^\mu[\partial_\mu+igG^0_\mu+igG^a_\mu\tau_a]\Psi+M\Psi=0. \tag{1.11.3}$$

2. The energy contributions of the N particles are indistinguishable, which implies the $SU(N)$ gauge invariance. Hence (1.11.3) are exactly the Dirac equations in the form of $SU(N)$ gauge fields $\{G^a_\mu \mid 1\leqslant a\leqslant N^2-1\}$ with a given external interaction field G^0_μ.

3. The $SU(N)$ gauge theory of N particles must obey PRI. Consequently there exists a constant $SU(N)$ tensor

$$\alpha_a^N = (\alpha_1^N, \cdots, \alpha_N^N),$$

such that the contraction field using PRI

$$G_\mu = \alpha_a^N G^a \tag{1.11.4}$$

is independent of the $SU(N)$ representation τ_a, and is the interaction field which can be experimentally observed.

With these three observations, it is natural for us to introduce the following postulate, which is also presented as Postulates 6.25-6.27 in Chapter 6:

Postulates for interacting multi-particle systems

1) *the Lagrangian action for an N-particle system satisfy the $SU(N)$ gauge invariance*;
2) gG_μ^a *represent the interaction potentials between the particles*; *and*
3) *for an N-particle system, only the interaction field G_μ in (1.11.4) can be measured, and is the interaction field under which this system interacts with other external systems.*

Field Equations for Multi-Particle System

Multi-particle systems are layered, and with the above postulates and basic symmetry principles, we are able to determine in a unique fashion field equations for different multi-particle systems.

For example, given an N-particle system consisting of N fermions with given charge g, the $SU(N)$ gauge symmetry dictates uniquely the Lagrangian density, given in two parts: 1) the sector of $SU(N)$ gauge fields $\mathscr{L}_G$, and 2) the Dirac sector of particle fields $\mathscr{L}_D$, as described earlier in the $SU(N)$ gauge theory. The combined action is 1) $SU(N)$ gauge invariant, 2) representation invariant (PRI), and 3) Lorentz invariant. The field equations are then derived by using PID:

$$\mathscr{G}_{ab}\left[\partial^\nu G^b_{\nu\mu} - g\lambda^b_{cd} g^{\alpha\beta} G^c_{\alpha\mu} G^d_\beta\right] - g\overline{\Psi}\gamma_\mu \tau_a \Psi = \left[\partial_\mu - \frac{k^2}{4} x_\mu + gG_\mu + gG^0_\mu\right]\phi_a, \tag{1.11.5}$$

$$i\gamma^\mu\left[\partial_\mu + igG^0_\mu + igG^a_\mu \tau_a\right]\Psi - M\Psi = 0, \tag{1.11.6}$$

for $1 \leqslant a \leqslant N^2 - 1$, where G^0_μ is the interaction field of external systems, and $G_\mu = \alpha_a G^a_\mu$. It is important to note that coupling to the external fields is achieved by the terms involving G^0_μ.

Unification and geometrization of matter fields

Also, we have established a unified field model coupling matter fields, which matches the vision of Einstein and Nambu, as stated in Nambu's Nobel lecture (Nambu, 2008):

> *Einstein used to express dissatisfaction with his famous equation of gravity*
>
> $$G_{\mu\nu} = 8\pi T_{\mu\nu}$$
>
> *His point was that, from an aesthetic point of view, the left hand side of the equation which describes the gravitational field is based on a beautiful geometrical principle, whereas the right hand side, which describes everything else,... looks arbitrary and ugly.*
>
> *... [today] Since gauge fields are based on a beautiful geometrical principle, one may shift them to the left hand side of Einstein s equation. What is left on the right are the matter fields which act as the source for the gauge fields ... Can one geometrize the matter fields and shift everything to the left*?

Basically, one needs to geometrize the energy-momentum tensor $T_{\mu\nu}$ appearing in the Einstein field equations. For example, for multi-particle system under gravity and electromagnetism, using the basic postulates as outlined above, a unified field model can be naturally derived so that the energy-momentum tensor $T_{\mu\nu}$ is derived from first principles and is geometrized.

The above vision of Einstein and Nambu is achieved by using the postulate for interacting multi-particle systems, the PID, the PRI, as well as the principle of symmetry-breaking (PSB):

1) *fundamental symmetries of Nature dictate the actions for both the subsystems as well as the global system,*
2) *the subsystems are coupled through PID, PRI and PSB.*

We believe that this is the essence of physical modeling and a unique route for unification.

1.12 Weakton Model of Elementary Particles

The weakton model of elementary particles was first introduced by (Ma and Wang, 2015b). Hereafter we address the basic ingredients of this theory.

Motivation and requirements of weaktons

The matter in the Universe is made up of a number of fundamental constituents, called elementary particles. Based on the current knowledge of particle physics, all forms of matter are made up of 6 leptons and 6 quarks, and their antiparticles, which are treated as elementary particles.

Great achievements and insights have been made for last 100 years on the understanding of the structure of subatomic particles and on their interactions; see among many others (Halzen and Martin, 1984; Griffiths, 2008; Kane, 1987; Quigg, 2013). However, there are still many longstanding open questions and challenges. Here are a few fundamental questions which are certainly related to the deepest secrets of our Universe. One such problem is that why leptons do not participate in strong interactions.

The starting point of the study is the puzzling decay and reaction behavior of subatomic particles. For example, the electron radiations and the electron-positron annihilation into photons or quark-antiquark pair clearly shows that there must be interior structure of electrons, and the constituents of an electron contribute to the making of photon or the quark in the hadrons formed in the process. In fact, all sub-atomic decays and reactions show clearly the following conclusion:

There must be interior structure of charged leptons, quarks and mediators.

This conclusion motivates us to propose a model for sub-lepton, sub-quark, and sub-mediators. It is clear that any such model should obey four basic requirements:

1) Mass generation mechanism.

2) Consistency of quantum numbers for both elementary and composite particles.

3) Exclusion of nonrealistic compositions of the elementary particles.

4) Weakton confinement.

The model should lead to consistency of masses for both elementary particles, which we call weaktons, and composite particles (the quarks, leptons and mediators). Since the mediators, the photon γ and the eight gluons are all massless, a natural requirement is that

the proposed elementary particles—weaktons—are massless. Namely, these proposed elementary particles must have zero rest mass.

Weaktons and their quantum numbers

Careful examinations of the above requirements and subatomic decays/reactions lead us to propose six elementary particles, which we call weaktons, and their anti-particles:

$$\begin{aligned} &w^*, \quad w_1, \quad w_2, \quad \nu_e, \quad \nu_\mu, \quad \nu_\tau, \\ &\overline{w}^*, \quad \overline{w}_1, \quad \overline{w}_2, \quad \overline{\nu}_e, \quad \overline{\nu}_\mu, \quad \overline{\nu}_\tau, \end{aligned} \tag{1.12.1}$$

where ν_e, ν_μ, ν_τ are the three generation neutrinos, and w^*, w_1, w_2 are three new particles, which we call w-weaktons. These weaktons in (1.12.1) are endowed with the quantum numbers: electric charge Q_e, weak charge Q_w, strong charge Q_s, weak color charge Q_c, baron number B, lepton numbers L_e, L_μ, L_τ, spin J, and mass m. The quantum numbers of weaktons are listed in Table 1.1.

Table 1.1 Weakton quantum numbers

Weakton	Q_e	Q_w	Q_s	Q_c	L_e	L_μ	L_τ	B	J	m
w^*	2/3	1	1	0	0	0	0	1/3	±1/2	0
w_1	−1/3	1	0	1	0	0	0	0	±1/2	0
w_2	−2/3	1	0	−1	0	0	0	0	±1/2	0
ν_e	0	1	0	0	1	0	0	0	−1/2	0
ν_μ	0	1	0	0	0	1	0	0	−1/2	0
ν_τ	0	1	0	0	0	0	1	0	−1/2	0

Weakton constituents and duality of mediators

The weakton constituents of charged leptons and quarks are given by

$$\begin{aligned} &e = \nu_e w_1 w_2, \quad \mu = \nu_\mu w_1 w_2, \quad \tau = \nu_\tau w_1 w_2, \\ &u = w^* w_1 \overline{w}_1, \quad c = w^* w_2 \overline{w}_2, \quad t = w^* w_2 \overline{w}_2, \\ &d = w^* w_1 w_2, \quad s = w^* w_1 w_2, \quad b = w^* w_1 w_2, \end{aligned} \tag{1.12.2}$$

where c,t and d,s,b are distinguished by the spin arrangements; see (5.3.20).

The weakton constituents of the mediators and their dual mediators are given by

$$\begin{aligned} &\gamma = \cos\theta_w w_1\overline{w}_1 - \sin\theta_w w_2\overline{w}_2\ (\uparrow\uparrow,\downarrow\downarrow), && \gamma_0 = \cos\theta_w w_1\overline{w}_1 - \sin\theta_w w_2\overline{w}_2\ (\uparrow\downarrow,\downarrow\uparrow), \\ &Z = \sin\theta_w w_1\overline{w}_1 + \cos\theta_w w_2\overline{w}_2\ (\uparrow\uparrow,\downarrow\downarrow), && H^0 = \sin\theta_w w_1\overline{w}_1 + \cos\theta_w w_2\overline{w}_2\ (\uparrow\downarrow,\downarrow\uparrow), \\ &W^- = w_1 w_2\ (\uparrow\uparrow,\downarrow\downarrow), && H^- = w_1 w_2(\uparrow\downarrow,\downarrow\uparrow), \\ &W^+ = \overline{w}_1\overline{w}_2(\uparrow\uparrow,\downarrow\downarrow), && H^+ = \overline{w}_1\overline{w}_2(\uparrow\downarrow,\downarrow\uparrow), \\ &g^k = w^*\overline{w}^*(\uparrow\uparrow,\downarrow\downarrow), && g_0^k = w^*\overline{w}^*(\uparrow\downarrow,\downarrow\uparrow). \end{aligned} \tag{1.12.3}$$

Remarkably, both the spin-1 mediators and the spin-0 dual mediators in the unified field theory have the *same* weakton constituents, but with different spins. The spin arrangements clearly demonstrate that there must be spin-0 particles with the same weakton constituents as the mediators. Consequently, there must be dual mediators with spin-0. This observation clearly supports the unified field model presented earlier. Conversely, the existence of the dual mediators makes the weakton constituents perfectly fit.

Also, a careful examination of weakton constituents predicts the existence of an additional mediator, which we call the ν-mediator:

$$\phi_\nu^0 = \sum_l \alpha_l \nu_l \overline{\nu}_l(\downarrow\uparrow), \quad \sum_l \alpha_l^2 = 1, \tag{1.12.4}$$

taking into consideration of neutrino oscillations. When examining decays and reactions of sub-atomic particles, it is apparent for us to predict the existence of this mediator.

With the above weakton constituents of charged leptons, quarks and mediators, we can then verify the four basic requirements as mentioned earlier, based in part on the weak interaction potentials.

1. *Mass generation.* One important conclusion of the aforementioned weakton model is that all particles—both matter particles and mediators—are made up of massless weaktons. A fundamental question is how the mass of a massive composite particle is generated. In fact, based on the Einstein formulas:

$$\frac{d}{dt}\vec{P} = \sqrt{1-\frac{v^2}{c^2}}\vec{F}, \qquad m = \sqrt{1-\frac{v^2}{c^2}}\frac{E}{c^2}, \tag{1.12.5}$$

we observe that a particle with an intrinsic energy E has zero mass $m = 0$ if it moves in the speed of light $v = c$, and possess nonzero mass if it moves with a velocity $v < c$. Hence by this mass generation mechanism, for a composite particle, the constituent massless weaktons can decelerate by the weak force, yielding a massive particle.

In principle, when calculating the mass of the composite particle, one should also consider the bounding and repelling energies of the weaktons, each of which can be very large. Fortunately, the constituent weaktons are moving in the "asymptotically-free" shell region of weak interactions as indicated by the weak interaction potential/force formulas, so that the bounding and repelling contributions to the mass are mostly canceled out. Namely, the mass of a composite particle is due mainly to the dynamic behavior of the constituent weaktons.

2. *Consistency and removal of unrealistic compositions.* The consistency can be easily checked. Also a few simple quantum rules can be devised so that unrealistic combinations of weaktons are easily excluded.

3. The weakton confinement is simply the direct consequence of the weak interaction potential/force formulas.

Mechanism of decays

Remarkably, the weakton model offers a perfect explanation for all sub-atomic decays. In particular, all decays are achieved by 1) exchanging weaktons and consequently exchanging newly formed quarks, producing new composite particles, and 2) separating the new composite particles by weak and/or strong forces.

One aspect of this decay mechanism is that we know now the precise constituents of particles involved in all decays/reactions both before and after the reaction. It is therefore believed that the new decay mechanism provides clear new insights for both experimental and theoretical studies.

Chapter 2

Fundamental Principles of Physics

The purposes of this chapter are 1) to provide an intuitive introduction to fundamental principles of Nature, and 2) to explain how the laws of Nature are derived based on these principles. The main focus is on the symbiotic interplay between advanced mathematics and the laws of Nature. For this purpose, we start with a brief overview on the perspective and the physical significance of a few fundamental principles.

Section 2.1 provides a basic intuitive introduction to fundamental principles, symmetries, the geometric interaction mechanism and the symmetry-breaking principle. We start with two guiding principles of theoretical physics, which can be synthesized as: the laws of Nature are represented by mathematical equations, are dictated by a few fundamental principles, and always take the simplest and aesthetic forms.

The geometric interaction mechanism was originally motivated by Albert Einstein's vision revealed in his principle of equivalence, and was first postulated in (Ma and Wang, 2014d).

Symmetry plays a fundamental role in physics. Many, if not all, physical systems obey certain symmetry. For this reason, many fundamental principles in physics address the underlying symmetries of physical systems.

However, a crucial component of the unification of four fundamental interactions as well as the modeling of multi-level physical systems is the symmetry-breaking mechanism. Consequently, we postulated in (Ma and Wang, 2014a) and in Section 2.1.7 a principle of symmetry-breaking 2.14, which states that for a system coupling different levels of physical laws, part of these symmetries of the subsystems must be broken.

Section 2.2 addresses essentials of the classical Lorentz invariance, and its applications to the derivation of Schrödinger equation, the Klein-Gordon equation, the Weyl equations and the Dirac equations in relativistic quantum mechanics.

Section 2.3 presents a brief introduction to the Einstein theory of general relativity, focusing on Einstein's two basic principles, the principle of equivalence and the principle of general relativity, and on the derivation of the Einstein gravitational field equations.

A brief introduction to both the $U(1)$ abelian and the $SU(N)$ non-abelian gauge theories are presented in Section 2.4. In particular, the $U(1)$ abelian gauge theory is defined on the complex spinor bundle $\mathscr{M}^4 \otimes_p \mathbb{C}^4$. The $SU(N)$ non-abelian gauge theory, also called the Yang-Mills theory, is the connection on the complex spinor bundle $\mathscr{M}^4 \otimes_p (\mathbb{C}^4)^N$. Note that the presentation here leads clearly to PRI, to be introduced in Chapter 4.

Section 2.5 introduces the principle of Lagrangian dynamics (PLD) and its many applications such as in classical mechanics and electrodynamics. The Noether Theorem, Theorem 2.38, is also proved.

Section 2.6 is on Hamiltonian dynamics and its connections to the Lagrangian dynamics.

2.1 Essence of Physics

2.1.1 General guiding principles

Physics studies fundamental interactions, motion and formation of matter in our Universe. The heart of fundamental physics is to seek experimentally verifiable, fundamental laws and principles of Nature. Namely, physical concepts and theories are transformed into mathematical models, and the predictions derived from these models can be verified experimentally and conform to the reality.

Modeling is a crucial step to understand a physical phenomena. A good model should be derived based on a few fundamental principles, and often presented in the form of differential equations. The first principles are often connected with symmetries, which also dictate specific forms of mathematical models, represented by differential equations.

In a nutshell, we have the following first principle of physics.

First Principle of Physics 2.1 *Physical systems obey laws and principles of Nature*:

1) *These laws and principles can be expressed using mathematical models*:

Model = Mathematical representation of physical laws and principles;

2) *These laws and principles are universal, and the universality is reflected by symmetries in physics*; *and*

3) *Physical symmetries dictate the precise forms of the mathematical models, which are often presented in the forms of ordinary or partial differential equations.*

In addition, the following principle provides the essence of fundamental physics, and is supported by known experimental facts.

Essence of Physics 2.2 *The essence of physics can be drawn in the following two important ingredients*:

1) *Theoretical physics is built upon a few fundamental principles of Nature*; *and*

2) *The laws of Nature always take the simplest and aesthetic forms.*

First Principle 2.1 and Essence of Physics 2.2 serve as guiding principles to understand and explore the physical world of our Universe and the laws of Nature.

2.1.2 Phenomenological methods

Theories of physics fall into two categories: first principle theories and phenomenological theories. The first principle theories are derived based on a few fundamental laws and principles of Nature, and phenomenological theories are conjectured and synthesized from observational data.

For example, the Newton's gravitational law is given by

$$F = -\frac{m_1 m_2 G}{r^2}, \tag{2.1.1}$$

which is essentially deduced by using the phenomenological technique based on a large number of astronomical data. First, one readily conjectures that the gravitational force is proportional to the masses m and M of the two bodies, and is a function of distance r:

$$F = -mM\Phi(r),$$

where $\Phi(r)$ is an undetermined function.

Then one considers two planets with masses m_1 and m_2, rotating around the Sun with velocities v_1 and v_2. By the balance between the gravitational and centrifugal forces, we have

$$\frac{v_i^2}{r_i} = M\Phi(r_i) \qquad \text{for } i = 1, 2,$$

where M is the mass of the Sun. Consequently,

$$\frac{\Phi(r_1)}{\Phi(r_2)} = \frac{r_2}{r_1}\frac{v_1^2}{v_2^2}. \tag{2.1.2}$$

By the Kepler's third law, the periods T_1 and T_2 of the two planets satisfy

$$\frac{T_1^2}{T_2^2} = \frac{r_1^3}{r_2^3}, \qquad T_i v_i = 2\pi r_i \qquad \text{for } i = 1, 2,$$

which implies that

$$\frac{v_1^2}{v_2^2} = \frac{r_2}{r_1}. \tag{2.1.3}$$

Then, it follows from (2.1.2) and (2.1.3) that

$$\Phi(r) = \frac{G}{r^2},$$

where G is a constant, called the gravitational constant. Thus, by the Kepler's third law, one can postulate naturally the Newton's gravitational law (2.1.1). In other words, the Newton's gravitational law can be regarded as a phenomenological theory.

Albert Einstein was the first who tried to deduce basic physical laws from the first principles. For example, the Einstein theory of general relativity and the Einstein gravitational field equations:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu}, \tag{2.1.4}$$

are derived based on the following three basic principles:

1) the principle of general relativity,

2) the principle of equivalence, and

3) the principle of Lagrangian dynamics.

The Schwarzschild solution of the Einstein gravitational field equations (2.1.4) offers a natural link between the field equations and the Newtonian gravitational law (2.1.1).

2.1.3 Fundamental principles in physics

Fundamental first principles refer to the laws of Nature that cannot be derived from other more basic laws.

Based on Essence of Physics 2.2, if we would like to better understand theoretical physics, it is crucial to know and find out all fundamental laws. In this subsection, we shall list the known and important principles in various physical fields, most of which will be introduced in later chapters.

We start with the introduction of Principle of Lagrangian Dynamics (PLD), which is of special importance.

In classical mechanics, we know the least action principle. For a mechanical system with the position and velocity variables

$$x = (x_1, \cdots, x_n), \qquad \dot{x} = (\dot{x}_1, \cdots, \dot{x}_n), \tag{2.1.5}$$

let the kinetic energy T and the potential energy V be functions of the position and velocity variables in (2.1.5). Then the states of the system are the extremum points of the functional

$$L(x) = \int_{t_0}^{t_1} \mathscr{L}(x,\dot{x})dt, \tag{2.1.6}$$

where the integrand $\mathscr{L}$

$$\mathscr{L}(x,\dot{x}) = T - V \tag{2.1.7}$$

is called the Lagrange density, and the functional (2.1.6) is called the Lagrange action. The extremum points x satisfy the variational equation, called the Euler-Lagrange equations of (2.1.6):

$$\delta L(x) = 0, \tag{2.1.8}$$

where δL is the variational derivative operator of L. The equation (2.1.8) can be equivalently expressed as

$$\frac{d}{dt}\frac{\partial \mathscr{L}}{\partial \dot{x}} - \frac{\partial \mathscr{L}}{\partial x} = 0.$$

The most important point is that the least action principle can be generalized to all physical systems describing motions, as well as to fundamental interactions of Nature. The generalization in a motion system is called the PLD, and in an interaction system is called the Principle of Interaction Dynamics (PID). PID will be introduced in detail In Chapter 4 of this book, and PLD is stated as follows.

Principle of Lagrangian Dynamics 2.3 *For a physical motion system, there are functions*

$$u = (u_1, \cdots, u_n), \tag{2.1.9}$$

which describe the states of this system, and there exists a functional of u, given by

$$L(u) = \int_\Omega \mathscr{L}(u, Du, \cdots, D^m u)dx, \tag{2.1.10}$$

where Ω is the domain of u, and $D^k u$ is the k-th derivative of u for any $0 \leqslant k \leqslant m$. Then the state functions of this system are the extremum points of (2.1.10). *Namely the state functions satisfy the variational equation of* (2.1.10):

$$\delta L(u) = 0. \tag{2.1.11}$$

The functional L is called the Lagrange action, and $\mathscr{L}$ is called the Lagrange density.

By PLD, to derive dynamical equations for a physical system, it suffices to find the corresponding Lagrange action. In the next subsection, we demonstrate that symmetries in physics dictate the precise forms of the Lagrange actions.

A list of known important physical principles and laws of Nature in various subfields of physics is given as follows.

1) Universal principles in all fields:

- Principle of Lagrangian Dynamics (PLD),
- Principle of Hamiltonian Dynamics (PHD),
- Principle of Interaction Dynamics (PID),
- Lorentz Invariance,

- Principle of General Relativity,
- Principle of Gauge Invariance,
- Principle of Representation Invariance (PRI),
- Principle of Symmetry-Breaking

2) Classical mechanics:

- Newton's Second Law,
- Principle of Least Action,
- Principle of Galilean Invariance,
- Fick and Fourier Diffusion Laws.

3) Quantum physics:

- Basic postulates of quantum physics,
- Pauli Exclusion Principle.

4) Statistical physics:

- Basic Laws of Thermodynamics,
- Le Châterlier Principle.

5) Nonlinear sciences:

- Principle of Phase Transition Dynamics.

A couple of remarks are now in order.

Remark 2.4 Because various conservation laws can be deduced by PHD, PLD and symmetries based on the Noether theorem, these laws are not listed in the fundamental principles. □

Remark 2.5 Mathematical models form the skeleton of theoretical physics, and most, if not all, mathematical models (differential equations) in theoretic physics can be derived based on the principles listed above. One of the ultimate goals of this book is to derive physical models based on a first principles. □

2.1.4 Symmetry

Symmetry plays an important role in physics. We start with intuitively examining how symmetry works, keeping in mind the relation between equations and laws of physics in First Principle 2.1, which can be simply recast as

$$\text{Laws of Physics} = \text{Differential Equations}. \tag{2.1.12}$$

The laws of Nature on the left hand side of (2.1.12) are often beyond words, and are best expressed by differential equations. It is this characteristic, together with the Noether theorem, that illustrates the importance of symmetry in physics, as illustrated below:

$$\begin{array}{ccc} \text{Invariance} & \text{Covariance} & \text{Symmetry} \\ \Downarrow & \Downarrow & \Downarrow \\ \text{Form of Equation} & \text{Space Structure} & \text{Conservation Law} \end{array}$$

Namely, symmetry dictates and determines

1) the explicit form of differential equations governing the underlying physical system,
2) the space-time structure of the Universe, and the mechanism of fundamental interactions of Nature, and
3) physical conservation laws.

We now give a simple example to demonstrate how symmetry determines the explicit form of equations. We begin with two basic implications of a symmetry:

a) Fundamental laws of Nature are universal, and their validity is independent of the space, time, and directions of experiments and observations, and
b) By (2.1.12), the universality of laws of Nature implies that the differential equations representing them are covariant. Equivalently the Lagrange actions are invariant under the corresponding coordinate transformations.

As an example, consider a physical system obeying PLD 2.3. For simplicity, we assume that the Lagrange action of this fictitious system is finite-dimensional. Namely, let

$$F : \mathbb{R}^n \to \mathbb{R}^1$$

be a finite dimensional function regarded as the action of the physical system, and we need to determine the explicit form of this function

$$F = F(x) \qquad \text{for } x = (x_1, \cdots, x_n) \in \mathbb{R}^n. \tag{2.1.13}$$

Consider the case where the system satisfies the following invariance principle, called the principle of rotational invariance.

Principle 2.6 (Rotational Invariance) *The laws that the system (2.1.13) obeys are the same under all orthogonal coordinate systems.*

Based on the implication b) of symmetry above, Principle of Rotational Invariance 2.6 says that the action (2.1.13) is invariant under orthogonal coordinate transformations:

$$\widetilde{x} = Ax \qquad \text{where } A \text{ is an arbitrary orthogonal matrix.} \tag{2.1.14}$$

We know that the two forms as

$$|x|^2 = x_1^2 + \cdots + x_n^2, \quad x \cdot y = x_1 y_1 + \cdots + x_n y_n$$

are invariant, where $y \in \mathbb{R}^n$ is given, and the invariant function $F(x)$ must be of the form:

$$F(x) = F(|x|,\ x \cdot y)$$

In addition, by the simplicity principle in Essence of Physics 2.2, the exponent of F in x is two. Thus the action (2.1.13) is defined in the form as

$$F(x) = \alpha |x|^2 + \beta x \cdot y, \tag{2.1.15}$$

where α, β are two constants, and y is a given vector.

Thus we deduce an explicit expression (2.1.15) for the function F, which are invariant under rotational symmetry, Principle 2.6. The graph of (2.1.15) is a sphere, and the invariance of (2.1.15) means that the sphere looks the same from different directions. This is the reason why an invariance is called a symmetric principle.

2.1.5 Invariance and tensors

Symmetries are characterized by three ingredients:

spaces, transformation groups, and tensors,

which are applicable to different physical fields. Here are the five currently known important symmetries and their corresponding ingredients:

1) Galileo Invariance:

- Space: Euclidean space $\mathbb{R}^3$,
- Transformation group: Galileo group and $SO(3)$,
- Tensor types: Cartesian tensors,
- Fields: classical mechanics, fluid dynamics, and astrophysics.

2) Lorentz Invariance:

- Space: Minkowski space,
- Transformation group: Lorentz groups,
- Tensor types: 4-dimensional tensors and spinors,
- Fields: Quantum physics and Interactions.

3) Einstein General Relativity:

- Space: 4-dimensional Riemann manifolds,
- Transformation group: $GL(4)$,
- Tensor types: General tensors and Riemann metric,
- Fields: Gravitation and Astrophysics.

4) Gauge Invariance:

- Space: complex vector bundle $\mathscr{M} \otimes_p \mathbb{C}^n$,
- Transformation groups: $U(1)$ and $SU(n)$,
- Tensor types: Wave functions and gauge fields,
- Fields: Quantum physics and Interactions.

5) Representation Invariance:

- Space: Tangent space of $SU(n)$,
- Transformation: $GL(\mathbb{C}^n)$,
- Tensor types: $SU(n)$ tensors,
- Fields: Quantum physics and Interactions.

We now use examples in classical mechanics to illustrate the main characteristics of these symmetries and to clarify how tensors describe invariances.

Principle 2.7 (Galilean Invariance) *Mechanical laws are invariant under the following transformations:*

1) *Galilean transformation*

$$\widetilde{t} = t, \qquad \widetilde{x} = x + vt \qquad \text{for } x \in \mathbb{R}^3, \tag{2.1.16}$$

where v is a constant velocity, and

2) *translational and rotational transformations*

$$\widetilde{t} = t + t_0, \qquad \widetilde{x} = Ax + b \qquad \text{for } x \in \mathbb{R}^3, \tag{2.1.17}$$

where $A \in SO(3)$ is an orthogonal matrix, $b \in \mathbb{R}^3$ and $t_0 \in \mathbb{R}^1$ are constant.

By (2.1.12) we know that the invariance of physical laws is equivalent to the covariance of the corresponding differential equations. We now demonstrate that the Newton's Second Law obeys the Galileo Invariance Principle 2.7.

The Newton's Second Law reads as

$$ma = F, \tag{2.1.18}$$

and the corresponding differential equation describing the motion of a particle is given by

$$m\frac{d^2x}{dt^2} = F. \tag{2.1.19}$$

When we investigate the motion in other reference frame $(\widetilde{x}, \widetilde{t})$ with the transformation

$$\widetilde{t} = t + t_0, \qquad \widetilde{x} = Ax + vt + b, \tag{2.1.20}$$

by (2.1.20) we derive

$$\frac{d^2\widetilde{x}}{d\widetilde{t}^2} = A\frac{d^2x}{dt^2}. \tag{2.1.21}$$

On the other hand, experiments show that the expressions $\widetilde{F}$ in $(\widetilde{x}, \widetilde{t})$ and F in (x,t) of force satisfy the relation:

$$\widetilde{F} = AF. \tag{2.1.22}$$

Hence, by (2.1.19) we deduce from (2.1.21) and (2.1.22) that

$$m\frac{d^2\widetilde{x}}{d\widetilde{t}^2} = \widetilde{F}. \tag{2.1.23}$$

It is clear that both forms of (2.1.19) and (2.1.23) are the same under the transformation (2.1.20), i.e. under the transformations (2.1.16) and (2.1.17).

Mathematically, the transformations corresponding to (2.1.16) and (2.1.17) are the Galileo group and $SO(3)$, and the functions satisfying (2.1.21) and (2.1.22) under the transformation (2.1.20) are called first-order Cartesian tensors, which are also the usual vectors in $\mathbb{R}^3$.

Also, this example demonstrates that invariance principles must be characterized by corresponding tensors. In the later sections, we shall give precise definitions of various types of tensors.

Remark 2.8 In classical mechanics, the laws were first discovered phenomenologically, and then were known to obey the Galileo Principle of Invariance. It is Einstein's vision that laws of physics can be derived from symmetries. □

Remark 2.9 The transformation (2.1.17) consists of three subclasses of transformations:

$$\begin{aligned} &\text{the rotation transformation:} && \widetilde{x} = Ax, \\ &\text{the space translation:} && \widetilde{x} = x + b, \\ &\text{the time translation:} && \widetilde{t} = t + t_0. \end{aligned} \tag{2.1.24}$$

Unlike the Galileo invariance of (2.1.16), which is only valid in the classical mechanics, the invariance for the transformations (2.1.24) are universally valid in all physical fields. In particular, based on the Noether theorem, Theorem 2.38, we can deduce the following corresponding conservation laws:

Time translation invariance	$\Rightarrow$	Energy conservation,
Space translation invariance	$\Rightarrow$	Momentum conservation,
Rotational invariance	$\Rightarrow$	Angular momentum conservation.

2.1.6 Geometric interaction mechanism

Albert Einstein was the first physicist who postulated that the gravitational force is caused by the time-space curvature. However, Yukawa's viewpoint, entirely different from Einstein's, is that the other three fundamental forces take place through exchanging intermediate bosons such as photons for the electromagnetic interaction, $W^{\pm}$ and Z intermediate vector bosons for the weak interaction, and gluons for the strong interaction.

Based on our recent studies on field theory of the four interactions, in the same spirit as the Einstein's mechanism of gravitational force, it is natural for us to postulate a mechanism for all four interactions different from that of Yukawa.

To proceed, we recall that in geometry, the square ds^2 of an infinitesimal arc-length in a flat space can be written as

$$ds^2 = dx_1^2 + \cdots + dx_n^2,$$

which is the well known Pythagorean theorem, and in a curved space ds^2 is given by

$$ds^2 = g_{ij}(x)dx_i dx_j \qquad \text{with } g_{ij} \neq \delta_{ij}, \tag{2.1.25}$$

and the Pythagorean theorem is in general not true in a curved space. Mathematically, a space $\mathscr{M}$ being flat indicates that one can choose properly a coordinate system so that the metric

$$g_{ij} = \delta_{ij},$$

and otherwise, $\mathscr{M}$ will be curved.

Regarding to the laws of Nature on our Universe $\mathscr{M}$, physical states are described by functions $u = (u_1, \cdots, u_n)$ defined on $\mathscr{M}$:

$$u : \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{R}^n \qquad \text{for non-quantum system,} \tag{2.1.26}$$

$$u : \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{C}^n \qquad \text{for quantum system,} \tag{2.1.27}$$

which are solutions of differential equations associated with the laws of the underlying physical system:

$$\delta L(Du) = 0, \tag{2.1.28}$$

where D is a derivative operator, and L is the Lagrange action. Consider two transformations for the two physical systems (2.1.26) and (2.1.27):

$$\widetilde{x} = Tx \qquad \text{for (2.1.26)}, \tag{2.1.29}$$

$$\widetilde{u} = e^{i\theta\tau}u \qquad \text{for (2.1.27)}, \tag{2.1.30}$$

where x is a coordinate system in $\mathscr{M}$, $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation, $e^{i\theta\tau} : \mathbb{C}^n \to \mathbb{C}^n$ is an $SU(n)$ transformation, θ is a function of x, and τ is a traceless Hermitian matrix; see Section 2.4 on gauge theory.

We now state two important symmetric principles: the Einstein principle of general relativity and the gauge invariance.

Principle 2.10 (General Relativity) *Laws of Physics are the same under all coordinate systems. Namely, equations* (2.1.28) *are covariant and equivalently the action L is invariant under all coordinate transformations* (2.1.29).

Principle 2.11 (Gauge Invariance) *A quantum system with electromagnetic, weak, and strong interactions is invariant under the corresponding $SU(n)$ gauge transformations* (2.1.30).

One important consequence of the invariance of (2.1.28) under the transformations (2.1.29) and (2.1.30) is that the derivatives D in (2.1.28) must take the following form (see Section 2.4 for detailed derivations):

$$D = \nabla + \Gamma \qquad \text{for (2.1.26)}, \tag{2.1.31}$$

$$D = \nabla + igA\tau \qquad \text{for (2.1.27)}, \tag{2.1.32}$$

where Γ depends on the metrics g_{ij}, A is a gauge field, representing the interaction potential, and g is the coupling constant, representing the interaction charge.

The derivatives defined in (2.1.31) and (2.1.32) are called connections respectively on $\mathscr{M}$ and on the complex vector bundle $\mathscr{M} \otimes_p \mathbb{C}^n$. For the connections, we have the following theorem, providing a mathematical basis for our interaction mechanism.

Theorem 2.12

1) *The space $\mathscr{M}$ is curved if and only if $\Gamma \neq 0$ in all coordinates, or equivalently $g_{ij} \neq \delta_{ij}$ under all coordinate systems*;

2) *The complex bundle $\mathscr{M} \otimes_p \mathbb{C}^n$ is geometrically nontrivial or twisted if and only if $A \neq 0$.*

Consequently, by Principle 2.10 the presence of the gravitational field implies that the space-time manifold is curved, and, by Principle 2.11, the presence of the electromagnetic, the weak and strong interactions indicates that the complex vector bundle $\mathscr{M} \otimes_p \mathbb{C}^n$ is twisted:

$$\begin{aligned} &\text{Principle 2.10} \quad \Rightarrow \quad g_{ij} \neq \delta_{ij} \quad \Rightarrow \quad \mathscr{M} \text{ is curved,} \\ &\text{Principle 2.11} \quad \Rightarrow \quad A \neq 0 \quad \quad \Rightarrow \quad \mathscr{M} \otimes_p \mathbb{C}^n \text{ is twisted.} \end{aligned} \tag{2.1.33}$$

This analogy, together with Einstein's vision on gravity as the curved effect of space-time manifold, it is natural for us to postulate the following mechanism for all four interactions.

Geometric Interaction Mechanism 2.13 *The gravitational force is the curved effect of the time-space, and the electromagnetic, weak, strong interactions are the twisted effects of the underlying complex vector bundles $\mathscr{M} \otimes_p \mathbb{C}^n$.*

As mentioned earlier, traditionally one adopts Yukawa's viewpoint that forces of the interactions of Nature are caused by exchanging the field mediators. Namely, the gravitation is due to exchanging gravitons, the electromagnetic force is due to exchanging photons, the weak force is due to exchanging the intermediate vector bosons, and the strong force is due to exchanging gluons. The point of view we are taking is the Interaction Mechanism 2.13, following Einstein's version.

2.1.7 Principle of symmetry-breaking

Different physical systems obey different physical principles. The four fundamental interactions of Nature, the quantum systems, the fluid dynamics and heat conduction obey the following symmetry principles:

$$\begin{aligned} &\text{the general relativity for gravity,} \\ &\text{the Lorentz and gauge invariances for the other three interactions,} \\ &\text{the Lorentz invariance for quantum systems, and} \\ &\text{the Galilean invariance for fluid and heat conductions.} \end{aligned} \tag{2.1.34}$$

The corresponding fields and systems in (2.1.34) are governed by the following physical laws and first principles:

$$\begin{aligned} &\text{PID and PRI for four fundamental interactions,} \\ &\text{PLD for quantum systems,} \\ &\text{the Newton Second Law for fluids,} \\ &\text{diffusion laws for heat conductivity,} \end{aligned} \tag{2.1.35}$$

Here PID stands for the principle of interaction dynamics, and PRI stands for the principle of representation invariance, both discovered recently by the authors. PLD stands for the principle of Lagrangian dynamics.

Astrophysics is the only field that involving all the fields in (2.1.34) and (2.1.35). Consequently, one needs to couple the basic laws in (2.1.35) to model astrophysical dynamics.

One difficulty we encounter now is that the Newtonian Second Law for fluid motion and the diffusion law for heat conduction are not compatible with the principle of general relativity. Also, there are no basic principles and rules for combining relativistic systems and the Galilean systems together to form a consistent system. The reason is that in a Galilean system, time and space are independent, and physical fields are 3-dimensional; while in a relativistic system, time and space are related, and physical fields are 4-dimensional.

The distinction between relativistic and Galilean systems gives rise to an obstacle for establishing a consistent model of astrophysical dynamics, coupling all the physical systems in (2.1.34) and (2.1.35).

In the unified field theory based on PID, the key ingredient for coupling gravity with the other three fundamental interactions is achieved through spontaneous gauge symmetry breaking. Here we propose that the coupling between the relativistic and the Galilean systems through relativistic-symmetry breaking.

In fact, the model given by (7.1.75)-(7.1.76) follows from this symmetry-breaking principle, where we have to chose the coordinate system

$$x^\mu = (x^0, x), \quad x^0 = ct \ \text{ and } x = (x^1, x^2, x^3),$$

such that the metric is in the form:

$$ds^2 = -\left(1 + \frac{2}{c^2}\phi(x,t)\right)c^2dt^2 + g_{ij}(x,t)dx^i dx^j. \tag{2.1.36}$$

Here g_{ij} $(1 \leqslant i, j \leqslant 3)$ are the spatial metric, and

$$\phi = \text{ the gravitational potential.} \tag{2.1.37}$$

With this metric (2.1.36)-(2.1.37), we can establish the fluid and heat equations as in (7.1.78). It is then clear that by fixing the coordinate system to ensure that the metric is in the form (2.1.36)-(2.1.37), the system breaks the symmetry of general coordinate transformations, and we call such symmetry-breaking as relativistic-symmetry breaking.

We believe the symmetry-breaking is a general phenomena when we deal with a physical system coupling different subsystems in different levels. The unified field theory for the four fundamental interactions is a special case, which couples the general relativity, the Lorentz and the gauge symmetries. Namely, the symmetry of general relativity needs to be linked to both the Lorentz invariance and the gauge invariance in two aspects as follows:

1) In the Dirac equations for the fermions:

$$i\gamma^\mu D_\mu \psi - \frac{c}{\hbar} m\psi = 0,$$

γ^μ have to obey two different transformations.

2) The gauge-symmetry breaks in the gravitational field equations coupling the other interactions:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} + D_\mu \Phi_\nu, \tag{2.1.38}$$

where

$$D_\mu = \nabla_\mu + \frac{k_1}{\hbar c}eA_\mu + \frac{k_2}{\hbar c}g_w W_\mu + \frac{k_3}{\hbar c}g_s S_\mu, \tag{2.1.39}$$

∇_μ is the covariant derivative, k_i $(1 \leqslant i \leqslant 3)$ are parameters, A_μ, W_μ, S_μ are the total electromagnetic, weak and strong interaction potentials. It is the terms A_μ, W_μ, S_μ in (2.1.39) that break the gauge symmetry of (2.1.38).

In summary, we are ready to postulate a general symmetry-breaking principle.

Principle of Symmetry-Breaking 2.14

1) *The three sets of symmetries,*

the general relativistic invariance,
the Lorentz and gauge invariances, and (2.1.40)
the Galileo invariance,

are mutually independent and dictate in part the physical laws in different levels of Nature; *and*

2) *for a system coupling different levels of physical laws, part of these symmetries must be broken.*

This principle of symmetry-breaking holds the key component for us to establish the PID unified field theory for four fundamental interactions in Chapter 4, the field equations for multi-particle systems in Chapter 6, and the astrophysical fluid dynamical equations in Chapter 7, resolving a number of important physical problems.

2.2 Lorentz Invariance

2.2.1 Lorentz transformation

In 1903, H. A. Lorentz discovered the Lorentz coordinate transformation, under which the laws of electromagnetism are invariant. Consequently, physicists discovered two different invariances:

the Galilean invariance in classical mechanics, (2.2.1)

the Lorentz invariance in electromagnetism. (2.2.2)

In 1905, Albert Einstein introduced the special theory of relativity, based on the following two first principles: the principle of special relativity and the principle of invariance of the speed of light.

Principle 2.15 (Special Relativity) *Physical laws are covariant in all inertial systems.*

Principle 2.16 (Invariance of speed of light) *The vacuum speed of light is a universal constant.*

A few remarks are now in order. First, a system of reference is always needed to describe the nature, and an inertial system is a system on which a freely moving object moves with constant velocity (Landau and Lifshitz, 1975). Second, it is clear that under the Galilean invariance, the speed of light changes in different inertial systems. Consequently, the Galilean invariance and the Lorentz invariance are incompatible. Third, the invariance of the vacuum speed of light was verified by the Michelson and Morley experiment. Einstein discovered his special theory of relativity by postulating, based on Principles 2.15-2.16, the following principle of Lorentz invariance.

Principle 2.17 (Lorentz Invariance) *Physical laws are invariant under Lorentz transformations.*

With this principle, the special theory of relativity was developed in two directions: a) the introduction of relativistic mechanics, replacing the classical mechanics, and b) the development of relativistic quantum mechanics.

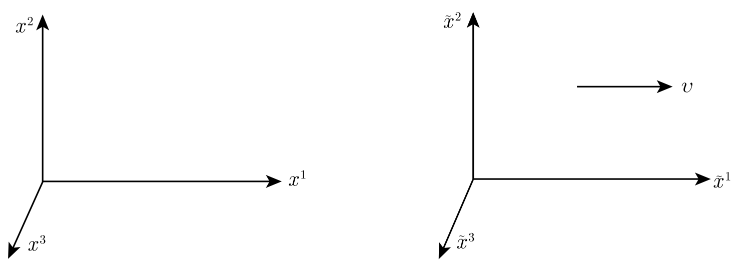


Figure 2.1

We now introduce typical Lorentz transformations [1]. Let (x,t) and $(\widetilde{x},\widetilde{t})$ be two inertial systems which are in relativistic motion with a constant velocity v, as shown in Figure 2.1, where v is in the x_1-axis direction. Then the Lorentz transformation for the two inertial systems is given by

$$(\widetilde{x}^1,\widetilde{x}^2,\widetilde{x}^3)=\left(\frac{x^1-vt}{\sqrt{1-v^2/c^2}},x^2,x^3\right), \qquad \widetilde{t}=\frac{t-x^1v/c^2}{\sqrt{1-v^2/c^2}}, \tag{2.2.3}$$

where c is the speed of light.

[1] General Lorentz transformations are linear transformations that preserves the Minkowski metric (2.2.6), and the transformation (2.2.3) is called a boost.

The vacuum speed of light is invariant under the above Lorentz transformation (2.2.3). To see this, notice that the speed of light is $\widetilde{c} = d\widetilde{x}^1/d\widetilde{t}$ measured in the $\widetilde{x}$-coordinate system, and is $c = dx^1/dt$ measured in the x-coordinate system. Consequently,

$$\widetilde{c} = \frac{d\widetilde{x}^1}{d\widetilde{t}} = \frac{dx^1 - vdt}{\sqrt{1-v^2/c^2}} \Big/ \frac{dt - dx^1 v/c^2}{\sqrt{1-v^2/c^2}} = \frac{dx^1/dt - v}{1-(dx^1/dt)(v/c^2)} = \frac{c-v}{1-v/c} = c.$$

2.2.2 Minkowski space and Lorentz tensors

We recall that each symmetry is characterized by three ingredients: space/manifold, transformation group, and tensors. For the Lorentz invariance, the transformation group is the Lorentz group, consisting of Lorentz transformations such as those given by (2.2.3), and the corresponding space is the Minkowski space introduced below.

Minkowski Space

We know that the Newtonian mechanics is defined in the Euclidean space $\mathbb{R}^3$, and the time t is regarded as a parameter. From the Lorentz transformation (2.2.3), we see that in relativistic physics, there is no absolute space and time, and the time t cannot be separated from the space $\mathbb{R}^3$. Namely the Minkowski space $\mathscr{M}^4$ is a 4-dimensional space-time manifold defined by

$$\mathscr{M}^4 = \{(x^0, x^1, x^2, x^3) \mid x^0 = ct,\ (x^1, x^2, x^3) \in \mathbb{R}^3\}, \tag{2.2.4}$$

where c is the speed of light, and $\mathscr{M}^4$ is endowed with the Riemannian metric:

$$(g_{\mu\nu}) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.2.5}$$

The doublet $\{\mathscr{M}^4, g_{\mu\nu}\}$ given by (2.2.4) and (2.2.5) is called the Minkowski space, and the metric (2.2.5) can be equivalently expressed in the form

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = -c^2dt^2 + dx^2 = -c^2dt^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2. \tag{2.2.6}$$

Lorentz transformations are linear transformations of the Minkowski space $\mathscr{M}^4$ that preserve the metric (2.2.6). The following coordinate transformation of the Minkowski space is a Lorentz transformation, called boost, corresponding to Figure 2.1:

$$\begin{pmatrix} \widetilde{x}^0 \\ \widetilde{x}^1 \\ \widetilde{x}^2 \\ \widetilde{x}^3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1-\beta^2}} & -\frac{\beta}{\sqrt{1-\beta^2}} & 0 & 0 \\ -\frac{\beta}{\sqrt{1-\beta^2}} & \frac{1}{\sqrt{1-\beta^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix}, \tag{2.2.7}$$

where $\beta = v/c$. Both transformations (2.2.3) and (2.2.7) are the same in form, but the mathematical implication is changed. Here, (2.2.7) represents the coordinate transformation for the Minkowski space $\mathscr{M}^4$. With Einstein's summation convention, (2.2.7) is often denoted by

$$\widetilde{x}^{\mu} = L^{\mu}_{\nu} x^{\nu}, \tag{2.2.8}$$

where

$$(L^{\mu}_{\nu}) = \begin{pmatrix} \frac{1}{\sqrt{1-\beta^2}} & -\frac{\beta}{\sqrt{1-\beta^2}} & 0 & 0 \\ -\frac{\beta}{\sqrt{1-\beta^2}} & \frac{1}{\sqrt{1-\beta^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.2.9}$$

is the Lorentz matrix, and its inverse $(l^{\mu}_{\nu}) = (L^{\mu}_{\nu})^{-1}$ is given by

$$(l^{\mu}_{\nu}) = \begin{pmatrix} \frac{1}{\sqrt{1-\beta^2}} & \frac{\beta}{\sqrt{1-\beta^2}} & 0 & 0 \\ \frac{\beta}{\sqrt{1-\beta^2}} & \frac{1}{\sqrt{1-\beta^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.2.10}$$

An important property of the Minkowski space is that its Riemannian metric is invariant under the Lorentz transformation (2.2.7).

Theorem 2.18 *The Minkowski metric (2.2.5) or (2.2.6) is invariant under the coordinate transformation (2.2.7). Namely the metric* $(\widetilde{g}_{\mu\nu})$ *in* $\{\widetilde{x}^{\mu}\}$ *is the same as that in* $\{x^{\mu}\}$:

$$(\widetilde{g}_{\mu\nu}) = (g_{\mu\nu}). \tag{2.2.11}$$

In other words, ds^2 *in* $\{\widetilde{x}^{\mu}\}$ *is also expressed as*

$$ds^2 = -c^2 d\widetilde{t}^2 + d\widetilde{x}^2, \quad \widetilde{x} = (\widetilde{x}^1, \widetilde{x}^2, \widetilde{x}^3).$$

The proof of Theorem 2.18 needs to use the properties of tensors. In fact, the Minkowski metric (2.2.5) is a second-order covariant Lorentz tensor, i.e. under the transformation (2.2.8), $g_{\mu\nu}$ transforms as

$$(\widetilde{g}_{\mu\nu}) = (l^{\alpha}_{\beta})(g_{\mu\nu})(l^{\alpha}_{\beta})^{\mathrm{T}}, \tag{2.2.12}$$

where (l^{μ}_{μ}) is the inverse of (L^{μ}_{ν}) given by (2.2.10). Then, a direct computation we can get (2.2.11) from (2.2.12).

Lorentz Transformation Group

Each transformation (2.2.8) corresponds to a Lorentz matrix (L^μ_ν) given by (2.2.9). These matrices constitute a group in the multiplication as

$$\begin{pmatrix} \frac{1}{\sqrt{1-\beta^2}} & -\frac{\beta}{\sqrt{1-\beta^2}} & 0 & 0 \\ -\frac{\beta}{\sqrt{1-\beta^2}} & \frac{1}{\sqrt{1-\beta^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{1-\gamma^2}} & -\frac{\gamma}{\sqrt{1-\gamma^2}} & 0 & 0 \\ -\frac{\gamma}{\sqrt{1-\gamma^2}} & \frac{1}{\sqrt{1-\gamma^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{\sqrt{1-\alpha^2}} & -\frac{\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 \\ -\frac{\alpha}{\sqrt{1-\alpha^2}} & \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where

$$\beta = \frac{v}{c}, \qquad \gamma = \frac{u}{c}, \qquad \alpha = \frac{w}{c}, \qquad w = \frac{u+v}{1+uv/c^2},$$

where w is the velocity composed by u and v by the theory of special relativity.

Thanks to Theorem 2.18, we can define Lorentz transformation group as all linear transformations of the Minkowski space $\mathscr{M}^4$, that preserve the Minkowski metric; see (2.2.12):

$$LG = \{(L^\mu_\nu) : \mathscr{M}^4 \to \mathscr{M}^4 \mid (g_{\mu\nu}) = (L^\alpha_\beta)(g_{\mu\nu})(L^\alpha_\beta)^{\mathrm{T}},\ \det(L^\mu_\nu) = 1\}. \tag{2.2.13}$$

Elements of LG are also called Lorentz matrices, and relativistic physics is referred to the invariance of action under the Lorentz group. The invariance of the Minkowski metric implies the Lorentz invariance of physical laws.

In this definition, we require $\det(L^\mu_\nu) = 1$, and such transformations are often called proper Lorentz transformation in physics literatures. The parity transformation and time reversal are two special linear transformations of the Minkowski space, which are not elements in LG as defined in (2.2.13) and are often dealt with separately:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \qquad T = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.2.14}$$

Lorentz Tensors

We now define Lorentz tensors, also called 4-dimensional (4-D) tensors, corresponding to the Lorentz invariance.

Definition 2.19 (Lorentz Tensors) *The following quantities are called Lorentz tensors, or 4-dimensional tensors*:

1) *A function T with 4^k components*

$$T = \{T_{\mu_1\cdots\mu_k}\}, \quad \mu_1, \cdots, \mu_k = 0,1,2,3,$$

is called a k-th order covariant tensor, if under the Lorentz transformation (2.2.13), the components of T transform as

$$\widetilde{T}_{\mu_1\cdots\mu_k} = l^{\nu_1}_{\mu_1}\cdots l^{\nu_k}_{\mu_k} T_{\nu_1\cdots\nu_k},$$

where $(l^\nu_\mu) = (L^\mu_\nu)^{-1}$ is the inverse transformation of the Lorentz transformation $(L^\mu_\nu) \in LG$.

2) *A tensor*

$$T = \{T^{\mu_1\cdots\mu_k}\}, \quad \mu_1, \cdots, \mu_k = 0,1,2,3,$$

is called a k-th order contra-variant tensor, if under the Lorentz transformation $(L^\mu_\nu) \in LG$, the components of T change as

$$\widetilde{T}^{\mu_1\cdots\mu_k} = L^{\mu_1}_{\nu_1}\cdots L^{\mu_k}_{\nu_k} T^{\nu_1\cdots\nu_k}.$$

3) *A function*

$$T = \{T^{\mu_1\cdots\mu_r}_{\nu_1\cdots\nu_s}\}, \quad r+s=k,$$

is called a k-th order (r,s)-type tensor, if under the Lorentz transformation $(L^\mu_\nu) \in LG$,

$$\widetilde{T}^{\mu_1\cdots\mu_r}_{\nu_1\cdots\nu_s} = l^{\alpha_1}_{\nu_1}\cdots l^{\alpha_s}_{\nu_s} L^{\mu_1}_{\beta_1}\cdots L^{\mu_r}_{\beta_r} T^{\beta_1\cdots\beta_r}_{\alpha_1\cdots\alpha_s}.$$

In physics, the most important tensors are first- and second-order tensors. The following is a list of commonly encountered 4-D tensors, and we always use $(g^{\mu\nu}) = (g_{\mu\nu})^{-1}$ with $(g_{\mu\nu})$ being the Minkowski metric given by (2.2.5):

1) Position vectors:

$$\begin{aligned} &x^\mu = (x^0, x^1, x^2, x^3), \qquad x^0 = ct, \\ &x_\mu = (x_0, x_1, x_2, x_3) = g_{\mu\nu}x^\nu = (-x^0, x^1, x^2, x^3). \end{aligned} \tag{2.2.15}$$

2) The 4-D electromagnetic potential:

$$\begin{aligned} &A_\mu = (A_0, A_1, A_2, A_3), \\ &A^\mu = (A^0, A^1, A^2, A^3) = g^{\mu\nu}A_\nu, \end{aligned} \tag{2.2.16}$$

where $A_0 = -A^0$ is the electric potential, $(A_1, A_2, A_3) = (A^1, A^2, A^3)$ is the magnetic vector potential.

3) The 4-D current density:

$$\begin{aligned} J_\mu &= (J_0, J_1, J_2, J_3), \\ J^\mu &= (J^0, J^1, J^2, J^3) = g^{\mu\nu} J_\nu, \end{aligned} \tag{2.2.17}$$

where $J^0 = -J_0 = c\rho$, ρ is the charge density, and $\vec{J} = (J_1, J_2, J_3)$ is the current density field.

4) The 4-D energy-momentum vector:

$$\begin{aligned} E_\mu &= (E, cP_1, cP_2, cP_3), \\ E^\mu &= g^{\mu\nu} E_\nu = (-E, cP_1, cP_2, cP_3), \end{aligned} \tag{2.2.18}$$

where E is the energy, and $P = (P_1, P_2, P_3)$ is the momentum vector,

5) The 4-D gradient operators:

$$\begin{aligned} \partial_\mu &= \left(\frac{\partial}{\partial x^0}, \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3}\right) = \left(\frac{1}{c}\frac{\partial}{\partial t}, \nabla\right), \\ \partial^\mu &= g^{\mu\nu}\nabla_\mu = \left(-\frac{\partial}{\partial x^0}, \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3}\right) = \left(-\frac{1}{c}\frac{\partial}{\partial t}, \nabla\right), \end{aligned} \tag{2.2.19}$$

transform as the first-order Lorentz tensors.

2.2.3 Relativistic invariants

All Lagrange actions in relativistic physics are Lorentz invariants. The most common Lorentz invariants are contractions of 4-D tensors. For example, let

$$A_\mu = (A_0, A_1, A_2, A_3), \quad B^\mu = (B^0, B^1, B^2, B^3)$$

be the covariant and contra-variant vectors. Then the contraction

$$A_\mu B^\mu = A_0 B^0 + A_1 B^1 + A_2 B^2 + A_3 B^3 \tag{2.2.20}$$

is a Lorentz invariant. In fact, under the Lorentz transformation (2.2.8), A_μ and B^μ satisfy

$$\widetilde{A}_\mu = L_\mu^\nu A_\nu, \quad \widetilde{B}^\mu = l_\nu^\mu B^\nu.$$

It follows that

$$\widetilde{A}_\mu \widetilde{B}^\mu = L_\mu^\alpha l_\beta^\mu A_\alpha B^\beta. \tag{2.2.21}$$

Since $(l_\mu^\nu) = (L_\mu^\nu)^{-1}$ is the inverse matrix of (L_μ^ν), then

$$L_\mu^\alpha l_\beta^\mu = \delta_\beta^\alpha.$$

Thus, (2.2.21) becomes

$$\widetilde{A}_\mu \widetilde{B}^\mu = A_\mu B^\mu,$$

which shows that the contraction (2.2.20) is a Lorentz invariant. In fact, the following theorem can be easily verified in the same fashion.

Theorem 2.20 (Lorentz Invariants)

1) *The contractions*

$$A_{\mu_1\cdots\mu_k}B^{\mu_1\cdots\mu_k}, \tag{2.2.22}$$

is a Lorentz invariant, and so are the following contractions:

$$g^{\alpha\beta}A_\alpha B_\beta, \quad g_{\alpha\beta}A^\alpha B^\beta. \tag{2.2.23}$$

2) *Let A^μ_ν be a (1,1) Lorentz tensor, then the contraction*

$$A^\mu_\mu = A^0_0 + A^1_1 + A^2_2 + A^3_3 \tag{2.2.24}$$

is a Lorentz invariant.

3) *The following differential operators*

$$\begin{aligned} A_\mu\partial^\mu &= \left(A_0\frac{\partial}{\partial x_0}, A_1\frac{\partial}{\partial x_1}, A_2\frac{\partial}{\partial x_2}, A_3\frac{\partial}{\partial x_3}\right), \\ A^\mu\partial_\mu &= \left(A^0\frac{\partial}{\partial x^0}, A^1\frac{\partial}{\partial x^1}, A^2\frac{\partial}{\partial x^2}, A^3\frac{\partial}{\partial x^3}\right), \\ \partial^\mu\partial_\mu &= -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \end{aligned} \tag{2.2.25}$$

are Lorentz invariant operators.

Theorem 2.20 provides a number of typical Lorentz invariants through contractions, and in fact, all Lagrange actions in relativistic physics are combinations of the Lorentz invariants in (2.2.22)-(2.2.25).

2.2.4 Relativistic mechanics

First, the 4-D velocity is defined by

$$u^\mu = (u^0, u^1, u^2, u^3) = \left(\frac{dx^0}{ds}, \frac{dx^1}{ds}, \frac{dx^2}{ds}, \frac{dx^3}{ds}\right),$$

where ds is the arc-length element in (2.2.6), and is given by

$$ds = c\sqrt{1 - v^2/c^2}\,dt, \quad v = (v^1, v^2, v^3), \quad v^k = \frac{dx^k}{dt}.$$

It is clear that the 4-D velocity u^μ can be expressed as

$$\begin{aligned} &u^\mu = (u^0, u^1, u^2, u^3), \\ &u^0 = \frac{1}{\sqrt{1 - v^2/c^2}}, \qquad u^k = \frac{v^k/c}{\sqrt{1 - v^2/c^2}} \qquad \text{for } 1 \leqslant k \leqslant 3. \end{aligned} \tag{2.2.26}$$

Second, the 4-D acceleration is defined by

$$
\begin{aligned}
a^\mu &= \frac{du^\mu}{ds} = (a^0, a^1, a^2, a^3), \\
a^0 &= \frac{1}{c\sqrt{1-v^2/c^2}} \frac{d}{dt}\left(\frac{1}{\sqrt{1-v^2/c^2}}\right), \\
a^k &= \frac{1}{c^2\sqrt{1-v^2/c^2}} \frac{d}{dt}\left(\frac{v^k}{\sqrt{1-v^2/c^2}}\right) \qquad \text{for } 1 \leqslant k \leqslant 3.
\end{aligned}
\tag{2.2.27}
$$

Third, the 4-D energy-momentum vector is

$$
\begin{aligned}
E^\mu &= (E, cP^1, cP^2, cP^3), \\
E &= \frac{mc^2}{\sqrt{1-v^2/c^2}}, \\
P^k &= \frac{mv^k}{\sqrt{1-v^2/c^2}}, \quad v^k = \frac{dx^k}{dt} \qquad \text{for } 1 \leqslant k \leqslant 3.
\end{aligned}
\tag{2.2.28}
$$

Fourth, by (2.2.28), we derive the most important formula in relativistic, called the Einstein energy-momentum relation:

$$E^2 = c^2P^2 + m^2c^4. \tag{2.2.29}$$

Remark 2.21 In 1905, Albert Einstein conjectured that a static object with mass m have energy E, and satisfy the relation, called the Einstein formula:

$$E = mc^2. \tag{2.2.30}$$

Thus, in a static coordinate system the energy-momentum is in the form

$$(E, cP) = (mc^2, 0) \tag{2.2.31}$$

Then, it follows from (2.2.31) that the energy-momentum (E, cP) of a moving object with velocity v is taken in the form of (2.2.28), which implies that the relation (2.2.29) holds true. Hence, the energy-momentum relation (2.2.29) is based on postulating (2.2.30), which was verified by numerous experiments. □

Fifth, in classical mechanics, the Newtonian second law takes the form

$$\frac{d}{dt}P = F, \qquad P = mv \text{ is the momentum.}$$

In the relativistic mechanics, the 4-D force is

$$F^\mu = \left(\frac{dE}{ds}, c\frac{dP}{ds}\right),$$

and the relativistic force is $F=(F^1,F^2,F^3)$:

$$F^k=\frac{1}{\sqrt{1-v^2/c^2}}\frac{d}{dt}P^k \quad \text{for} 1\leqslant k\leqslant 3,$$

where $P=(P^1,P^2,P^3)$ is as in (2.2.28). Thus, it follows that the relativistic motion law is given by

$$\frac{d}{dt}P^k=\sqrt{1-v^2/c^2}F^k, \qquad P^k=\frac{mv^k}{\sqrt{1-v^2/c^2}} \qquad \text{for } k=1,2,3. \tag{2.2.32}$$

2.2.5 Lorentz invariance of electromagnetism

The Maxwell equations for electromagnetic fields are

$$\text{curl } E=-\frac{1}{c}\frac{\partial H}{\partial t}, \tag{2.2.33}$$

$$\text{div } H=0, \tag{2.2.34}$$

$$\text{curl } H=\frac{1}{c}\frac{\partial E}{\partial t}+\frac{4\pi}{c}J, \tag{2.2.35}$$

$$\text{div } E=4\pi\rho, \tag{2.2.36}$$

where E,H are the electric and magnetic fields, and J is the current density, and ρ the charge density.

To show the Lorentz invariance of the Maxwell equations (2.2.33)-(2.2.36), we need to express them in the form of the 4-D electromagnetic potential and current density.

The electromagnetic potential and current density are briefly introduced in (2.2.16) and (2.2.17):

$$\begin{aligned} &A_\mu=(A_0,A_1,A_2,A_3),\\ &J_\mu=(J_0,J_1,J_2,J_3), \quad J_0=-c\rho. \end{aligned} \tag{2.2.37}$$

Using the Lorentz tensor operator

$$\partial_\mu=\frac{\partial}{\partial x^\mu}=\left(\frac{\partial}{\partial x^0},\frac{\partial}{\partial x^1},\frac{\partial}{\partial x^2},\frac{\partial}{\partial x^3}\right),$$

we can construct two second-order Lorentz tensors:

$$\begin{aligned} &F_{\mu\nu}=\frac{\partial A_\nu}{\partial x^\mu}-\frac{\partial A_\mu}{\partial x^\nu},\\ &G_{\mu\nu}=\frac{1}{2}\varepsilon_{\mu\nu\alpha\beta}g^{\alpha\kappa}g^{\beta\lambda}F_{\kappa\lambda}, \end{aligned} \tag{2.2.38}$$

where $g^{\alpha\beta}$ is the Minkowski metric, and

$$\varepsilon_{\mu\nu\alpha\beta}=\begin{cases} 1 & (\mu,\nu,\alpha,\beta) \text{ is an even permutation of } (0123),\\ -1 & (\mu,\nu,\alpha,\beta) \text{ is an odd permutation of } (0123),\\ 0 & \text{otherwise,} \end{cases}$$

is a 4-th order Lorentz tensor. Thanks to the relations:

$$H = \text{ curl}\,\vec{A}, \quad E = \nabla A_0 - \frac{1}{c}\frac{\partial \vec{A}}{\partial t}, \quad \vec{A} = (A_1, A_2, A_3),$$

or equivalently,

$$\begin{aligned} H_1 &= \frac{\partial A_3}{\partial x^2} - \frac{\partial A_2}{\partial x^3}, & H_2 &= \frac{\partial A_1}{\partial x^3} - \frac{\partial A_3}{\partial x^1}, & H_3 &= \frac{\partial A_2}{\partial x^1} - \frac{\partial A_1}{\partial x^2}, \\ E_1 &= \frac{\partial A_0}{\partial x^1} - \frac{\partial A_1}{\partial x^0}, & E_2 &= \frac{\partial A_0}{\partial x^2} - \frac{\partial A_2}{\partial x^0}, & E_3 &= \frac{\partial A_0}{\partial x^3} - \frac{\partial A_3}{\partial x^0}, \end{aligned}$$

the first pair of the Maxwell equations (2.2.33) and (2.2.34) are in the form:

$$\frac{\partial G_{\mu\nu}}{\partial x_\nu} = 0 \qquad \text{for } \mu = 0, 1, 2, 3. \tag{2.2.39}$$

The second pair of the Maxwell equations (2.2.35) and (2.2.36) are

$$\frac{\partial F_{\mu\nu}}{\partial x_\nu} = \frac{4\pi}{c} J_\mu \qquad \text{for } \mu = 0, 1, 2, 3. \tag{2.2.40}$$

It is clear that the Maxwell equations (2.2.39) and (2.2.40) are covariant under the Lorentz transformations. In fact, (2.2.39) and (2.2.40) can be equivalently written as

$$\partial^\nu G_{\mu\nu} = 0, \qquad \partial^\nu F_{\mu\nu} = \frac{4\pi}{c} J_\mu.$$

Also, two electromagnetic dynamic equations and the charge conservation law are written as

$$\frac{\partial \rho}{\partial t} + \text{ div}\, J = 0. \tag{2.2.41}$$

The motion equation in an electromagnetic field is given by

$$m\frac{dv}{dt} = eE + \frac{e}{c} v \times H. \tag{2.2.42}$$

It is clear that (2.2.41) can be written as

$$\partial_\mu J^\mu = 0, \tag{2.2.43}$$

which is Lorentz invariant.

The Lorentz covariance of the motion equation in an electromagnetic field follows from the following Lorentz covariant formulation of (2.2.42):

$$m\frac{du_\mu}{ds} = \frac{e}{c^2} F_{\mu\nu} u^\nu \qquad \text{for } \mu = 0, 1, 2, 3, \tag{2.2.44}$$

where u_μ is the 4-D velocity given by (2.2.26), and $F_{\mu\nu}$ is as in (2.2.38).

In summary, all electromagnetic equations can be written in the Lorentz covariant forms given by (2.2.39)-(2.2.40), (2.2.43) and (2.2.44), and the Lorentz invariance of these equations follows.

2.2.6 Relativistic quantum mechanics

Quantum physics is based on several fundamental principles, also called basic postulates of quantum mechanics, which will be introduced systematically in Chapter 6.

For our purpose, we introduce hereafter two of these basic postulates.

Basic Postulate 2.22 *An observable physical quantity can be represented by a Hermitian linear operator. In particular, the energy E, momentum $\vec{P}$, scalar-valued momentum P are represented by the operators given by*

$$E = i\hbar\frac{\partial}{\partial t}, \qquad \vec{P} = -i\hbar\nabla, \tag{2.2.45}$$

$$\begin{aligned} &P_0 = i\hbar(\vec{\sigma}\cdot\nabla) && \text{for massless fermions},\\ &P_1 = -i\hbar(\vec{\alpha}\cdot\nabla) && \text{for massive fermions}, \end{aligned} \tag{2.2.46}$$

where $\hbar$ is the Plank constant, $\vec{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$, $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$, σ_k $(1 \leqslant k \leqslant 3)$ are the Pauli matrices defined by

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{2.2.47}$$

and α_k $(1 \leqslant k \leqslant 3)$ are 4-th order matrices given by

$$\alpha_1 = \begin{pmatrix} 0 & \sigma_1 \\ \sigma_1 & 0 \end{pmatrix}, \qquad \alpha_2 = \begin{pmatrix} 0 & \sigma_2 \\ \sigma_2 & 0 \end{pmatrix}, \qquad \alpha_3 = \begin{pmatrix} 0 & \sigma_3 \\ \sigma_3 & 0 \end{pmatrix}. \tag{2.2.48}$$

Basic Postulate 2.23 *For a quantum system with observable Hermitian operators $L_1, \cdots, L_m$, if the physical quantities l_k corresponding to L_k $(1 \leqslant k \leqslant m)$ satisfy a relation*

$$H(l_1, \cdots, l_m) = 0, \tag{2.2.49}$$

then the following equation induced by (2.2.49)

$$H(L_1, \cdots, L_m)\psi = 0 \tag{2.2.50}$$

may give a model describing this system provided the operator $H(L_1, \cdots, L_m)$ in (2.2.49) *is Hermitian.*

Remark 2.24 If the operator $H(L_1, \cdots, L_m)$ in (2.2.49) is irreducible, then (2.2.49) must describe the system. □

Based on Postulates 2.22 and 2.23, we derive three basic equations of relativistic quantum mechanics: the Klein-Gordon equation describing the bosons, the Weyl equations describing massless free fermions, and the Dirac equations describing massive fermions.

1. *Klein-Gordon equation.* By the Einstein energy m omentum relation for energy E, momentum P and the rest mass m:

$$E^2 - c^2\vec{P}^2 - m^2c^4 = 0,$$

and by (2.2.45), we derive from Postulate 2.23 an relativistic equation, called the Klein-Gordon (KG) equation:

$$\left[\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2 + \left(\frac{mc}{\hbar}\right)^2\right]\psi = 0. \tag{2.2.51}$$

The KG equation describes a free boson with spin $J = 0$.

When an electromagnetic field presents, the operators in (2.2.45) and (2.2.46) are rewritten as[1]

$$\begin{aligned} &E = i\hbar\frac{\partial}{\partial t} - eA_0, \quad &&\vec{P} = -i\hbar\nabla - \frac{e}{c}\vec{A}, \\ &P_0 = i\hbar(\vec{\sigma}\cdot\vec{D}), \quad &&P_1 = -i\hbar(\vec{\alpha}\cdot\vec{D}), \end{aligned} \tag{2.2.52}$$

where $A_\mu = (A_0, \vec{A})$ is the electromagnetic potential, and

$$\vec{D} = \nabla + \frac{ie}{\hbar c}\vec{A}.$$

Thus, as electromagnetic field presents, the Klein-Gordon equation (2.2.51) is expressed as

$$\left(D_\mu D^\mu - \left(\frac{mc}{\hbar}\right)^2\right)\psi = 0, \tag{2.2.53}$$

where

$$D_\mu = \partial_\mu + i\frac{e}{\hbar c}A_\mu, \tag{2.2.54}$$

is a 4-dimensional vector operator. The expression (2.2.53) is clearly Lorentz invariant.

2. *Weyl equations.* Based on the de Broglie relation

$$E = \hbar\omega, \quad P = \hbar/\lambda, \quad c = \omega\lambda,$$

we obtain

$$E = cP. \tag{2.2.55}$$

The relation (2.2.55) is valid to a massless free fermion, e.g. as neutrinos. Inserting E and P_0 in (2.2.45) and (2.2.46) into (2.2.55), we obtain that

$$\frac{\partial\psi}{\partial t} = c(\vec{\sigma}\cdot\nabla)\psi, \tag{2.2.56}$$

which are called the Weyl equations describing the free massless neutrinos. Here $\psi = (\psi_1, \psi_2)^T$ is a two component Weyl spinor.

3. *Dirac equations.* For a massive fermion, the de Broglie matter-wave duality relation can be generalized in the form

$$E = \hbar\omega \pm mc^2, \quad P = \hbar/\lambda, \quad c = \omega\lambda.$$

[1] Here e is the electric charge of an electron, and is negative.

Then we have

$$E = cP \pm mc^2. \tag{2.2.57}$$

Inserting the operators E and P_1 in (2.2.45) and (2.2.46) into (2.2.57), and taking the mass operator as

$$\pm mc^2 = mc^2\alpha_0, \qquad \alpha_0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \qquad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

we derive the following Dirac equations:

$$i\hbar\frac{\partial\psi}{\partial t} = -i\hbar c(\vec{\alpha}\cdot\nabla)\psi + mc^2\alpha_0\psi, \tag{2.2.58}$$

where $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)^{\mathrm{T}}$ is a four-component Dirac spinor, and $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ is as defined by (2.2.48).

Usually, we multiply both sides of (2.2.58) by the matrix α_0, and denote

$$\gamma^\mu = (\gamma^0, \gamma^1, \gamma^2, \gamma^3), \tag{2.2.59}$$

where

$$\gamma^0 = \alpha_0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \qquad \gamma^k = \alpha_0\alpha_k = \begin{pmatrix} 0 & \sigma_k \\ -\sigma_k & 0 \end{pmatrix} \qquad \text{for } 1 \leqslant k \leqslant 3.$$

The matrices γ^μ are the Dirac matrices. Then the Dirac equations (2.2.58) are in the form

$$\left(i\gamma^\mu\partial_\mu - \frac{mc}{\hbar}\right)\psi = 0. \tag{2.2.60}$$

When we consider the case under an electromagnetic field, we need to replace ∂_μ in (2.2.60) by D_μ given by (2.2.54).

Remark 2.25 The reason why the scalar-valued momentum operators P are taken in the form of (2.2.46) is due to the following Einstein energy-momentum formulas:

$$\begin{aligned} &E^2 = c^2P_0^2 = c^2\vec{P}^2 && \text{for mass } m = 0, \\ &E^2 = c^2(P_1 + mc\alpha_0)^2 = c^2\vec{P}^2 + m^2c^4 && \text{for } m \neq 0, \end{aligned}$$

where P_0, P_1 are as in (2.2.46), and $\vec{P}$ is as in (2.2.45). □

2.2.7 Dirac spinors

The Dirac matrices (2.2.59) are not invariant under the Lorentz transformations, i.e. $\gamma^\mu = (\gamma^0, \gamma^1, \gamma^2, \gamma^3)$ is not a 4-D vector operator. Hence, the covariance of the Dirac equations (2.2.60) requires that under the transformation

$$\widetilde{x}_\mu = Lx_\nu, \quad L = (L_\nu^\mu) \text{ as in (2.2.9)}, \tag{2.2.61}$$

the left-hand side of (2.2.60) should be

$$\left(i\gamma^\mu\widetilde{\partial}_\mu - \frac{mc}{\hbar}\right)\widetilde{\psi} = R\left(i\gamma^\mu\partial_\mu - \frac{mc}{\hbar}\right)\psi, \tag{2.2.62}$$

where $\widetilde{\partial}_\mu = \partial/\partial\widetilde{x}^\mu$, and

$$\widetilde{\psi} = R\psi. \tag{2.2.63}$$

Here R is a 4×4 matrix depending on the Lorentz matrix L in (2.2.61). To determine the matrix R, we note that

$$\frac{\partial}{\partial\widetilde{x}^\mu} = L^\nu_\mu\frac{\partial}{\partial x^\nu}, \qquad (L^\nu_\mu) = L. \tag{2.2.64}$$

Inserting (2.2.63) and (2.2.64) into (2.2.62), we deduce that

$$\left(iR^{-1}\gamma^\mu L^\nu_\mu R\frac{\partial}{\partial x^\nu} + \frac{mc}{\hbar}\right)\psi = \left(i\gamma^\nu\frac{\partial}{\partial x^\nu} + \frac{mc}{\hbar}\right)\psi.$$

It follows that

$$R^{-1}\gamma^\mu L^\nu_\mu R = \gamma^\nu,$$

which is equivalent to

$$R^{-1}\gamma^\mu R = L^\mu_\nu\gamma^\nu \qquad \text{for } \mu = 0,1,2,3. \tag{2.2.65}$$

Hence the covariance of (2.2.62) is equivalent to the transformation matrix R in (2.2.63) obeying equations (2.2.65).

To derive an explicit form of R in (2.2.65), we need to write the Lorentz matrix L in the form

$$L^\mu_\nu = \begin{pmatrix} \cosh\theta & -\sinh\theta & 0 & 0 \\ -\sinh\theta & \cosh\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where $\cosh\theta$ and $\sinh\theta$ are the hyperbolic functions and θ satisfies

$$\cosh\theta = \frac{1}{\sqrt{1-v^2/c^2}}, \quad \sinh\theta = \frac{v/c}{\sqrt{1-v^2/c^2}}.$$

In this form, the equations (2.2.65) can be written as

$$\begin{aligned} &R^{-1}\gamma^0 R = \cosh\theta\gamma^0 - \sinh\theta\gamma^1, \\ &R^{-1}\gamma^1 R = -\sinh\theta\gamma^0 + \cosh\theta\gamma^1, \\ &R^{-1}\gamma^2 R = \gamma^2, \\ &R^{-1}\gamma^3 R = \gamma^3. \end{aligned} \tag{2.2.66}$$

By the expressions of γ^μ, we infer from (2.2.66) that

$$R = \cosh\frac{\theta}{2} I - \sinh\frac{\theta}{2}\gamma^0\gamma^1 = \begin{pmatrix} \cosh\frac{\theta}{2} & 0 & 0 & -\sinh\frac{\theta}{2} \\ 0 & \cosh\frac{\theta}{2} & -\sinh\frac{\theta}{2} & 0 \\ 0 & -\sinh\frac{\theta}{2} & \cosh\frac{\theta}{2} & 0 \\ -\sinh\frac{\theta}{2} & 0 & 0 & \cosh\frac{\theta}{2} \end{pmatrix}. \tag{2.2.67}$$

The four-component function ψ satisfying (2.2.63) under the Lorentz transformation (2.2.61) is called the Dirac spinor, which ensures the Lorentz covariance for the Dirac equations.

Noting that

$$\cosh\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta}), \qquad \sinh\theta = \frac{1}{2}(e^{i\theta} - e^{-i\theta}),$$

and $\cosh^2\theta - \sinh^2\theta = 1$, we infer from (2.2.67) that

$$R^{-1} = R^\dagger. \tag{2.2.68}$$

2.3 Einstein's Theory of General Relativity

2.3.1 Principle of general relativity

First Principle of Physics 2.1 amounts to saying that

$$\begin{aligned} &\text{Laws of Physics} &&= &&\text{Differential Equations} \\ &\text{Universality of Laws} &&= &&\text{Covariance of Equations.} \end{aligned} \tag{2.3.1}$$

In retrospect, Albert Einstein must have followed the spirit of this principle for his discovery of the general theory of relativity. As mentioned earlier, coordinate systems, also called reference systems, are just an indispensable tool to express the laws of physics in the form of differential equations. Consequently the validity of laws of physics is independent of coordinate systems. Hence Einstein proposed the following principle of general relativity.

Principle 2.26(General Relativity) *Laws of physics are the same under all coordinate systems, both inertial and non-inertial. In other words, the models describing the laws of physics are invariant under general coordinate transformations.*

In Section 2.1.5, we mentioned that each symmetry in physics is characterized by three ingredients:

space, transformation group, tensors.

The special theory of relativity or the Lorentz invariance is dictated by

Minkowski space, Lorentz group, and Lorentz tensors.

The three ingredients of the Theory of General Relativity are

- the space-time Riemannian space,
- general coordinate transformations, and
- general tensors,

under which the theory of general relativity is developed.

First, the Minkowski space is now replaced by the Riemannian space. In Theorem 2.18, we see that the Minkowski metric

$$ds^2 = -c^2dt^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \tag{2.3.2}$$

is invariant under the Lorentz transformations. However, when we consider a non-inertial reference system

$$(c\widetilde{t}, \widetilde{x}^1, \widetilde{x}^2, \widetilde{x}^3),$$

rotating with constant angular velocity Ω around the x^3-axis of an inertial system (ct, x^1, x^2, x^3), the coordinate transformation is given by

$$\begin{aligned}
x^1 &= \widetilde{x}^1 \cos\Omega\widetilde{t} - \widetilde{x}^2 \sin\Omega\widetilde{t}, \\
x^2 &= \widetilde{x}^1 \sin\Omega\widetilde{t} + \widetilde{x}^2 \cos\Omega\widetilde{t}, \\
x^3 &= \widetilde{x}^3, \\
t &= \widetilde{t}.
\end{aligned}$$

Under this transformation, the metric (2.3.2) becomes

$$\begin{aligned}
ds^2 = &- [c^2 - (\widetilde{x}^1)^2\Omega^2 - (\widetilde{x}^2)^2\Omega^2]d\widetilde{t}^2 - 2\Omega\widetilde{x}^2 d\widetilde{x}^1 d\widetilde{t} \\
&+ 2\Omega\widetilde{x}_1 d\widetilde{x}^2 d\widetilde{t} + (d\widetilde{x}^1)^2 - (d\widetilde{x}^2)^2 + (d\widetilde{x}^3)^2.
\end{aligned} \tag{2.3.3}$$

Hence in a general non-inertial system, the metric ds^2 takes the form

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \tag{2.3.4}$$

where $\{g_{\mu\nu}\}$ is a Riemannian metric different from the Minkowski metric given by (2.2.5).

In mathematics, the space $\mathscr{M}$ endowed with the metric (2.3.4), denoted by $\{\mathscr{M}, g_{\mu\nu}\}$, is called a Riemannian space, or a Riemannian manifold, and (2.3.4) or $\{g_{\mu\nu}\}$ is the Riemannian metric.

In Theorem 2.12, we know that the Minkowski space $\mathscr{M}^4$ is flat, and the Riemannian space $\{\mathscr{M}, g_{\mu\nu}\}$ is curved provided that the metric $g_{\mu\nu}$ is not the same as the Minkowski metric in any coordinate systems.

2.3.2 Principle of equivalence

In the last subsection we see that the underlying space for the general theory of relativity is the 4-dimensional Riemannian space $\{\mathscr{M}, g_{\mu\nu}\}$, instead of the Minkowski space. Now, a crucial step is that we have to make sure the physical significance of the Riemannian metric $\{g_{\mu\nu}\}$.

Consider a free particle moving in a Riemannian space $\{\mathscr{M}, g_{\mu\nu}\}$, which satisfies the motion equations

$$D\left(\frac{dx^k}{ds}\right) = 0 \qquad \text{for } k = 1,2,3, \tag{2.3.5}$$

where D is the covariant derivative, and

$$ds = \sqrt{g_{\mu\nu}dx^\mu dx^\nu}.$$

According to the theory of Riemannian geometry, the motion equations (2.3.5) in the Minkowski space are in the form

$$\frac{d^2x^k}{ds^2} = 0 \qquad \text{for } k = 1,2,3,$$

which are the motion equations of special relativity, and (2.3.5) in the Riemannian space become

$$\frac{d^2x^k}{ds^2} + \Gamma^k_{\mu\nu}\frac{dx^\mu}{ds}\frac{dx^\nu}{ds} = 0 \qquad \text{for } k = 1,2,3. \tag{2.3.6}$$

where $\Gamma^\alpha_{\mu\nu}$ is the Levi-Civita connection, given by

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2}g^{\alpha\beta}\left(\frac{\partial g_{\mu\beta}}{\partial x^\nu} + \frac{\partial g_{\nu\beta}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\alpha}\right), \tag{2.3.7}$$

and $(g^{\alpha\beta})$ is the inverse of $(g_{\alpha\beta})$.

The equations (2.3.6) are the generalized Newtonian Second Law, the first term in the left-hand side of (2.3.6) is the acceleration, and the second terms represent the force acting on the particle. In view of (2.3.7) we see that the force is caused by the curvature of space-time, i.e. by the non-flat metric

$$\partial_\mu g_{\alpha\beta} \neq 0.$$

In addition, from this fact we can also think that the gravitation results in a non-flat Riemann metric. However, in (2.3.3) we obviously see that the inertial force also lead to the non-flat metrics. Thus, we encounter a difficulty that the curved Riemann space can be caused by both gravitational and inertial forces.

Einstein proposed the principle of equivalence overcoming this difficulty.

Principle 2.27(Principle of Equivalence) *One cannot distinguish the gravitational and the inertial forces at any space-time point by experiments. In other words, any non-inertial system can be equivalently regarded as an inertial system located in a gravitational field.*

With the principle of equivalence, the Riemann metric is regarded as the effects caused only by the gravitation. Furthermore, by the classical gravitational theory, the acting force

$$F = -\nabla\varphi = -\left(\frac{\partial\varphi}{\partial x^1}, \frac{\partial\varphi}{\partial x^2}, \frac{\partial\varphi}{\partial x^3}\right), \tag{2.3.8}$$

where φ is the gravitational potential. In comparing the force (2.3.8) with (2.3.7) which contains first-order derivative terms $\partial_\alpha g_{\mu\nu}$, the connections (2.3.7) also play a role of acting forces in (2.3.6). Hence we need to regard the Riemann metric $\{g_{\mu\nu}\}$ as the gravitational potential.

Thus, the principle of general relativity requires that the underlying space-time manifold be a 4-dimensional Riemann space $\{\mathscr{M}, g_{\mu\nu}\}$, and the principle of equivalence defines the Riemann metric $g_{\mu\nu}$ as the gravitational potential.

In conclusion, by Principles 2.26 and 2.27, we derive the following crucial physical conclusion of the general theory of relativity, and the second crucial physical conclusion is the Einstein field equations for the gravitational potential:

Physical Conclusion 2.28 (General Theory of Relativity) *The physical space of our Universe is a 4-dimensional Riemannian space $\{\mathscr{M}, g_{\mu\nu}\}$, and the Riemann metric $g_{\mu\nu}$ represents the gravitational potential. The physical laws associated with gravitation are covariant under the general coordinate transformations in $\{\mathscr{M}, g_{\mu\nu}\}$.*

The principle of equivalence have received many supports by experiments. In fact, the principle is based on the fact that the inertial mass is the same as the gravitational mass. By the Newton second law,

$$F = am,$$

where m is called the inertial mass, denoted by m_I, and the gravitational law provides

$$F = \frac{Gm_1m_2}{r^2},$$

where m_1 and m_2 are called the gravitational masses, and denoted by m_g.

Theoretically both masses m_I and m_g are different physical quantities, and we cannot *a prior* claim that they are the same. It must be verified by experiments. Newton was the first man to check them, resulting

$$\frac{m_g}{m_I} = 1 + o(10^{-3}),$$

i.e., $m_g = m_I$ in the error of 10^{-3}. In 1890, Eötvös obtained the precision to 10^{-8}, and in 1964, Dicke reached at 10^{-11}.

2.3.3 General tensors and covariant derivatives

In order to obtain the second crucial conclusion of general theory of relativity, i.e., the Einstein gravitational field equations for the gravitational potential $g_{\mu\nu}$, we introduce, in this subsection, the concept of general tensors, general invariants, and covariant derivatives.

1. *General tensors.* Let $\{\mathscr{M}, g_{\mu\nu}\}$ be an n-dimensional Riemannian space, and $x = (x^1, \cdots, x^n)$ be a local coordinate system of $\mathscr{M}$. We call the following transformation

$$\widetilde{x}^k = \varphi^k(x) \qquad \text{for } 1 \leqslant k \leqslant n, \tag{2.3.9}$$

a general coordinate transformation if the functions $\varphi^k(x)$ $(1 \leqslant k \leqslant n)$ in (2.3.9) satisfy that the Jacobian

$$\left(a^i_j\right) = \left(\frac{\partial \varphi^i}{\partial x^j}\right) \qquad \text{for } x \in \mathscr{M} \tag{2.3.10}$$

is continuous and non-degenerate. The inverse of (2.3.10) is denoted by

$$(b^i_j) = (a^i_j)^{-1}. \tag{2.3.11}$$

We remark that if the matrices in (2.3.10) and (2.3.11) are not continuous, then the transformation (2.3.8) are not permitted.

Definition 2.29(General Tensors) *Let T be a set of quantities defined on the Riemann space $\{\mathscr{M}, g_{ij}\}$, and T have n^{r+s} components in each coordinate system $x \in \mathscr{M}$:*

$$T = \{T^{i_1 \cdots i_r}_{j_1 \cdots j_s}(x)\} \qquad \text{for } x \in \mathscr{M}.$$

If under the coordinate transformation (2.3.9), the components of T transform as

$$\widetilde{T}^{i_1 \cdots i_r}_{j_1 \cdots j_s} = a^{i_1}_{k_1} \cdots a^{i_r}_{k_r} b^{l_1}_{j_1} \cdots b^{l_s}_{j_s} T^{k_1 \cdots k_r}_{l_1 \cdots l_s}, \tag{2.3.12}$$

then T is called a (r,s) type of kth-order general tensor with $k = r + s$, where a^i_j and b^m_l are as in (2.3.10) and (2.3.11). The $(r,0)$ type tensors are called contra-variant tensors, and the $(0,s)$ type tensors are called covariant tensors.

On a Riemannian space $\{\mathscr{M}, g_{ij}\}$, the metric $\{g_{ij}\}$ and its inverse $\{g^{ij}\}$ are second-order covariant and contra-variant symmetric tensors, i.e. they satisfy

$$\begin{aligned} &g_{ij} = g_{ji}, && g^{ij} = g^{ji}, \\ &\widetilde{g}_{ij} = b^l_i b^k_j g_{lk}, && \widetilde{g}^{ij} = a^i_l a^j_k g^{lk}, \end{aligned} \tag{2.3.13}$$

under the transformation (2.3.9). For first-order tensors, we have

$$\begin{pmatrix} \widetilde{A}_1 \\ \vdots \\ \widetilde{A}_n \end{pmatrix} = (b^i_j)^{\mathrm{T}} \begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix}, \quad \begin{pmatrix} \widetilde{A}^1 \\ \vdots \\ \widetilde{A}^n \end{pmatrix} = (a^i_j) \begin{pmatrix} A^1 \\ \vdots \\ A^n \end{pmatrix} \tag{2.3.14}$$

and for second-order tensors, we have

$$(\widetilde{T}_{ij}) = (b^k_l)^{\mathrm{T}} (T_{ij}) (b^k_l), \quad (\widetilde{T}^{ij}) = (a^k_l)(T^{ij})(a^k_l)^{\mathrm{T}}. \tag{2.3.15}$$

2. *General Invariants.* General Invariants are derived from contractions of general tensors. We infer from (2.3.14) that

$$\begin{aligned}\widetilde{A}_k\widetilde{A}^k &= (\widetilde{A}_1,\cdots,\widetilde{A}_n)\begin{pmatrix}\widetilde{A}_1\\ \vdots\\ \widetilde{A}_n\end{pmatrix} = (A_1,\cdots,A_n)(b^i_j)(a^i_j)\begin{pmatrix}A^1\\ \vdots\\ A^n\end{pmatrix}\\ &= (A_1,\cdots,A_n)\begin{pmatrix}A^1\\ \vdots\\ A^n\end{pmatrix} = A_kA^k, \quad \text{by } (b^i_j) = (a^i_j)^{-1},\end{aligned}$$

and by (2.3.15) we have

$$\begin{aligned}\widetilde{T}_{ij}\widetilde{T}^{ij} &= \text{tr}\left[(\widetilde{T}_{ij})(\widetilde{T}^{ij})^{\text{T}}\right]\\ &= \text{tr}\left[(b^k_l)^{\text{T}}(T_{ij})(b^k_l)(a^k_l)(T^{ij})^{\text{T}}(a^k_l)^{\text{T}}\right]\\ &= \text{tr}\left[(b^k_l)^{\text{T}}(T_{ij})(T^{ij})^{\text{T}}(a^k_l)^{\text{T}}\right]\\ &= \text{tr}[(T_{ij})(T^{ij})^{\text{T}}]\\ &= T_{ij}T^{ij}.\end{aligned}$$

Here we have used the following property for matrices:

$$\text{tr}[ABA^{-1}] = \text{tr}\, B, \tag{2.3.16}$$

which can be shown by the fact that the eigenvalues $\lambda_1,\cdots,\lambda_n$ of B are the same as those of ABA^{-1}. Indeed we have

$$\text{tr}\, B = \lambda_1 + \cdots + \lambda_n = \text{tr}[ABA^{-1}].$$

Moreover, all general invariants are in contraction form:

$$A^{l_1\cdots l_s}_{k_1\cdots k_r}B^{k_1\cdots k_r}_{l_1\cdots l_s}, \tag{2.3.17}$$

where the invariance holds true as well if the indices are exchanged. For example, $A_{ij}B^{ji}$ is also invariant.

By the invariant form (2.3.17), the metric form ds^2 of a Riemannian space is a general invariant, i.e.

$$ds^2 = g_{ij}dx^idx^j = \widetilde{g}_{ij}d\widetilde{x}^id\widetilde{x}^j.$$

3. *Covariant Derivatives.* The covariance of differential equations requires that the derivative operator ∇ be also covariant, i.e. ∇ is a general tensor operator.

Let us consider the usual derivatives

$$\partial_k = \partial/\partial x_k \qquad \text{for } 1 \leqslant k \leqslant n. \tag{2.3.18}$$

For a vector field $A = (A^1, \cdots, A^k)$, in the transformation (2.3.9) we have

$$\widetilde{A}^k = a^k_l A^l. \tag{2.3.19}$$

Differentiating both sides of (2.3.19), we have

$$\frac{\partial \widetilde{A}^k}{\partial \widetilde{x}^j} = a^k_l \frac{\partial A^l}{\partial x^i}\frac{\partial x^i}{\partial \widetilde{x}^j} + \frac{\partial a^k_l}{\partial x^i}\frac{\partial x^i}{\partial \widetilde{x}^j} A^l = a^k_l b^i_j \frac{\partial A^l}{\partial x^i} + \frac{\partial a^k_l}{\partial x^i} b^i_j A^l, \tag{2.3.20}$$

where $b^i_j = \partial x^i/\partial \widetilde{x}^j$ are as in (2.3.11). In view of Definition 2.29, we infer from (2.3.20) that the usual derivative operators (2.3.18) are not tensor operators. Namely, $\left\{\dfrac{\partial A^i}{\partial x^j}\right\}$ is not a tensor, and the transformation formula (2.3.20) contains an extra term: $\dfrac{\partial a^k_l}{\partial x^i} b^i_j A^l$.

To solve this problem, as shown in (2.1.31), we need to add a term Γ to the derivative operator ∂_k, resulting a new derivative operator ∇:

$$\nabla_j A^k = \frac{\partial A^k}{\partial x^j} + \Gamma^k_{ij} A^i,$$

such that $\nabla = \{\nabla_j\}$ is a tensor operator. Namely, $\{\nabla_j A^k\}$ is a (1,1) type tensor, and transforms as

$$\widetilde{\nabla}_j \widetilde{A}^k = \frac{\partial \widetilde{A}^k}{\partial \widetilde{x}^j} + \widetilde{\Gamma}^k_{ij}\widetilde{A}^i = a^k_l b^i_j \left(\frac{\partial A^l}{\partial x^i} + \Gamma^l_{ri} A^r\right) = a^k_l b^i_j \nabla_i A^l. \tag{2.3.21}$$

By (2.3.20), it follows from (2.3.21) that

$$\widetilde{\Gamma}^k_{ij}\widetilde{A}^i = a^k_l b^i_j \Gamma^l_{ri} A^r - b^i_j \frac{\partial a^k_l}{\partial x^i} A^l. \tag{2.3.22}$$

By (2.3.22) we deduce the transformation rule for Γ as

$$\widetilde{\Gamma}^k_{ij} = a^k_l b^r_i b^s_j \Gamma^l_{rs} - b^r_i b^s_j \frac{\partial a^k_r}{\partial x^s}. \tag{2.3.23}$$

Fortunately, for a Riemannian space $\{\mathscr{M}, g_{ij}\}$, there exists a set of functions

$$\Gamma = \{\Gamma^k_{ij}\}, \tag{2.3.24}$$

called the Levi-Civita connection, which satisfies the transformation given by (2.3.23), and are given by

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl}\left(\frac{\partial g_{il}}{\partial x^j} + \frac{\partial g_{jl}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l}\right), \tag{2.3.25}$$

which we have seen in (2.3.7).

Based on the connection (2.3.24)–(2.3.25), we now define the covariant derivatives in the Riemannian space $\{\mathscr{M}, g_{ij}\}$ as follows:

$$\begin{aligned}
&\nabla_k u = \frac{\partial u}{\partial x^k} && \text{for a scalar field } u,\\
&\nabla_k u^j = \frac{\partial u^j}{\partial x^k} + \Gamma^j_{kl} u^l && \text{for a vector field } \{u^j\},\\
&\nabla_k u_j = \frac{\partial u_j}{\partial x^k} - \Gamma^l_{kj} u_l && \text{for a covector field } \{u_j\},
\end{aligned}$$

and for a (r,s) type general tensor field $\{u^{i_1\cdots i_r}_{j_1\cdots j_s}\}$,

$$\begin{aligned}
\nabla_k u^{i_1\cdots i_r}_{j_1\cdots j_s} =& \frac{\partial u^{i_1\cdots i_r}_{j_1\cdots j_s}}{\partial x^k} - \Gamma^l_{kj_1} u^{i_1\cdots i_r}_{lj_2\cdots j_s} - \cdots - \Gamma^l_{kj_s} u^{i_1\cdots i_r}_{j_1\cdots j_{s-1}l} \\
&+ \Gamma^{i_1}_{kl} u^{li_2\cdots i_r}_{j_1\cdots j_s} + \cdots + \Gamma^{i_r}_{kl} u^{i_1\cdots i_{s-1}l}_{j_1\cdots j_s}.
\end{aligned} \tag{2.3.26}$$

Consequently, physical laws obeying Principle 2.26 of Einstein general relativity must be in a form of covariant partial differential equations:

$$L(u, \nabla u, \cdots, \nabla^m u) = 0.$$

Now, what remains is to establish the field equations governing the Riemann metric $\{g_{ij}\}$ (the gravitational potential), which will be introduced in the next two subsections.

2.3.4 Einstein-Hilbert action

In view of Principle 2.3 (PLD), to derive the gravitational field equations it suffices to derive the Lagrange action for the gravitational potential $\{g_{\mu\nu}\}$.

To this end, we first introduce the Ricci curvature tensor $R_{\mu\nu}$ and the scalar curvature R in a Riemannian manifold.

1. *Ricci tensor.* It is natural to conjecture that the Lagrange density $\mathscr{L}$ for the field equations depends on the terms $\partial g_{\mu\nu}$, i.e.

$$\mathscr{L} = \mathscr{L}(g_{\mu\nu}, \cdots, \partial^m g_{\mu\nu}). \tag{2.3.27}$$

However, it is known that all covariant derivatives of the Riemann metric are zero:

$$\nabla g_{\mu\nu} = 0, \quad \nabla g^{\mu\nu} = 0. \tag{2.3.28}$$

Hence, we are not able to directly use the terms $\nabla g_{\mu\nu}, \cdots, \nabla^m g$ to construct the density (2.3.27), and have to look for invariants depending on $\partial g_{\mu\nu}, \cdots, \partial^m g_{\mu\nu}$ $(m \geqslant 1)$ in a different fashion.

The Ricci curvature tensor provides a natural way for us to find the Lagrange action. For a covector field $A = \{A_k\}$ we have

$$\nabla_\mu \nabla_\nu A_\alpha = \frac{\partial}{\partial x^\mu}(\nabla_\nu A_\alpha) - \Gamma^\beta_{\mu\nu}(\nabla_\beta A_\alpha) - \Gamma^\beta_{\mu\alpha}(\nabla_\nu A_\beta).$$

Note that

$$\nabla_\nu A_\alpha = \frac{\partial A_\alpha}{\partial x^\nu} - \Gamma^\gamma_{\nu\alpha} A_\gamma.$$

Then we can deduce that

$$[\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu] A_\alpha = R^\beta_{\alpha\mu\nu} A_\beta, \tag{2.3.29}$$

where

$$R^\beta_{\alpha\mu\nu} = \frac{\partial \Gamma^\beta_{\alpha\mu}}{\partial x^\nu} - \frac{\partial \Gamma^\beta_{\alpha\nu}}{\partial x^\mu} + \Gamma^\gamma_{\alpha\mu}\Gamma^\beta_{\gamma\nu} - \Gamma^\gamma_{\alpha\gamma}\Gamma^\beta_{\gamma\mu}. \tag{2.3.30}$$

The tensor in (2.3.30) is called the curvature tensor, and its self-contraction given by

$$R_{\mu\nu} = R^\alpha_{\mu\alpha\nu} \tag{2.3.31}$$

is the Ricci tensor.

2. *Scalar curvature.* Again by contraction of the Ricci tensor with the metric tensor, we derive an invariant, called scalar curvature:

$$R = g^{\mu\nu} R_{\mu\nu}. \tag{2.3.32}$$

3. *Lagrangian action.* In the Riemannian space $\{\mathscr{M}, g_{\mu\nu}\}$, the scalar curvature (2.3.32) is a unique invariant which contains up to second-order derivatives of $g_{\mu\nu}$. Hence it is natural to choose the scalar curvature R as the main part of the Lagrange density. Namely, $\mathscr{L}$ should be in the form

$$\mathscr{L} = R + S, \tag{2.3.33}$$

where S is the energy-momentum term of baryonic matter in the Universe. Physically, the energy-momentum density term S is taken as

$$S = \frac{8\pi G}{c^4} g^{\alpha\beta} S_{\alpha\beta}, \tag{2.3.34}$$

where G is the gravitational constant, and $S_{\alpha\beta}$ is the energy-momentum stress tensor. Therefore we obtain the Lagrange action of gravitational fields:

$$L_{EH} = \int_{\mathscr{M}} \left[R + \frac{8\pi G}{c^4} g^{\alpha\beta} S_{\alpha\beta} \right] \sqrt{-g}\, dx, \tag{2.3.35}$$

where $g = \det(g_{\alpha\beta})$, and $\sqrt{-g}\, dx$ is the volume element.

The functional (2.3.35) is called the Einstein-Hilbert action or Einstein-Hilbert functional. Historically, the functional was first introduced by David Hilbert in 1915 after he listened to the lecture on the general theory of relativity by Einstein. In fact, it is not easy to determine the expression of the variational derivative operator δL_R for the functional

$$L_R = \int_{\mathscr{M}} R\sqrt{-g}dx. \tag{2.3.36}$$

2.3.5 Einstein gravitational field equations

The gravitational field equations based on PLD are the variational equations of the Einstein-Hilbert action L_{EH} in (2.3.35):

$$\delta L_{EH} = 0. \tag{2.3.37}$$

By (2.3.35), the equation (2.3.37) can be explicitly expressed as

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu}, \tag{2.3.38}$$

where $R_{\mu\nu}$ is the Ricci tensor, R the scalar curvature, and $T_{\mu\nu}$ is the energy-momentum tensor. The equations (2.3.38) are the well-known Einstein gravitational field equations.

Remark 2.30 The terms in (2.3.38) are

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \delta L_R, \qquad L_R \text{ as in (2.3.36)},$$
$$T_{\mu\nu} = \delta L_s, \qquad L_s = \int_M g^{\mu\nu}S_{\mu\nu}\sqrt{-g}dx.$$

In Section 3.3, we shall give the derivation of δL_R. □

Einstein derived the field equations (2.3.38) in 1915 using his great physical intuition. According to the classical gravity theory, the Newton potential φ satisfies the Laplace equation

$$\Delta\varphi = 4\pi G\rho, \tag{2.3.39}$$

where ρ is the mass density. Einstein thinks the gravitational field equations of general relativity should be in the form

$$G_{\mu\nu} = \beta T_{\mu\nu}, \quad \beta \text{ is a constant}, \tag{2.3.40}$$

where $T_{\mu\nu}$ represents the energy-momentum tensor, which are provided by physical observations, and $G_{\mu\nu}$ represents gravitational potential which are the generalization of $\Delta\varphi$ in (2.3.39). Hence $G_{\mu\nu}$ contains the derivatives of $g_{\mu\nu}$ up to the second order. In Riemannian geometry, only the curvature tensors satisfy the needed properties. Thus $G_{\mu\nu}$ must be in the form

$$G_{\mu\nu} = R_{\mu\nu} + \lambda_1 g_{\mu\nu}R + \lambda_2 g_{\mu\nu}.$$

where λ_1, λ_2 are constants.

By the conservation of energy-momentum:

$$\nabla^\mu T_{\mu\nu} = 0,$$

the second-order tensor $G_{\mu\nu}$ should be divergence-free, i.e.

$$\nabla^\mu G_{\mu\nu} = 0. \tag{2.3.41}$$

By the Bianchi identity, $G_{\mu\nu}$ satisfying (2.3.41) are uniquely determined in the form up to a constant λ,

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \lambda g_{\mu\nu}, \tag{2.3.42}$$

where $\lambda g_{\mu\nu}$ are divergence-free due to (2.3.28). Furthermore, Einstein determined the constant β in (2.3.40) by comparing with (2.3.39), and β is given by

$$\beta = -\frac{8\pi G}{c^4}.$$

Thus, by (2.3.40) and (2.4.42), Einstein deduced the field equations in the general form as follows

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \lambda g_{\mu\nu} = -\frac{8\pi G}{c^4}T_{\mu\nu}, \tag{2.3.43}$$

and the constant λ is usually called the cosmological constant. Today, a large number of physical experiments manifest that $\lambda = 0$. However, the recent discovery of the acceleration of our universe leads to some physicists to think that $\lambda \neq 0$.

In Chapter 7, based on the unified field theory developed by the authors, we present a theory of dark matter and dark energy, which clearly explains the phenomena of dark matter and dark energy, and shows that the constant λ should be zero, i.e. $\lambda = 0$.

2.4 Gauge Invariance

2.4.1 $U(1)$ gauge invariance of electromagnetism

Gauge symmetry is one of fundamental invariance principles of physics, and determines the Lagrange actions of the electromagnetic, the weak and the strong interactions. In order to show the origin of gauge theory, we first introduce the $U(1)$ gauge invariance of electromagnetic fields.

Fermions under an electromagnetic field with potential A_μ obey the Dirac equations:

$$i\gamma^\mu D_\mu \psi - \frac{mc}{\hbar}\psi = 0, \tag{2.4.1}$$

$$D_\mu = (\partial_\mu + ieA_\mu), \tag{2.4.2}$$

where the electromagnetic potential A_μ satisfies the Maxwell equations (2.2.40):

$$\partial_\nu F^{\mu\nu} = \frac{4\pi}{c} J^\mu, \tag{2.4.3}$$

$$F^{\mu\nu} = g^{\mu\alpha} g^{\nu\beta} \left(\frac{\partial A_\alpha}{\partial x^\beta} - \frac{\partial A_\beta}{\partial x^\alpha} \right). \tag{2.4.4}$$

It is easy to see that the system of equations (2.4.1)-(2.4.4) is invariant under the following transformation:

$$\widetilde{\psi} = e^{i\theta}\psi, \qquad \widetilde{A}_\mu = A_\mu - \frac{1}{e}\partial_\mu \theta. \tag{2.4.5}$$

Since $e^{i\theta} \in U(1)$, (2.4.5) is called a $U(1)$-gauge transformation.

Now, we consider the gauge invariance from another point of view. Let ψ be a Dirac spinor describing a fermion:

$$\psi : \mathscr{M}^4 \to \mathscr{M}^4 \otimes_p \mathbb{C}^4.$$

Experimentally, we cannot observe the phase angles of ψ. Namely, under a phase rotation transformation

$$\widetilde{\psi} = e^{i\theta}\psi, \quad \theta = \theta(x^\mu), \tag{2.4.6}$$

we cannot distinguish the two states $\widetilde{\psi}$ and ψ experimentally. Mathematically speaking, the phenomenon amounts to saying that the Dirac equations are covariant under the transformation (2.4.6). This covariance requires that the derivative be covariant:

$$\widetilde{D}_\mu \widetilde{\psi} = e^{i\theta} D_\mu \psi. \tag{2.4.7}$$

However, we note that

$$\partial_\mu \widetilde{\psi} = \partial_\mu (e^{i\theta}\psi) = e^{i\theta}(\partial_\mu + i\partial_\mu \theta)\psi, \tag{2.4.8}$$

indicating that ∂_μ is not covariant as defined by (2.4.7). In view of (2.4.8), to obtain a covariant derivative, we have to add a term G_μ to ∂_μ. Namely, we need to define D_μ by

$$D_\mu = \partial_\mu + igG_\mu \qquad \text{with } g \text{ being a coupling constant}, \tag{2.4.9}$$

where G_μ is a 4-dimensional vector field and for (2.4.6) G_μ transforms as

$$\widetilde{G}_\mu = G_\mu - \frac{1}{g}\partial_\mu \theta. \tag{2.4.10}$$

Thus, it is readily to check that

$$\widetilde{D}_\mu \widetilde{\psi} = (\partial_\mu + ig\widetilde{G}_\mu)(e^{i\theta}\psi) = e^{i\theta}(\partial_\mu + igG_\mu)\psi = e^{i\theta} D_\mu \psi.$$

Hence, the derivative operator defined by (2.4.9) is covariant under the gauge transformation (2.4.6) and (2.4.10):

$$\widetilde{\psi} = e^{i\theta}\psi, \qquad \widetilde{G} = G_\mu - \frac{1}{g}\partial_\mu \theta. \tag{2.4.11}$$

In view of (2.4.11) and (2.4.5), the field G_μ and the coupling constant g in (2.4.9) are the electromagnetic potential A_μ and electric charge e:

$$\begin{aligned} &G_\mu = A_\mu \quad && \text{the electromagnetic potential,} \\ &g = e && \text{the electric charge.} \end{aligned}$$

The above process illustrates that the electromagnetic potential A_μ and the electric charge e are the outcomes from the $U(1)$ gauge invariance. In other words, from the phase angle symmetry of particles ($U(1)$ gauge symmetry) we deduce the following conclusions:

$$\begin{aligned} &U(1) \text{ gauge symmetry} && \Rightarrow && U(1) \text{ gauge field } A_\mu, \\ &U(1) \text{ gauge field } A_\mu && = && \text{electromagnetic potential,} \\ &U(1) \text{ gauge coupling constant } g && = && \text{electric charge.} \end{aligned} \tag{2.4.12}$$

Recall that in the Einstein general theory of relativity, the principle of general relativity not only leads to the equivalence of gravitational potential to the Riemannian metric $g_{\mu\nu}$, but also determines the field action—the Einstein-Hilbert action (2.3.35). Now, by the $U(1)$ gauge invariance, we are able to deduce both the electromagnetic potential A_μ as (2.4.12) and the $U(1)$ gauge action as follows.

In fact, in the same fashion as used in (2.3.29) for gravity, by the covariance of (2.4.7) we derive the covariant field $F_{\mu\nu}$ as

$$\left[D_\mu D_\nu - D_\nu D_\mu\right]\psi = ieF_{\mu\nu}\psi, \tag{2.4.13}$$

where $F_{\mu\nu}$ is given by (2.4.4). The left-hand side of (2.4.13) is covariant under the gauge transformation (2.4.5), and so does $F_{\mu\nu}$. Consequently the contraction

$$F = F_{\mu\nu}F^{\mu\nu}, \tag{2.4.14}$$

is invariant under both the gauge and the Lorentz transformations.

Thus, the invariant F in (2.4.14) is the main part of the action of the $U(1)$ gauge field A_μ, which is given by

$$L_{EM} = \int_{M^4}\left[-\frac{1}{4}F + \frac{4\pi}{c}A_\mu J^\mu\right]dxdt. \tag{2.4.15}$$

This is the action of the Maxwell field.

The form of (2.4.15) is similar to that of the Einstein-Hilbert action (2.3.35). In mathematics, the gauge field A_μ in (2.4.2) is the connection of the complex vector bundle $\mathscr{M}^4 \otimes_p \mathbb{C}^4$, with which the Dirac spinor ψ is defined:

$$\psi : \mathscr{M}^4 \to \mathscr{M}^4 \otimes_p \mathbb{C}^4.$$

The tensor $F_{\mu\nu}$ in (2.4.13) is the curvature tensor of the bundle $\mathscr{M}^4 \otimes_p \mathbb{C}^4$, and $F = F_{\mu\nu}F^{\mu\nu}$ in (2.4.15) is the scalar curvature. In addition, the current energy $A_\mu J^\mu$ in (2.4.15) corresponds to the energy-momentum $g^{\mu\nu}S_{\mu\nu}$ in (2.3.35).

In the same fashion as the electromagnetism, we can derive the gauge theories for both the weak and the strong interactions from the $SU(2)$ and $SU(3)$ gauge invariances. In the next subsections we shall introduce the mathematical theory of $SU(N)$ gauge fields and the principle of gauge invariance in physics.

2.4.2 Generator representations of $SU(N)$

Both the weak and strong interactions are described by the $SU(N)$ gauge theory, which is a generalization of the $U(1)$ gauge theory introduced in the last subsection.

In an $SU(N)$ gauge theory, there are N wave functions, representing N fermions:

$$\Psi = (\psi^1, \cdots, \psi^N)^{\mathrm{T}},$$

where each ψ^k $(1 \leqslant k \leqslant N)$ is a 4-component Dirac spinor. We seek for a set of gauge fields G^k_μ such that the Dirac equations

$$\left[i\gamma^\mu D_\mu - \frac{cm}{\hbar}\right]\Psi = 0 \tag{2.4.16}$$

are invariant under the $SU(N)$ gauge transformation

$$\widetilde{\Psi} = \Omega\Psi, \quad \forall\Omega = e^{i\theta^k(x)\tau_k} \in SU(N), \tag{2.4.17}$$

where m is the mass matrix, and θ^k are real parameters,

$$D_\mu = \partial_\mu + igG^k_\mu\tau_k, \tag{2.4.18}$$

g is a coupling constant, τ_k $(1 \leqslant k \leqslant N^2 - 1)$ are the generators of $SU(N)$.

We recall that $SU(N)$ is the group consisting of $N \times N$ unitary matrices with unit determinant:

$$SU(N) = \{\Omega |\ \Omega \text{ the } N \times N \text{ matrix}, \ \Omega^\dagger = \Omega^{-1}, \ \det \Omega = 1\}. \tag{2.4.19}$$

In the $SU(N)$ gauge theory for (2.4.16)-(2.4.19), we encounter a mathematical concept, generators τ_k $(1 \leqslant k \leqslant N^2 - 1)$ of $SU(N)$. We now introduce the representation of $SU(N)$.

By (2.4.19) we see that each matrix $\Omega \in SU(N)$ satisfies

$$\Omega^\dagger = \Omega^{-1}, \quad \det \Omega = 1, \quad \Omega \in SU(N), \tag{2.4.20}$$

where $\Omega^\dagger = (\Omega^T)^*$ is the complex conjugate of the transport of Ω. Note that an exponent e^{iA} of a matrix iA is an $N \times N$ complex matrix satisfying

$$(e^{iA})^\dagger = e^{-iA^\dagger}.$$

It follows that

$$(e^{iA})^\dagger(e^{iA}) = e^{-iA^\dagger}e^{iA} = e^{i(A-A^\dagger)}.$$

Hence we see that

$$(e^{iA})^\dagger = (e^{iA})^{-1} \text{ if and only if } A = A^\dagger \text{ is Hermitian.} \tag{2.4.21}$$

In addition, we have

$$\det e^A = e^{\mathrm{tr}A}, \tag{2.4.22}$$

which will be proved at the end of this subsection.

Based on (2.4.20)-(2.4.22), it is not difficult to understand the representation of $SU(N)$ in a neighborhood of the unit matrix, stated in the following theorem.

Theorem 2.31 ($SU(N)$ Representation) *The matrices Ω in a neighborhood $U \subset SU(N)$ of the unit matrix can be expressed as*

$$\Omega = e^{iA}, \quad A^\dagger = A, \quad \mathrm{tr}A = 0. \tag{2.4.23}$$

Furthermore, the tangent space $T_eSU(N)$ of $SU(N)$ at the unit matrix $e = I$ is a K-dimensional linear space ($K = N^2 - 1$), generated by K linear independent traceless Hermitian matrices τ_k $(1 \leqslant k \leqslant K)$:

$$T_eSU(N) = \mathrm{span}\{\tau_1, \cdots, \tau_K\} \qquad \text{with } \tau_k^\dagger = \tau_k, \quad \mathrm{tr}\ \tau_k = 0. \tag{2.4.24}$$

In particular, the matrices Ω in (2.4.23) can be written as

$$\Omega = e^{i\theta^k\tau_k}, \quad \tau_k\ (1 \leqslant k \leqslant K) \text{ as in (2.4.24)}, \tag{2.4.25}$$

where θ^k $(1 \leqslant k \leqslant K)$ are real numbers, and

$$[\tau_k, \tau_l] = \tau_k\tau_l - \tau_l\tau_k = i\lambda_{kl}^j\tau_j, \tag{2.4.26}$$

where λ_{kl}^j are called the structure constants of $SU(N)$.

We remark here that the basis $\{\tau_k \mid 1 \leqslant k \leqslant K\}$ of the tangent space (2.4.24) is usually called the generators of $SU(N)$, and (2.4.25) is called the generator representation of $SU(N)$.

We are now in position to give a proof of (2.4.22). By the classical Jordan theorem, for a given matrix A there is a non-degenerate matrix B such that

$$BAB^{-1} = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}, \tag{2.4.27}$$

is an upper triangular matrix, where $\lambda_1, \cdots, \lambda_N$ are the eigenvalues of A. Recall that e^A is defined by

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

By (2.4.27) we have

$$Be^A B^{-1} = \sum_{k=0}^{\infty} \frac{1}{k!}(BAB^{-1})^k = \sum_{k=0}^{\infty} \frac{1}{k!} \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}^k = \begin{pmatrix} e^{\lambda_1} & & * \\ & \ddots & \\ 0 & & e^{\lambda_N} \end{pmatrix}.$$

Hence we get

$$\det e^A = \det (Be^A B^{-1}) = e^{\lambda_1 + \cdots + \lambda_N}. \tag{2.4.28}$$

It is known that

$$\operatorname{tr} A = \lambda_1 + \cdots + \lambda_N.$$

Thus, the formula (2.4.22) follows from (2.4.28).

2.4.3 Yang-Mills action of $SU(N)$ gauge fields

An $SU(N)$ gauge theory mainly deals with the invariance of the Dirac equations for N spinors under $SU(N)$ gauge transformations. The physical meaning of $SU(N)$ gauge invariance is that in a system of N fermions we cannot distinguish one particle from others by the weak or strong interaction.

Based on the $SU(N)$ representation Theorem 2.31, we now introduce the $SU(N)$ gauge theory.

Consider N Dirac spinors ψ^k $(1 \leqslant k \leqslant N)$ and $K = (N^2 - 1)$ Lorentz vector fields G^a_μ $(1 \leqslant a \leqslant K)$, called the gauge fields, given by

$$\Psi = (\psi^1, \cdots, \psi^N)^{\mathrm{T}}, \quad G_\mu = (G^1_\mu, \cdots, G^K_\mu). \tag{2.4.29}$$

The Dirac equations for the fermions are:

$$\left[i\gamma^\mu D_\mu - \frac{mc}{\hbar}\right]\Psi = 0, \tag{2.4.30}$$

where

$$m = \begin{pmatrix} m_1 & & 0 \\ & \ddots & \\ 0 & & m_N \end{pmatrix} \quad \text{is the the mass matrix}, \tag{2.4.31}$$

$$D_\mu = \partial_\mu + igG^a_\mu \tau_a, \tag{2.4.32}$$

and τ_a $(1 \leqslant a \leqslant K)$ are a set of $SU(N)$ generators as in Theorem 2.31.

For the N spinors Ψ in (2.4.29), consider the following $SU(N)$ gauge transformations

$$\widetilde{\Psi} = \Omega\Phi \qquad \forall \Omega = e^{i\theta^a \tau_a} \in SU(N), \tag{2.4.33}$$

where θ^a $(1 \leqslant a \leqslant K)$ are functions of $x^\mu \in \mathscr{M}^4$.

To ensure that (2.4.30)-(2.4.32) are covariant under the transformation (2.4.33), we need to determine

(1) the transformation form of the gauge fields G_μ, and

(2) the action for the gauge fields G_μ.

First, consider (1). The covariance of Dirac equations (2.4.30) is equivalent to the covariance of the derivative $D_\mu \Psi$ as given by

$$\widetilde{D}_\mu \widetilde{\Psi} = \Omega D_\mu \Psi \qquad \forall \Omega \text{ as in } (2.4.33). \tag{2.4.34}$$

The left-hand side of (2.4.34) can be directly computed as

$$\widetilde{D}_\mu \widetilde{\Psi} = (\partial_\mu + ig\widetilde{G}^a_\mu \tau_a)\Omega\Psi = \Omega\partial_\mu\Psi + (\partial_\mu\Omega)\Psi + ig\widetilde{G}^a_\mu \tau_a \Omega\Psi,$$

and the right-hand side is

$$\Omega D_\mu \Psi = \Omega(\partial_\mu + igG^a_\mu \tau_a)\Psi = \Omega\partial_\mu\Psi + igG^a_\mu \Omega\tau_a\Psi.$$

Hence, it follows from (2.4.34) that

$$\widetilde{G}^a \tau_a = G^a_\mu \Omega \tau_a \Omega^{-1} + \frac{i}{g}(\partial_\mu \Omega)\Omega^{-1}, \tag{2.4.35}$$

which is called the $SU(N)$ gauge transformation of the gauge fields G^a_μ.

In addition, the mass matrix (2.4.31) satisfies that

$$\widetilde{m} = \Omega m \Omega^{-1}. \tag{2.4.36}$$

In summary, the transformations (2.4.33), (2.4.35) and (2.4.36) constitute the $SU(N)$ gauge transformations defined as

$$\begin{aligned} &\widetilde{\Psi} = \Omega\Psi \qquad\qquad \forall \Omega = e^{i\theta^a \tau_a} \in SU(N), \\ &\widetilde{G}^a_\mu \tau_a = G^a_\mu \Omega \tau_a \Omega^{-1} + \frac{i}{g}(\partial_\mu \Omega)\Omega^{-1}, \\ &\widetilde{m} = \Omega m \Omega^{-1}. \end{aligned} \tag{2.4.37}$$

Usually, (2.4.37) is taken as an infinitesimal transformation as follows

$$\Omega = \mathrm{I} + i\theta^a \tau_a, \quad \theta^a \text{ are infinitesimal.}$$

In this case, (2.4.37) can be written as

$$\begin{aligned} &\widetilde{\Psi} = \Psi + i\theta^a \tau_a \Psi, \\ &\widetilde{G}^a_\mu = G^a_\mu - \lambda^a_{bc}\theta^b G^c_\mu - \frac{1}{g}\partial_\mu \theta^a, \\ &\widetilde{m}^a = m + i\theta^a(\tau_a m - m\tau_a), \end{aligned} \tag{2.4.38}$$

where λ^a_{bc} are the structure constants as defined in (2.4.26).

We have seen that the Dirac equations (2.4.30)-(2.4.32) are covariant under the $SU(N)$ gauge transformations (2.4.37) or (2.4.38).

What remains to do is to find out the action of the gauge fields G^a_μ, which is invariant under both the Lorentz transformation and the $SU(N)$ gauge transformation (2.4.37).

Recall that in (2.4.13), the $U(1)$ gauge field action (2.4.15) is derived by using the commutator of the covariant derivative operator:

$$[D_\mu, D_\nu] = D_\mu D_\nu - D_\nu D_\mu.$$

We now derive in the same fashion the $SU(N)$ action, called the Yang-Mills action.

By (2.4.34),

$$\widetilde{D}_\mu(\widetilde{D}_\nu\widetilde{\Psi}) = \widetilde{D}_\mu(\Omega D_\nu\Psi) = \Omega(D_\mu D_\nu\Psi).$$

It follows that

$$\left[\widetilde{D}_\mu, \widetilde{D}_\nu\right]\widetilde{\Psi} = \Omega\left[D_\mu, D_\nu\right]\Psi. \tag{2.4.39}$$

On the other hand, by (2.4.32), we have

$$\frac{i}{g}\left[D_\mu, D_\nu\right] = \frac{i}{g}(\partial_\mu + igG^a_\mu\tau_a)(\partial_\nu + igG^a_\nu\tau_a) - \frac{i}{g}(\partial_\nu + igG^a_\nu\tau_a)(\partial_\mu + igG^a_\mu\tau_a). \tag{2.4.40}$$

Notice that

$$\partial_\mu\partial_\nu = \partial_\nu\partial_\mu, \quad \partial_\nu(G^a_\mu\tau_a) = \partial_\nu G^a_\beta\tau_a + G^a_\beta\tau_a\partial_\nu.$$

Then (2.4.40) becomes

$$\begin{aligned}\frac{i}{g}\left[D_\mu, D_\nu\right] =&(\partial_\mu G^a_\nu - \partial_\nu G^a_\mu)\tau_a - ig\left[G^a_\mu\tau_a, G^b_\nu\tau_b\right] \\ =&(\partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g\lambda^a_{bc}G^b_\mu G^c_\nu)\tau_a \quad (\text{by } (2.4.26))\end{aligned} \tag{2.4.41}$$

Therefore we define

$$F^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g\lambda^a_{bc}G^b_\mu G^c_\nu. \tag{2.4.42}$$

Then, thanks to (2.4.41), we deduce that

$$\begin{aligned}\frac{i}{g}\left[\widetilde{D}_\mu, \widetilde{D}_\nu\right]\widetilde{\Psi} =&\widetilde{F}^a_{\mu\nu}\tau_a\Omega\Psi \\ =&\frac{i}{g}\Omega\left[D_\mu, D_\nu\right]\Psi \quad (\text{by } (2.4.39)) \\ =&F^a_{\mu\nu}\Omega\tau_a\Psi,\end{aligned}$$

which implies that

$$\widetilde{F}^a_{\mu\nu}\tau_a = F^a_{\mu\nu}\Omega\tau_a\Omega^{-1}. \tag{2.4.43}$$

As μ, ν are the indices of 4-D tensors, a Lorentz invariant can be constructed using the following contraction

$$\widetilde{F}^a_{\mu\nu}\tau_a\widetilde{F}^{\mu\nu b}\tau_b^\dagger = \widetilde{F}^a_{\mu\nu}\widetilde{F}^{\mu\nu b}\tau_a\tau_b^\dagger, \tag{2.4.44}$$

where $F^{\mu\nu b} = g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}$. By (2.4.43) we have

$$\begin{aligned}\widetilde{F}^a_{\mu\nu}\tau_a\widetilde{F}^{\mu\nu b}\tau_b^\dagger =& F^a_{\mu\nu}(\Omega\tau_a\Omega^{-1})F^{\mu\nu b}(\Omega\tau_b\Omega^{-1})^\dagger \\ =& F^a_{\mu\nu}F^{\mu\nu b}\Omega\tau_a\tau_b^\dagger\Omega^{-1} \quad (\text{by } \Omega^\dagger = \Omega^{-1}).\end{aligned} \tag{2.4.45}$$

Therefore we deduce from (2.4.44) and (2.4.45) that

$$\widetilde{F}^a_{\mu\nu}\widetilde{F}^{\mu\nu b}\tau_a\tau_b^\dagger = F^a_{\mu\nu}F^{\mu\nu b}\Omega\tau_a\tau_b^\dagger\Omega^{-1}. \tag{2.4.46}$$

In (2.3.16) we have verified that

$$\text{tr}\, A = \text{tr}\, (BAB^{-1}), \quad \forall \text{ matrices } A \text{ and } B.$$

Hence it follows from (2.4.46) that

$$\mathscr{G}_{ab}\widetilde{F}^a_{\mu\nu}\widetilde{F}^{\mu\nu b} = \mathscr{G}_{ab}F^a_{\mu\nu}F^{\mu\nu b}, \tag{2.4.47}$$

where $\mathscr{G}_{ab} = \frac{1}{2}\text{tr}(\tau_a\tau_b^\dagger)$.

The equality (2.4.47) shows that

$$F = \mathscr{G}_{ab}F^a_{\mu\nu}F^{\mu\nu b} = \mathscr{G}_{ab}g^{\mu\alpha}g^{\nu\beta}F^a_{\mu\nu}F^b_{\alpha\beta} \tag{2.4.48}$$

is invariant under both the Lorentz transformations and the $SU(N)$ gauge transformations, where $F^a_{\mu\nu}$ are given by (2.4.42).

The function (2.4.48) is a unique form which is both Lorentz and the $SU(N)$ gauge invariant, and contains up to first-order derivatives of the gauge fields G^a_μ. Hence, F in (2.4.48) is a unique candidate to be the Lagrange density.

In Section 3.5, we have shown that $\{\mathscr{G}_{ab}\}$ in (2.4.48) is a Riemannian metric of $SU(N)$, and

$$\mathscr{G}_{ab} = \frac{1}{2}\text{tr}(\tau_a\tau_b^\dagger) = \frac{1}{4N}\lambda^c_{ad}\lambda^d_{cb}, \tag{2.4.49}$$

where λ^c_{ab} are the structure constants of $SU(N)$.

In the classical $SU(N)$ gauge theory, the $SU(N)$ generator τ_k $(1 \leqslant k \leqslant K)$ are taken to be Hermitian and traceless, and satisfy

$$\frac{1}{2}\text{tr}\,(\tau_a\tau_b^\dagger) = \delta_{ab}.$$

In this case, the Lagrange density F in (2.4.48) becomes

$$F = F^a_{\mu\nu}F^{\mu\nu a}.$$

The Lagrange action of an $SU(N)$ gauge theory is usually taken in the following form, called the Yang-Mills action:

$$L_{YM} = \int_{\mathscr{M}^4} \left[-\frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} + \overline{\Psi} \left(i\gamma^\mu D_\mu - \frac{cm}{\hbar} \right) \Psi \right] dx, \tag{2.4.50}$$

where $\overline{\Psi} = \Psi^\dagger \gamma^0$,

$$\begin{aligned} F^a_{\mu\nu} &= \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g\lambda^a_{bc} G^b_\mu G^c_\nu, \\ D_\mu &= \partial_\mu + igG^a_\mu \tau_a. \end{aligned} \tag{2.4.51}$$

The second term on the right-hand side of (2.4.50) is the action for the Dirac equations (2.4.30)-(2.4.32).

2.4.4 Principle of gauge invariance

In Sections 2.4.2-2.4.3, we introduced the mathematical framework of $SU(N)$ gauge fields, leading to the following principle of gauge invariance.

Principle 2.32 (Gauge Invariance) *The electromagnetic, the weak, and the strong interactions obey gauge invariance. Namely, the motion equations involved in the three interactions are gauge covariant and the actions of the interaction fields are gauge invariant.*

A few remarks are now in order.

Remark 2.33 The Standard Model in particle physics is currently a prevailing theory describing all, except the gravity, fundamental interactions. It consists of the Glashow-Weinberg-Salam (GWS) electroweak theory, the transition theory of weak interaction decay, the quark model, and the Quantum Chromodynamics (QCD). Based on the Standard Model, the strong interaction is described by an $SU(3)$ gauge theory, and the electromagnetic and weak interactions are unified in an action of $U(1) \times SU(2)$ gauge fields, combing with the Higgs mechanism and the Yukawa coupling. □

Remark 2.34 All up-to-date experiments illustrate that the electromagnetic, weak, strong (EWS) interactions obey Principle of Gauge Invariance 2.32. □

Remark 2.35 All current theories about interactions, concluding the Standard Model and the String Theory, have a remarkable drawback: they cannot provide a set of acceptable field equations for the weak and strong interactions from which we can deduce the weak and strong interaction potentials. It is known that, the variational equations of the Yang-Mills action (2.4.50)-(2.4.51) are in the form

$$\begin{aligned} &\partial^\mu F^a_{\mu\nu} - g\lambda^a_{bc} g^{\alpha\beta} F^b_{\alpha\nu} G^c_\beta - gJ^a_\nu = 0, \\ &\left(i\gamma^\mu D_\mu - \frac{cm}{\hbar} \right) \Psi = 0, \end{aligned} \tag{2.4.52}$$

where

$$J^a_\nu = \overline{\Psi} \gamma_\nu \tau_a \Psi, \qquad \gamma_\nu = g_{\nu\beta} \gamma^\beta.$$

However, it appears that we are not able to derive from (2.4.52) any weak and strong interaction potentials in agreement with experiments. An important reason is that the $SU(N)$ gauge theory has N^2-1 fields:

$$G^1_\mu,\cdots,G^k_\mu \qquad \text{for } K=N^2-1, \tag{2.4.53}$$

and we don't know which or what combination of the potentials in (2.4.53) gives rise to an interaction force formula. □

Remark 2.36 The authors have developed a new unified field theory for the four fundamental interactions recently , based only on the following fundamental principles: the Einstein Principle of General Relativity, the Principle of Lorentz Invariance, the Principle of Gauge Invariance, and the two newly developed principles by the authors: the Principle of Interaction Dynamics (PID) and the Principle of Representation Invariance (PRI). The unified field theory will be presented in detail in Chapter 4, and we shall see that the theory provides sound explanations and resolutions to the following problems:

1) dark matter and the dark energy,

2) spontaneous symmetry breaking based on first principles,

3) quark confinement,

4) asymptotically freedom,

5) strong interaction potential of nucleons, and

6) layered formula for the weak and strong interaction potentials.

2.5 Principle of Lagrangian Dynamics (PLD)

2.5.1 Introduction

PLD has been briefly introduced in Subsection 2.1.3, and we now address PLD in detail.

In classical mechanics, a physical motion system can be described by three dynamical principles: the Newtonian Dynamics, the Lagrangian Dynamics, and the Hamiltonian Dynamics. Both the Lagrangian dynamics and the Hamiltonian dynamics remain valid in other physical fields such as the electrodynamics and quantum physics.

The three dynamical principles are equivalent in describing the motion of an N-body system. Consider an N-body system of planets, with masses $m_1,\cdots,m_N$ and coordinates $x_k=(x^1_k,x^2_k,x^3_k)$.

1. *Newtonian dynamics.* The motion equations governing the N planets are the Newtonian second law:

$$m_k\frac{d^2x_k}{dt^2}=F_k \qquad \text{for } 1\leqslant k\leqslant N, \tag{2.5.1}$$

where F_k is the gravitational force acting on the k-th planet by the other planets, which can be expressed as

$$F_k = -\sum_{j\neq k} \frac{m_k m_j G}{|x_j - x_k|^3}(x_j - x_k), \tag{2.5.2}$$

where G is the gravitational constant.

2. *Lagrangian dynamics.* Based on the least action principle, the Lagrange density $\mathscr{L}$ of the N-body system is

$$\mathscr{L} = T - V,$$

where T is the total kinetic energy, and V is the potential energy:

$$T = \frac{1}{2}\sum_{k=1}^{N} m_k \left(\frac{dx_k}{dt}\right)^2, \quad V = -\sum_{i,j=1,i\neq j} \frac{m_i m_j G}{2|x_i - x_j|}, \tag{2.5.3}$$

Hence, the Lagrange action is

$$L = \int_{t_0}^{t_1} \mathscr{L} dt = \frac{1}{2}\int_{t_0}^{t_1} \left[\sum_{k=1}^{N} m_k \dot{x}_k^2 + \sum_{i\neq j} \frac{m_i m_j G}{|x_i - x_j|}\right] dt. \tag{2.5.4}$$

The variational derivative operator δL of L is

$$\delta L = -m_k \ddot{x}_k - \sum_{j\neq k} \frac{m_k m_j G}{|x_j - x_k|^3}(x_j - x_k), \tag{2.5.5}$$

and the motion equations of the Lagrangian dynamics are given by

$$\delta L = 0. \tag{2.5.6}$$

It is clear that the motion equations (2.5.6) derived from the Lagrangian dynamics is the same as the equations (2.5.1)-(2.5.2) of the Newtonian dynamics.

3. *Hamiltonian dynamics.* In the next section we shall introduce the principle of Hamiltonian dynamics (PHD), and we derive here the Hamiltonian system for the N-body motion.

The total energy H of the N-body system is given by

$$H(x,y) = \sum_{k=1}^{N} \frac{1}{2m_k} y_k^2 + V(x), \tag{2.5.7}$$

where y_k is the momentum of the k-th planet, and $V(x)$ is the potential energy as in (2.5.3). Then the motion equations derived from the PHD are as follows

$$\begin{aligned} \frac{c\partial x_k}{\partial t} &= \frac{\partial H}{\partial y_k}, \\ \frac{\partial y_k}{\partial t} &= -\frac{\partial H}{\partial x_k}. \end{aligned} \tag{2.5.8}$$

where $H(x,y)$ is defined by (2.5.7). It is easy to check that the equations (2.5.8) are equivalent to those in (2.5.1)-(2.5.2).

2.5.2 Elastic waves

In an elastic continuous medium, the wave vibration is described by the PLD.

Let $\Omega \subset \mathbb{R}^3$ be a domain of elastic continuous medium, and the function $u(x,t)$ represent the displacement of the medium at time t and $x \in \Omega$. Then the total kinetic energy is

$$T = \int_\Omega \frac{1}{2}\rho \left|\frac{\partial u}{\partial t}\right|^2 dx, \tag{2.5.9}$$

where ρ is the density of the medium.

For an elastic material, the deformation potential energy V is taken in the general form

$$V = \int_\Omega \left[\frac{1}{2}k|\nabla u|^2 + F(x,u)\right] dx, \tag{2.5.10}$$

where k is a constant.

The Lagrange action for the elastic wave is written as

$$L = \int_{t_0}^{t_1}(T-V)dt = \int_{t_0}^{t_1}\int_\Omega \left[\frac{1}{2}\rho\dot{u}^2 - \frac{1}{2}k|\nabla u|^2 - F(x,u)\right] dxdt. \tag{2.5.11}$$

By the PLD, the wave equation is derived by

$$\delta L = 0,$$

and it follows from (2.5.11) that

$$\rho\frac{\partial^2 u}{\partial t^2} - k\Delta u = f(x,u) \qquad \text{for } x \in \Omega. \tag{2.5.12}$$

where

$$f(x,z) = \frac{\partial F(x,z)}{\partial z}.$$

The equation (2.5.12) is the usual wave equation describing elastic vibration in a continuous medium.

2.5.3 Classical electrodynamics

Electrodynamics consists of two parts: the Maxwell field equations and the motion of charged particles, each described by the related Lagrange actions.

Lagrange density of electromagnetic fields

In classical Maxwell theory, we usually take the electric field E, the magnetic field H, the current density $\vec{J}$ and the electric charge density ρ as state functions, because these physical quantities are observable. However, the fields E and H are not fundamental physical quantities, and the basic fields describing electromagnetism are the 4-D electromagnetic potential A_μ and the current density J_μ:

$$\begin{aligned} A_\mu &= (A_0, A_1, A_2, A_3), \\ J_\mu &= (J_0, J_1, J_2, J_3). \end{aligned} \tag{2.5.13}$$

Hence the Lagrange density for the Maxwell theory should be constructed with the fields in (2.5.13).

Based on the Lorentz Invariance and the $U(1)$ Gauge Invariance, the action of (2.5.13) is uniquely determined and is given in the form (2.4.15):

$$L = \int_0^T \int_\Omega \mathscr{L}(A_\mu, J_\mu) dxdt, \tag{2.5.14}$$

where

$$\mathscr{L} = \frac{1}{16\pi} F_{\mu\nu} F^{\mu\nu} + \frac{1}{c} A_\mu J^\mu \qquad F_{\mu\nu} \text{ as in } (2.4.4). \tag{2.5.15}$$

We now derive the variational equation, also called the Euler-Lagrange equation, of (2.5.14)-(2.5.15). Since J_μ is an applied external field, we only take variation with respect to the field A_μ.

It is known that δL is a 4-dimensional field

$$\delta L = (\delta L^0, \delta L^1, \delta L^2, \delta L^3),$$

and satisfies that for any $\widetilde{A}_\mu$ with $\widetilde{A}_\mu|_{\partial Q_T} = 0$, we have

$$\int_{Q_T} (\delta L)^\mu \widetilde{A}_\mu dxdt = \frac{d}{d\lambda}\Big|_{\lambda=0} L(A_\mu + \lambda \widetilde{A}_\mu), \tag{2.5.16}$$

where $Q_T = \Omega \times (0,T)$ and λ is a real parameter. By (2.5.14) we see that

$$\begin{aligned} &L = L_1 + L_2, \\ &L_1 = \int_{Q_T} \frac{1}{16\pi} g^{\mu a} g^{\nu\beta} F_{\mu\nu} F_{\alpha\beta} dxdt, \qquad F_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu, \\ &L_2 = \int_{Q_T} \frac{1}{c} A_\mu J^\mu dxdt. \end{aligned}$$

It is clear that

$$\frac{d}{d\lambda}\Big|_{\lambda=0} L_2(A_\mu + \lambda \widetilde{A}_\mu) = \frac{1}{c} \int_{Q_T} \frac{d}{d\lambda} (A_\mu + \lambda \widetilde{A}_\mu) J^\mu dxdt = \frac{1}{c} \int_{Q_T} J^\mu \widetilde{A}_\mu dxdt.$$

We infer from (2.5.16) that

$$\delta L_2^\mu = \frac{1}{c} J^\mu. \tag{2.5.17}$$

Noting that $g^{\mu\nu} = g^{\nu\mu}$, we have

$$\begin{aligned}\frac{d}{d\lambda}\Big|_{\lambda=0} L_1(A_\mu + \lambda\widetilde{A}_\mu) =& \frac{1}{8\pi}\int_{Q_T} g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}\frac{d}{d\lambda}\Big|_{\lambda=0}(F_{\mu\nu}+\lambda\widetilde{F}_{\mu\nu})dxdt \\ =& \frac{1}{8\pi}\int_{Q_T} g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}\left(\frac{\partial\widetilde{A}_\mu}{\partial x^\nu} - \frac{\partial\widetilde{A}_\nu}{\partial x^\mu}\right)dxdt\end{aligned}$$

By the Gauss formula,

$$\begin{aligned}\int_{Q_T} g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}\frac{\partial\widetilde{A}_\mu}{\partial x^\nu}dxdt =& -\int_{Q_T} g^{\mu\alpha}g^{\nu\beta}\frac{\partial F_{\alpha\beta}}{\partial x^\nu}\widetilde{A}_\mu dxdt, \\ \int_{Q_T} g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}\frac{\partial\widetilde{A}_\nu}{\partial x^\mu}dxdt =& -\int_{Q_T} g^{\mu\alpha}g^{\nu\beta}\frac{\partial F_{\alpha\beta}}{\partial x^\mu}\widetilde{A}_\nu dxdt \\ =& (\text{by the permutation of } \mu \text{ and } \nu) \\ =& -\int_{Q_T} g^{\nu\alpha}g^{\mu\beta}\frac{\partial F_{\alpha\beta}}{\partial x^\nu}\widetilde{A}_\mu dxdt \\ =& \int_{Q_T} g^{\mu\alpha}g^{\nu\beta}\frac{\partial F_{\alpha\beta}}{\partial x^\nu}\widetilde{A}_\mu dxdt,\end{aligned}$$

where a permutation on α and β is performed, and $F_{\beta\alpha} = -F_{\alpha\beta}$.

Thus, we obtain that

$$\frac{d}{d\lambda}\Big|_{\lambda=0} L_1(A_\mu + \lambda\widetilde{A}_\mu) = -\frac{1}{4\pi}\int_{Q_T} g^{\mu\alpha}g^{\nu\beta}\frac{\partial F_{\alpha\beta}}{\partial x^\nu}\widetilde{A}_\mu dxdt.$$

We infer then from (2.5.16) that

$$\delta L_1 = -\frac{1}{4\pi}g^{\mu\alpha}g^{\nu\beta}\frac{\partial F_{\alpha\beta}}{\partial x^\nu} = -\frac{1}{4\pi}\frac{\partial F^{\mu\nu}}{\partial x^\nu}. \tag{2.5.18}$$

Hence it follows from (2.5.17) and (2.5.18) that

$$\delta L = -\frac{1}{4\pi}\frac{\partial F^{\mu\nu}}{\partial x^\nu} + \frac{1}{c}J^\mu,$$

which implies that the equation $\delta L = 0$ takes the form:

$$\frac{\partial F^{\mu\nu}}{\partial x^\nu} = \frac{4\pi}{c}J^\mu. \tag{2.5.19}$$

This is the second pair of the Maxwell equations (2.2.35) and (2.2.36). Since the first pair of the Maxwell equations (2.2.33) and (2.2.34) are direct consequence of (2.4.4) and

$$H = \text{curl}\,\vec{A}, \quad E = \nabla A_0 - \frac{1}{c}\frac{\partial\vec{A}}{\partial t}, \quad \vec{A} = (A_1, A_2, A_3), \tag{2.5.20}$$

the action (2.5.14)-(2.5.15) completely determines the Maxwell equations (2.2.33)-(2.2.36).

Lagrangian for dynamics of charged particles

A particle with an electric charge e moving in an external electromagnetic field (E,H) is governed by

$$\frac{dp}{dt} = eE + \frac{e}{c} v \times H, \tag{2.5.21}$$

where p is the momentum of the particle, v is the velocity, and $f = \frac{e}{c} v \times H$ is the Lorentz force.

The action for the motion equation (2.5.21) is taken as

$$L = \int_0^T -mcds + \frac{e}{c} A_\mu dx^\mu. \tag{2.5.22}$$

It is clear that (2.5.22) is Lorentz invariant. By

$$ds = \sqrt{1 - v^2/c^2}cdt, \quad dx^\mu = (1, v^1, v^2, v^3)cdt,$$

the action (2.5.22) can be written as

$$\begin{aligned} &L = \int_0^T \mathscr{L}(v, A_\mu)dt, \\ &\mathscr{L} = -mc^2\sqrt{1 - v^2/c^2} + \frac{e}{c} A_k v^k + eA_0. \end{aligned} \tag{2.5.23}$$

All physical properties of electromagnetic kinematics can be derived from the action (2.5.23).

We shall deduce (2.5.21) from (2.5.23). The Euler-Lagrange equation of (2.5.23) is given by

$$\frac{d}{dt}\left(\frac{\partial \mathscr{L}}{\partial v}\right) = \nabla \mathscr{L}, \qquad \nabla \mathscr{L} = \frac{\partial \mathscr{L}}{\partial x} \quad \text{for } x \in \mathbb{R}^3. \tag{2.5.24}$$

By (2.5.23), we have

$$\frac{\partial \mathscr{L}}{\partial v} = \frac{mv}{\sqrt{1 - v^2/c^2}} + \frac{e}{c}\vec{A}.$$

Then (2.5.24) becomes

$$\begin{aligned} &\frac{dP}{dt} + \frac{e}{c}\frac{d\vec{A}}{dt} = \nabla \mathscr{L}, \\ &P = \frac{mv}{\sqrt{1 - v^2/c^2}} \quad \text{is the momentum.} \end{aligned} \tag{2.5.25}$$

By (2.5.23), we have

$$\nabla \mathscr{L} = \frac{e}{c}\nabla\vec{A} \cdot v + e\nabla A_0.$$

Thanks to

$$\begin{aligned}\nabla\vec{A}\cdot v =&\nabla(\vec{A}\cdot v) \quad (\text{by } \partial v = 0)\\ =&(\vec{A}\cdot\nabla)v + (v\cdot\nabla)\vec{A} + v\times \text{curl}\,\vec{A} + \vec{A}\text{curl}\, v\\ =&(v\cdot\nabla)\vec{A} + v\times\text{curl}\,\vec{A},\end{aligned}$$

the equation (2.5.25) reads

$$\frac{dP}{dt} + \frac{e}{c}\frac{d\vec{A}}{dt} = \frac{e}{c}(v\cdot\nabla)\vec{A} + \frac{e}{c}v\times\text{curl}\,\vec{A} + e\nabla A_0. \tag{2.5.26}$$

It is known that

$$\frac{d\vec{A}}{dt} = \frac{\partial\vec{A}}{\partial t} + \frac{\partial\vec{A}}{\partial x^k}\frac{dx^k}{dt} = \frac{\partial\vec{A}}{\partial t} + (v\cdot\nabla)\vec{A}.$$

Hence (2.5.26) is rewritten as

$$\frac{dP}{dt} = -\frac{e}{c}\frac{\partial\vec{A}}{\partial t} + e\nabla A_0 + \frac{e}{c}v\times\text{curl}\,\vec{A}. \tag{2.5.27}$$

Then, by (2.5.20), the equation (2.5.27) takes the form

$$\frac{dP}{dt} = eE + \frac{e}{c}v\times\text{curl}\,\vec{A}.$$

Thus, we have deduced (2.5.21) from the action (2.5.23).

Now, we deduce the Einstein energy-momentum formula for the 4-dimensional energy-momentum under an electromagnetic field. Corresponding to the least action principle of classical mechanics, the momentum P and energy E of a charged particle are given by

$$P = \frac{\partial\mathscr{L}}{\partial v}, \quad E = v\cdot\frac{\partial\mathscr{L}}{\partial v} - \mathscr{L}. \tag{2.5.28}$$

By (2.5.23), we refer from (2.5.28) that

$$\begin{aligned}P &= \frac{mv}{\sqrt{1-v^2/c^2}} + \frac{e}{c}\vec{A}, \quad \vec{A} \text{ the magnetic potential},\\ E &= \frac{mc^2}{\sqrt{1-v^2/c^2}} + eA_0, \quad A_0 \text{ the electric potential}.\end{aligned} \tag{2.5.29}$$

It follows from (2.5.29) that

$$(E - eA_0)^2 = c^2\left(\vec{P} - \frac{e}{c}\vec{A}\right)^2 + m^2c^4, \tag{2.5.30}$$

which is the Einstein energy-momentum relation under with an electromagnetic field.

It is the formula (2.5.30) that makes us to take the energy and momentum operators in quantum mechanics in the following form:

$$i\hbar\frac{\partial}{\partial t} - eA_0, \quad -i\hbar\nabla - \frac{c}{e}\vec{A}, \tag{2.5.31}$$

which were also given in (2.2.52). In particular, the particular form of (2.5.31) leads to the origin of gauge theory.

2.5.4 Lagrangian actions in quantum mechanics

PLD is also valid in quantum physics. In this subsection we shall introduce the actions for basic equations of quantum mechanics: the Schrödinger equation, the Klein-Gordon equation, and the Dirac equations.

Action for the Schrödinger equation

A particle moving at lower velocity can be approximatively described by the Schrödinger equation, given by

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi, \tag{2.5.32}$$

where m is the mass of the particle, $V(x)$ is the potential energy, and $\psi = \psi_1 + i\psi_2$ is a complex valued wave function. By the Basic Postulates 2.22-2.23, the equation (2.5.32) is derived using the following the non-relativistic energy momentum relation:

$$E = \frac{1}{2m}\vec{P}^2 + V.$$

Hence the Schrödinger equation (2.5.32) is a basic equation in non-relativistic quantum mechanics.

The Lagrange action for the Schrödinger equation is

$$\begin{aligned} L &= \int_0^T \int_{\mathbb{R}^3} \mathscr{L}(\psi,\psi^*)dxdt, \\ \mathscr{L} &= i\hbar\psi^*\frac{\partial\psi}{\partial t} - \frac{1}{2}\left[\frac{\hbar^2}{2m}|\nabla\psi|^2 + V|\psi|^2\right] \end{aligned} \tag{2.5.33}$$

We now compute δL to show that (2.5.33) is indeed the action of (2.5.32). Take the variation for ψ^* for (2.5.33):

$$\begin{aligned} \int_{Q_T}(\delta L)\widetilde{\psi}^* dxdt =& \frac{d}{d\lambda}\Big|_{\lambda=0} L(\psi,\psi^* + \lambda\widetilde{\psi}^*) \\ =& \int_{Q_T}\frac{d}{d\lambda}\Big|_{\lambda=0}\mathscr{L}(\psi,\psi^* + \lambda\widetilde{\psi}^*)dxdt \\ =& \int_{Q_T}\left[i\hbar\frac{\partial\psi}{\partial t}\widetilde{\psi}^* - \frac{\hbar^2}{2m}\nabla\psi\nabla\widetilde{\psi}^* + V\psi\widetilde{\psi}^*\right]dxdt, \end{aligned}$$

where $Q_T = \mathbb{R}^3\times(0,T)$, and $\widetilde{\psi}^*$ satisfies

$$\begin{aligned} &\widetilde{\psi}^*(0,x) = \widetilde{\psi}^*(T,x) = 0 && \forall x\in\mathbb{R}^3, \\ &\widetilde{\psi}^* \to 0 && \text{as } |x|\to\infty. \end{aligned}$$

Then by the Gauss formula, we have

$$\int_{Q_T}\nabla\psi\cdot\nabla\widetilde{\psi}^* dxdt = -\int_{Q_T}\Delta\psi\widetilde{\psi}^* dxdt,$$

which implies that

$$\int_{Q_T} \delta L \widetilde{\psi}^* dxdt = \int_{Q_T} \left[i\hbar \frac{\partial \psi}{\partial t} + \frac{\hbar^2}{2m} \Delta \psi - V\psi \right] \widetilde{\psi}^* dxdt.$$

Since $\widetilde{\psi}^*$ is arbitrary, we derive that

$$\delta L = i\hbar \frac{\partial \psi}{\partial t} + \frac{\hbar^2}{2m} \Delta \psi - V\psi.$$

Hence we have

$$\frac{\delta L}{\delta \psi^*} = 0 \qquad \Leftrightarrow \qquad i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \Delta \psi + V(x)\psi. \tag{2.5.34}$$

We can derive in the same fashion that

$$\frac{\delta L}{\delta \psi} = 0 \qquad \Leftrightarrow \qquad -i\hbar \frac{\partial \psi^*}{\partial t} = -\frac{\hbar^2}{2m} \Delta \psi^* + V(x)\psi^*. \tag{2.5.35}$$

It follows from (2.5.34) and (2.5.35) that

$$\frac{\delta L}{\delta \psi^*} = \left(\frac{\delta L}{\delta \psi} \right)^*.$$

In other words, (2.5.34) and (2.5.35) are equivalent, and are exactly the Schrödinger equation.

Action for Klein-Gordon fields

The field equations governing the spin-0 bosons are the Klein-Gordon equations:

$$\frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} - \Delta \psi + \left(\frac{mc}{\hbar} \right)^2 \psi = 0. \tag{2.5.36}$$

It is easy to introduce the Lagrange action for (2.5.36):

$$L = \int_{\mathscr{M}^4} \left[\nabla_\mu \psi \nabla^\mu \psi^* + \left(\frac{mc}{\hbar} \right)^2 |\psi|^2 \right] \sqrt{-g} dx, \tag{2.5.37}$$

where ∇_μ and ∇^μ are 4-dimensional gradient operators as defined by (2.2.19), $\mathscr{M}^4$ is the Minkowski space, and $g = \det(g_{\mu\nu})$.

It is readily to see that

$$\frac{\delta L}{\delta \psi^*} = \left(\frac{\delta L}{\delta \psi} \right)^* = 0 \qquad \Leftrightarrow \qquad \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} - \Delta \psi + \left(\frac{mc}{\hbar} \right)^2 \psi = 0.$$

Hence, the Klein-Gordon equation is a variational equation of the action (2.5.37).

Action for Dirac spinor fields

To introduce an action for the Dirac equations introduced in Subsections 2.2.5 and 2.2.6, we first recall the Dirac equations:

$$i\gamma^\mu \frac{\partial \psi}{\partial x^\mu} - \frac{mc}{\hbar}\psi = 0, \tag{2.5.38}$$

where $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)^T$ is the Dirac spinor, and $\gamma^\mu = (\gamma^0, \gamma^1, \gamma^2, \gamma^3)$ is the Dirac matrices, defined by

$$\gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \gamma^k = \begin{pmatrix} 0 & \sigma_k \\ -\sigma_k & 0 \end{pmatrix} \quad \text{for } k = 1,2,3, \tag{2.5.39}$$

where the Pauli matrices are given by

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The Lagrange action of (2.5.38) is

$$L = \int_{\mathcal{M}^4} \overline{\psi}(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar})\psi\sqrt{-g}dx, \tag{2.5.40}$$

where $\overline{\psi} = \psi^\dagger \gamma^0$.

The Lagrange density of (2.5.40) is

$$\mathscr{L} = \overline{\psi}\left(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar}\right)\psi$$

is Lorentz invariant. To see this, recall the spinor transformation defined by (2.2.63):

$$\widetilde{x}_\mu = L^\mu_\nu x^\nu \;\Rightarrow\; \widetilde{\psi} = R\psi,$$

which satisfies that (2.2.62):

$$(i\gamma^\mu \widetilde{\partial}_\mu - \frac{mc}{\hbar})\widetilde{\psi} = R(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar})\psi, \tag{2.5.41}$$

where $\widetilde{\partial}_\mu = \partial/\partial\widetilde{x}^\mu, R$ is as in (2.2.67). By (2.2.68) and

$$R^\dagger \gamma^0 = \gamma^0 R^\dagger, \quad (\gamma^0)^\dagger = \gamma^0. \tag{2.5.42}$$

It follows from (2.5.41) and (2.4.2) that

$$\begin{aligned}\overline{\widetilde{\psi}}(i\gamma^\mu \widetilde{\partial}_\mu - \frac{mc}{\hbar})\widetilde{\psi} =& \psi^\dagger R^\dagger \gamma^0 R\left(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar}\right)\psi \\ =& (\text{by } \psi^\dagger\gamma^0 = \overline{\psi}) \\ =& \overline{\psi}(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar})\psi.\end{aligned}$$

It implies that the Lagrange action (2.5.40) is Lorentz invariant.

It is easy to see that

$$\frac{\delta L}{\delta \psi^*} = 0 \qquad \Leftrightarrow \qquad (i\gamma^\mu \partial_\mu - \frac{mc}{\hbar})\psi = 0.$$

2.5.5 Symmetries and conservation laws

The importance of PLD lies in the following three points:

1) Physics are established based on a few universal principles, which provides a solid foundation for physics;

2) Based on PLD, many physical problems become simpler. In particular, by means of invariance it is easier to find the Lagrange actions than to seek for the differential equations; and

3) Lagrange actions contain more physical information than the differential equations. In fact, a conservation law of a physical system can be derived from the invariance of the Lagrange action under the associated symmetric transformation.

The correspondence between symmetries and conservation laws are revealed by the Noether theorem, to be introduced below. For this purpose, we need to introduce some related concepts on group action and symmetry.

We begin with a simple example. A circle with radius r is described by

$$x^2+y^2=r^2. \tag{2.5.43}$$

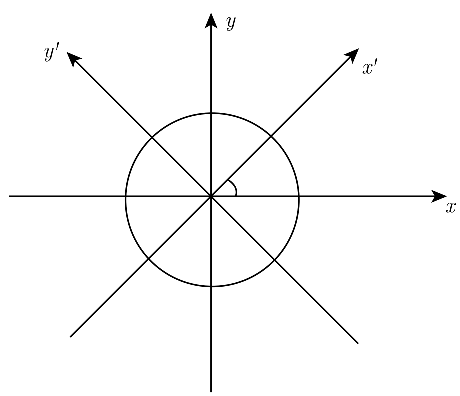


Figure 2.2

We can clearly see the symmetry of the circle shown in Figure 2.2—the graph is the same from whatever the direction we look at it. This phenomenon is expressed in mathematics as the invariance of the equation (2.5.43) under the following coordinate transformation

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{2.5.44}$$

Namely, in the coordinate system (x',y'), the equation describing the circle is invariant and still takes form:

$$x'^2 + y'^2 = r^2. \tag{2.5.45}$$

Now, we discuss the symmetry from the viewpoint of action functional and group action. The key feature is then characterized by the Noether theorem. Let $F : \mathbb{R}^2 \to \mathbb{R}^1$ be a function (a finite dimensional functional) defined by

$$F(u) = x^2 + y^2 \qquad \text{for } u = (x,y) \in \mathbb{R}^2. \tag{2.5.46}$$

All the transformation matrices in (2.5.44) constitute a group, denoted by $SO(2)$, called the orthogonal group:

$$SO(2) = \left\{ A_\theta \,\middle|\, A_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix},\ \theta \in \mathbb{R}^1 \right\}. \tag{2.5.47}$$

The symmetry here is the invariance of the function (2.5.46) under the $SO(2)$ group action:

$$F(A_\theta u) = F(u) \qquad \forall A_\theta \in SO(2).$$

The generalization to the Lagrangian dynamics is what we shall introduce in this subsection.

Definition 2.37 *Let G be a group, X be a Banach space, and $F : X \to \mathbb{R}^1$ a continuous functional. Let G be a group acting on X:*

$$\begin{aligned} &Au \in X && \forall u \in X \text{ and } A \in G, \\ &A(Bu) = (AB)u && \forall u \in X,\ A,B \in G. \end{aligned} \tag{2.5.48}$$

The functional F is called invariant under the action (2.5.48) of G, if

$$F(Au) = F(u) \qquad \forall A \in G.$$

We consider a Lagrange action defined on a function space X:

$$L(u,\dot{u}) = \int_0^T \mathscr{L}(u,\dot{u})dt \qquad \text{for } u \in X, \tag{2.5.49}$$

where the Lagrange density $\mathscr{L}$ is defined by

$$\mathscr{L}(u,\dot{u}) = \begin{cases} \mathscr{L}(x,\dot{x}) & \text{for an N-body system,} \\ \displaystyle\int_\Omega g(u,\dot{u},Du,\cdots,D^m u)dx & \text{otherwise.} \end{cases}$$

The following theorem is the well-known Noether theorem, which provides a correspondence between symmetries and conservation laws in the Lagrangian system (2.5.49).

Theorem 2.38 (Noether Theorem) *Let $G = \{A_\lambda \mid \lambda \in \mathbb{R}^1\}$ be a parameterized group, and $L(u,\dot{u})$ be the Lagrange action given by (2.5.49). If L is invariant under the group action of G:*

$$\mathscr{L}(u,\dot{u}) = \mathscr{L}(A_\lambda u, A_\lambda \dot{u}), \qquad \forall A_\lambda \in G, \tag{2.5.50}$$

then the system has a conserved quantity induced by G, expressed as

$$I(u,\dot{u}) = \left\langle \frac{\delta L}{\delta \dot{u}}, \frac{dA_\lambda(u)}{d\lambda}\Big|_{\lambda=0} \right\rangle \qquad \text{for } A_\lambda \in G. \tag{2.5.51}$$

In other words,

$$\frac{d}{dt} I(u,\dot{u}) = 0 \qquad \text{for any solutions } u \text{ of } \delta L = 0. \tag{2.5.52}$$

Remark 2.39 The correspondence between symmetries and conservation laws in the Noether Theorem holds true as well for discrete transformation groups, such as the reflections of time and space. □

Proof of Theorem 2.38 Let (u,u_t) be a solution of the variational equation of L given by (2.5.49):

$$\frac{d}{dt}\left(\frac{\delta \mathscr{L}(u,\dot{u})}{\delta \dot{u}}\right) = \frac{\delta}{\delta u}\mathscr{L}(u,\dot{u}). \tag{2.5.53}$$

By the invariance (2.5.50), $(A_\lambda u, A_\lambda \dot{u})$ are also solutions of (2.5.53). Then it follows from (2.5.50) that

$$0 = \frac{\partial \mathscr{L}(A_\lambda u, A_\lambda \dot{u})}{\partial \lambda} = \left\langle \frac{\delta \mathscr{L}}{\partial \Phi}, \frac{d\Phi}{d\lambda} \right\rangle + \left\langle \frac{\delta \mathscr{L}}{\delta \dot{\Phi}}, \frac{d\dot{\Phi}}{d\lambda} \right\rangle, \tag{2.5.54}$$

where $\Phi = A_\lambda u$. Since $(A_\lambda u, A_\lambda \dot{u})$ satisfies (2.5.53), we have

$$\frac{\delta}{\delta \Phi}\mathscr{L}(A_\lambda u, A_\lambda \dot{u}) = \frac{d}{dt}\left(\frac{\delta \mathscr{L}(A_\lambda u, A_\lambda \dot{u})}{\delta \dot{\Phi}}\right). \tag{2.5.55}$$

Inserting (2.5.55) in (2.5.54) we deduce that for any $\Phi = A_\lambda u$ and any $\lambda \in \mathbb{R}^1$,

$$0 = \left\langle \frac{d}{dt}\left(\frac{\delta \mathscr{L}}{\delta \dot{\Phi}}\right), \frac{d\Phi}{d\lambda} \right\rangle + \left\langle \frac{\partial \mathscr{L}}{\partial \dot{\Phi}}, \frac{d}{dt}\left(\frac{d\Phi}{d\lambda}\right) \right\rangle = \frac{d}{dt}\left\langle \frac{\delta \mathscr{L}}{\delta \dot{\Phi}}, \frac{d\Phi}{d\lambda} \right\rangle,$$

which implies that

$$\frac{d}{dt} I(u,\dot{u}) = \frac{d}{dt}\left\langle \frac{\delta \mathscr{L}}{\delta \dot{\Phi}}, \frac{d\Phi}{d\lambda} \right\rangle\Big|_{\lambda=0} = 0,$$

where $A_\lambda u = u$ if $\lambda = 0$. The proof is complete. □

We now give two examples to show how to apply the Noether theorem to a specific physical problem.

Example 2.40 Consider an N-body motion, such as a system of N planets. Let m_k and x_k be the mass and the coordinates of the k-th body. The Lagrange density is

$$\mathscr{L} = \frac{1}{2}\sum_{k=1}^{N} m_k |\dot{x}_k|^2 - \sum_{i \neq j} V(|x_i - x_j|). \tag{2.5.56}$$

Note that x is the u in Theorem 2.38. Let G be the translation group:

$$G = \{A_\lambda \,|\, A_\lambda : \mathbb{R}^3 \to \mathbb{R}^3,\ A_\lambda x = x + \lambda \vec{r}\},$$

where $\vec{r}$ is a given vector. It clear that (2.5.56) is invariant under the transformation of G. By

$$\frac{\delta \mathscr{L}}{\delta \dot{x}} = \left(\frac{\partial \mathscr{L}}{\partial \dot{x}_1}, \cdots, \frac{\partial \mathscr{L}}{\partial \dot{x}_N}\right) = (m_1 \dot{x}_1, \cdots, m_N \dot{x}_N),$$

$$\frac{d}{d\lambda}\Big|_{\lambda=0} A_\lambda x = \frac{d}{d\lambda}\Big|_{\lambda=0} (x + \lambda \vec{r}) = (\underbrace{\vec{r}, \cdots, \vec{r}}_{N}).$$

Hence we derive from the Noether Theorem that

$$I = \left\langle \frac{\delta \mathscr{L}}{\delta \dot{x}}, \frac{d}{d\lambda}\Big|_{\lambda=0} A_\lambda x \right\rangle = \sum_{k=1}^{N} m_k \langle \dot{x}_k, \vec{r} \rangle$$

is conserved, and is the total momentum in the direction $\vec{r}$. Thus we have shown that the translation invariance corresponds to momentum conservation. □

Example 2.41 Consider the Schrödinger equation, the action is given by (2.5.33), and

$$\mathscr{L}(\psi, \dot{\psi}) = \int_{\mathbb{R}^3} \left[-i\hbar \psi^* \dot{\psi} + \frac{1}{2} \left(\frac{\hbar^2}{2m} |\nabla \psi|^2 + V|\psi|^2 \right) \right] dx. \tag{2.5.57}$$

Let G be the phase rotation group

$$G = \{A_\lambda = e^{i\lambda} \,|\, \lambda \in \mathbb{R}^1\}.$$

It is clear that (2.5.57) is invariant under the phase rotation:

$$\psi \to A_\lambda \psi = e^{i\lambda} \psi.$$

We see that

$$\frac{\delta \mathscr{L}}{\delta \dot{\psi}} = -i\hbar \psi^*, \quad \frac{d}{d\lambda} A_\lambda \psi|_{\lambda=0} = i\psi.$$

Thus, the quantity (2.5.51) reads as

$$I(\psi, \dot{\psi}) = \left\langle \frac{\delta \mathscr{L}}{\delta \dot{\psi}}, \frac{d}{d\lambda} A_\lambda \psi \right\rangle = \int_{\mathbb{R}^3} \hbar |\psi|^2 dx.$$

Hence, the modulus of ψ

$$\int_{\mathbb{R}^3} |\psi|^2 dx \text{ is conserved.} \tag{2.5.58}$$

The property (2.5.58) is just what we need because in quantum mechanics, we have

$$\int_{\mathbb{R}^3} |\psi|^2 dx = 1. \tag{2.5.59}$$

Here, the conservation (2.5.59) corresponds to the invariance of phase rotation of the wave function ψ. □

Remark 2.42 In classical mechanics, the energy conservation can be deduced from the time translation invariance

$$t \to t + \lambda \qquad \text{for } \lambda \in \mathbb{R}^1. \tag{2.5.60}$$

However, instead of the formula (2.5.51), it is derived using following relation:

$$dH = \frac{\partial H}{\partial p} dp + \frac{\partial H}{\partial q} dq - \frac{\partial \mathscr{L}}{\partial t} dt, \tag{2.5.61}$$

where $H = H(q, p, t)$ is the total energy, $\mathscr{L}$ is the Lagrangian, and q, p satisfy the Hamilton equations

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q}. \tag{2.5.62}$$

If $\mathscr{L} = \mathscr{L}(q, \dot{q}, t)$ is invariant under the translation (2.5.60), then $\mathscr{L}$ does not explicitly contain time t, i.e.

$$\frac{\partial \mathscr{L}}{\partial t} = 0.$$

Then it follows from (2.5.61) and (2.5.62) that

$$\frac{dH}{dt} = -\frac{\partial \mathscr{L}}{\partial t} = 0.$$

Hence the energy conservation is deduced using the time translation invariance. □

We end this section with some relations of symmetries and conservation laws:

energy	⇔	time translation,
momentum	⇔	space translation,
angular Momentum	⇔	space rotation,
particle number	⇔	phase rotation,
parity	⇔	space reflection.

2.6 Principle of Hamiltonian Dynamics (PHD)

2.6.1 Hamiltonian systems in classical mechanics

In classical mechanics, the principle of Hamiltonian dynamics (PHD) consists of the following three main ingredients:

1) for an isolated (conserved) mechanical system, its states are described by a set of state variables given by

$$q_1, \cdots, q_N, \quad p_1, \cdots, p_N; \tag{2.6.1}$$

2) its total energy H is a function of (2.6.1), i.e.

$$H = H(q,p), \tag{2.6.2}$$

3) the state variables q and p satisfy

$$\begin{aligned} \frac{dq_k}{dt} &= \frac{\partial H}{\partial p_k}, \\ \frac{dp_k}{dt} &= -\frac{\partial H}{\partial q_k}, \end{aligned} \qquad \text{for } 1 \leqslant k \leqslant N. \tag{2.6.3}$$

The state variables q_k $(1 \leqslant k \leqslant N)$ represent positions, and p_k $(1 \leqslant k \leqslant N)$ represent momentums. The system (2.6.1)-(2.6.3) is called Hamiltonian system.

In physics, PLD and PHD are two independent fundamental principles. However, the two sets of equations derived from PLD and PHD are usually equivalent.

In classical mechanics, Hamiltonian systems and Lagrange systems can be transformed to each other by the Legendre transformation. We start with a simple example. Consider a particle with mass m in a force field F. The Lagrange action for this system is given by

$$\begin{aligned} L &= \int_0^T \mathscr{L}(q,\dot{q})dt, \\ \mathscr{L} &= T - V = \frac{1}{2}m\dot{q}^2 - Fq, \end{aligned} \tag{2.6.4}$$

where q stands for position, and the Euler-Lagrange equation of (2.6.4) is the Newtonian Second Law, written as

$$\frac{d}{dt}\frac{\delta L}{\delta \dot{q}} = \frac{\delta L}{\delta q} \qquad \Rightarrow \qquad m\ddot{q} = F. \tag{2.6.5}$$

Based on (2.6.1)-(2.6.3), the state variables q,p of PHD are

$$q \text{ as in (2.6.4)} \qquad \text{and} \qquad p = m\dot{q}. \tag{2.6.6}$$

The total energy is

$$H = \frac{1}{2m}p^2 + Fq, \tag{2.6.7}$$

and the Hamilton equations of (2.6.7) are given by

$$\begin{aligned} \frac{dp}{dt} &= \frac{\partial H}{\partial q} = F, \\ \frac{dq}{dt} &= -\frac{\partial H}{\partial p} = \frac{1}{m}p. \end{aligned} \tag{2.6.8}$$

By (2.6.6), it is clear that

$$\text{Hamilton Eqs } (2.6.8) = \text{Lagrange Eq } (2.6.5) = \text{Newton 2nd Law}. \tag{2.6.9}$$

The equivalences in (2.6.9) only manifest that the three principles have an intrinsic relation. In particular, by (2.6.4)-(2.6.7), the variables $q,\dot{q}$ of PLD and the variables q,p of PHD have the relation:

$$p=\frac{\delta L}{\delta\dot{q}} \qquad (\text{by } \frac{\delta L}{\delta\dot{q}}=\frac{\partial\mathscr{L}}{\partial\dot{q}}=m\dot{q}), \tag{2.6.10}$$

and the Lagrange density $\mathscr{L}(q,\dot{q})$ and the Hamilton energy $H(q,p)$ are related by

$$p\dot{q}-\mathscr{L}(q,\dot{q})=H(q,p). \tag{2.6.11}$$

In this example, the two relations (2.6.10) and (2.6.11) are obtained directly by their physical meanings, and by which we can deduce one by another.

In fact, the relations (2.6.10) and (2.6.11) are also valid in general.We discuss this problem as follows.

Consider a mechanical system. For PLD, the state variables are positions q_k and velocities $\dot{q}_k$:

$$q_1,\cdots,q_N \qquad \text{and} \qquad \dot{q}_1,\cdots,\dot{q}_N, \tag{2.6.12}$$

The Lagrange action is given by

$$L=\int_0^T\mathscr{L}(q,\dot{q})dt. \tag{2.6.13}$$

For PHD, the state variables are

$$q_1,\cdots,q_N \qquad \text{and} \qquad p_1,\cdots,p_N, \tag{2.6.14}$$

The Hamilton energy is

$$H=H(q,p). \tag{2.6.15}$$

The two systems (2.6.12)-(2.6.13) and (2.6.14)-(2.6.15) of PLD and PHD satisfy the following relations, which implies the equivalence of PLD and PHD in classical mechanics.

Dynamical Relation 2.43 (PLD and PHD) *For the two systems of PLD and PHD in classical mechanics, the following conclusions hold true:*

1) *The two sets (2.6.12) and (2.6.14) of variables satisfy the following relation:*

$$p_k=\frac{\partial\mathscr{L}(q,\dot{q})}{\partial\dot{q}_k} \qquad \text{for } 1\leqslant k\leqslant N. \tag{2.6.16}$$

2) *The two functions (2.6.13) and (2.6.15) satisfy the following relation, which is usually called the Legendre transformation:*

$$p_k\dot{q}_k-\mathscr{L}(q,\dot{q})=H(q,p). \tag{2.6.17}$$

By (2.6.16)-(2.6.17), if we have obtained the PLD system (2.6.12)-(2.6.13), then by the implicit function theorem we can solve from (2.6.16) the functions

$$\dot{q}_k = f_k(q,p). \tag{2.6.18}$$

Then, inserting (2.6.18) in the left-hand side of (2.6.17) we deduce the expression of the Hamilton energy:

$$H(q,p) = p_k f_k(q,p) - \mathscr{L}(q, f(q,p)), \tag{2.6.19}$$

which gives rise to the Hamiltonian system (2.6.3). In other words, we can derive the Hamiltonian dynamics from the Lagrangian dynamics by the relations (2.6.16) and (2.6.17).

Conversely, if we know the PHD system (2.6.14) and (2.6.15), then it follows from (2.6.16) and (2.6.17) that the Lagrange density $\mathscr{L}$ satisfies

$$H\left(x, \frac{\partial \mathscr{L}}{\partial y}\right) - y_k \frac{\partial \mathscr{L}}{\partial y_k} + \mathscr{L}(x,y) = 0. \tag{2.6.20}$$

Theoretically we can solve the differential equation (2.6.20), and obtain the solution $\mathscr{L} = \mathscr{L}(x,y)$. Let $x = q$ and $y = \dot{q}$, then we deduce the expression for the Lagrange density from the Hamilton function $H(q,p)$.

Remark 2.44 We remark that both PLD and PHD are independent to each other. However, the PLD dynamical system (2.1.5)-(2.1.8) and the PHD dynamical system (2.6.1)-(2.6.3) are equivalent in mechanics by (2.6.16) and (2.6.17). But we shall see later that the relations (2.6.16) and (2.6.17) are not valid in electromagnetism where PLD and PHD still hold true.

In addition, the differential dH is given by

$$dH = \frac{\partial H}{\partial q_k} dq_k + \frac{\partial H}{\partial p_k} dp_k + \frac{\partial H}{\partial t} dt. \tag{2.6.21}$$

By (2.6.17),

$$dH = \dot{q}_k dp_k - \frac{\partial \mathscr{L}}{\partial q_k} dq_k - \frac{\partial \mathscr{L}}{\partial t} dt. \tag{2.6.22}$$

It follows from (2.6.21) and (2.6.22) that

$$\dot{q}_k = \frac{\partial H}{\partial p_k}, \tag{2.6.23}$$

$$\frac{\partial H}{\partial q_k} = -\frac{\partial \mathscr{L}}{\partial q_k}, \tag{2.6.24}$$

$$\frac{\partial H}{\partial t} = -\frac{\partial \mathscr{L}}{\partial t}. \tag{2.6.25}$$

From (2.6.21) and (2.6.25) we deduce (2.5.61). Then by (2.6.16) and the Lagrange equation (2.1.8), we have

$$\frac{dp_k}{dt} = \frac{d}{dt}\left(\frac{\delta L}{\delta \dot{q}_k}\right) = \frac{\partial \mathscr{L}}{\partial q_k}.$$

Hence (2.6.24) becomes

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}. \tag{2.6.26}$$

Thus, by the relations (2.6.16) and (2.6.17) we deduce the equivalence of PLD dynamics and PHD dynamics in another fashion.

2.6.2 Dynamics of conservative systems

Energy conservation is a universal law in physics. It implies that the PHD is a universal principle to describe conservative physical systems. In other words, the Hamilton principle introduced in the last subsection can be generalized to all physical fields. The most remarkable characteristic of PHD is that the total energy H of the physical system is conserved:

$$\frac{d}{dt}H(q(t),p(t)) = 0, \tag{2.6.27}$$

where (q,p) are the solutions of the Hamiltonian system.

Now, we introduce the PHD.

Principle 2.45 (Hamiltonian Dynamics) *For any conservative physical system, there are two sets of state functions*

$$u = (u_1,\cdots,u_N) \quad \text{and} \quad v = (v_1,\cdots,v_N), \tag{2.6.28}$$

such that the energy density $\mathscr{H}$ *is a function of (2.6.28)*:

$$\mathscr{H} = \mathscr{H}(u,v,\cdots,D^m u,D^m v), \qquad m \geqslant 0. \tag{2.6.29}$$

The total energy of the system is

$$H = \int_\Omega \mathscr{H}(u,v,\cdots,D^m u,D^m v)dx, \qquad \Omega \subset \mathbb{R}^3, \tag{2.6.30}$$

provided that the system is described by continuous fields. Moreover the state functions u *and* v *satisfy the equations*

$$\begin{aligned} \frac{\partial u}{\partial t} &= \alpha \frac{\delta H}{\delta v}, \\ \frac{\partial v}{\partial t} &= -\alpha \frac{\delta H}{\delta u}, \end{aligned} \tag{2.6.31}$$

where α *is a constant.*

In general, the energy density (2.6.29) for a continuous field system depends on u,v only up to the first-order derivatives Du and Dv:

$$\mathscr{H} = \mathscr{H}(u,v,Du,Dv). \tag{2.6.32}$$

Then the Hamilton equations (2.6.31) can be expressed in the following form:

$$\begin{aligned}\frac{\partial u_k}{\partial t} &= \alpha\left[-\partial_j\left(\frac{\partial \mathscr{H}}{\partial \zeta_{jk}}\right)+\frac{\partial \mathscr{H}}{\partial v_k}\right],\\ \frac{\partial v_k}{\partial t} &= \alpha\left[\partial_j\left(\frac{\partial \mathscr{H}}{\partial \xi_{jk}}\right)-\frac{\partial \mathscr{H}}{\partial u_k}\right],\end{aligned} \tag{2.6.33}$$

where ξ_{jk}, ζ_{jk} are variables corresponding to $\partial_j u_k$ and $\partial_j v_k$.

Hereafter, we always assume that $\mathscr{H}$ is in the form (2.6.32). The following theorem shows that the total energy H of the Hamiltonian system (2.6.31) (or (2.6.33)) is conserved.

Theorem 2.46 (Energy Conservation) *Let (u,v) be the solutions of (2.6.33), and $\mathscr{H}$ does not explicitly contain time t. Then the energy $H(u,v)$ is conserved, i.e. $H(u,v)$ satisfies (2.6.27) with $q=u$ and $p=v$, if $\Omega=\mathbb{R}^3$. In addition, if $\Omega\neq\mathbb{R}^3$, then we have*

$$\frac{dH}{dt}=\int_{\partial\Omega}\left[\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\frac{\partial u_j}{\partial t}+\frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\frac{\partial v_j}{\partial t}\right]n_i ds, \tag{2.6.34}$$

where $n=(n_1,n_2,n_3)$ is the unit outward normal at $\partial\Omega$.

Proof For (2.6.32), we have

$$\frac{dH}{dt}=\int_{\Omega}\left[\frac{\partial \mathscr{H}}{\partial u_k}\frac{\partial u_k}{\partial t}+\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\partial_i\left(\frac{\partial u_j}{\partial t}\right)+\frac{\partial \mathscr{H}}{\partial v_k}\frac{\partial v_k}{\partial t}+\frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\partial_i\left(\frac{\partial v_j}{\partial t}\right)\right]dx. \tag{2.6.35}$$

By the Gauss formula,

$$\begin{aligned}\int_{\Omega}\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\partial_i\left(\frac{\partial u_j}{\partial t}\right)dx &= \int_{\partial\Omega}\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\frac{\partial u_j}{\partial t}n_i ds-\int_{\Omega}\partial_i\left(\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\right)\frac{\partial u_j}{\partial t}dx,\\ \frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\partial_i\left(\frac{\partial v_j}{\partial t}\right)dx &= \int_{\partial\Omega}\frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\frac{\partial v_j}{\partial t}n_i ds-\int_{\Omega}\partial_i\left(\frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\right)\frac{\partial v_j}{\partial t}dx.\end{aligned}$$

Hence (2.6.35) is rewritten as

$$\begin{aligned}\frac{d}{dt}H=&\int_{\Omega}\left[\left(\frac{\partial \mathscr{H}}{\partial u_k}-\partial_i\left(\frac{\partial \mathscr{H}}{\partial \xi_{ik}}\right)\right)\frac{\partial u_k}{\partial t}+\left(\frac{\partial \mathscr{H}}{\partial v_k}-\partial_i\left(\frac{\partial \mathscr{H}}{\partial \zeta_{ik}}\right)\right)\frac{\partial v_k}{\partial t}\right]dx\\ &+\int_{\partial\Omega}\left[\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\frac{\partial u_j}{\partial t}+\frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\frac{\partial v_j}{\partial t}\right]n_i ds.\end{aligned} \tag{2.6.36}$$

Since (u,v) is a solution of (2.6.33), then (2.6.36) becomes

$$\frac{dH}{dt}=\int_{\partial\Omega}\left[\frac{\partial \mathscr{H}}{\partial \xi_{ij}}\frac{\partial u_j}{\partial t}+\frac{\partial \mathscr{H}}{\partial \zeta_{ij}}\frac{\partial v_j}{\partial t}\right]n_i ds,$$

and (2.6.34) follows.

If $\Omega=\mathbb{R}^3$ (i.e. $\partial\Omega=\varnothing$) or $u=v=0$ on $\partial\Omega$, then

$$\frac{d}{dt}H(u,v)=0 \qquad \text{for } (u,v) \text{ satisfy (2.6.33).}$$

The proof is complete. □

Remark 2.47 The integral functions in (2.6.34):

$$P=(P_1,P_2,P_3),$$
$$P_k=-\left[\frac{\partial\mathscr{H}}{\partial\xi_{kj}}\frac{\partial u_j}{\partial t}+\frac{\partial\mathscr{H}}{\partial\zeta_{kj}}\frac{\partial v_j}{\partial t}\right], \tag{2.6.37}$$

are the energy fluxes. Hence (2.6.34) can be expressed as

$$\frac{d}{dt}H(u,v)=-\int_{\partial\Omega}P\cdot nds,$$

which means that the rate of energy change in Ω equals to the difference of the input and output of energy flow crossing the boundary of Ω per unit time. Consequently (2.6.34) is equivalent to energy conservation. □

In the Lagrangian dynamics, Noether Theorem 2.38 gives a correspondence between symmetries and conservations, and provides a way to seek the conservation laws. Likewise, the Hamiltonian dynamics provides another criterion to find conservation laws.

Let $S(u,v)$ be a functional given by

$$S(u,v)=\int_\Omega S(u,v,Du,Dv)dx. \tag{2.6.38}$$

The following theorem provides a condition for $S(u,v)$ to be a conserved quantity.

Theorem 2.48 (Conservation Laws of Hamiltonian System) *Let $S(u,v)$ be a functional as given by (2.6.38). If S and the Hamilton energy H satisfy the following relation*

$$\begin{aligned}&\int_\Omega\left[\frac{\partial S}{\partial u_k}-\partial_i\left(\frac{\partial S}{\partial\xi_{ik}}\right)\right]\left[\frac{\partial\mathscr{H}}{\partial v_k}-\partial_i\left(\frac{\partial\mathscr{H}}{\partial\zeta_{ik}}\right)\right]dx\\&=\int_\Omega\left[\frac{\partial S}{\partial v_k}-\partial_i\left(\frac{\partial S}{\partial\zeta_{ik}}\right)\right]\left[\frac{\partial\mathscr{H}}{\partial u_k}-\partial_i\left(\frac{\partial\mathscr{H}}{\partial\xi_{ik}}\right)\right]dx,\end{aligned} \tag{2.6.39}$$

for solutions (u,v) of (2.6.33), then S is a conserved quantity of the Hamiltonian system. Namely, S satisfies that

$$\frac{dS}{dt}=-\int_{\partial\Omega}P_s\cdot nds, \tag{2.6.40}$$

where $P_s=(P_s^1,P_s^2,P_s^3)$ is the flux given by

$$P_s^k=-\left[\frac{\partial S}{\partial\xi_{kj}}\frac{\partial u_j}{\partial t}+\frac{\partial S}{\partial\zeta_{kj}}\frac{\partial v_j}{\partial t}\right].$$

Proof The proof is similar to that of Theorem 2.46. By

$$\begin{aligned}\frac{dS}{dt}=&\int_\Omega\left[\frac{\partial S}{\partial u_j}\frac{\partial u_j}{\partial t}+\frac{\partial S}{\partial\xi_{ij}}\partial_i\left(\frac{\partial u_j}{\partial t}\right)+\frac{\partial S}{\partial v_j}\frac{\partial v_j}{\partial t}+\frac{\partial S}{\partial\zeta_{ij}}\partial_i\left(\frac{\partial v_j}{\partial t}\right)\right]dx\\=&\int_\Omega\left[\left(\frac{\partial S}{\partial u_j}-\partial_i\left(\frac{\partial S}{\partial\xi_{ij}}\right)\right)\frac{\partial u_j}{\partial t}+\left(\frac{\partial S}{\partial v_j}-\partial_i\left(\frac{\partial S}{\partial\zeta_{ij}}\right)\right)\frac{\partial v_j}{\partial t}\right]dx\\&+\int_{\partial\Omega}P_s^k\cdot n_k dS.\end{aligned}$$

By (2.6.33) and (2.6.39) we deduce (2.6.40). The proof is complete. □

In fact, Theorem 2.46 is a special case of Theorem 2.48 for $S = H$. Theorem 2.48 is useful for Hamiltonian systems.

2.6.3 PHD for Maxwell electromagnetic fields

PHD is also valid in the Maxwell electrodynamics. To establish the Hamiltonian dynamics for electromagnetism, we first determine the conjugate field functions as follows

$$\begin{aligned} u &= E = (E_1, E_2, E_3), \\ v &= A = (A_1, A_2, A_3), \end{aligned} \tag{2.6.41}$$

where E is the electric field, and A is the magnetic potential. The Hamilton energy H is given by

$$\begin{aligned} H &= \int_\Omega \mathscr{H}(E,A)dx, \\ \mathscr{H} &= \frac{1}{8\pi}(E^2 + |\text{curl}\, A|^2) + \frac{1}{4\pi}\nabla\varphi \cdot E - \frac{1}{c}J \cdot A, \end{aligned} \tag{2.6.42}$$

where φ is the electric potential, and J is the current. Using

$$\begin{aligned} \left\langle \frac{\delta H}{\delta E}, \widetilde{E} \right\rangle &= \frac{d}{d\lambda}\Big|_{\lambda=0} H(E + \lambda\widetilde{E}, A), \\ \left\langle \frac{\delta H}{\delta A}, \widetilde{A} \right\rangle &= \frac{d}{d\lambda}\Big|_{\lambda=0} H(E, A + \lambda\widetilde{A}), \end{aligned}$$

we can compute the derivatives of (2.6.42) as follows:

$$\frac{\delta H}{\delta E} = \frac{1}{4\pi}(E + \nabla\varphi), \quad \frac{\delta H}{\delta A} = \frac{1}{4\pi}\text{curl}^2 A - \frac{1}{c}J.$$

Thus, the Hamilton equations (2.6.31) are in the form:

$$\begin{aligned} \frac{1}{c}\frac{\partial E}{\partial t} &= 4\pi\frac{\delta H}{\delta A} = \text{curl}^2 A - \frac{4\pi}{c}J, \\ \frac{1}{c}\frac{\partial A}{\partial t} &= -4\pi\frac{\delta H}{\delta E} = -E - \nabla\varphi, \end{aligned} \tag{2.6.43}$$

which are the classical Maxwell equations (2.2.33) and (2.2.35).

Remark 2.49 We note that the Lagrange action of electromagnetism defined by (2.4.15) can be expressed as

$$L_{EM} = \int_{\mathscr{M}^4} \left[-\frac{1}{2}|\text{curl}\, A|^2 + \frac{1}{2}E^2 + \frac{4\pi}{c}A_\mu J^\mu \right] dxdt, \tag{2.6.44}$$

where

$$E = -\frac{1}{c}\frac{\partial A}{\partial t} - \nabla A_0, \quad A_\mu = (A_0, A_1, A_2, A_3).$$

The fields of PLD are given by

$$A_0, A_1, A_2, A_3 \qquad \text{and} \qquad \dot{A}_0, \dot{A}_1, \dot{A}_2, \dot{A}_3 \quad (A_0 = \varphi),$$

and the fields of PHD are

$$A_1, A_2, A_3 \qquad \text{and} \qquad E_1, E_2, E_3.$$

It is clear that

$$\frac{1}{4\pi} E(-\dot{A}) - \frac{1}{4\pi} \mathscr{L}_{EM} = \mathscr{H}(A, E) - \frac{1}{c} \varphi J^0,$$

and $J^0 = \rho$. Hence, the relation (2.6.17) does not hold true in general. $\square$

Now, we consider the energy conservation. When $\mathscr{H}$ contains $\nabla\varphi$ and J, which depend on t, the Hamilton $\mathscr{H} = \mathscr{H}(A, E, t)$ contains explicitly t. It implies that the Maxwell fields have energy exchange with other charged particles. Then H is not conserved.

As there is no charged particles in Ω, then

$$(\rho, J) = 0 \quad \text{in } \Omega.$$

In this case,

$$\int_\Omega \nabla\varphi \cdot E dx = -\int_\Omega \varphi \text{div } E dx = -\int_\Omega 4\pi\varphi\rho dx = 0.$$

Hence, the Hamilton energy (2.6.42) becomes

$$H = \frac{1}{8\pi} \int_\Omega (E^2 + |\text{curl } A|^2) dx. \tag{2.6.45}$$

By (2.6.37), the energy flux P defined in (2.6.45) reads

$$P = \frac{1}{4\pi} \text{curl } A \times \frac{\partial A}{\partial t}$$

Note that

$$\frac{\partial A}{\partial t} = -cE, \quad \text{curl } A = B \quad (B \text{ the magnetic field}),$$

Then, by (2.6.34) we have that

$$\frac{dH}{dt} = -\int_{\partial\Omega} P \cdot n ds = -\int_{\partial\Omega} \left(\frac{c}{4\pi} E \times B \right) \cdot n ds.$$

The field

$$P = \frac{c}{4\pi} E \times B$$

is the Poynting vector, represents the energy flux density of an electromagnetic field.

2.6.4 Quantum Hamiltonian systems

In quantum physics, the state function is a set of complex valued wave functions:

$$\psi = (\psi_1, \cdots, \psi_N)^{\mathrm{T}}, \quad N \geqslant 1,$$

and ψ_k are as

$$\psi_k = \psi_k^1 + i\psi_k^2 \qquad \text{for } 1 \leqslant k \leqslant N. \tag{2.6.46}$$

In view of PHD for a quantum system, the conjugate fields are taken as real and imaginary parts of the wave functions in (2.6.46):

$$\psi_1^1, \cdots, \psi_N^1 \quad \text{and} \quad \psi_1^2, \cdots, \psi_N^2. \tag{2.6.47}$$

Let $H = H(\psi)$ be the Hamilton energy. Then the Hamilton equations for the quantum system are as follows:

$$\begin{aligned} \alpha \frac{\partial \psi_k^1}{\partial t} &= \frac{\delta}{\delta \psi_k^2} H, \\ \alpha \frac{\partial \psi_k^2}{\partial t} &= -\frac{\delta}{\delta \psi_k^1} H, \end{aligned} \qquad \text{for } 1 \leqslant k \leqslant N, \tag{2.6.48}$$

where α is a constant.

We now introduce quantum Hamiltonian systems for the Schrödinger equation, the Weyl equation, the Dirac equations, the Klein-Gordon equation, and the BEC equation.

1. *Schrödinger equation*:

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi, \tag{2.6.49}$$

where $\psi = \psi^1 + i\psi^2$. The Hamilton energy of (2.6.49) is given by

$$H(\psi) = \frac{1}{2}\int_\Omega \left[\frac{\hbar^2}{2m}|\nabla\psi|^2 + V(x)|\psi|^2\right] dx.$$

It is easy to derive that

$$\begin{aligned} \frac{\delta}{\delta\psi^1} H &= -\frac{\hbar^2}{2m}\Delta\psi^1 + V\psi^1, \\ \frac{\delta}{\delta\psi^2} H &= -\frac{\hbar^2}{2m}\Delta\psi^2 + V\psi^2. \end{aligned}$$

Hence, the Hamiltonian system is

$$\begin{aligned} \hbar\frac{\partial\psi^1}{\partial t} &= -\frac{\hbar^2}{2m}\Delta\psi^2 + V\psi^2, \\ \hbar\frac{\partial\psi^2}{\partial t} &= \frac{\hbar^2}{2m}\Delta\psi^2 - V\psi^2. \end{aligned} \tag{2.6.50}$$

It is clear that (2.6.50) and (2.6.49) are equivalent.

2. *Weyl equations*:

$$\frac{1}{c}\frac{\partial\psi}{\partial t}=(\vec{\sigma}\cdot\nabla)\psi, \tag{2.6.51}$$

where $\psi=(\psi_1,\psi_2)^{\mathrm{T}}$, and $\psi_k=\psi_k^1+i\psi_k^2$ $(1\leqslant k\leqslant 2)$. The Hamilton energy H of (2.6.51) is in the form

$$\begin{aligned}H=&\int_{\mathbb{R}^3}i\psi^\dagger(\vec{\sigma}\cdot\nabla)\psi dx \\ =&\int_{\mathbb{R}^3}\left[\frac{\partial\psi_1^2}{\partial x^1}\psi_2^1+\frac{\partial\psi_2^2}{\partial x^1}\psi_1^1+\frac{\partial\psi_1^2}{\partial x^2}\psi_2^2+\frac{\partial\psi_1^1}{\partial x^2}\psi_2^1+\frac{\partial\psi_1^2}{\partial x^3}\psi_1^1+\frac{\partial\psi_2^1}{\partial x^3}\psi_1^1+\frac{\partial\psi_2^1}{\partial x^3}\psi_2^2\right]dx.\end{aligned} \tag{2.6.52}$$

The Hamilton equations are

$$\left.\begin{aligned}\frac{1}{c}\frac{\partial\psi_k^1}{\partial t}&=\frac{\delta}{\delta\psi_k^2}H,\\ \frac{1}{c}\frac{\partial\psi_k^2}{\partial t}&=-\frac{\delta}{\delta\psi_k^1}H,\end{aligned}\right.\qquad \text{for } k=1,2,$$

which, in view of for (2.6.52), are in the form:

$$\begin{aligned}\frac{1}{c}\frac{\partial\psi_1^1}{\partial t}&=\left(\frac{\partial\psi_2^1}{\partial x^1}+\frac{\partial\psi_2^2}{\partial x^2}+\frac{\partial\psi_1^1}{\partial x^3}\right),\\ \frac{1}{c}\frac{\partial\psi_1^2}{\partial t}&=\left(\frac{\partial\psi_2^2}{\partial x^1}-\frac{\partial\psi_2^1}{\partial x^2}+\frac{\partial\psi_1^2}{\partial x^3}\right),\\ \frac{1}{c}\frac{\partial\psi_2^1}{\partial t}&=\left(\frac{\partial\psi_1^1}{\partial x^1}-\frac{\partial\psi_1^2}{\partial x^2}-\frac{\partial\psi_2^1}{\partial x^3}\right),\\ \frac{1}{c}\frac{\partial\psi_2^2}{\partial t}&=\left(\frac{\partial\psi_1^2}{\partial x^1}+\frac{\partial\psi_1^1}{\partial x^2}-\frac{\partial\psi_2^2}{\partial x^3}\right).\end{aligned} \tag{2.6.53}$$

It is readily to check that (2.6.53) and (2.6.51) are equivalent.

3. *Dirac equations*:

$$i\hbar\frac{\partial\psi}{\partial t}=-i\hbar c(\vec{\alpha}\cdot\nabla)\psi+mc^2\alpha_0\psi, \tag{2.6.54}$$

where $\vec{\alpha},\alpha_0$ are as in (2.2.58), $\psi=(\psi_1,\psi_2,\psi_3,\psi_4)^{\mathrm{T}}$, and

$$\psi_k=\psi_k^1+i\psi_k^2 \qquad \text{for } 1\leqslant k\leqslant 4.$$

The Hamilton energy of (2.6.54) is defined by

$$H=\int_{\mathbb{R}^3}\left[-i\hbar c\psi^\dagger(\vec{\alpha}\cdot\nabla)\psi+mc^2\psi^\dagger\alpha_0\psi\right]dx. \tag{2.6.55}$$

It is clear that the expansions of (2.6.54) in terms of real and imaginary parts are in the form

$$\begin{aligned} \hbar\frac{\partial \psi_k^1}{\partial t} &= -\frac{\delta}{\delta \psi_k^2}H, \\ \hbar\frac{\partial \psi_k^2}{\partial t} &= \frac{\delta}{\delta \psi_k^1}H, \end{aligned} \qquad \text{for } 1 \leqslant k \leqslant 4,$$

where H is given by (2.6.55).

4. *Klein-Gordon equation*:

$$\frac{1}{c^2}\frac{\partial^2 \psi}{\partial t^2} - \Delta\psi + \left(\frac{mc}{\hbar}\right)^2 \psi = 0. \tag{2.6.56}$$

The conjugate fields are

$$\psi^1 = \psi, \quad \psi^2 = \frac{\partial \psi}{\partial t}.$$

The Hamilton energy is

$$H = \frac{1}{2}\int_\Omega \left[|\psi^2|^2 + |\nabla\psi^1|^2 + \left(\frac{mc}{\hbar}\right)^2 |\psi^1|^2\right] dx. \tag{2.6.57}$$

Then (2.6.56) can be equivalently written as

$$\begin{aligned} \frac{1}{c}\frac{\partial \psi^1}{\partial t} &= \frac{\delta}{\delta \psi^2}H, \\ \frac{1}{c}\frac{\partial \psi^2}{\partial t} &= -\frac{\delta}{\delta \psi^1}H. \end{aligned}$$

5. *Bose-Einstein condensation (BEC) equation*:

$$i\hbar\frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\Delta\psi + V(|\psi|^2)\psi, \tag{2.6.58}$$

where $V(|\psi|^2)$ is a function of $|\psi|^2$, and $\psi = \psi^1 + i\psi^2$. The Hamilton energy of (2.6.58) is given by

$$\begin{aligned} H &= \int_\Omega \left[\frac{\hbar^2}{4m}|\nabla\psi|^2 + G(|\psi|^2)\right] dx, \\ G(z) &= \frac{1}{2}\int_0^z V(s)ds. \end{aligned} \tag{2.6.59}$$

For (2.6.59), the equation (2.6.58) can be rewritten as

$$\begin{aligned} \hbar\frac{\partial \psi^1}{\partial t} &= \frac{\delta}{\delta \psi^2}H, \\ \hbar\frac{\partial \psi^2}{\partial t} &= -\frac{\delta}{\delta \psi^1}H. \end{aligned}$$

The examples given above show that PHD holds true in general in quantum physics. In particular, we now show that the relations (2.6.16) and (2.6.17) also valid for quantum Hamiltonian systems.

In fact, the Lagrange action of a quantum system with the Hamilton energy $H(\psi)$ is given by

$$\begin{aligned} L &= \int_0^T \int_{\mathbb{R}^3} \mathscr{L}(\psi, \dot{\psi}) dx dt, \\ \mathscr{L} &= i\hbar \psi^* \dot{\psi} - \mathscr{H}(\psi, D\psi), \end{aligned} \tag{2.6.60}$$

where H reads

$$H = \int_{\mathbb{R}^3} \mathscr{H}(\psi, D\psi) dx.$$

When taking the variation, the functional

$$\begin{aligned} \widetilde{L} &= \int_{Q_T} i\psi^* \dot{\psi} dx dt \\ &= \int_{Q_T} i(\psi^1 - i\psi^2)(\dot{\psi}^1 + i\dot{\psi}^2) dx dt \\ &= \int_{Q_T} \left[\frac{i}{2} \frac{\partial}{\partial t} |\psi|^2 - \frac{\partial}{\partial t}(\psi^1 \psi^2) + 2\psi^2 \dot{\psi}_1 \right] dx dt \\ &= \int_0^T \int_{\mathbb{R}^3} 2\psi^2 \dot{\psi}^1 dx dt, \end{aligned}$$

where $Q_T = \mathbb{R}^3 \times (0, T)$. Hence, the density $\mathscr{L}$ in (2.6.60) is equivalent to

$$\mathscr{L} = 2\hbar \psi^2 \dot{\psi}^1 - \mathscr{H}. \tag{2.6.61}$$

It follows from (2.6.61) that

$$\begin{aligned} \psi^2 &= \frac{1}{2\hbar} \partial \mathscr{L} / \partial \dot{\psi}^1, \\ \psi^2 \dot{\psi}^1 - \frac{1}{2\hbar} \mathscr{L} &= \frac{1}{2\hbar} \mathscr{H}. \end{aligned} \tag{2.6.62}$$

The relations (2.6.62) are what we expected.

Chapter 3

Mathematical Foundations

The aim of this chapter is to provide mathematical foundations for the remaining part of the book on the unified field theory, elementary particles, quantum physics, astrophysics and cosmology.

As addressed in Chapter 2, Nature speaks the language of mathematics. Thanks to Einstein's principle of equivalence and the geometric interaction mechanism, the space-time is a 4D Riemannian manifold $\mathscr{M}$, and physical fields are regarded as functions on vector bundles over the base manifold $\mathscr{M}$. For example, the Riemannian metric $g_{\mu\nu}$, representing the gravitational potential, is a function on the second-order cotangent bundle, $g_{\mu\nu}: \mathscr{M} \to T^*\mathscr{M} \otimes T^*\mathscr{M}$; the electromagnetic field A_μ is a function on the cotangent bundle $A_\mu : \mathscr{M} \to T^*\mathscr{M}$; the $SU(2)$ gauge fields for the weak interaction is a function $\{W^a_\mu\} : \mathscr{M} \to (T^*\mathscr{M})^3$; and the $SU(3)$ gauge fields for the strong interaction is a function $\{S^k_\mu\} : \mathscr{M} \to (T^*\mathscr{M})^8$. Also, the Dirac spinor field is defined a complex bundle: $\Psi : \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{C}^4$.

The three most fundamental symmetries of Nature are the principle of Lorentz invariance, the principle of general relativity, and the principle of gauge invariance. They lead to transformations in either the base space-time manifold $\mathscr{M}$, or the corresponding vector bundles.

In summary, we have the following

1) *The space-time is a 4D Riemannian manifold, and all physical fields are functions on vector bundles over the space-time manifold* $\mathscr{M} \otimes_p E^N$; *and*
2) *Fundamental symmetries are invariances of physical field equations under the underlying transformations on the corresponding vector bundles.*

The basic concepts in Section 3.1 include Riemannian manifolds such as the space-time manifold, vector bundles, tensor fields, connections and linear transformations on vector bundles. It is particularly important that we identify all physical fields as functions on proper vector bundles defined on the space-time manifold, and symmetries correspond then to linear transformations on the corresponding vector bundles.

Section 3.2 is on basic functional analysis and partial differential equations on manifolds, which are needed for rigorous proofs of theorems and concepts used in later chapters of the book. The readers may consult other references such as (Ma, 2011).

Section 3.3 proves the orthogonal decomposition theorem for tensors on Riemannian manifolds, which forms the mathematical foundation for the principle of interaction dynamics (PID), postulated by (Ma and Wang, 2014e, 2015a). Basically, PID takes the variation of the Lagrangian action $L(u)$ under the generalized energy-momentum conservation constraints, which we call div_A-free constraints with A being the gauge potentials:

$$\langle \delta L(u_0), X\rangle = \frac{d}{d\lambda}\Big|_{\lambda=0} L(u_0+\lambda X) = 0, \qquad \forall\ \text{div}_A X \stackrel{\text{def}}{=} \text{div} X - X\cdot A = 0. \tag{3.0.1}$$

Here $\text{div}_A X = 0$ represents a generalized energy-momentum conservation. The study of the constraint variation (3.0.1) requires the decomposition of all tensor fields into the space of div_A-free fields and its orthogonal complements. Such a decomposition is reminiscent to the classical Helmholtz decomposition of a vector into the sum of an irrotational (curl-free) vector field and a solenoidal (divergence-free) vector field. In particular, we show in this chapter that (see Theorem 3.17):

$$\begin{aligned} & L^2(T^k_r\mathscr{M}) = G(T^k_r\mathscr{M}) \oplus L^2_D(T^k_r\mathscr{M}), \\ & G(T^k_r\mathscr{M}) = \left\{ v\in L^2(T^k_r\mathscr{M}) \mid v = D\varphi + A\otimes\varphi,\ \varphi\in H^1(T^k_{r-1}\mathscr{M})\right\}, \\ & L^2_D(T^k_r\mathscr{M}) = \{v\in L^2(T^k_r\mathscr{M}) \mid \text{div}_A v = 0\}, \end{aligned} \tag{3.0.2}$$

where D is the connection on the space-time manifold.

The first part of Section 3.4 deals with classical variations of the Lagrangian actions, and gives detailed calculations for the variations of the Einstein-Hilbert functional and the Yang-Mills action. The second part studies variations under div_A-free constraints, where A stands for gauge potentials. These constraints represent generalized energy-momentum conservation and provide basic mathematical theorems for deriving the unified field model for the fundamental interactions. This section is based entirely on (Ma and Wang, 2014e, 2015a).

Section 3.5 explores the inner/hidden symmetry behind the $SU(N)$ representation for the non-abelian gauge theory. Basically, we have realized in (Ma and Wang, 2014h) that the set of generators $SU(N)$ plays exactly the role of a coordinate system, leading to a new invariance, which we call the principle of representation invariance (PRI), first discovered in (Ma and Wang, 2014h).

Section 3.6 addresses the spectral analysis of the Dirac and Weyl operators, which will play an important role in studying the energy-levels of subatomic particles in Section 6.4, which is based entirely on (Ma and Wang, 2014g).

3.1 Basic Concepts

3.1.1 Riemannian manifolds

The n-sphere S^n and the Euclidean space $\mathbb{R}^n$ are two typical examples of n-dimensional manifolds. The most common manifolds, which we can visually see, are one-dimensional curves and two-dimensional surfaces. It is, however, difficult for us to tell whether the three-dimensional space we live in is curved or flat by our common sense. The Riemannian geometry provides a theory with which the intelligent beings living in an n-dimensional manifold are able to determine the curvature of this space from its metric.

A plane in Figure 3.1 (a) is expressed by

$$\vec{r}(x^1,x^2) = (x^1,x^2,x^3(x^1,x^2)), \tag{3.1.1}$$

where $x_3 = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3$, α_j $(1 \leqslant j \leqslant 3)$ are constants, and the metric of the plane (3.1.1) are given by

$$\begin{aligned} ds^2 = d\vec{r}\cdot d\vec{r} =&(dx^1)^2 + (dx^2)^2 + (\alpha_1 dx^1 + \alpha_2 dx^2)^2 \\ =&(1+\alpha_1^2)(dx^1)^2 + 2\alpha_1\alpha_2 dx^1 dx^2 + (1+\alpha_2^2)(dx^2)^2. \end{aligned} \tag{3.1.2}$$

A sphere in Figure 3.1 (b) is expressed by

$$\vec{r}(x^1,x^2) = (x^1,x^2,x^3(x^1,x^2)), \tag{3.1.3}$$

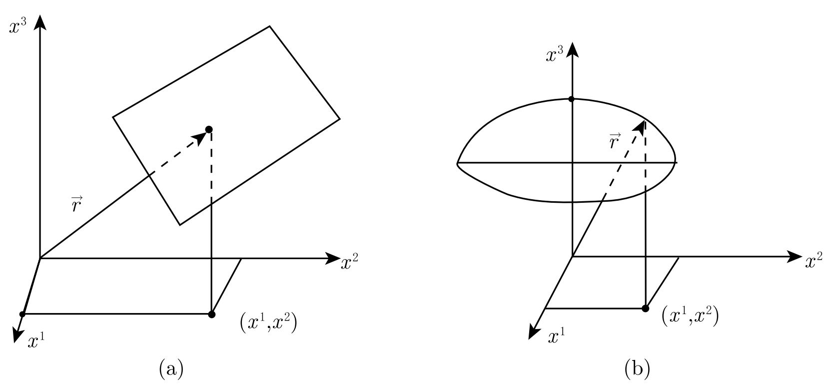


Figure 3.1 (a) a plane, and (b) a sphere with radius R

where $x^3 = \sqrt{R^2 - (x^1)^2 - (x^2)^2}$, and R is the radius. The metric of this sphere (3.1.3) is given by

$$\begin{aligned} ds^2 =&(dx^1)^2 + (dx^2)^2 + \left(\frac{\partial x^3}{\partial x^1}dx^1 + \frac{\partial x^3}{\partial x^2}dx^2\right)^2 \\ =&(1+\varphi)(dx^1)^2 + 2\varphi dx^1 dx^2 + (1+\varphi)(dx^2)^2, \end{aligned} \tag{3.1.4}$$

where $\varphi = 1/(R^2 - (x^1)^2 - (x^2)^2)$.

It is clear that the metric

$$ds^2 = g_{ij}dx^i dx^j$$

defined on a surface dictates its curvature. We see that the metric (3.1.2) of the plane (3.1.1), i.e.

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} 1+\alpha_1^2 & \alpha_1\alpha_2 \\ \alpha_1\alpha_2 & 1+\alpha_2^2 \end{pmatrix} \tag{3.1.5}$$

is a constant metric, and the metric (3.1.4) of the sphere (3.1.3), i.e.

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} 1+\varphi(x^1,x^2) & \varphi(x^1,x^2) \\ \varphi(x^1,x^2) & 1+\varphi(x^1,x^2) \end{pmatrix} \tag{3.1.6}$$

is not constant. In fact, for the plane shown by Figure 3.1 (a), we can find a coordinate transformation

$$\begin{pmatrix} \widetilde{x}^1 \\ \widetilde{x}^2 \end{pmatrix} = \begin{pmatrix} a_1^1 & a_2^1 \\ a_1^2 & a_2^2 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix},$$

such that in the coordinate system $(\widetilde{x}^1, \widetilde{x}^2)$, the metric (3.1.2) is expressed in the diagonal form:

$$ds^2 = (d\widetilde{x}^1)^2 + (d\widetilde{x}^2)^2. \tag{3.1.7}$$

In other words, the metric is $g_{ij} = \delta_{ij}$. However, it is impossible to achieve this for the metric (3.1.4) of the sphere.

The conclusion in this example holds true as well for all Riemannian manifolds. Namely, for a Riemannian manifold $\{\mathscr{M}, g_{ij}\}$, $\mathscr{M}$ is flat if and only if there is a coordinate system x, such that the metric $\{g_{ij}\}$ can be expressed as $g_{ij} = \delta_{ij}$ under the x-coordinate system.

The following are a few general properties of an n-dimensional Riemannian manifold $\{\mathscr{M}, g_{ij}\}$.

1. The Riemannian metric

$$ds^2 = g_{ij}dx^i dx^j \tag{3.1.8}$$

is invariant. In fact, under a coordinate transformation

$$\widetilde{x} = \varphi(x), \quad x = \varphi^{-1}(\widetilde{x}), \tag{3.1.9}$$

the second-order covariant tensor field $\{g_{ij}\}$ satisfies that

$$(\widetilde{g}_{ij}) = (b_i^k)^{\mathrm{T}}(g_{kl})(b_j^l), \qquad \begin{pmatrix} d\widetilde{x}^1 \\ \vdots \\ d\widetilde{x}^n \end{pmatrix} = (a_j^i) \begin{pmatrix} dx^1 \\ \vdots \\ dx^n \end{pmatrix},$$

$$(a_j^i) = \left(\frac{\partial \varphi^i}{\partial x^j} \right), \qquad (b_j^i) = (a_j^i)^{-1}.$$

Hence we have

$$\begin{aligned}\widetilde{g}_{ij}d\widetilde{x}^i d\widetilde{x}^j =&(d\widetilde{x}^1,\cdots,d\widetilde{x}^n)(\widetilde{g}_{ij})\begin{pmatrix}d\widetilde{x}^1\\ \vdots\\ d\widetilde{x}^n\end{pmatrix} \\ =&(dx^1,\cdots,dx^n)(a_i^k)^{\mathrm{T}}(b_i^k)^{\mathrm{T}}(g_{kl})(b_j^l)(a_j^l)\begin{pmatrix}dx^1\\ \vdots\\ dx^n\end{pmatrix} \\ =&(dx^1,\cdots,dx^n)(g_{ij})\begin{pmatrix}dx^1\\ \vdots\\ dx^n\end{pmatrix} \\ =&g_{ij}dx^i dx^j.\end{aligned} \tag{3.1.10}$$

It follows that the metric (3.1.8) is invariant.

2. The length of a curve on $\mathscr{M}$ is determined by the metric $\{g_{ij}\}$. Let $\gamma(t)$ be a curve connecting two points $p,q\in M$, and $\gamma(t)$ be expressed by

$$x(t)=(x^1(t),\cdots,x^n(t)),\quad 0\leqslant t\leqslant T,\quad x(0)=p,\quad x(T)=q. \tag{3.1.11}$$

By (3.1.8) the infinitesimal arc-length is given by

$$ds=\sqrt{g_{ij}dx^i dx^j}=\sqrt{g_{ij}(x(t))\dot{x}^i\dot{x}^j}dt.$$

Hence the length L of $\gamma(t)$ is given by

$$L=\int_p^q ds=\int_0^T\sqrt{g_{ij}(x(t))\frac{dx^i}{dt}\frac{dx^j}{dt}}dt. \tag{3.1.12}$$

The invariance of ds as shown in (3.1.10) implies that the length L in (3.1.12) is independent of the coordinate systems.

3. The volume of a bounded domain $U\subset\mathscr{M}$ is invariant. The reason why $\{g_{ij}\}$ is called a metric is that such quantities as the length, area, volume and angle are all determined by the metric $\{g_{ij}\}$.

Given a Riemannian manifold $\{\mathscr{M},g_{ij}\}$, let $U\subset\mathscr{M}$ be a bounded domain. Then the volume of U is written as

$$V=\int_U\Omega(x)dx. \tag{3.1.13}$$

where the volume element Ωdx reads

$$\Omega(x)dx=\sqrt{-g}dx,\quad g=\det(g_{ij}). \tag{3.1.14}$$

To derive (3.1.14), let $\mathscr{M} \subset \mathbb{R}^{n+1}$ be an embedding, and

$$\vec{r}(x) = \{r_1(x), \cdots, r_{n+1}(x)\},$$

be the embedding function. Consider the vector product of n vectors in $\mathbb{R}^n$:

$$\left[\frac{\partial \vec{r}}{\partial x^1}, \cdots, \frac{\partial \vec{r}}{\partial x^n}\right] = \begin{vmatrix} \vec{e}_1 & \cdots & \vec{e}_{n+1} \\ \partial_1 r_1 & \cdots & \partial_1 r_{n+1} \\ \vdots & & \vdots \\ \partial_n r_1 & \cdots & \partial_n r_{n+1} \end{vmatrix},$$

where $\{\vec{e}_1, \cdots, \vec{e}_{n+1}\}$ is an orthogonal basis of $\mathbb{R}^{n+1}$. Then the volume element Ωdx is

$$\Omega dx = \left|\left[\frac{\partial \vec{r}}{\partial x^1}, \cdots, \frac{\partial \vec{r}}{\partial x^n}\right]\right| dx.$$

By $g_{ij} = \frac{\partial \vec{r}}{\partial x^i} \cdot \frac{\partial \vec{r}}{\partial x^j}$, the norm of the vector $[\partial_1 \vec{r}, \cdots, \partial_n \vec{r}]$ is

$$|[\partial_1 \vec{r}, \cdots, \partial_n \vec{r}]| = \sqrt{-g}, \quad g = \det\,(g_{ij}).$$

Thus (3.1.14) follows.

We now verify the invariance of the volume element. Under the transformation (3.1.9),

$$\begin{aligned} d\widetilde{x} &= d\widetilde{x}^1 \wedge \cdots \wedge d\widetilde{x}^n \\ &= \left(\frac{\partial \varphi^1}{\partial x^1} dx^1 + \cdots + \frac{\partial \varphi^1}{\partial x^n} dx^n\right) \wedge \cdots \wedge \left(\frac{\partial \varphi^n}{\partial x^1} dx^1 + \cdots + \frac{\partial \varphi^n}{\partial x^n} dx^n\right) \\ &= \det\left(\frac{\partial \varphi^i}{\partial x^j}\right) dx^1 \wedge \cdots \wedge dx^n. \end{aligned} \tag{3.1.15}$$

On the other hand,

$$(\widetilde{g}_{ij}) = \left(\frac{\partial \psi^i}{\partial y^j}\right) (g_{ij}) \left(\frac{\partial \psi^i}{\partial y^j}\right)^{\mathrm{T}}, \quad \psi = \varphi^{-1}. \tag{3.1.16}$$

Hence

$$\sqrt{\det(\widetilde{g}_{ij})} = \det\left(\frac{\partial \varphi^i}{\partial x^j}\right)^{-1} \sqrt{\det(g_{ij})}. \tag{3.1.17}$$

We deduce from (3.1.15) and (3.1.17) that

$$\sqrt{\det(\widetilde{g}_{ij})} d\widetilde{x} = \sqrt{\det(g_{ij})} dx.$$

Namely, both the volume and the volume element in (3.1.13)-(3.1.14) are invariant.

4. The metric $\{g_{ij}\}$ gives rise to an inner product structure on the tangent space of a Riemann manifold $\mathscr{M}$, and defines the angle between two tangent vectors.

Let $p \in \mathscr{M}$ be a given point, and $T_p\mathscr{M}$ be the tangent space at $p \in M$. For two vectors $X, Y \in T_p\mathscr{M}$,

$$X = \{X^1, \cdots, X^n\}, \quad Y = \{Y^1, \cdots, Y^n\},$$

the inner product of X and Y is defined by

$$\langle X, Y \rangle = g_{ij}(p) X^i Y^j. \tag{3.1.18}$$

It is clear that the inner product (3.1.18) is an invariant. The angle between X and Y is defined as

$$\cos\theta = \frac{\langle X, Y \rangle}{|X||Y|}, \quad |Z| = \sqrt{g_{ij} Z^i Z^j}, \qquad \text{for } Z = X, Y.$$

5. The inverse of (g_{ij}), denoted by

$$(g^{ij}) = (g_{ij})^{-1},$$

is a second order contra-variant tensor. In fact, by (3.1.16) we have that

$$I = (\widetilde{g}_{ij})(\widetilde{g}^{ij}) = \left(\frac{\partial \varphi}{\partial x}\right)^{-1} (g_{ij}) \left[\left(\frac{\partial \varphi}{\partial x}\right)^{\mathrm{T}}\right]^{-1} (\widetilde{g}^{ij}),$$

where I is unit matrix. It follows that

$$(\widetilde{g}^{ij}) = \left(\frac{\partial \varphi}{\partial x}\right)^{\mathrm{T}} (g^{ij}) \left(\frac{\partial \varphi}{\partial x}\right).$$

Hence $\{g^{ij}\}$ is a second-order contra-variant tensor.

3.1.2 Physical fields and vector bundles

Let $\mathscr{M}$ be a manifold. A vector bundle on $\mathscr{M}$ is obtained by gluing an N-dimensional linear space E^N at each point $p \in \mathscr{M}$, denoted by

$$\mathscr{M} \otimes_p E^N \stackrel{\text{def}}{=} \bigcup_{p \in \mathscr{M}} \{p\} \times E^N_p, \tag{3.1.19}$$

where $E^N = \mathbb{R}^N$, or $\mathbb{C}^N$, or the Minkowski space.

In a vector bundle (3.1.19), the geometric position of E^N_p is related with $p \in \mathscr{M}$. For example, the set of all tangent spaces is a vector bundle on $\mathscr{M}$, called the tangent bundle of $\mathscr{M}$, denoted by

$$T\mathscr{M} = \mathscr{M} \otimes_p T_p\mathscr{M}.$$

For (3.1.19), if the positions of bundle spaces E_p^N $(p \in \mathscr{M})$ are independent of p, then it is called a geometrically trivial vector bundle. Vector bundles on a flat manifold are always geometrically trivial.

The physical background of vector bundles are very clear. Each physical field must be a mapping from the base space $\mathscr{M}$ to a vector bundle $\mathscr{M} \otimes_p E^N$. Actually, all physical events occur in the space-time universe $\mathscr{M}$, and the physical fields describing these events are defined on vector bundles. This point of view is well illustrated in the following examples.

Example 3.1 The motion of the air or seawater can be ideally considered as a fluid motion on a two-dimensional sphere S^2, and the velocity field u is defined on the tangent planes TS^2:

$$u: \ S^2 \to TS^2 \qquad \text{with } u(p) \in T_pS^2, \quad \forall p \in S^2.$$

Example 3.2 The electromagnetic interaction takes place on the 4D space-time manifold $\mathscr{M}$, and the electromagnetic field A^μ is defined on the tangent space:

$$A^\mu: \ \mathscr{M} \to T\mathscr{M} \qquad \text{with } A^\mu(p) \in T_p\mathscr{M}, \quad \forall p \in \mathscr{M},$$

where at each point $p \in \mathscr{M}$, $T_p\mathscr{M}$ is the Minkowski space. □

Example 3.3 An electron moves in the space-time manifold $\mathscr{M}$, and the field describing the electron state is the Dirac spinor ψ, which is defined on a 4D complex space $\mathbb{C}^4$:

$$\psi: \ \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{C}_p^4 \qquad \text{with } \psi(p) \in \mathbb{C}_p^4, \quad \forall p \in \mathscr{M}.$$

Examples 3.1-3.3 clearly illustrate that all physical fields are defined on a vector bundle on $\mathscr{M}$. Namely, a physical field F is a mapping:

$$F: \ \mathscr{M} \to \mathscr{M} \otimes_p E_p^N \qquad \text{with } F(p) \in E_p^N, \quad \forall p \in \mathscr{M}. \tag{3.1.20}$$

The physical field F takes its values in the N-dimensional linear space E^N, and can be written as N components,

$$F = (F_1, \cdots, F_N)^{\mathrm{T}}. \tag{3.1.21}$$

Considering invariance under certain symmetry, the transformation group always acts on the bundle space E^N, and induces a corresponding transformation for the field F in (3.1.21). Also, in order to ensure the covariance of the field equations for F, a differential operator D acting F must also be covariant, leading to the introduction of connections on the vector bundle $\mathscr{M} \otimes_p E^N$. This process is shown as follows:

1. Symmetric group G act on E^N:

$$G_p: \ E_p^N \to E_p^N. \tag{3.1.22}$$

2. The fields F induced to be transformed:

$$F \to T_p F. \tag{3.1.23}$$

3. The vector bundle $\mathscr{M} \otimes_p E^N$ is endowed with connections Γ_μ:

$$D_\mu = \partial_\mu + \Gamma_\mu. \tag{3.1.24}$$

Finally, we say that the vector bundle $\mathscr{M} \otimes_p E^N$ is geometrically trivial if and only if the connections Γ_μ in (3.1.24) are zero, i.e. $\Gamma_\mu = 0$ on $\mathscr{M}$. It implies that the transformations (3.1.22)-(3.1.23) determines whether $\mathscr{M} \otimes_p E_p^N$ is geometrically trivial, and if G_p, T_p in (3.1.22) and (3.1.23) are independent of $p \in \mathscr{M}$, then $\mathscr{M} \otimes_p E^N$ is geometrically trivial, and otherwise it's not. This viewpoint is important for the unified field theory introduced in Chapter 4, because it implies that the geometry of $\mathscr{M} \otimes_p E^N$ is determined by symmetry principles.

The other reason to adopt vector bundles as the mathematical framework to describe physical fields F is that the types of F can be directly reflected by the bundle space E^N. Hereafter we list a few useful physical fields:

1) For a real (complex) scalar field ϕ, the associated vector bundle is $\mathscr{M} \otimes_p \mathbb{R}^1$ $(\mathscr{M} \otimes_p \mathbb{C})$:

$$\phi : \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{R}^1 \quad (\mathscr{M} \otimes_p \mathbb{C}).$$

2) Let A^μ be a 4-dimensional vector field. Then

$$A^\mu : \mathscr{M} \to T\mathscr{M},$$

and for any $p \in \mathscr{M}$, $T_p\mathscr{M}$ is the Minkowski space.

3) For a 4-dimensional covector field A_μ, we have

$$A_\mu : \mathscr{M} \to T^*\mathscr{M},$$

and for any $p \in \mathscr{M}$, $T_p^*\mathscr{M}$ is the dual space of $T_p\mathscr{M}$.

4) A (k,r)-tensor field T on $\mathscr{M}$:

$$T = \{T^{\mu_1\cdots\mu_k}_{\nu_1\cdots\nu_r}\},$$

is expressed by

$$T : \mathscr{M} \to T^k_r\mathscr{M},$$

where $T^k_r\mathscr{M}$ is the (k,r)-tensor bundle on $\mathscr{M}$, denoted by

$$T^k_r\mathscr{M} = \underbrace{T\mathscr{M} \otimes \cdots \otimes T\mathscr{M}}_{k} \otimes \underbrace{T^*\mathscr{M} \otimes \cdots \otimes T^*\mathscr{M}}_{r}, \tag{3.1.25}$$

and $\otimes$ represents the tensor product.

5) The Dirac spinor field Ψ is defined by

$$\Psi: \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{C}^4.$$

In particular, for N Dirac spinor fields $\Psi = (\psi^1, \cdots, \psi^N)^{\mathrm{T}}$,

$$\Psi: \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^N.$$

6) The Riemann metric $g_{\mu\nu}$ defined on a 4D space-time manifold $\mathscr{M}$, representing the gravitational potential, is a mapping:

$$g_{\mu\nu}: \mathscr{M} \to T_2^0\mathscr{M},$$

where $T_2^0\mathscr{M} = T^*\mathscr{M} \otimes T^*\mathscr{M}$ is a (0,2)-tensor bundle as defined by (3.1.25).

The fields given by 1)-6) above include all types of physical fields, and the associated vector bundles are physically significant.

In classical theories of interactions, the physical fields and the associated bundle spaces are given as follows:

1) Gravity:

$$g_{\mu\nu}: \mathscr{M} \to T^*\mathscr{M} \otimes T^*\mathscr{M}. \tag{3.1.26}$$

2) Electromagnetism: $U(1)$ gauge and fermion fields,

$$A_\mu: \mathscr{M} \to T^*\mathscr{M}, \qquad \Psi: \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{C}^4. \tag{3.1.27}$$

3) Weak interaction: $SU(2)$ gauge fields

$$W_\mu^a: \mathscr{M} \to (T^*\mathscr{M})^3, \qquad (\Psi^1, \Psi^2)^{\mathrm{T}}: \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^2. \tag{3.1.28}$$

4) Strong interaction: $SU(3)$ gauge fields and fermion fields,

$$S_\mu^k: \mathscr{M} \to (T^*\mathscr{M})^8, \qquad (\Psi^1, \Psi^2, \Psi^3)^{\mathrm{T}}: \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^3. \tag{3.1.29}$$

In the unified field theory to be introduced in Chapter 4, each of the four interactions (3.1.26)-(3.1.29) possesses corresponding dual fields as follows:

$$\phi_\mu^G \leftrightarrow g_{\mu\nu}, \quad \phi^E \leftrightarrow A_\mu, \quad \phi_a^w \leftrightarrow W_\mu^a, \quad \phi_k^s \leftrightarrow S_\mu^k.$$

The corresponding vector bundles for the dual fields are as follows

$$\begin{aligned} &\phi_\mu^G: \mathscr{M} \to T^*\mathscr{M}, \\ &\phi^E: \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{R}^1, \\ &\phi_a^w: \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{R}^3, \\ &\phi_k^s: \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{R}^8. \end{aligned} \tag{3.1.30}$$

The fields (3.1.26)-(3.1.30) are all physical fields in the unified field theory. The dual fields (3.1.30) are introduced only in the unified field equations using PID, but they do not appear in the actions of the unified field theory.

3.1.3 Linear transformations on vector bundles

In the last subsection, a field on a manifold $\mathscr{M}$ can be regarded as a mapping from the base manifold $\mathscr{M}$ to some vector bundle $\mathscr{M}\otimes_p E^N$:

$$F:\ \mathscr{M}\to\mathscr{M}\otimes_p E^N. \tag{3.1.31}$$

Let G be a transformation group acting on the fiber space E^N of the bundle. This group action induces naturally a transformation on the bundle $\mathscr{M}\otimes_p E^N$ as follows:

$$X\to gX,\qquad \forall X\in E^N_p,\quad g\in G,\quad p\in\mathscr{M}. \tag{3.1.32}$$

For the field (3.1.31), we know that

$$F(p)\in E^N_p,\qquad \forall p\in\mathscr{M}.$$

Hence the transformation (3.1.32) induces a natural transformation on the field F:

$$F\to gF,\qquad \forall g\in G. \tag{3.1.33}$$

As discussed in Subsection 2.1.5, each symmetry possesses three ingredients: space (manifold), transformation group, and tensors. Consequently, the spaces M are different for different symmetries. However, in the fashion provided by (3.1.31)-(3.1.33), the base space $\mathscr{M}$ is fixed, and all symmetric transformations as groups act on the bundle spaces E^N as in (3.1.32), and the fields of (3.1.31) are automatically transformed as in (3.1.33). Thus, the physical invariances are referred to a few group actings on bundle spaces, i.e. the linear transformation on E^N.

To illustrate this idea, we start with general tensors on an n-dimensional Riemannian manifold $\mathscr{M}$. Let two coordinate systems $x=(x^1,\cdots,x^n)$ and $\widetilde{x}=(\widetilde{x}^1,\cdots,\widetilde{x}^n)$ are transforming under the following coordinate transformation:

$$\widetilde{x}^k=\varphi^k(x),\qquad 1\leqslant k\leqslant n. \tag{3.1.34}$$

Then (k,r)-tensors transform under

$$\widetilde{T}^{i_1\cdots i_k}_{j_1\cdots j_r}=b^{l_1}_{j_1}\cdots b^{l_r}_{j_r}a^{i_1}_{s_1}\cdots a^{i_k}_{s_k}T^{s_1\cdots s_k}_{l_1\cdots l_r}, \tag{3.1.35}$$

where $a^i_j=\partial\widetilde{x}^i/\partial x^j$, $b^i_j=\partial x^i/\partial\widetilde{x}^j$.

Now, we consider the general tensors from the viewpoint of vector bundles. A (k,r) tensor field is a mapping:

$$T:\ \mathscr{M}\to T^k_r\mathscr{M}=\mathscr{M}\otimes_p E_p, \tag{3.1.36}$$

and the bundle space E_p is given by

$$E_p = \underbrace{T_p\mathscr{M} \otimes \cdots \otimes T_p\mathscr{M}}_{k} \otimes \underbrace{T_p^*\mathscr{M} \otimes \cdots \otimes T_p^*\mathscr{M}}_{r}. \tag{3.1.37}$$

In the x-coordinate system, the bases of $T_p\mathscr{M}$ and $T_p^*\mathscr{M}$ are:

$$\begin{aligned} T_p\mathscr{M}:\ & e_i = \partial/\partial x^i \quad \text{for } 1 \leqslant i \leqslant n, \\ T_p^*\mathscr{M}:\ & e^i = dx^i \quad\quad \text{for } 1 \leqslant i \leqslant n. \end{aligned} \tag{3.1.38}$$

These bases induce a basis of the linear space E_p in (3.1.37):

$$e^{j_1\cdots j_r}_{i_1\cdots i_k} = e_{i_1} \otimes \cdots \otimes e_{i_k} \otimes e^{j_1} \otimes \cdots \otimes e^{j_r},$$

and the field (3.1.36) can be expressed as

$$T = T^{i_1\cdots i_k}_{j_1\cdots j_r} e^{j_1\cdots j_r}_{i_1\cdots i_k}. \tag{3.1.39}$$

When we take linear transformations on $T_p\mathscr{M}$ and $T_p^*\mathscr{M}$:

$$\begin{aligned} A:\ & T_p\mathscr{M} \to T_p\mathscr{M}, \quad A = (a^i_j), \\ B:\ & T_p^*\mathscr{M} \to T_p^*\mathscr{M}, \quad B = (b^i_j) = (A^T)^{-1}, \end{aligned} \tag{3.1.40}$$

then the field T of (3.1.36) (i.e. (3.1.39)) will transform in the same fashion as (3.1.35).

The above two ways to define general tensors are equivalent. However, in the second fashion we replace the coordinate transformation (3.1.39) by (3.1.40). This approach is very convenient to uniformly treat the transformations in the unified field theory.

Hereafter we list a few typical linear transformations of fiber spaces for various physical fields.

1. *Lorentz transformations.* A (k,r) type Lorentz field is a mapping

$$F:\ \mathscr{M} \to \mathscr{M} \otimes_p E^N, \tag{3.1.41}$$

where the fiber space E_p^N is

$$E_p^N = \underbrace{T_p\mathscr{M} \otimes \cdots \otimes T_p\mathscr{M}}_{k} \otimes \underbrace{T_p^*\mathscr{M} \otimes \cdots \otimes T_p^*\mathscr{M}}_{r}.$$

When $T_p\mathscr{M}$ undergoes a Lorentz transformation

$$L: T_p\mathscr{M} \to T_p\mathscr{M}, \quad L \text{ is a Lorentz matrix}, \tag{3.1.42}$$

the dual space $T_p^*\mathscr{M}$ transforms as

$$\widetilde{L}:\ T_p^*\mathscr{M} \to T_p^*\mathscr{M}, \quad \widetilde{L} = (L^{\mathrm{T}})^{-1}, \tag{3.1.43}$$

This leads to a natural linear transformation for the fiber space E_p^N, which induces a transformation for the field F in (3.1.41).

2. *$SU(N)$ gauge transformation.* A set of N Dirac spinor fields Ψ are referred to the mapping:

$$\Psi:\ \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^N. \tag{3.1.44}$$

When we take the bundle space transformation

$$\Omega:\ (\mathbb{C}_p^4)^N \to (\mathbb{C}_p^4)^N \qquad \text{for } \Omega \in SU(N),$$

then the mapping Ψ of (3.1.44) transforms as

$$\Psi \to \Omega\Psi \qquad \text{for } \Omega \in SU(N).$$

This is the $SU(N)$ gauge transformation for the Dirac spinors.

3. *Spinor transformation.* The $SU(N)$ gauge fields are the set of functions

$$G_\mu^a\ (1 \leqslant a \leqslant N^2-1) \quad \text{and} \quad \Psi = (\psi^1, \cdots, \psi^N)^{\mathrm{T}},$$

and for each a, G_μ^a is a 4-dimensional vector field

$$G_\mu^a:\ \mathscr{M} \to T\mathscr{M}.$$

Hence, the $SU(N)$ gauge fields (G_μ^a, Ψ) are the mapping

$$(G_\mu^a, \Psi):\ \mathscr{M} \to \mathscr{M} \otimes_p \left[(T_p\mathscr{M})^K \times (\mathbb{C}^4)^N\right] \qquad \text{for } K = N^2-1. \tag{3.1.45}$$

By the definition of spinors in Subsection 2.2.6, under the Lorentz transformation of (3.1.42), we have

$$\begin{aligned} &G_\mu^a \to \widetilde{L} G_\mu^a && \text{for } 1 \leqslant a \leqslant N^2-1, \\ &\psi^i \to R\psi^i && \text{for } 1 \leqslant i \leqslant N, \end{aligned} \tag{3.1.46}$$

and $\widetilde{L} = (L^{\mathrm{T}})^{-1}$, and R is the spinor representation matrix determined by the Lorentz matrix L, as given by (2.2.67).

Hence, for $SU(N)$ gauge fields (3.1.45), if we take the linear transformations for the bundle spaces of (3.1.45) as

$$\begin{aligned} &\widetilde{L}:\ T_p\mathscr{M} \to T_p\mathscr{M}, \\ &R:\ \mathbb{C}_p^4 \to \mathbb{C}_p^4, \end{aligned}$$

then the gauge fields (3.1.45) transform as (3.1.46).

Remark 3.4 The Lorentz transformations (3.1.42) and (3.1.43) are independent of $p \in \mathscr{M}$. Hence, the associated vector bundles in (3.1.41) are trivial. In other words, the vector bundles corresponding only to the Lorentz transformations are trivial. But, other vector bundles given above are in general nontrivial. □

3.1.4 Connections and covariant derivatives

In the last subsection, we see that a field

$$F:\ \mathscr{M}\to\mathscr{M}\otimes_p E^N \tag{3.1.47}$$

undergoes a transformation:

$$\widetilde{F}=T_pF,\quad p\in\mathscr{M}, \tag{3.1.48}$$

if the bundle space E_p^N undergoes a linear transformation

$$T_p:\ E_p^N\to E_p^N. \tag{3.1.49}$$

By PLD, for a physical field F as defined by (3.1.47), there is a Lagrangian action

$$L=\int\mathscr{L}(F,DF,\cdots,D^mF)dx. \tag{3.1.50}$$

Physical symmetry principles amount to saying that the action (3.1.50) is invariant under the transformation (3.1.48). This requires that the derivative DF in (3.1.50) be covariant. Namely, for (3.1.48) we have

$$\widetilde{D}\widetilde{F}=NDF,\qquad N\text{ is a matrix depends on }T_p. \tag{3.1.51}$$

If the transformation T_p in (3.1.48) is independent of $p\in\mathscr{M}$, then

$$D_\mu=\partial_\mu\quad(\partial_\mu=\partial/\partial x^\mu).$$

However, if T_p depends on $p\in M$, then we have

$$\widetilde{\partial}_\mu(T_pF)=T_p\widetilde{\partial}_\mu F+\widetilde{\partial}_\mu T_pF, \tag{3.1.52}$$

which violates the covariance of (3.1.51), because it has a superfluous term $\widetilde{\partial}_\mu T_pF$ in the right-hand side of (3.1.52). Hence, the derivative operator D_μ must be in the form

$$D_\mu=\partial_\mu+\Gamma_\mu, \tag{3.1.53}$$

such that (3.1.51) holds true. The field Γ_μ is the connection of the vector bundle $\mathscr{M}\otimes_p E^N$ under the transformation (3.1.49).

To make explicit of $\widetilde{\Gamma}_\mu$, we assume that $\widetilde{\partial}_\mu$ and ∂_μ have the following relation

$$\widetilde{\partial}_\mu=A\partial_\mu, \tag{3.1.54}$$

and A is a matrix in the following form

$$A=\begin{cases}\text{identity} & \text{for } SU(N)\text{ gauge fields},\\ \text{Lorentz matrix} & \text{for Lorentz tensors},\\ \text{affine matrix} & \text{for general tensors}.\end{cases} \tag{3.1.55}$$

By (3.1.48) and (3.1.53)-(3.1.54), we have

$$\widetilde{D}_\mu \widetilde{F} = (\widetilde{\partial}_\mu + \widetilde{\Gamma}_\mu)(T_p F) = AT_p \partial_\mu F + A\partial_\mu T_p F + \widetilde{\Gamma}_\mu T_p F.$$

Thanks to (3.1.51),

$$N(\partial_\mu + \Gamma_\mu)F = AT_p \partial_\mu F + A\partial_\mu T_p F + \widetilde{\Gamma}_\mu T_p F.$$

It follows that

$$N = AT_p, \qquad \widetilde{\Gamma}_\mu = AT_p \Gamma_\mu T_p^{-1} - A\partial_\mu T_p T_p^{-1}.$$

In other words, under the linear transformation (3.1.49), the system transforms as follows

$$\begin{aligned} &\widetilde{D}\widetilde{F} = (AT_p)DF, \\ &\widetilde{\Gamma} = AT_p \Gamma T_p^{-1} - A(\partial T_p)T_p^{-1}. \end{aligned} \tag{3.1.56}$$

The following summarizes the connections of all symmetry transformations:

1. *Connection for Lorentz group.* As the Lorentz transformation T_p is independent of $p \in \mathscr{M}$, the connections Γ_μ are zero:

$$\Gamma_\mu = 0 \qquad \text{for the Lorentz action.}$$

2. *Connection for $SU(N)$ group.* For the $SU(N)$ group action, (3.1.49) is

$$\Omega_p : (\mathbb{C}^4)^N \to (\mathbb{C}^4)^N, \quad \Omega_p \in SU(N), \quad p \in \mathscr{M}.$$

By (3.1.55), $A = I$. The connection of $SU(N)$ group is the gauge fields G^a_μ:

$$\Gamma_\mu = \left\{ igG^a_\mu \tau_a \mid \{\tau_a\}_{a=1}^{N^2-1} \text{ is a set of generators of } SU(N) \right\}.$$

Hence, relations (3.1.56) for the $SU(N)$ gauge fields G^a_μ are written as

$$\begin{aligned} &\widetilde{D}\widetilde{\Psi} = \Omega\Psi, && \Psi : \mathscr{M} \to M \otimes_p (\mathbb{C}^4)^N, \\ &\widetilde{G}^a_\mu \tau_a = \Omega G^a_\mu \tau_a \Omega^{-1} - \frac{1}{ig}\partial\Omega\Omega^{-1}, && \Omega \in SU(N). \end{aligned}$$

3. *Connections for general linear group $GL(n)$.* For (k,r)-tensors:

$$F : \mathscr{M} \to T^k_r \mathscr{M}, \tag{3.1.57}$$

the $GL(n)$ group action

$$A_p : T\mathscr{M} \to T\mathscr{M}, \quad A_p \in GL(n), \tag{3.1.58}$$

induces a linear transformation:

$$T_p = \underbrace{A_p \otimes \cdots \otimes A_p}_{k} \otimes \underbrace{A_p^{-1} \otimes \cdots \otimes A_p^{-1}}_{r} : \ T_r^k \mathscr{M} \to T_r^k \mathscr{M}, \tag{3.1.59}$$

which can be equivalently expressed a $K \times K$ matrix with $K = n^{k+r}$, and $\otimes$ is the tensor product of matrices defined by (3.1.67); see Remark 3.5. In this case, the matrix A of (3.1.54)-(3.1.55) is precisely the A_p in (3.1.58). Hence, by (3.1.56) we have

$$\widetilde{D}\widetilde{F} = (A \otimes T)DF, \tag{3.1.60}$$

where F is as in (3.1.57), T is as in (3.1.59), and A is as in (3.1.58).

The covariant derivative operator D depends on the indices k and r of bundle spaces $T_r^k M$, and are derived by induction.

4. *Derivative on $T\mathscr{M}$.* For $F = (F^1, \cdots, F^n)$,

$$D_j F^i = \partial_j F^i + \Gamma^i_{jl} F^l, \tag{3.1.61}$$

where Γ^i_{jl} are connections defined on $T\mathscr{M}$. As $\mathscr{M}$ is a Riemannian manifold, $\{\Gamma^i_{jk}\}$ are the Levi-Civita connection as given by (2.3.25). It follows from (3.1.56) that the connection of (3.1.61) transforms as

$$\widetilde{\Gamma} = A \otimes A \otimes (A^{\mathrm{T}})^{-1} \Gamma - A \otimes \partial A \otimes A^{-1}, \tag{3.1.62}$$

which is the the same as those of (2.3.23).

5. *Derivative on $T^*\mathscr{M}$.* Consider $F = (F_1, \cdots, F_n)$. The derivative operators satisfy that

$$\begin{aligned} &D(A \cdot B) = DA \cdot B + A \cdot DB, \\ &D(A \otimes B) = DA \otimes B + A \otimes DB. \end{aligned} \tag{3.1.63}$$

It is known that

$$F_k G^k = \text{ a scalar field, and } D(F_k G^k) = \partial(F_k G^k). \tag{3.1.64}$$

We infer then from (3.1.63) and (3.1.64) that

$$D_k F_i G^i + F_i D_k G^i = \partial_k F_i G^i + F_i \partial_k G^i. \tag{3.1.65}$$

Inserting (3.1.61) in (3.1.65) we deduce that

$$D_k F_i = \partial_k F_i - \Gamma^j_{ki} F_j.$$

6. *Derivative on $T_r^k\mathscr{M}$.* For two vector fields A^k and B^k, their tensor product $A \otimes B = \{A^i B^j\}$ is a second order tensor. By (3.1.63) we have

$$\begin{aligned} D_k(A \otimes B) =& D_k(A^i B^j) = D_k A^i B^j + A^i D_k B^j \\ =& (\partial_k A^i + \Gamma^i_{kl} A^l) B^j + A^i (\partial_k B^j + \Gamma^j_{kl} B^l) \\ =& \partial_k (A^i B^j) + \Gamma^i_{kl} A^l B^j + \Gamma^j_{kl} A^i B^l. \end{aligned}$$

Replacing $A \otimes B$ by T^{ij}, we obtain

$$D_k T^{ij} = \partial_k T^{ij} + \Gamma^i_{kl} T^{lj} + \Gamma^j_{kl} T^{il}.$$

In the same fashion, for general (k,r)-tensors

$$T = T^{i_1 \cdots i_k}_{j_1 \cdots j_r} : \mathscr{M} \to T^k_r \mathscr{M},$$

its covariant derivative can be expressed as

$$\begin{aligned} D_k T^{i_1 \cdots i_k}_{j_1 \cdots j_r} = & \partial_k T^{i_1 \cdots i_k}_{j_1 \cdots j_l} + \Gamma^{i_1}_{kl} T^{l i_2 \cdots i_k}_{j_1 \cdots j_r} + \cdots \\ & + \Gamma^{i_k}_{kl} T^{i_1 \cdots i_{k-1} l}_{j_1 \cdots j_r} - \Gamma^l_{kj_1} T^{i_1 \cdots i_k}_{j_2 \cdots j_r} - \cdots - \Gamma^l_{kj_r} T^{i_1 \cdots i_k}_{j_1 \cdots j_{k-1} l}. \end{aligned} \tag{3.1.66}$$

The derivative (3.1.66) were given in (2.3.26).

Remark 3.5 We have encountered tensor products in (3.1.59) and (3.1.62). A further explanation of this considered is now in order. Let A and B be two matrices with orders n and m respectively. Then $A \otimes B$ is a matrix of order $N = nm$, defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}, \qquad a_{ij}B = \begin{pmatrix} a_{ij}b_{11} & \cdots & a_{ij}b_{1m} \\ \vdots & & \vdots \\ a_{ij}b_{m1} & \cdots & a_{ij}b_{mm} \end{pmatrix}. \tag{3.1.67}$$

In (3.1.60) and (3.1.62), the components of $DF = \{D_k F^{i_1 \cdots i_k}_{j_1 \cdots j_r}\}$ and $\Gamma = \{\Gamma^k_{ij}\}$ are arranged to be in two vectorial forms. □

3.2 Analysis on Riemannian Manifolds

3.2.1 Sobolev spaces of tensor fields

The gravitational field equations are defined on a Riemannian manifold. To study these equations, it is necessary to introduce various types of function spaces on manifolds, which possess different differentiability. In particular, we need to introduce the concept of weak derivatives.

Let $\mathscr{M}$ be an n-dimensional manifold with metric $\{g_{ij}\}$. The following functions are defined on $\mathscr{M}$.

1. *L^p spaces.* For a real number p $(1 \leqslant p < \infty)$, we denote

$$L^p(\mathscr{M} \otimes_p E^N) = \left\{ u : \mathscr{M} \to \mathscr{M} \otimes_p E^N \;\middle|\; \int_{\mathscr{M}} |u|^p \sqrt{-g} dx < \infty \right\}, \tag{3.2.1}$$

where $\sqrt{-g}dx$ is the volume element, and $|u|$ is the modulus of u. For example, if $u : \mathscr{M} \to T\mathscr{M}$ is a vector field, then

$$|u| = |g_{ij} u^i u^j|^{1/2}.$$

The space (3.2.1) is endowed with L^p-norm as

$$||u||_{L^p} = \left[\int_{\mathscr{M}} |u|^p \sqrt{-g} dx\right]^{1/p}.$$

By Functional Analysis, the spaces $L^p(\mathscr{M} \otimes_p E^N)$ are Banach spaces, and for $1 < p < \infty$, $L^p(\mathscr{M} \otimes_p E^N)$ are reflective and separable. The dual spaces of $L^p(\mathscr{M} \otimes_p E^N)$ $(1 < p < \infty)$ are

$$\begin{aligned} &L^p(\mathscr{M} \otimes_p E^N)^* = L^q(\mathscr{M} \otimes_p (E^N)^*), \\ &L^q(\mathscr{M} \otimes_p E^N)^* = L^p(\mathscr{M} \otimes_p (E^N)^*), \end{aligned} \qquad \frac{1}{p} + \frac{1}{q} = 1,$$

where $(E^N)^*$ is the dual space of E^N.

For $p = \infty$, we define that

$$L^\infty(\mathscr{M} \otimes_p E^N) = \{u: \ M \to \mathscr{M} \otimes_p E^N | \ u \text{ is bounded almost everywhere}\}.$$

The norm of $L^\infty(\mathscr{M} \otimes_p E^N)$ is defined by

$$||u||_{L^\infty} = \sup_{\mathscr{M}} |u|.$$

The space $L^\infty(\mathscr{M} \otimes_p E^N)$ is a Banach space, but not reflective and separable. The space $L^\infty(\mathscr{M} \otimes_p E^N)$ is the dual space of $L^1(\mathscr{M} \otimes_p E^N)$, i.e.

$$L^1(\mathscr{M} \otimes_p E^N)^* = L^\infty(\mathscr{M} \otimes_p (E^N)^*).$$

2. *Weakly differentiable functions.* A field $u \in L^p(\mathscr{M} \otimes_p E^N)$ is called k-th order weakly differentiable, if each component u_j is k-th order weakly differentiable, i.e. for each u_j there exists uniquely a function φ such that for all $v \in C_0^\infty(\mathscr{M})$ we have

$$\int_{\mathscr{M}} \varphi v \sqrt{-g} dx = (-1)^k \int_{\mathscr{M}} u_j \partial^k v dx. \tag{3.2.2}$$

In this case, φ is called the k-th weak derivative of u_j, denoted by

$$\varphi = \partial^k u_j. \tag{3.2.3}$$

The space $C_0^\infty(\mathscr{M})$ consists of infinitely differentiable functions, and

$$C_0^\infty(\mathscr{M}) = \begin{cases} C^\infty(\mathscr{M}) & \text{if } \mathscr{M} \text{ is compact and } \partial\mathscr{M} = \varnothing, \\ \{u \in C^\infty(\mathscr{M}) | \ u \neq 0 \text{ in a compact set of } \mathscr{M}\}. \end{cases} \tag{3.2.4}$$

The definition (3.2.2)-(3.2.3) for weak derivatives is very abstract. In the following, we discuss the distinction between continuity and weak differentiability in an intuitive fashion.

Let u be a function defined on $\mathbb{R}^n$. It is known that if u is differentiable at $x=0$, then u can be Taylor expanded as

$$u(x) = ax + o(|x|), \tag{3.2.5}$$

where $a = \nabla u(0)$ is the first order derivative of u at $x=0$, i.e. the gradient of u at $x=0$.

However, if u is weakly but not continuously differentiable at $x=0$, then in a neighborhood of $x=0, u$ must contain at least a term as $|x_i|^\alpha$ $(\alpha \leqslant 1)$. Without loss of generality, we express u in the form

$$u = a|x|^\alpha + \text{continuously differentiable terms}, \tag{3.2.6}$$

where $a \neq 0$ is a constant, and $1-n < \alpha \leqslant 1$. The index α in (3.2.6) determines the regularity of u as follows:

$$\begin{aligned} u &= \text{Lipschitz}, && \text{if } \alpha = 1, \\ u &= \text{Hölder}, && \text{if } 0 < \alpha < 1, \\ u &= \text{singularity}, && \text{if } \alpha < 1. \end{aligned} \tag{3.2.7}$$

Expressions (3.2.5) and (3.2.6) exhibit the essential difference between continuity and weak differentiability.

Remark 3.6 For a weakly differentiable function as (3.2.6), its index α has to satisfy $1-n < \alpha \leqslant 1$, which is crucial in the Sobolev embedding theorems in the next subsection. □

3. $W^{k,p}$ *spaces* (*Sobolev spaces*). Let $u : \mathscr{M} \to \mathscr{M} \otimes_p E^N$. Then, $u = \{u_j |\ 1 \leqslant j \leqslant N\}$, and each component u_j of u is a function on M. We denote

$$\partial^\alpha u_j = \frac{\partial^k u_j}{(\partial x^1)^{\alpha_1} \cdots (\partial x^n)^{\alpha_n}} \qquad \text{for } \alpha = (\alpha_1, \cdots, \alpha_n),$$

and $k = |\alpha| = \sum_{i=1}^n \alpha_i$ $(\alpha_i \geqslant 0)$. Then, we define $W^{k,p}$ spaces as

$$W^{k,p}(\mathscr{M} \otimes_p E^N) = \{u \in L^p(\mathscr{M} \otimes_p E^N) |\ \partial^\beta u_j \in L^p(\mathscr{M}),\ 1 \leqslant j \leqslant n,\ 0 \leqslant \beta \leqslant k\}, \tag{3.2.8}$$

and $\partial^\beta u_j$ in (3.2.8) are weak derivatives of u_j.

The spaces $W^{k,p}(\mathscr{M} \otimes_p E^N)$ are called Sobolev spaces, and the norms are defined by

$$||u||_{W^{k,p}} = \sum_{\beta=0}^k \left[\int_{\mathscr{M}} |D^\beta u|^p \sqrt{-g} dx \right]^{1/p}. \tag{3.2.9}$$

It is clear that

$$u \in W^{k,p}(\mathscr{M} \otimes_p E^N) \ \Rightarrow\ u \text{ is } k\text{-th weakly differentiable.}$$

In addition, we introduce the spaces $W_0^{k,p}(\mathscr{M} \otimes_p E^N)$ as

$$W_0^{k,p}(\mathscr{M}\otimes_p E^N) = \text{Closure of } C_0^\infty(\mathscr{M}\otimes_p E^N) \text{ under } W^{k,p} \text{ norm (3.2.9)}.$$

Here $C_0^\infty(\mathscr{M}\otimes_p E^N)$ is as defined in (3.2.4).

If $\partial M = \varnothing$, then $W_0^{k,p}(\mathscr{M}\otimes_p E^N) = W^{k,p}(\mathscr{M}\otimes_p E^N)$, and if $\partial\mathscr{M} \neq \varnothing$ then for $u \in W_0^{k,p}(\mathscr{M}\otimes_p E^N)$ we have

$$u|_{\partial\mathscr{M}} = 0, \ \cdots, \ \partial^\beta u|_{\partial\mathscr{M}} = 0, \qquad \forall |\beta| \leqslant k-1.$$

4. *H^k spaces.* As $\mathscr{M}$ is a Riemannian manifold, $\mathscr{M}\otimes_p E^N$ and its dual bundle $\mathscr{M}\otimes_p (E^N)^*$ are isomorphic. In this case, the spaces $W^{k,2}(\mathscr{M}\otimes_p E^N)$ are Hilbert spaces, denoted by

$$\begin{aligned} H^k(\mathscr{M}\otimes_p E^N) &= W^{k,2}(\mathscr{M}\otimes_p E^N), \\ H_0^k(\mathscr{M}\otimes_p E^N) &= W_0^{k,2}(\mathscr{M}\otimes_p E^N). \end{aligned} \tag{3.2.10}$$

The inner products of (3.2.10) are defined by

$$\langle u,v\rangle_{H^k} = \int_{\mathscr{M}} \sum_{|\beta|=0}^{k} D^\beta u \cdot D^\beta v^* \sqrt{-g}dx$$

where $v^* \in H^k(\mathscr{M}\otimes_p (E^N)^*)$ is the dual field of $v \in H^k(\mathscr{M}\otimes_p E^N)$.

5. *Lipschitz spaces.* Let $k \geqslant 0$ be integrals. The Lipschitz space $C^{k,1}(\mathscr{M}\otimes_p E^N)$ consists of all k-th order continuously differentiable functions u with $D^k u$ being Lipschitz continuous:

$$C^{k,1}(\mathscr{M}\otimes_p E^N) = \{u \in C^k(\mathscr{M}\otimes_p E^N) | \ [\partial^k u]_{\text{Lip}} < \infty\},$$

where $[\partial^k u]_{\text{Lip}}$ is the Lipschitz modulus, defined by

$$[v]_{\text{Lip}} = \sup_{x,y\in\mathscr{M},\ x\neq y} \frac{|v(x)-v(y)|}{|x-y|}.$$

A Lipschitz continuous function $u \in C^{0,1}(\mathscr{M}\otimes_p E^N)$ is as shown in (3.2.6)-(3.2.7) with $\alpha = 1$.

6. *Hölder spaces.* The Hölder space $C^{k,\alpha}(\mathscr{M}\otimes_p E^N)$ $(0 < \alpha < 1)$ consists of all k-th order continuously differentiable functions u with $D^k u$ being Hölder continuous:

$$C^{k,\alpha}(\mathscr{M}\otimes_p E^N) = \{u \in C^k(\mathscr{M}\otimes_p E^N) | \ [D^k u]_\alpha < \infty\},$$

and $[v]_\alpha$ is the Hölder modulus:

$$[v]_\alpha = \sup_{x,y\in\mathscr{M},\ x\neq y} \frac{|v(x)-v(y)|}{|x-y|^\alpha}, \quad 0 < \alpha < 1.$$

A Hölder continuous function $u \in C^{0,\alpha}(\mathscr{M}\otimes_p E^N)$ is as shown in (3.2.6)–(3.2.7) with $0 < \alpha < 1$.

The norm of $C^{k,\alpha}(\mathscr{M}\otimes_p E^N)(0<\alpha\leqslant 1)$ is given by

$$||u||_{C^{k,\alpha}}=||u||_{C^k}+[D^k u]_\alpha \quad (0<\alpha\leqslant 1),$$

where $||\cdot||_{C^k}$ is the norm of $C^k(\mathscr{M}\otimes_p E^N)$:

$$||u||_{C^k}=\sum_{|\beta|=0}^{k}\sup_{\mathscr{M}}|\partial^\beta u|.$$

3.2.2 Sobolev embedding theorem

In (3.2.6)–(3.2.7) we see that for $\Omega\subset\mathbb{R}^n$, a function $u\in W^{1,p}(\Omega)$ does't imply that $u\in L^q(\Omega)$ for any $q>p$. For example, for the function

$$u=|x|^{-\alpha},\quad 0<\alpha<\frac{1}{2},\quad x\in\Omega\subset\mathbb{R}^3,\quad \Omega \text{ bounded},$$

it is known that

$$\nabla u=|x|^{-\frac{\alpha+4}{2}}x\in L^2(\Omega).$$

Obviously we have

$$u\in W^{1,2}(\Omega),\quad u\notin L^q(\Omega),\qquad \forall q>\frac{n}{\alpha}.$$

The following embedding problem of Sobolev spaces provides a solution for this problem.

Theorem 3.7 (Sobolev Embedding Theorem) *Let $\mathscr{M}$ be an n-dimensional compact manifold. Then we have the embeddings:*

$$W^{1,p}(\mathscr{M}\otimes_p E^N)\hookrightarrow\begin{cases} L^q(\mathscr{M}\otimes_p E^N) & \text{for } 1\leqslant q\leqslant\dfrac{np}{n-p}, & \text{if } n>p,\\ L^q(\mathscr{M}\otimes_p E^N) & \text{for } 1\leqslant q<\infty, & \text{if } n=p,\\ C^{0,\alpha}(\mathscr{M}\otimes_p E^N) & \text{for } \alpha=1-n/p, & \text{if } n<p.\end{cases} \tag{3.2.11}$$

Here $C^{0,\alpha}(\mathscr{M}\otimes_p E^N)$ are the Hölder spaces. Moreover we have the following inequalities for the norms:

$$\begin{aligned} &||u||_{L^q}\leqslant C||u||_{W^{1,p}} && \text{for } 1\leqslant q\leqslant\frac{np}{n-p}, && \text{if } n>p,\\ &||u||_{L^q}\leqslant C||u||_{W^{1,p}} && \text{for } 1\leqslant q<\infty, && \text{if } n=p,\\ &||u||_{C^{0,\alpha}}\leqslant C||u||_{W^{1,p}} && \text{for } \alpha=1-n/p, && \text{if } n<p,\end{aligned} \tag{3.2.12}$$

where $C>0$ are constants depending on n,p and M.

Remark 3.8 When $\mathscr{M}$ is non-compact, the conclusions (3.2.11) and (3.2.12) are also valid only for q satisfying

$$\begin{aligned} &p\leqslant q\leqslant\frac{np}{n-p} && \text{if } n>p,\\ &p\leqslant q<\infty && \text{if } n=p.\end{aligned}$$

Remark 3.9 By the recurrence relations

$$W^{k,p}(\mathscr{M}\otimes_p E^N)\hookrightarrow W^{k-1,q}(\mathscr{M}\otimes_p E^N),$$

we readily deduce from (3.2.11) that

$$W^{k,p}(\mathscr{M}\otimes_p E^N)\hookrightarrow\begin{cases}L^q(\mathscr{M}\otimes_p E^N) & \text{for } 1\leqslant q\leqslant \dfrac{np}{n-kp}, \quad \text{if } n>kp,\\ L^q(\mathscr{M}\otimes_p E^N) & \text{for } 1\leqslant q<\infty, \quad \text{if } n=kp,\\ C^{\mathscr{M},\alpha}(M\otimes_p E^N) & \text{for } m+\alpha=k-n/p, \quad \text{if } n<kp.\end{cases}$$

Based on weakly differentiability properties (3.2.6) and (3.2.7), essence of Theorem 3.7 can be seen from embeddings (3.2.11) using the following function:

$$u(x)=\frac{|x|^{\alpha}}{|\ln|x||^{\beta}},\quad x\in\mathbb{R}^n, \tag{3.2.13}$$

where $\beta>1$ is given, and the exponent $\alpha<1$ is to reflect the critical embedding index q^* in (3.2.11).

The derivatives of u given by (3.2.13) are as follows

$$\begin{aligned}&\nabla u=(\partial_1 u,\cdots,\partial_n u),\\ &\partial_i u=\left(\frac{\alpha}{|\ln|x||^{\beta}}-\frac{\beta}{|\ln|x||^{\beta+1}}\right)|x|^{\alpha-2}x_i.\end{aligned} \tag{3.2.14}$$

Let $B_R=\{x\in\mathbb{R}^n|\ |x|<R\},\ 0<R<1$. Assume that

$$u\in W^{1,p}(B_R)\quad \text{for some } 1\leqslant p<\infty.$$

Then, by (3.2.14) we have

$$\int_{B_R}|\nabla u|^p dx\leqslant C\int_{B_R}\frac{|x|^{(\alpha-1)p}}{|\ln|x||^{\beta p}}dx. \tag{3.2.15}$$

In the spherical coordinate system,

$$dx=r^{n-1}drds,$$

and ds is the area element of the unit sphere, (3.2.15) becomes

$$\int_{B_R}|\nabla u|^p dx\leqslant C\int_0^R\int_{S^{n-1}}\frac{r^{n+(\alpha-1)p-1}}{|\ln r|^{\beta p}}drds\leqslant C\int_0^R\frac{r^k}{|\ln|r||^{\beta p}}dr, \tag{3.2.16}$$

for $0<R<1$ and $\beta p>1$. It follows from (3.2.16) that

$$\int_{B_R}|\nabla u|^p dx<\infty\ \Leftrightarrow\ k=n+(\alpha-1)p-1\geqslant -1.$$

Hence we obtain that

$$u = \frac{|x|^\alpha}{|\ln|x||^\beta} \in W^{1,p}(B_R) \ \Leftrightarrow \ \alpha \geqslant 1 - \frac{n}{p}. \tag{3.2.17}$$

On the other hand, in the same fashion we see that

$$u = \frac{|x|^\alpha}{|\ln|x||^\beta} \in L^q(B_R) \ \Leftrightarrow \ \alpha \geqslant -\frac{n}{q}. \tag{3.2.18}$$

Hence, by (3.2.17) we can see that as $p > n$,

$$u \in C^{0,\alpha}(B_R), \quad \alpha = 1 - \frac{n}{p},$$

and as $p \leqslant n$, then from (3.2.17) and (3.2.18), at the critical embedding exponent q^* we have

$$\alpha = 1 - \frac{n}{p}, \quad \alpha = -\frac{n}{q^*}.$$

It follows that

$$q^* \begin{cases} = \dfrac{np}{n-p} & \text{for } n > p, \\ < \infty & \text{for } n = p. \end{cases}$$

Thus, we deduce that

$$u \in W^{1,p}(B_R) \ \Rightarrow \ \begin{cases} u \in L^q(B_R) & \text{for } 1 \leqslant q \leqslant \dfrac{np}{n-p}, & \text{if } n > p, \\ u \in L^q(B_R) & \text{for } 1 \leqslant q < \infty, & \text{if } n = p, \\ u \in C^{0,\alpha}(B_R) & \text{for } \alpha = 1 - \dfrac{n}{p}, & \text{if } p > n. \end{cases} \tag{3.2.19}$$

The relations (3.2.19) are the embeddings given by (3.2.11).

3.2.3 Differential operators

A differential operator defined on a manifold $\mathscr{M}$ is a mapping given by

$$G: \ W^{k,p}(\mathscr{M} \otimes_p E_1^{N_1}) \ \to \ L^p(\mathscr{M} \otimes_p E_2^{N_2}) \qquad \text{for some } k \geqslant 1,$$

and is called a k-th order differential operator.

The most important operators in physics are:

1) the gradient operator: ∇,

2) the divergent operator: div,

3) the Laplace operator: $D^k D_k = \text{div} \cdot \nabla$,

4) the Laplace-Beltrami operator: $\Delta = d\delta + \delta d$,

5) the wave operator: $\Box = D^\mu D_\mu$, with D^μ being the 4-D gradient operator.

We now give a detailed account on the above operators.

1. *Gradient operator.* The gradient operator ∇ is a mapping of the following types:

$$\begin{aligned}&\nabla^k: \ W^{1,p}(T^k_r\mathscr{M}) \to L^p(T^{k+1}_r\mathscr{M}),\\ &\nabla_k: \ W^{1,p}(T^k_r\mathscr{M}) \to L^p(T^k_{r+1}\mathscr{M}),\end{aligned} \tag{3.2.20}$$

and ∇^k and ∇_k have the relation

$$\nabla_k = g_{kl}\nabla^l, \quad \nabla^k = g^{kl}\nabla_l,$$

and $\{g_{kl}\}$ the Riemann metric of $\mathscr{M}$. ∇ is expressed as

$$\begin{aligned}&\nabla_k = (D_1, \cdots, D_n),\\ &D_j \text{ the covariant derivative operators.}\end{aligned} \tag{3.2.21}$$

2. *Divergence operator.* The divergence operator div is a mapping of the following types:

$$\begin{aligned}&\text{div}: \ W^{1,p}(T^{k+1}_r\mathscr{M}) \to L^p(T^k_r\mathscr{M}),\\ &\text{div}: \ W^{1,p}(T^k_{r+1}\mathscr{M}) \to L^p(T^k_r\mathscr{M}).\end{aligned} \tag{3.2.22}$$

For $T = \{T^{i_1\cdots i_{k+1}}_{j_1\cdots j_r}\} \in W^{1,p}(T^{k+1}_r\mathscr{M})$ and $T = \{T^{i_1\cdots i_k}_{j_1\cdots j_{r+1}}\} \in W^{1,p}(T^k_{r+1}M)$,

$$\begin{aligned}&\text{div } T = D_{i_l} T^{i_1\cdots i_l \cdots i_{k+1}}_{j_1\cdots j_r}, \quad \text{and}\\ &\text{div } T = D^{j_l} T^{i_1\cdots i_k}_{j_1\cdots j_l\cdots j_{k+1}}\end{aligned} \tag{3.2.23}$$

As $u \in W^{1,p}(T\mathscr{M})$ and $u \in W^{1,p}(T^*\mathscr{M})$, we can give the expressions of div u in the following.

Let $u \in W^{1,p}(T\mathscr{M}), u = (u^1, \cdots, u^n)$. Then by (3.2.23),

$$\text{div } u = D_k u^k = \frac{\partial u^k}{\partial x^k} + \Gamma^k_{kj} u^j.$$

By the Levi-Civita connections (2.3.25), the contraction

$$\Gamma^k_{kj} = \frac{1}{2} g^{kl} \frac{\partial g_{kl}}{\partial x^j} = \frac{1}{2g}\frac{\partial g}{\partial x^j} = \frac{1}{\sqrt{-g}}\frac{\partial\sqrt{-g}}{\partial x^j},$$

here $g = \det\,(g_{ij})$. Thus we have

$$\text{div } u = \frac{\partial u^k}{\partial x^k} + \frac{1}{\sqrt{-g}}\frac{\partial\sqrt{-g}}{\partial x^k} u^k = \frac{1}{\sqrt{-g}}\frac{\partial(\sqrt{-g}u^k)}{\partial x^k}. \tag{3.2.24}$$

Let $u \in W^{1,p}(T^*\mathscr{M}), u = (u_1, \cdots, u_n)$. Then,

$$\begin{aligned}\text{div } u =& D^k u_k = g^{kl} D_l(u_k) = D_l(g^{lk} u_k) \\ =& \frac{1}{\sqrt{-g}} \frac{\partial}{\partial x^l}(\sqrt{-g} g^{lk} u_k) \quad (\text{by } (3.2.24)).\end{aligned} \tag{3.2.25}$$

The formula (3.2.24) and (3.2.25) give expression of div u as:

$$\text{div } u = \begin{cases} \dfrac{1}{\sqrt{-g}} \dfrac{\partial}{\partial x^k}(\sqrt{-g} u^k) & \text{for} u \in W^{1,p}(T\mathscr{M}), \\ \dfrac{1}{\sqrt{-g}} \dfrac{\partial}{\partial x^k}(\sqrt{-g} g^{kl} u_l) & \text{for} u \in W^{1,p}(T^*\mathscr{M}). \end{cases} \tag{3.2.26}$$

3. *Laplace operators.* The Laplace operator $D^k D_k$ in $\mathbb{R}^n$ is in the familiar form

$$D^k D_k = \frac{\partial}{\partial x^k} \frac{\partial}{\partial x^k}.$$

However, the Laplace operators $D^k D_k$ defined on a Riemannian manifold are usually very complex.

By (3.2.20) and (3.2.22), we have

$$\begin{aligned}&W^{2,p}(T_r^k \mathscr{M}) \xrightarrow{\nabla_k} W^{1,p}(T_{r+1}^k \mathscr{M}) \xrightarrow{\text{div}} L^p(T_r^k \mathscr{M}), \\ &W^{2,p}(T_r^k \mathscr{M}) \xrightarrow{\nabla^k} W^{1,p}(T_r^{k+1} \mathscr{M}) \xrightarrow{\text{div}} L^p(T_r^k \mathscr{M}).\end{aligned}$$

Hence, the Laplace operator div $\cdot \nabla$ is the mapping:

$$\text{div} \cdot \nabla : \ W^{2,p}(T_r^k \mathscr{M}) \to L^p(T_r^k \mathscr{M}). \tag{3.2.27}$$

By (3.2.21) and (3.2.23), div $\cdot \nabla$ is written as

$$\text{div} \cdot \nabla = D^k D_k = g^{kl} D_k D_l.$$

4. *Expression of* div $\cdot \nabla$ *on* $\mathscr{M} \otimes_p \mathbb{R}^1$. A scalar field u on M can be regarded as $u : \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{R}^1$. In this case, ∇u is written as

$$\nabla u = \left(\frac{\partial u}{\partial x^1}, \cdots, \frac{\partial u}{\partial x^n} \right).$$

By (3.2.26), we get that

$$\text{div} \cdot \nabla u = \frac{1}{\sqrt{-g}} \frac{\partial}{\partial x^k} \left(\sqrt{-g} g^{kl} \frac{\partial u}{\partial x^l} \right). \tag{3.2.28}$$

5. *Expression of* $\text{div}\cdot\nabla$ *on* $T\mathscr{M}$. A vector field $u:\mathscr{M}\to T\mathscr{M}$ can be written as

$$u=(u^1,\cdots,u^n),$$

and its gradient (i.e. covariant derivatives) read as

$$\begin{aligned}&\nabla u=\{D_k u^i\},\\&D_k u^i=\frac{\partial u^i}{\partial x^k}+\Gamma^i_{kl}u^l,\end{aligned}\tag{3.2.29}$$

and the divergence of ∇u is

$$\text{div}\cdot\nabla u^i=D^kD_ku^i=g^{kl}D_l(D_ku^i)=g^{kl}\left[\frac{\partial}{\partial x^l}(D_ku^i)+\Gamma^i_{lj}D_ku^j-\Gamma^j_{kl}D_ju^i\right].$$

Then, by (3.2.29) we obtain that

$$\begin{aligned}\text{div}\cdot\nabla u=g^{kl}&\left[\frac{\partial}{\partial x^l}\left(\frac{\partial u^i}{\partial x^k}+\Gamma^i_{kj}u^j\right)+\Gamma^i_{lj}\left(\frac{\partial u^j}{\partial x^k}+\Gamma^j_{ks}u^s\right)\right.\\&\left.-\Gamma^j_{kl}\left(\frac{\partial u^i}{\partial x^j}+\Gamma^i_{js}u^s\right)\right].\end{aligned}\tag{3.2.30}$$

6. *Laplace-Beltrami operators.* The Laplace-Beltrami operators are defined by

$$\Delta=d\delta+\delta d,$$

where d is the differential operator, and δ the Hodge operator, which are defined on the spaces of all differential forms. For a scalar field, Δ is the same as $\text{div}\cdot\nabla$, i.e.

$$\Delta u=\text{div}\cdot\nabla u,\quad \text{as } u\in W^{2,p}(\mathscr{M}\otimes_p\mathbb{R}^1).$$

For vector fields and covector fields, we have

$$\begin{aligned}&\Delta u^i=-\text{div}(\nabla u^i)-g^{ij}R_{jk}u^k,\\&\Delta u_i=-\text{div}(\nabla u_j)-g^{kj}R_{ij}u_k,\end{aligned}\tag{3.2.31}$$

where R_{ij} is the Ricci curvature tensor defined by (2.3.31), and in (3.2.30) we give the formula of div (∇u^k). The expression of div (∇u_i) is written as

$$\text{div }(\nabla u_i)=g^{kl}\left\{\frac{\partial}{\partial x^l}\left[\frac{\partial u_i}{\partial x^k}-\Gamma^j_{ik}u_j\right]-\Gamma^j_{lk}\left[\frac{\partial u_i}{\partial x^j}-\Gamma^r_{ij}u_r\right]-\Gamma^j_{li}\left[\frac{\partial u_j}{\partial x^k}-\Gamma^r_{jk}u_r\right]\right\}.\tag{3.2.32}$$

Physically, it suffices to introduce the Laplace-Beltrami operators Δ only for vector and covector fields, i.e. the formula in (3.2.31).

Remark 3.10 The Navier-Stokes equations governing a fluid flow on a sphere (e.g. the surface of a planet) are expressed in the spherical coordinate system, and the viscosity differential operators are the Laplace-Beltrami operators given by (3.2.31); see Chapter 7 for details.

3.2.4 Gauss formula

If $\mathscr{M} = \mathbb{R}^n$, it is known that the Gauss formula is

$$\int_{\mathbb{R}^n} u \cdot \nabla f dx = -\int_{\mathbb{R}^n} f \text{div } u dx, \qquad \forall u \in H^1\ (\mathbb{R}^n \times \mathbb{R}^n),\ f \in H^1(\mathbb{R}^n),$$

which can be generalized to tensor fields on Riemann manifolds.

Let $\mathscr{M}$ be a Riemannian manifold with $\partial\mathscr{M} = \varnothing$. Then, there is an inner product defined on $T\mathscr{M}$ defined by

$$(X,Y) = g_{ij}X^iX^j, \qquad \forall X,Y \in T\mathscr{M}.$$

Also, there is an inner product on T^*M:

$$(X^*,Y^*) = g^{ij}X_i^*Y_j^*, \qquad \forall X^*,Y^* \in T^*\mathscr{M}.$$

The tensor bundle $T_r^k\mathscr{M}$ is

$$T_r^k\mathscr{M} = \underbrace{T\mathscr{M} \otimes \cdots \otimes T\mathscr{M}}_{k} \otimes \underbrace{T^*\mathscr{M} \otimes \cdots \otimes T^*\mathscr{M}}_{r}.$$

The inner products on $T\mathscr{M}$ and $T^*\mathscr{M}$ induce a natural inner product on $T_r^k\mathscr{M}$:

$$(u,v) = g_{i_1j_1} \cdots g_{i_kj_k} g^{l_1s_1} \cdots g^{l_rs_r} u^{i_1\cdots i_k}_{l_1\cdots l_r} v^{j_1\cdots j_r}_{s_1\cdots s_r}, \qquad \forall u,v \in T_r^k\mathscr{M}. \tag{3.2.33}$$

Hence we can define the inner product $\langle\cdot,\cdot\rangle$ on $L^2(T_r^k\mathscr{M})$ by

$$\langle u,v\rangle_{L^2} = \int_{\mathscr{M}} (u,v)\sqrt{-g}dx, \qquad \forall u,v \in L^2(T_r^k\mathscr{M}), \tag{3.2.34}$$

where (u,v) is as in (3.2.33).

Furthermore, we can also define an inner product on the spaces $H^m(T_r^k\mathscr{M})$ as follows

$$\langle u,v\rangle_{H^m} = \int_{\mathscr{M}} [(-1)^m(\Delta^m u,v) + (u,v)]\sqrt{-g}dx, \qquad \forall u,v \in H^m(T_r^k\mathscr{M}), \tag{3.2.35}$$

where $\Delta = \text{div} \cdot \nabla$.

In the following, we give the Gauss formula on $T_r^k\mathscr{M}$, which are crucial for the orthogonal decomposition theory in the next section.

Theorem 3.11 (Gauss Formula) *For any $u \in H^1(T_r^k\mathscr{M})$ and $v \in H^1(T_r^{k+1}\mathscr{M})$ (or $v \in H^1(T_{r+1}^k\mathscr{M})$),*

$$\int_{\mathscr{M}} (\nabla u,v)\sqrt{-g}dx = -\int_{\mathscr{M}} (u,\text{div}v)\sqrt{-g}dx. \tag{3.2.36}$$

Formula (3.2.36) is a corollary of the classical Gauss formula

$$\int_{\mathscr{M}} \text{div}\, w\sqrt{-g}dx = \int_{\partial\mathscr{M}} w\cdot nds. \tag{3.2.37}$$

In fact, let $w^k = (u, v^k)$, then by (2.3.28) we have

$$\text{div}\, w = D_k(u, v^k) = (\nabla u, v) + (u, \text{div}\, v).$$

If $\partial\mathscr{M} = \varnothing$, then we derive from (3.2.37) that

$$\int_{\mathscr{M}} \text{div}\, w\sqrt{-g}dx = \int_{\mathscr{M}} [(\nabla u, v) + (u, \text{div}\, v)]\sqrt{-g}dx = 0,$$

which is the formula (3.2.36).

The Gauss formula (3.2.36) can be generalized to more general gradient and divergent operators, denoted by D_A and div_A.

Let A be a vector field or a covector field, and $u \in L^2(T^k_r\mathscr{M})$. We define the operators D_A and div_A by

$$\begin{aligned} &D_A u = Du + u\otimes A,\\ &\text{div}_A u = \text{div} u - u\cdot A. \end{aligned} \tag{3.2.38}$$

Based on (3.2.36), it is readily to verify that the following formula holds true for the operators (3.2.38).

$$\int_{\mathscr{M}} (D_A u, v)\sqrt{-g}dx = -\int_{\mathscr{M}} (u, \text{div}_A v)\sqrt{-g}dx. \tag{3.2.39}$$

Remark 3.12 The motivation to generalize the Gauss formulas (3.2.36) to the operators D_A and div_A is to develop a new unified field theory for the fundamental interactions. The vector fields A in (3.2.38) and (3.2.39) represent gauge fields in the interaction field equations, and lead to a mass generation mechanism based on a first principle, called PID, different from the famous Higgs mechanism.

3.2.5 Partial differential equations on Riemannian manifolds

To develop an orthogonal decomposition theory for general (k, r)-tensor fields, we need to introduce the existence theorems for linear elliptic and hyperbolic equations on closed Riemann and Minkowski manifolds. The existence results are well-known. In the following, we give the definition of weak solutions and the basic existence theorems for PDEs without proofs.

Linear elliptic equations

Consider the following PDEs defined on a Riemannian manifold $\{\mathscr{M}, g_{ij}\}$ with $\partial\mathscr{M} = \varnothing$:

$$g^{ij}D_iD_ju = \text{div}\, f + g, \tag{3.2.40}$$

where D_i are the covariant derivative operators, the unknown function $u : \mathscr{M} \to T_r^k \mathscr{M}$ is a (k,r) tensor field, $g : \mathscr{M} \to T_r^k \mathscr{M}$ and $f : \mathscr{M} \to T_r^{k+1} \mathscr{M}$ (or $f : \mathscr{M} \to T_{r+1}^k \mathscr{M}$) are given.

We need to introduce the concept of weak solutions for (3.2.40).

Definition 3.13 *Let $f \in L^2(T_r^{k+1}\mathscr{M})$(or $f \in L^2(T_{r+1}^k \mathscr{M})$) and $g \in L^2(T_r^k \mathscr{M})$. A field $u \in H^1(T_r^k \mathscr{M})$ is called a weak solution of (3.2.40), if for all $v \in H^1(T_r^k \mathscr{M})$ the following equality holds true,*

$$\int_{\mathscr{M}} (\nabla u, \nabla v)\sqrt{-g}dx = \int_{\mathscr{M}} [(f,\nabla v) - (g,v)]\sqrt{-g}dx,$$

where $(\cdot,\cdot)$ is the inner product as defined in (3.2.33).

The following existence theorem is a classical result, which is a corollary of the well-known Fredholm Alternative Theorem.

Theorem 3.14 *Let the metric g_{ij} be $W^{2,\infty}$, and f,g be L^2. If*

$$\int_{\mathscr{M}} [(f,\nabla\phi) - (g,\phi)]\sqrt{-g}dx = 0 \tag{3.2.41}$$

holds for all ϕ satisfying

$$g^{ij} D_i D_j \phi = 0, \tag{3.2.42}$$

then the equation (3.2.40) possesses a weak solution $u \in H^1(T_r^k \mathscr{M})$. In particular, if g_{ij}, f, g are C^∞, then $u \in C^\infty(T_r^k \mathscr{M})$.

Two remarks are now in order. First, the solutions $\phi \neq 0$ of (3.2.42) are the eigenfunctions of the Laplace operator $\text{div} \cdot \nabla = D^i D_i$ corresponding to the zero eigenvalue $\lambda = 0$, and the condition (3.2.41) represents that $\text{div}\, f + g$ is orthogonal to all eigenfunctions for $\lambda = 0$ of $\text{div} \cdot \nabla$.

Second, by the L^p-estimate theorem, if $g_{ij} \in W^{m+2,p}, f \in W^{m+1,p}$, and $g \in W^{m,p}$, then $u \in W^{m,p}(T_r^k \mathscr{M})$. Therefore, by the Sobolev Embedding Theorem 3.7, if g_{ij}, f, g are C^∞, then the solution u is also C^∞.

Linear hyperbolic equations

Let $\{\mathscr{M}, g_{\mu\nu}\}$ be a Minkowski manifold, i.e., its metric $g_{\mu\nu}$ can be written in some coordinate system as

$$(g_{\mu\nu}) = \begin{pmatrix} -1 & 0 \\ 0 & G \end{pmatrix}, \quad G = (g_{ij}) \tag{3.2.43}$$

where G is an $(n-1)\times(n-1)$ positive definite symmetric matrix. Then, the Laplace operator

$$D^\mu D_\mu = g^{\mu\nu} D_\mu D_\nu = -\frac{\partial^2}{\partial t^2} + g^{ij} D_i D_j$$

is a hyperbolic operator.

Let $\mathscr{M} = S^1 \times \widetilde{\mathscr{M}}$ with the metric (3.2.43), and $\widetilde{\mathscr{M}}$ is a Riemannian manifold with metric $\{g_{ij}\}$ as in (3.2.43), and $\partial\widetilde{\mathscr{M}} = \varnothing$. Then, for the Minkowski manifold $S^1 \times \widetilde{\mathscr{M}}$, the equation

$$g^{\mu\nu} D_\mu D_\nu u = \text{div } f, \tag{3.2.44}$$

is a hyperbolic equation, written as

$$-\frac{\partial^2 u}{\partial t^2} + g^{ij} D_i D_j u = \text{div } f, \tag{3.2.45}$$

with the periodic condition

$$u(t+T) = u(t), \qquad \forall t \in \mathbb{R}^1. \tag{3.2.46}$$

The following existence theorem is classical.

Theorem 3.15 *Let $\mathscr{M} = S^1 \times \widetilde{\mathscr{M}}$ is a Minkowski manifold with metric (3.2.43), and $\partial\widetilde{\mathscr{M}} = \varnothing$. Assume that $g_{\mu\nu}$ and f are C^∞, then the problem (3.2.45)-(3.2.46) has a C^∞ solution u.*

Remark 3.16 If $\mathscr{M} = \mathbb{R}^1 \times \widetilde{\mathscr{M}}$ with metric (3.2.43) and $\partial\widetilde{M} = \varnothing$, then the problem (3.2.45)-(3.2.46) is replaced by the following initial value problem:

$$\begin{aligned} & -\frac{\partial^2 u}{\partial t^2} + g^{ij} D_i D_j u = \text{div } f, \\ & u(0) = \varphi, \\ & u_t(0) = \psi. \end{aligned} \tag{3.2.47}$$

The same existence theorem holds true as well for problem (3.2.47).

3.3 Orthogonal Decomposition for Tensor Fields

3.3.1 Introduction

Let $\mathscr{M}$ be a Riemannian manifold (or a Minkowski manifold), $\partial\mathscr{M} = \varnothing$, and u is a tensor field on $\mathscr{M}$:

$$u : \mathscr{M} \to T_r^k \mathscr{M}. \tag{3.3.1}$$

The orthogonal decomposition of tensor fields means that the field u given by (3.3.1) can be decomposed as

$$u = \nabla\phi + v \qquad \text{and} \qquad \text{div } v = 0, \tag{3.3.2}$$

for some $\phi : \mathscr{M} \to T_r^{k-1}\mathscr{M}$ (or $\phi : \mathscr{M} \to T_{r-1}^k \mathscr{M}$). Moreover $\nabla\phi$ and v are orthogonal in the following sense:

$$\langle \nabla\phi, v \rangle = \int_{\mathscr{M}} (\nabla\phi, v)\sqrt{-g}\,dx = 0. \tag{3.3.3}$$

In this section, we shall show that all tensor fields as given by (3.3.1) can be decomposed into the form (3.3.2) satisfying (3.3.3).

In order to understand the problem well, we first introduce some classical results: the Helmholtz decomposition and the Leray decomposition on $\mathscr{M} = \mathbb{R}^n$.

1. *Helmholtz decomposition.* Let $u \in L^2(T\mathbb{R}^3)$ be a 3-dimensional vector field, i.e.

$$u(x) = (u^1(x),\ u^2(x),\ u^3(x)) \qquad \text{for } x \in \mathbb{R}^3,$$

then there exist a function $\phi \in H^1(\mathbb{R}^3)$ and a vector field $A \in H^1(T\mathbb{R}^3)$, such that u can be decomposed as

$$\begin{aligned} & u = \nabla\phi + \operatorname{curl} A, \\ & \int_{\mathbb{R}^3} \nabla\phi \cdot \operatorname{curl} A dx = 0. \end{aligned} \tag{3.3.4}$$

Note that $\operatorname{div}(\operatorname{curl} A) = 0$. Hence, the Helmholtz decomposition is an important initial result on orthogonal decompositions.

2. *Leray decomposition.* Let $\Omega \in \mathbb{R}^n$ be a domain, and $u \in L^2(T\Omega)$ be an n-dimensional vector field. Then u can be decomposed as

$$\begin{aligned} & u = \nabla\phi + v, \\ & v \cdot n|_{\partial\Omega} = 0, \qquad \operatorname{div} v = 0, \quad \phi \in H^1(\Omega), \\ & \int_\Omega \nabla\phi \cdot v dx = 0. \end{aligned} \tag{3.3.5}$$

The Leray decomposition (3.3.5) is crucial in fluid dynamics.

The decompositions (3.3.4) and (3.3.5) can be generalized to more general tensor fields as shown in (3.3.1)-(3.3.3). Now we discuss the simplest case to illustrate the main idea.

Let $u:\ \mathbb{R}^n \to T\mathbb{R}^n$ be a given vector field. Then div u is a known function. It is known that the Poisson equation

$$\Delta\phi = \operatorname{div} u \qquad \text{for } x \in \mathbb{R}^n \tag{3.3.6}$$

has a weak solution $\phi \in H^1(\mathbb{R}^n)$, enjoying

$$\int_{\mathbb{R}^n} [\nabla\phi - u] \cdot \nabla\varphi dx = 0 \qquad \forall\varphi \in H^1(\mathbb{R}^n). \tag{3.3.7}$$

Let $v = u - \nabla\phi$. Then, by (3.3.7) we have

$$\int_R v \cdot \nabla\phi dx = 0,$$

which means that div $v = 0$. Thus we obtain the orthogonal decomposition $u = \nabla\phi + v$ with div $v = 0$.

3.3.2 Orthogonal decomposition theorems

The aim of this subsection is to derive an orthogonal decomposition for (k,r)-tensor fields, with $k+r\geqslant 1$, into divergence-free and gradient parts. This decomposition plays a crucial role for the unified field theory coupling four fundamental interactions to be introduced in Chapter 4 of this book.

Let $\mathscr{M}$ be a closed Riemannian manifold or $\mathscr{M}=S^1\times\widetilde{\mathscr{M}}$ be a closed Minkowski manifold with metric (3.2.43), and $v\in L^2(T^k_r\mathscr{M})$ $(k+r\geqslant 1)$. We say that v is div_A-free, denoted by $\text{div}_A v=0$, if

$$\int_{\mathscr{M}}(\nabla_A\psi,v)\sqrt{-g}dx=0,\qquad \forall\nabla_A\psi\in L^2(T^k_r\mathscr{M}). \tag{3.3.8}$$

Here $\psi\in H^1(T^{k-1}_r\mathscr{M})$ or $H^1(T^k_{r-1}\mathscr{M}),\nabla_A$ and div_A are as in (3.2.38).

We remark that if $v\in H^1(T^k_r\mathscr{M})$ satisfies (3.3.8), then v is weakly differentiable, and div $v=0$ in L^2-sense. If $v\in L^2(T^k_r\mathscr{M})$ is not differentiable, then (3.3.8) means that v is div_A-free in the distribution sense.

Theorem 3.17 (Orthogonal Decomposition Theorem) *Let A be a given vector field or covector field, and $u\in L^2(T^k_r\mathscr{M})$. Then the following assertions hold true*:

1) *The tensor field u can be orthogonally decomposed into*

$$u=\nabla_A\varphi+v\quad\text{with } \text{div}_A v=0, \tag{3.3.9}$$

where $\varphi\in H^1(T^{k-1}_r\mathscr{M})$ or $\varphi\in H^1(T^k_{r-1}\mathscr{M})$.

2) *If $\mathscr{M}$ is a compact Riemannian manifold, then u can be orthogonally decomposed into*

$$u=\nabla_A\varphi+v+h, \tag{3.3.10}$$

where φ and v are as in (3.3.9), and h is a harmonic field, i.e.

$$\text{div}_A h=0,\quad \nabla_A h=0.$$

In particular, the subspace of all harmonic tensor fields in $L^2(T^k_r\mathscr{M})$ is of finite dimensional:

$$\begin{aligned}&H(T^k_r\mathscr{M})=\{h\in L^2(T^k_r\mathscr{M})|\ \nabla_A h=0,\ \text{div}_A h=0\},\ \text{and}\\ &\dim H(T^k_r\mathscr{M})<\infty.\end{aligned} \tag{3.3.11}$$

Remark 3.18 The above orthogonal decomposition theorem implies that $L^2(T^k_r\mathscr{M})$ can be decomposed into

$$\begin{aligned}&L^2(T^k_r\mathscr{M})=G(T^k_r\mathscr{M})\oplus L^2_D(T^k_r\mathscr{M}) &&\text{for general case,}\\ &L^2(T^k_r\mathscr{M})=G(T^k_r\mathscr{M})\oplus H(T^k_r\mathscr{M})\oplus L^2_N(T^k_r\mathscr{M}) &&\text{for } \mathscr{M} \text{ compact Riemainnian.}\end{aligned} \tag{3.3.12}$$

Here H is as in (3.3.11), and

$$\begin{aligned}
&G(T_r^k\mathscr{M}) = \{v \in L^2(T_r^k\mathscr{M}) |\ v = \nabla_A\varphi,\ \varphi \in H^1(T_{r-1}^k\mathscr{M})\},\\
&L_D^2(T_r^k\mathscr{M}) = \{v \in L^2(T_r^k\mathscr{M}) |\ \mathrm{div}_A v = 0\},\\
&L_N^2(T_r^k\mathscr{M}) = \{v \in L_D^2(T_r^k\mathscr{M}) |\ \nabla_A v \neq 0\}.
\end{aligned}$$

They are orthogonal to each other:

$$L_D^2(T_r^k\mathscr{M}) \perp G(T_r^k\mathscr{M}), \quad L_N^2(T_r^k\mathscr{M}) \perp H(T_r^k\mathscr{M}), \quad G(T_r^k\mathscr{M}) \perp H(T_r^k\mathscr{M}).$$

Remark 3.19 The orthogonal decomposition (3.3.12) of $L^2(T_r^k\mathscr{M})$ implies that if a tensor field $u \in L^2(T_r^k\mathscr{M})$ satisfies that

$$\langle u, v\rangle_{L^2} = \int_{\mathscr{M}} (u,v)\sqrt{-g}dx = 0, \qquad \forall \mathrm{div}_A v = 0,$$

then u must be a gradient field, i.e.

$$u = \nabla_A\varphi \quad \text{for some } \varphi \in H^1(T_r^{k-1}\mathscr{M}) \text{ or } H^1(T_{r-1}^k\mathscr{M}).$$

Likewise, if $u \in L^2(T_r^k\mathscr{M})$ satisfies that

$$\langle u, v\rangle_{L^2} = 0, \qquad \forall v \in G(T_r^k\mathscr{M}),$$

then $u \in L_D^2(T_r^k\mathscr{M})$. It is the reason why we define a div_A-free field by (3.3.8). □

Proof of Theorem 3.17 We proceed in several steps as follows.

STEP 1. PROOF OF ASSERTION (1). Let $u \in L^2(E), E = T_r^k\mathscr{M}$ $(k + r \geqslant 1)$. Consider the equation

$$\Delta\varphi = \mathrm{div}_A u \quad \text{in } \mathscr{M}, \tag{3.3.13}$$

where Δ is the Laplace operator defined by

$$\Delta = \mathrm{div}_A \cdot \nabla_A. \tag{3.3.14}$$

Without loss of generality, we only consider the case where $\mathrm{div}_A u \in \widetilde{E} = T_r^{k-1}\mathscr{M}$. It is clear that if (3.3.13) has a solution $\varphi \in H^1(\widetilde{E})$, then by (3.3.14), the following vector field must be div_A-free

$$v = u - \nabla_A\varphi \in L^2(E). \tag{3.3.15}$$

Moreover, by (3.3.8), we have

$$\langle v, \nabla_A\psi\rangle_{L^2} = \int_{\mathscr{M}} (v, \nabla_A\psi)\sqrt{-g}dx = 0, \qquad \forall \nabla_A\psi \in L^2(T_r^k\mathscr{M}). \tag{3.3.16}$$

Namely v and $\nabla_A\varphi$ are orthogonal. Therefore, the orthogonal decomposition $u = v + \nabla_A\varphi$ follows from (3.3.15) and (3.3.16).

It suffices then to prove that (3.3.13) has a weak solution $\varphi \in H^1(\widetilde{E})$:

$$\langle \nabla_A \varphi - u, \nabla_A \psi \rangle_{L^2} = 0, \qquad \forall \psi \in H^1(\widetilde{E}). \tag{3.3.17}$$

Obviously, if ϕ satisfies

$$\Delta \phi = 0, \tag{3.3.18}$$

where Δ is as in (3.3.14), then, by (3.2.39),

$$\int_{\mathscr{M}} (\Delta \phi, \phi) \sqrt{-g} dx = -\int_{\mathscr{M}} (\nabla_A \phi, \nabla_A \phi) \sqrt{-g} dx = 0.$$

Hence (3.3.18) is equivalent to

$$\nabla_A \phi = 0. \tag{3.3.19}$$

Therefore, for all ϕ satisfying (3.3.18) we have

$$\int_{\mathscr{M}} (u, \nabla_A \phi) \sqrt{-g} dx = 0.$$

By Theorem 3.14, we derive that the equation (3.3.13) has a unique weak solution $\varphi \in H^1(\widetilde{E})$.

For Minkowski manifolds, by Theorem 3.15, the equation (3.3.13) also has a solution. Thus Assertion (1) is proved.

STEP 2. PROOF OF ASSERTION (2). Based on Assertion (1), we have

$$H^k(E) = H^k_D \oplus G^k, \quad L^2(E) = L^2_D \oplus G,$$

where

$$H^k_D = \{u \in H^k(E) |\ \mathrm{div}_A u = 0\},$$
$$G^k = \{u \in H^k(E) |\ u = \nabla_A \psi\}.$$

Define an operator $\widetilde{\Delta}: \ H^2_D(E) \to L^2_D(E)$ by

$$\widetilde{\Delta} u = P \Delta u, \tag{3.3.20}$$

where $P: \ L^2(E) \to L^2_D(E)$ is the canonical orthogonal projection.

We known that the Laplace operator Δ can be expressed as

$$\Delta = \mathrm{div}_A \cdot \nabla_A = g^{kl} \frac{\partial^2}{\partial x^k \partial x^l} + B, \tag{3.3.21}$$

where B is a lower-order differential operator. Since M is compact, the Sobolev embeddings

$$H^2(E) \hookrightarrow H^1(E) \hookrightarrow L^2(E)$$

are compact. Hence the lower-order differential operator

$$B:\ H^2(\mathscr{M},\mathbb{R}^N)\to L^2(\mathscr{M},\mathbb{R}^N)$$

is a linear compact operator. Therefore the operator in (3.3.21) is a linear completely continuous field

$$\Delta:\ H^2(E)\to L^2(E),$$

which implies that the operator of (3.3.20) is also a linear completely continuous field

$$\widetilde{\Delta}=P\Delta:\ H^2_D(E)\to L^2_D(E).$$

By the spectrum theorem of completely continuous fields (Ma and Wang, 2005), the space

$$\widetilde{H}=\{u\in H^2_D(E)|\ \widetilde{\Delta}u=0\}$$

is finite dimensional, and is the eigenspace of the eigenvalue $\lambda=0$. By (3.2.39), for $u\in\widetilde{H}$ we have

$$\begin{aligned}\int_{\mathscr{M}}(\widetilde{\Delta}u,u)\sqrt{-g}dx&=\int_{\mathscr{M}}(\Delta u,u)\sqrt{-g}dx\quad(\text{by }\mathrm{div}_A u=0)\\&=-\int_{\mathscr{M}}(\nabla_A u,\nabla_A u)\sqrt{-g}dx\\&=0\quad(\text{by }\widetilde{\Delta}u=0).\end{aligned}$$

It follows that

$$u\in\widetilde{H}\Leftrightarrow\nabla_A u=0,$$

which implies that $\widetilde{H}$ is the same as the harmonic space H of (3.3.11), i.e. $\widetilde{H}=H$. Thus we have

$$\begin{aligned}&L^2_D(E)=H\oplus L^2_N(E),\\&L^2_N(E)=\{u\in L^2_D(E)|\ \nabla_A u\neq 0\}.\end{aligned}$$

The proof of Theorem 3.17 is complete. □

3.3.3 Uniqueness of orthogonal decompositions

In this subsection we only consider the case where $\mathscr{M}$ is a closed manifold with zero first Betti number.

In Theorem 3.17, a tensor field $u\in L^2(T^k_r\mathscr{M})$ with $k+r\geqslant 1$ can be orthogonally decomposed into

$$\begin{aligned}&u=\nabla\varphi+v&&\text{for general closed manifolds,}\\&u=\nabla\varphi+v+h&&\text{for compact Riemannian manifolds.}\end{aligned}\tag{3.3.22}$$

Now we address the uniqueness problem of the decomposition (3.3.22). In fact, if u is a vector field or a covector field:

$$u \in L^2(T\mathscr{M}) \quad \text{or} \quad u \in L^2(T^*\mathscr{M}),$$

then the decomposition (3.3.22) is unique.

We see that if $u \in L^2(T^k_r\mathscr{M})$ with $k+r \geqslant 2$, then there are different types of decompositions of (3.3.22). For example, for $u \in L^2(T^0_2\mathscr{M})$, in a local coordinate system, u is given by

$$u = \{u_{ij}(x)\}.$$

In this case, u admits two types of decompositions

$$u_{ij} = D_i\varphi_j + v_{ij}, \qquad D^i v_{ij} = 0, \tag{3.3.23}$$

$$u_{ij} = D_j\psi_i + w_{ij}, \qquad D^j w_{ij} = 0. \tag{3.3.24}$$

It is easy to see that if $u_{ij} \neq u_{ji}$, then (3.3.23) and (3.3.24) can be two different decomposition of u_{ij}. Namely

$$\{v_{ij}\} \neq \{w_{ij}\}, \quad (\varphi_1, \cdots, \varphi_n) \neq (\psi_1, \cdots, \psi_n).$$

The reason is that the two differential equations generating the two decompositions (3.3.23) and (3.3.24) as

$$D^iD_i\varphi_j = D^iu_{ij} \quad \text{and} \quad D^iD_i\psi_j = D^iu_{ji} \tag{3.3.25}$$

are different because $D^iu_{ij} \neq D^iu_{ji}$.

However for a symmetric tensor field $u_{ij} = u_{ji}$, as

$$D^iu_{ij} = D^iu_{ji},$$

the two equations in (3.3.25) are the same. By the uniqueness of solutions of (3.3.25), the two solutions φ_j and ψ_j are the same:

$$\varphi_i = \psi_i \qquad \text{for } 1 \leqslant i \leqslant n.$$

Thus (3.3.23) and (3.3.24) can be expressed as

$$u_{ij} = D_i\varphi_j + v_{ij}, \qquad D^i v_{ij} = 0, \tag{3.3.26}$$

$$u_{ij} = D_j\varphi_i + w_{ij}, \qquad D^j w_{ij} = 0. \tag{3.3.27}$$

From (3.3.26) and (3.3.27) we can deduce the following theorem.

Theorem 3.20 *Let $u \in L^2(T^0_2\mathscr{M})$ be symmetric, i.e. $u_{ij} = u_{ji}$, and the first Betti number $\beta_1(\mathscr{M}) = 0$ for $\mathscr{M}$. Then the following assertions hold true:*

1) *u has a unique orthogonal decomposition if and only if there is a scalar function* $\varphi \in H^2(\mathscr{M})$ *such that u can be expressed as*

$$\begin{aligned} u_{ij} &= v_{ij} + D_i D_j \varphi, \\ v_{ij} &= v_{ij}, \qquad D^i v_{ij} = 0. \end{aligned} \tag{3.3.28}$$

2) *If* v_{ij} *in* (3.3.26) *is symmetric:* $v_{ij} = v_{ji}$*, then u can be expressed by* (3.3.28).

3) *u can be orthogonally decomposed in the form* (3.3.28) *if and only if the following differential equations have a solution* $\varphi \in H^2(\mathscr{M})$*, and* φ *is the scalar field in* (3.3.28):

$$\frac{\partial}{\partial x^i}\Delta\varphi + R_i^k \frac{\partial \varphi}{\partial x^k} = -D^j u_{ji} \qquad \text{for } 1 \leqslant i \leqslant n, \tag{3.3.29}$$

where $R_i^k = g^{kj} R_{ij}$ *and* R_{ij} *are the Ricci curvature tensors, and* Δ *is the Laplace operator for scalar fields as defined by* (3.2.28).

Proof We only need to prove Assertions (2) and (3).

We first prove Assertion (2). Since v_{ij} in (3.3.26) is symmetric, then we have

$$D_i \varphi_j = D_j \varphi_i. \tag{3.3.30}$$

Note that

$$D_i \varphi_j = \frac{\partial \varphi_j}{\partial x^i} - \Gamma_{ij}^k \varphi_k,$$

and $\Gamma_{ij}^k = \Gamma_{ji}^k$. We infer then from (3.3.30) that

$$\frac{\partial \varphi_j}{\partial x^i} = \frac{\partial \varphi_i}{\partial x^j}. \tag{3.3.31}$$

By assumption, the 1-dimensional homology of $\mathscr{M}$ is zero,

$$H_1(\mathscr{M}) = 0,$$

and by the de Rham theorem (Ma, 2010), it follows that all closed 1-forms are complete differentials, i.e. for any

$$\omega = \psi_i dx^i, \qquad d\omega = 0,$$

there is a scalar function ψ such that

$$\psi_i = \partial \psi / \partial x_i \qquad \text{for } 1 \leqslant i \leqslant n.$$

In view of (3.3.31), it implies that the 1-form

$$\omega = \varphi_k dx^k$$

is closed. Hence it follows that there is a $\varphi \in H^1(\mathscr{M})$ such that

$$\varphi_k = \frac{\partial \varphi}{\partial x^k} \qquad \text{for } 1 \leqslant k \leqslant n.$$

Assertion (2) is proved.

Now we prove Assertion (3). Taking the divergence on both sides of (3.3.26), we obtain that

$$D^i D_i \varphi_j = D^i u_{ij}. \tag{3.3.32}$$

By the Laplace-Beltrami operator in (3.2.31),

$$-\widetilde{\Delta}\varphi_j = D^i D_i \varphi_j + R^k_j \varphi_k, \tag{3.3.33}$$

where $\widetilde{\Delta} = \delta d + d\delta$. By the Hodge theory, for $\omega = \varphi_i dx^i$ we have

$$\begin{aligned} d\omega = 0 \qquad &\Leftrightarrow \quad \varphi_i = \partial\varphi/\partial x^i, \\ \delta\omega = \Delta\varphi \quad &\Leftrightarrow \quad \varphi_i = \partial\varphi/\partial x^i. \end{aligned} \tag{3.3.34}$$

Here ∇ is the gradient operator, and Δ the Laplace operator as in (3.2.28). It follows from (3.3.34) that

$$\widetilde{\Delta}\omega = (\delta d + d\delta)\omega = d\delta\omega \ \Leftrightarrow \ \varphi_i = \partial\varphi/\partial x^i,$$

and

$$d\delta\omega = \frac{\partial}{\partial x^i}(\Delta\varphi) dx^i.$$

Hence we deduce from (3.3.33) that

$$D^i D_j \varphi_j = -\frac{\partial}{\partial x^j}(\Delta\varphi) - R^k_j \frac{\partial\varphi}{\partial x^k} \Leftrightarrow \varphi_i = \partial\varphi/\partial x^i. \tag{3.3.35}$$

Inserting (3.3.32) in (3.3.35) we obtain that the equations

$$\frac{\partial}{\partial x^j}(\Delta\varphi) + R^k_j \frac{\partial\varphi}{\partial x^k} = -D^i u_{ij}$$

have a solution φ if and only if φ_j in (3.3.26) is a gradient field of φ, i.e. $\varphi_j = \partial\varphi/\partial x^j$.

Assertion (3) is proven, and the proof of the theorem is complete. □

Remark 3.21 The conclusions of Theorem 3.20 are also valid for second-order contravariant symmetric tensors $u = \{u^{ij}\}$, and the decomposition is given as follows:

$$\begin{aligned} &u^{ij} = v^{ij} + g^{ik} g^{jl} D_k D_l \varphi \\ &D_i v^{ij} = 0, \quad v^{ij} = v^{ji}, \quad \varphi \in H^2(\mathscr{M}). \end{aligned}$$

3.3.4 Orthogonal decomposition on manifolds with boundary

In the above subsections, we mainly consider the orthogonal decomposition of tensor fields on the closed Riemannian and Minkowski manifolds. In this section we discuss the problem on manifolds with boundary.

1. *Orthogonal decomposition on Riemannian manifolds with boundaries.* The Leray decomposition (3.3.5) is for the vector fields on a domain $\Omega \subset \mathbb{R}^n$ with $\partial\Omega \neq \varnothing$. This result can be also generalized to general (k,r)-tensor fields defined on manifolds with boundaries.

Theorem 3.22 *Let $\mathscr{M}$ be a Riemannian manifold with boundary $\partial\mathscr{M} \neq \varnothing$, and*

$$u : \mathscr{M} \to T_r^k \mathscr{M} \tag{3.3.36}$$

be a (k,r)-tensor field. Then we have the following orthogonal decomposition:

$$\begin{aligned} & u = \nabla_A \varphi + v, \\ & \text{div}_A v = 0, \qquad v \cdot n|_{\partial\mathscr{M}} = 0, \qquad \int_{\mathscr{M}} (\nabla_A \varphi, v)\sqrt{-g}dx = 0, \end{aligned} \tag{3.3.37}$$

where $\partial v/\partial n = \nabla_A v \cdot n$ is the derivative of v in the direction of outward normal vector n on $\partial\Omega$.

Proof For the tensor field u in (3.3.36), consider

$$\begin{aligned} & \text{div}_A \cdot \nabla_A \varphi = \text{div}_A u, && \forall x \in \mathscr{M}, \\ & \frac{\partial \varphi}{\partial n} = u \cdot n, && \forall x \in \partial\mathscr{M}. \end{aligned} \tag{3.3.38}$$

This Neumann boundary problem possesses a solution provided the following condition holds true:

$$\int_{\partial\mathscr{M}} \frac{\partial \varphi}{\partial n} ds = \int_{\partial\mathscr{M}} u \cdot n ds, \tag{3.3.39}$$

which is ensured by the boundary condition in (3.3.38). Hence by (3.3.38) the field

$$v = u - \nabla_A \varphi \tag{3.3.40}$$

is div_A-free, and satisfies the boundary condition

$$v \cdot n|_{\partial\mathscr{M}} = 0. \tag{3.3.41}$$

Then it follows from (3.3.40) and (3.3.41) that the tensor field u in (3.3.36) can be orthogonally decomposed into the form of (3.3.37). The proof is complete. □

2. *Orthogonal decomposition on Minkowski manifolds.* Let $\mathscr{M}$ be a Minkowski manifold in the form

$$\mathscr{M} = \widetilde{\mathscr{M}} \times (0,T), \tag{3.3.42}$$

with the metric

$$(g_{\mu\nu}) = \begin{pmatrix} -1 & 0 \\ 0 & G \end{pmatrix}. \tag{3.3.43}$$

Here $\widetilde{\mathscr{M}}$ is a closed Riemannian manifold, and $G = (g_{ij})$ is the Riemann metric of $\widetilde{\mathscr{M}}$.

In view of the Minkowski metric (3.3.43), we see that the operator $\text{div}_A \cdot \nabla_A$ is a hyperbolic differential operator expressed as

$$\text{div}_A \cdot \nabla_A = -\left(\frac{\partial}{\partial t} + A_0\right)^2 + g^{ij} D_{Ai} D_{Aj}. \tag{3.3.44}$$

Now a tensor field $u \in L^2(T_r^k \mathscr{M})$ has an orthogonal composition if the following hyperbolic equation

$$\text{div}_A \cdot \nabla_A \varphi = \text{div}_A u, \qquad \text{in } \mathscr{M} \tag{3.3.45}$$

has a weak solution $\varphi \in H^1(T_r^{k-1}\mathscr{M})$ in the following sense:

$$\int_{\mathscr{M}} (D_A \varphi, D_A \psi)\sqrt{-g}dx = \int_{\mathscr{M}} (u, D_A \psi)\sqrt{-g}dx, \qquad \forall \psi \in H^1(T_r^{k-1}\mathscr{M}). \tag{3.3.46}$$

Theorem 3.23 *Let $\mathscr{M}$ be a Minkowski manifold as defined by (3.3.42)-(3.3.43), and $u \in L^2(T_r^l \mathscr{M})$ $(k + r \geqslant 1)$ be an (k,r)-tensor field. Then u can be orthogonally decomposed into the following form*

$$\begin{aligned} &u = \nabla_A \varphi + v, \quad \text{div}_A v = 0, \\ &\int_M (\nabla_A \varphi, v)\sqrt{-g}dx = 0, \end{aligned} \tag{3.3.47}$$

if and only if equation (3.3.45) has a weak solution $\varphi \in H^1(T^{k-1}\mathscr{M})$ in the sense of (3.3.46).

3.4 Variations with div_A-Free Constraints

3.4.1 Classical variational principle

Variational approach originates from the minimization problem of a functional. Let X be a Banach space, and F be a functional on X:

$$F: X \to \mathbb{R}^1. \tag{3.4.1}$$

The minimization problem of F is to find a point $u \in X$, which is a minimal point of F. Namely, there is a neighborhood $U \subset X$ of u, such that F is minimal at u in U:

$$F(u) = \min_{v \in U} F(v). \tag{3.4.2}$$

In the classical variational principle we know that the minimal point u of F in (3.4.2) is a solution of the variational equation of F:

$$\delta F(u) = 0, \tag{3.4.3}$$

where δF is the derivative operator of F.

Given a variational problem, it is important to compute δF for a given functional F. Hereafter we give a brief introduction of general methods to compute the derivative operators from F.

Let F be the functional given by (3.4.1), and X^* be the dual space of X. The derivative operator $\delta F(u)$ of F at $u \in X$ is a linear functional on X, i.e. $\delta F(u) \in X^*$ for each $u \in X$. In other words, δF is a mapping from X to X^*:

$$\delta F : X \to X^*. \tag{3.4.4}$$

Denote $\langle \cdot, \cdot \rangle$ the product between X and X^*, i.e.

$$\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{R}.$$

Then the derivative operator δF in (3.4.4) satisfies the relation:

$$\langle \delta F(u), v \rangle = \frac{d}{d\lambda}\Big|_{\lambda=0} F(u + \lambda v) \qquad \text{for } u, v \in X, \tag{3.4.5}$$

where $\lambda \in \mathbb{R}^1$ is real number.

Based on (3.4.5), it is easy to see that the minimal point u satisfying (3.4.2) is a solution of the variational equation (3.4.3). In fact, by (3.4.2) for any given $v \in X$ the function $f(\lambda) = F(u + \lambda v)$ is minimal at $\lambda = 0$:

$$\frac{df(0)}{d\lambda} = 0 \Rightarrow \frac{d}{d\lambda}\Big|_{\lambda=0} F(u + \lambda v) = 0, \qquad \forall v \in X.$$

It follows then by (3.4.5) that

$$\langle \delta F(u), v \rangle = 0, \qquad \forall v \in X,$$

which means that u satisfies (3.4.3).

In the following, we give a simple example to show how to compute δF using formula (3.4.5).

Let $X = H^1(\mathbb{R}^n)$, and $F : H^1(\mathbb{R}^n) \to \mathbb{R}^1$ be given by

$$F(u) = \int_{\mathbb{R}^n} \left[\frac{1}{2} |\nabla u|^2 + f(u) \right] dx \qquad \text{for } u \in H^1(\mathbb{R}^n). \tag{3.4.6}$$

We see that

$$\begin{aligned} \frac{d}{d\lambda} F(u + \lambda v) =& \frac{d}{d\lambda} \int_{\mathbb{R}^n} \left[\frac{1}{2} |\nabla u + \lambda \nabla v|^2 + f(u + \lambda v) \right] dx \\ =& \int_{\mathbb{R}^n} \left[(\nabla u + \lambda \nabla v) \cdot \nabla v + f'(u + \lambda v) v \right] dx. \end{aligned}$$

Hence we have

$$\begin{aligned}\frac{d}{d\lambda}\Big|_{\lambda=0}F(u+\lambda v) &= \int_{\mathbb{R}^n}\left[\nabla u\cdot\nabla v+f'(u)v\right]dx\\ &= \int_{\mathbb{R}^n}\left[-\text{div }(\nabla u)+f'(u)\right]vdx. \quad (\text{by}(3.2.36))\end{aligned}$$

On the other hand, by (3.4.5) and

$$\langle\delta F(u),v\rangle=\int_{R^n}\delta F(u)vdx,$$

we deduce that

$$\int_{\mathbb{R}^n}\delta F(u)vdx=\int_{\mathbb{R}^n}(-\Delta u+f'(u))vdx, \qquad \forall v\in H^1(\mathbb{R}^n).$$

Hence we obtain the derivative operator $\delta F(u)$ of (3.4.6) as

$$\delta F(u)=-\Delta u+f'(u).$$

3.4.2 Derivative operators of the Yang-Mills functionals

Let G_μ^a $(1\leqslant a\leqslant N^2-1)$ be the $SU(N)$ gauge fields. The Yang-Mills functional for G_μ^a is defined by

$$F=\int_{\mathscr{M}}\left[-\frac{1}{4}F_{\mu\nu}^aF^{\mu\nu a}\right]dx, \tag{3.4.7}$$

where $\mathscr{M}$ is the 4-dimensional Minkowski space, and

$$F_{\mu\nu}^a=\partial_\mu G_\nu^a-\partial_\nu G_\mu^a+g\lambda^{abc}G_\mu^bG_\nu^c. \tag{3.4.8}$$

In Section 2.4.3, we have seen that the functional (3.4.7) is the scalar curvature part in the Yang-Mills action (2.4.50).

Referring to derivative operators of functionals for the electromagnetic potential deduced in Subsection 2.5.3, we now derive the derivative operator for the Yang-Mills functional (3.4.7):

$$\begin{aligned}\frac{d}{d\lambda}\Big|_{\lambda=0}F(G+\lambda\widetilde{G}) &= -\frac{1}{2}\int_{\mathscr{M}}g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}^a\frac{d}{d\lambda}\Big|_{\lambda=0}F_{\mu\nu}^a(G+\lambda\widetilde{G})dx\\ &= -\frac{1}{2}\int_{\mathscr{M}}F^{\mu\nu a}\left(\frac{\partial\widetilde{G}_\nu^a}{\partial x^\mu}-\frac{\partial\widetilde{G}_\mu^a}{\partial x^\nu}\right)dx\\ &\quad -\frac{1}{2}\int_{\mathscr{M}}F^{\mu\nu a}g\lambda^{abc}(G_\mu^b\widetilde{G}_\nu^c+\widetilde{G}_\mu^bG_\nu^c)dx \quad (\text{by}(3.4.8))\\ &= \int_{\mathscr{M}}g^{\mu\alpha}g^{\nu\beta}\left(\frac{\partial F_{\alpha\beta}^a}{\partial x^\mu}-gF_{\alpha\beta}^c\lambda^{cba}G_\mu^b\right)\widetilde{G}_\nu^adx.\end{aligned}$$

By (3.4.5) we deduce the derivative operator δF of (3.4.7) as

$$\delta F = \partial^{\alpha} F^a_{\alpha\beta} - g g^{\alpha\mu} \lambda^a_{cb} F^c_{\alpha\beta} G^b_{\mu}, \quad \beta = 0,1,2,3, \tag{3.4.9}$$

where $\lambda^a_{cb} = \lambda^{cba}$.

For the general form of Yang-Mills functional given by

$$F = \int_{\mathscr{M}} \left[-\frac{1}{4} \mathscr{G}_{ab} F^a_{\mu\nu} F^{\mu\nu b} \right] dx, \tag{3.4.10}$$

where $(\mathscr{G}_{ab})$ is the Riemann metric on $SU(N)$ given by (2.4.49). The derivative operator of F in (3.4.10) is as follows

$$\delta F = \mathscr{G}_{ab} \partial^{\alpha} F^b_{\alpha\beta} - g g^{\alpha\mu} \mathscr{G}_{bc} \lambda^c_{da} F^b_{\alpha\beta} G^d_{\mu}. \tag{3.4.11}$$

3.4.3 Derivative operator of the Einstein-Hilbert functional

The Einstein-Hilbert functional is in the form

$$F = \int_{\mathscr{M}} R\sqrt{-g} dx, \tag{3.4.12}$$

where $\mathscr{M}$ is an n-dimensional Riemannian manifold with metric g_{ij}, and $R = g^{ij} R_{ij}$ is the scalar curvature of $\mathscr{M}$, and R_{ij} is the Ricci curvature tensor:

$$R_{ij} = \frac{1}{2} g^{kl} \left(\frac{\partial^2 g_{kl}}{\partial x^i \partial x^j} + \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} - \frac{\partial^2 g_{ik}}{\partial x^j \partial x^l} - \frac{\partial^2 g_{jl}}{\partial x^i \partial x^k} \right) + g^{kl} g_{rs} (\Gamma^r_{kl} \Gamma^s_{ij} - \Gamma^r_{ik} \Gamma^s_{jl}), \tag{3.4.13}$$

and the Levi-Civita connections Γ^r_{kl} are written as

$$\Gamma^r_{kl} = \frac{1}{2} g^{rs} \left(\frac{\partial g_{ks}}{\partial x^l} + \frac{\partial g_{ls}}{\partial x^k} - \frac{\partial g_{kl}}{\partial x^s} \right). \tag{3.4.14}$$

First we verify the following derivative operator δF of the Einstein-Hilbert functional (3.4.12)-(3.4.14):

$$\delta F = R_{ij} - \frac{1}{2} g_{ij} R. \tag{3.4.15}$$

Note that g_{ij} and g^{ij} have the relations

$$g^{ij} = \frac{1}{g} \times ||g^{ij}||, \qquad ||g^{ij}|| \text{ the cofactor of } g_{ij}, \tag{3.4.16}$$

$$g_{ij} = \frac{1}{g} \times ||g_{ij}||, \qquad ||g_{ij}|| \text{ the cofactor of } g^{ij}. \tag{3.4.17}$$

Hence by (3.4.16), we have

$$\frac{d}{d\lambda}\bigg|_{\lambda=0} \det\,(g_{ij} + \lambda \widetilde{g}_{ij}) = \widetilde{g}_{ij} \times ||g^{ij}|| = \widetilde{g}_{ij} g^{ij} g. \tag{3.4.18}$$

In addition, by $g_{ik}g^{kj}=\delta_i^j$, we obtain

$$\frac{d}{d\lambda}\Big|_{\lambda=0}(g_{ik}+\lambda\widetilde{g}_{ik})(g^{kj}+\lambda\widetilde{g}^{kj})=0.$$

It follows that

$$\widetilde{g}_{ij}=-g_{ik}g_{jl}\widetilde{g}^{kl}. \tag{3.4.19}$$

Thus, (3.4.18) is rewritten as

$$\frac{d}{d\lambda}\Big|_{\lambda=0}\det\,(g_{ij}+\lambda\widetilde{g}_{ij})=-gg_{ij}\widetilde{g}^{ij}. \tag{3.4.20}$$

For the Einstein-Hilbert functional (3.4.12), we have

$$\begin{aligned}&\frac{d}{d\lambda}\Big|_{\lambda=0}F(g_{ij}+\lambda\widetilde{g}_{ij})\\ =&\int_{\mathscr{M}}\left[R_{ij}\widetilde{g}^{ij}\sqrt{-g}-\frac{1}{2\sqrt{-g}}R\frac{d}{d\lambda}\det\,(g_{ij}+\lambda\widetilde{g}_{ij})\right.\\ &\left.+g^{ij}\frac{d}{d\lambda}R_{ij}(g_{ij}+t\widetilde{g}_{ij})\sqrt{-g}\right]dx\Big|_{\lambda=0}\\ =&\int_{\mathscr{M}}(R_{ij}-\frac{1}{2}g_{ij}R)\widetilde{g}^{ij}\sqrt{-g}dx\\ &+\int_{\mathscr{M}}g^{ij}\frac{d}{d\lambda}\Big|_{\lambda=0}R_{ij}(g_{ij}+\lambda\widetilde{g}_{ij})\sqrt{-g}dx\quad(\text{by }(3.4.20)).\end{aligned}$$

In view of (3.4.5) and

$$\begin{aligned}\langle\delta F,\widetilde{g}_{ij}\rangle&=\int_{\mathscr{M}}\delta F\widetilde{g}^{ij}\sqrt{-g}dx,\\ \frac{d}{d\lambda}\Big|_{\lambda=0}R_{ij}(g_{ij}+\lambda\widetilde{g}_{ij})&=\frac{\partial R_{ij}}{\partial g_{kl}}\widetilde{g}_{kl},\end{aligned}$$

we arrive at

$$\int_{\mathscr{M}}\left[\delta F-\left(R_{ij}-\frac{1}{2}g_{ij}R\right)\right]\widetilde{g}^{ij}\sqrt{-g}dx=\int_{\mathscr{M}}g^{ij}\frac{\partial R_{ij}}{\partial g_{kl}}\widetilde{g}_{kl}\sqrt{-g}dx.$$

To verify (3.4.15), it suffices to prove that

$$\int_{\mathscr{M}}g^{ij}\delta R_{ij}\sqrt{-g}dx=0, \tag{3.4.21}$$

where δR_{ij} is the variational element

$$\delta R_{ij}=R_{ij}(g_{kl}+\delta g_{kl})-R_{ij}(g_{kl}),$$

which are equivalent to the following directional derivative:

$$\frac{d}{d\lambda}\Big|_{\lambda=0}R_{ij}(g_{ij}+\lambda\widetilde{g}_{ij}).$$

To get (3.4.21), we take R_{ij} in the form

$$R_{ij} = \frac{\partial \Gamma^k_{ki}}{\partial x^j} - \frac{\partial \Gamma^k_{ij}}{\partial x^k} + \Gamma^l_{ik}\Gamma^k_{lj} - \Gamma^l_{ij}\Gamma^k_{lk}. \tag{3.4.22}$$

By the Riemannian Geometry, for each point $x_0 \in M$ there exists a coordinate system under which

$$\Gamma^k_{ij}(x_0) = 0, \qquad \forall 1 \leqslant i,j,k \leqslant n. \tag{3.4.23}$$

It is known that the covariant derivatives of g_{ij} and g^{ij} are zero, i.e. $Dg_{ij} = 0$ and $Dg^{ij} = 0$. Hence we infer from (3.4.33) that

$$\frac{\partial g_{ij}(x_0)}{\partial x^k} = 0, \quad \frac{\partial g^{ij}(x_0)}{\partial x^k} = 0, \qquad \forall 1 \leqslant i,j,k \leqslant n.$$

By (3.4.22), at x_0, we have

$$g^{ij}\delta R_{ij} = g^{ij}\left(\frac{\partial}{\partial x^j}\delta\Gamma^k_{ik} - \frac{\partial}{\partial x^k}\delta\Gamma^k_{ij}\right) = \frac{\partial}{\partial x^k}\left(g^{ik}\delta\Gamma^l_{il} - g^{ij}\delta\Gamma^k_{ij}\right). \tag{3.4.24}$$

Although Γ^k_{ij} are not tensor fields, the variations

$$\delta\Gamma^k_{ij}(x) = \Gamma^k_{ij}(x+\delta x) - \Gamma^k_{ij}(x)$$

are (1,2)-tensor fields. Therefore at x_0, (3.4.24) can be rewritten as

$$g^{ij}\delta R_{ij} = \frac{\partial u^k}{\partial x^k} = \text{div } u, \quad \text{at } x_0 \in M. \tag{3.4.25}$$

where

$$u^k = g^{ik}\delta\Gamma^l_{il} - g^{ij}\delta\Gamma^k_{ij}$$

is a vector field. Since (3.4.25) is independent of the coordinate systems, in a general coordinate system the relation (3.4.25) becomes

$$g^{ij}\delta R_{ij} = \text{div } u = \frac{1}{\sqrt{-g}}\frac{\partial}{\partial x^k}(\sqrt{-g}u^k), \quad \text{at } x_0 \in \mathscr{M}. \tag{3.4.26}$$

As $x_0 \in \mathscr{M}$ is arbitrary, the formula (3.4.26) holds true on $\mathscr{M}$. Hence we have

$$\int_M g^{ij}\delta R_{ij}\sqrt{-g}dx = \int_{\mathscr{M}} \text{div } u\sqrt{-g}dx.$$

Since $\mathscr{M}$ is closed, i.e. $\partial\mathscr{M} = \varnothing$, it follows from (3.2.37) that

$$\int_{\mathscr{M}} g^{ij}\delta R_{ij}\sqrt{-g}dx = 0.$$

Thus we derive (3.4.21), and the derivative operator of the Einstein-Hilbert functional (3.4.12) is as given by (3.4.15).

3.4.4 Variational principle with div_A-free constraint

Let $\mathscr{M}$ be a closed manifold. A Riemannian metric G on $\mathscr{M}$ is a mapping

$$G : \mathscr{M} \to T_2^0\mathscr{M} = T^*\mathscr{M} \otimes T^*\mathscr{M},$$

which is symmetric and nondegenerate. Namely, in a local coordinate system, G can be expressed as

$$G = \{g_{ij}\} \quad \text{with} \quad g_{ij} = g_{ji}, \tag{3.4.27}$$

and the matrix (g_{ij}) is invertible on $\mathscr{M}$:

$$G^{-1} = (g^{ij}) = (g_{ij})^{-1} : \mathscr{M} \to T_0^2\mathscr{M} = T\mathscr{M} \otimes T\mathscr{M}.$$

If we regard a Riemannian metric $G = \{g_{ij}\}$ as a tensor field on the manifold $\mathscr{M}$, then the set of all metrics $G = \{g_{ij}\}$ on $\mathscr{M}$ constitute a topological space, called the space of Riemannian metrics on $\mathscr{M}$. The space of Riemannian metrics on $\mathscr{M}$ is defined by

$$\begin{aligned} W^{m,2}(\mathscr{M}, g) =& \{G |\ G \in W^{m,2}(T_2^0\mathscr{M}),\ G^{-1} \in W^{m,2}(T_0^2\mathscr{M}), \\ & G \text{ is the Riemannian metric on } \mathscr{M} \text{ as in (3.4.27)}\}. \end{aligned}$$

The space $W^{m,2}(\mathscr{M}, g)$ is a metric space, but not a Banach space. However, it is a subspace of the direct sum of two Sobolev spaces $W^{m,2}(T_2^0\mathscr{M})$ and $W^{m,2}(T_0^2\mathscr{M})$:

$$W^{m,2}(\mathscr{M}, g) \subset W^{m,2}(T_2^0\mathscr{M}) \oplus W^{m,2}(T_0^2\mathscr{M}).$$

A functional defined on $W^{m,2}(\mathscr{M}, g)$:

$$F :\ W^{m,2}(\mathscr{M}, g) \to \mathbb{R} \tag{3.4.28}$$

is called the functional of Riemannian metric. In general, the functional (3.4.28) can be expressed as

$$F(g_{ij}) = \int_{\mathscr{M}} f(g_{ij}, \cdots, \partial^m g_{ij})\sqrt{-g}dx. \tag{3.4.29}$$

Since (g^{ij}) is the inverse of (g_{ij}), we have

$$g_{ij} = \frac{1}{g} \times \text{the cofactor of } g^{ij}. \tag{3.4.30}$$

Therefore, $F(g_{ij})$ in (3.4.29) also depends on g^{ij}, i.e. putting (3.4.30) in (3.4.29) we get

$$F(g^{ij}) = \int_{\mathscr{M}} \widetilde{f}(g^{ij}, \cdots, \partial^m g^{ij})\sqrt{-g}dx. \tag{3.4.31}$$

We note that although $W^{m,2}(\mathscr{M},g)$ is not a linear space, but for a given element $g_{ij}\in W^{m,2}(\mathscr{M},g)$ and any symmetric tensor fields X_{ij},X^{ij}, there is a number $\lambda_0>0$ such that

$$\begin{aligned} &g_{ij}+\lambda X_{ij}\in W^{m,2}(\mathscr{M},g), \qquad \forall 0\leqslant|\lambda|<\lambda_0,\\ &g^{ij}+\lambda X^{ij}\in W^{m,2}(\mathscr{M},g), \qquad \forall 0\leqslant|\lambda|<\lambda_0. \end{aligned} \tag{3.4.32}$$

With (3.4.32), we can define the following derivative operators of the functional F:

$$\begin{aligned} &\delta_*F:\ W^{m,2}(\mathscr{M},g)\to W^{-m,2}(T^2_0\mathscr{M}),\\ &\delta^*F:\ W^{m,2}(\mathscr{M},g)\to W^{-m,2}(T^0_2\mathscr{M}), \end{aligned}$$

where $W^{-m,2}(E)$ is the dual space of $W^{m,2}(E)$, and δ_*F,δ^*F are defined by

$$\langle\delta_*F(g_{ij}),X\rangle=\frac{d}{d\lambda}\Big|_{\lambda=0}F(g_{ij}+\lambda X_{ij}), \tag{3.4.33}$$

$$\langle\delta^*F(g^{ij}),X\rangle=\frac{d}{d\lambda}\Big|_{\lambda=0}F(g^{ij}+\lambda X^{ij}). \tag{3.4.34}$$

For any give metric $g_{ij}\in W^{m,2}(\mathscr{M},g)$, the value of δ_*F and δ^*F at g_{ij} are second-order contra-variant and covariant tensor fields:

$$\begin{aligned} &\delta_*F(g_{ij}):\mathscr{M}\to T\mathscr{M}\times T\mathscr{M},\\ &\delta^*F(g_{ij}):\mathscr{M}\to T^*\mathscr{M}\times T^*\mathscr{M}. \end{aligned} \tag{3.4.35}$$

Theorem 3.24 *Let F be the functionals defined by (3.4.29) and (3.4.31). Then the following assertions hold true*:

1) *For any $g_{ij}\in W^{m,2}(\mathscr{M},g),\delta_*F(g_{ij})$ and $\delta^*F(g_{ij})$ are symmetric tensor fields.*

2) *If $\{g_{ij}\}\in W^{m,2}(\mathscr{M},g)$ is an extremum point of F, i.e. $\delta F(g_{ij})=0$, then $\{g^{ij}\}$ is also an extremum point of F.*

3) *δ_*f and δ^*F have the following relation*

$$(\delta^*F(g_{ij}))^{kl}=-g^{kr}g^{ls}(\delta_*F(g_{ij}))_{rs},$$

where $(\delta^*F)^{kl}$ and $(\delta_*F)_{kl}$ are the components of δ^*F and δ_*F.

Proof We only need to verify Assertion (3). In view of $g_{ik}g^{kj}=\delta^j_i$, we have the variational relation

$$\delta(g_{ik}g^{kj})=g_{ik}\delta g^{kj}+g^{kj}\delta g_{ik}=0.$$

It implies that

$$\delta g^{kl}=-g^{ki}g^{lj}\delta g_{ij}. \tag{3.4.36}$$

In addition, in (3.4.33) and (3.4.34),

$$\lambda X_{ij} = \delta g_{ij}, \quad \lambda X^{ij} = \delta g^{ij}, \quad \lambda \neq 0 \text{ small}.$$

Therefore, by (3.4.36) we get

$$\langle (\delta_* F)_{kl}, \delta g^{kl} \rangle = -\langle (\delta_* F)_{kl}, g^{ki} g^{lj} \delta g_{ij} \rangle = -\langle g^{ki} g^{lj} (\delta_* F)_{kl}, \delta g_{ij} \rangle = \langle (\delta^* F)^{ij}, \delta g_{ij} \rangle.$$

Hence we have

$$(\delta^* F)^{ij} = -g^{ki} g^{lj} (\delta_* F)_{kl}.$$

Thus Assertion (3) follows and the proof is complete. □

We are now in position to consider the variation with div_A-free constraints. We know that an extremum point g_{ij} of a metric functional is a solution of the equation

$$\delta F(g_{ij}) = 0, \tag{3.4.37}$$

in the sense that for any $X_{kl} = X_{lk} \in L^2(T_2^0 \mathscr{M})$,

$$\langle \delta F(g_{ij}), X \rangle = \frac{d}{d\lambda}\Big|_{\lambda=0} F(g_{ij} + \lambda X_{ij})|_{\lambda=0} = \int_{\mathscr{M}} (\delta^* F(g_{ij}))^{kl} X_{kl} \sqrt{-g} dx = 0. \tag{3.4.38}$$

Note that the solution g_{ij} of (3.4.37) in the usual sense should satisfy

$$\langle \delta F(g_{ij}), X \rangle = 0, \qquad \forall X \in L^2(T_2^0 \mathscr{M}). \tag{3.4.39}$$

Notice that (3.4.38) has a symmetric constraint on the variational elements X_{ij}: $X_{ij} = X_{ji}$. Therefore, comparing (3.4.38) with (3.4.39), we may wonder if a solution g_{ij} satisfying (3.4.38) is also a solution of (3.4.39). Fortunately, note that $L^2(T_2^0 \mathscr{M})$ can be decomposed into a direct sum of symmetric and anti-symmetric spaces as follows

$$\begin{aligned} &L^2(T_2^0 \mathscr{M}) = L_s^2(T_2^0 \mathscr{M}) \oplus L_c^2(T_2^0 \mathscr{M}), \\ &L_s^2(T_2^0 \mathscr{M}) = \{u \in L^2(T_2^0 \mathscr{M}) |\ u_{ij} = u_{ji}\}, \\ &L_c^2(T_2^0 \mathscr{M}) = \{u \in L^2(T_2^0 \mathscr{M}) |\ u_{ij} = -u_{ji}\}, \end{aligned}$$

and $L_s^2(T_2^0 \mathscr{M})$ and $L_c^2(T_2^0 \mathscr{M})$ are orthogonal:

$$\begin{aligned} \int_M g^{ik} g^{jl} u_{ij} v_{kl} \sqrt{-g} dx = & -\int_{\mathscr{M}} g^{ik} g^{jl} u_{ij} v_{lk} \sqrt{-g} dx \\ = & \ 0, \qquad \forall u \in L_s^2(T_2^0 \mathscr{M}), \quad v \in L_c^2(T_2^0 \mathscr{M}). \end{aligned}$$

Thus, due to the symmetry of $\delta F(g_{ij})$, the solution g_{ij} of (3.4.37) satisfying (3.4.38) must also satisfy (3.4.39). Hence the solutions of (3.4.37) in the sense of (3.4.38) are the solutions in the usual sense.

However, if we consider the variations of F under the div_A-free constraint, then the extremum points of F are not solutions of (3.4.37) in the usual sense. Motived by physical considerations, we now introduce variations with div_A-free constraints.

Definition 3.25 *Let $F(u)$ be a functional of tensor fields u. We say that u_0 is an extremum point of $F(u)$ under the* dev$_A$*-free constraint, if*

$$\langle \delta F(u_0), X\rangle = \frac{d}{d\lambda}\Big|_{\lambda=0} F(u_0+\lambda X) = 0, \qquad \forall \mathrm{div}_A X = 0, \tag{3.4.40}$$

where div$_A$ *is as defined in (3.2.38).*

In particular, if F is a functional of Riemannian metrics, and the solution $u_0 = g_{ij}$ is a Riemannian metric, then the differential operator D_A in div$_A X$ *in (3.4.40) is given by*

$$D_A = D + A, \quad D = \partial + \Gamma, \tag{3.4.41}$$

and the connection Γ is taken at the extremum point $u_0 = g_{ij}$.

We have the following theorems for div$_A$-free constraint variations.

Theorem 3.26 *Let $F: W^{m,2}(\mathscr{M}, g) \to \mathbb{R}^1$ be a functional of Riemannian metrics. Then there is a vector field $\Phi \in H^1(T\mathscr{M})$ such that the extremum points $\{g_{ij}\}$ of F with the* div$_A$*-free constraint satisfy the equation*

$$\delta F(g_{ij}) = D\Phi + A \otimes \Phi, \tag{3.4.42}$$

where D is the covariant derivative operator as in (3.4.41).

Theorem 3.27 *Let $F: H^m(T\mathscr{M}) \to \mathbb{R}^1$ be a functional of vector fields. Then there is a scalar function $\varphi \in H^1(\mathscr{M})$ such that for a given vector field A, the extremum points u of F with the* div$_A$*-free constraint satisfy the equation*

$$\delta F(u) = (\partial + A)\varphi. \tag{3.4.43}$$

Proof of Theorems 3.26 and 3.27 First we prove Theorem 3.26. By (3.4.40), the extremum points $\{g_{ij}\}$ of F with the div$_A$-free constraint satisfy

$$\int_{\mathscr{M}} \delta F(g_{ij}) \cdot X \sqrt{-g}\, dx = 0, \qquad \forall X \in L^2(T_0^2\mathscr{M}) \text{ with } \mathrm{div}_A X = 0.$$

It implies that

$$\delta F(g_{ij}) \perp L_D^2(T_2^0\mathscr{M}) = \{v \in L^2(T_2^0\mathscr{M}) |\ \mathrm{div}_A v = 0\}. \tag{3.4.44}$$

By Theorem 3.17, $L^2(T_2^0\mathscr{M})$ can be orthogonally decomposed into

$$L^2(T_2^0\mathscr{M}) = L_D^2(T_2^0\mathscr{M}) \oplus G^2(T_2^0\mathscr{M}),$$
$$G^2(T_2^0\mathscr{M}) = \{D_A\Phi |\ \Phi \in H^1(T_1^0\mathscr{M})\}.$$

Hence it follows from (3.4.44) that

$$\delta F(g_{ij}) \in G^2(T_2^0\mathscr{M}),$$

which means that the equality (3.4.42) holds true.

To prove Theorem 3.27, for an extremum vector field u of F with the div_A-free constraint, we derive in the same fashion that u satisfies the following equation

$$\int_{\mathscr{M}} \delta F(u)\cdot X\sqrt{-g}dx = 0, \qquad \forall X\in L^2(T\mathscr{M}) \text{ with } \text{div}_A x = 0. \tag{3.4.45}$$

In addition, Theorem 3.17 means that

$$\begin{aligned}
&L^2(T\mathscr{M}) = L^2_D(T\mathscr{M})\oplus G^2(T\mathscr{M}),\\
&L^2_D(T\mathscr{M}) = \{v\in L^2(T\mathscr{M})|\ \text{div}_A v = 0\},\\
&G^2(T\mathscr{M}) = \{D_A\varphi|\ \varphi\in H^1(\mathscr{M})\}.
\end{aligned}$$

Then we infer from (3.4.45) that

$$\delta F(u)\in G^2(T\mathscr{M}).$$

Thus we deduce the equality (3.4.43).

The proofs of Theorems 3.26 and 3.27 are complete. □

3.4.5 Scalar potential theorem

In Theorem 3.26, if the vector field A in D_A is zero, and the first Betti number $\beta_1(\mathscr{M}) = 0$ for $\mathscr{M}$, then we have the following scale potential theorem. This result is also important for the gravitational field equations and the theory of dark matter and dark energy introduced in Chapter 7.

Theorem 3.28 (Scalar Potential Theorem) *Assume that the first Betti number of $\mathscr{M}$ is zero, i.e. $\beta_1(\mathscr{M}) = 0$. Let F be a functional of Riemannian metrics. Then there is a scalar field $\varphi\in H^2(\mathscr{M})$ such that the extremum points $\{g_{ij}\}$ of F with divergence-free constraint satisfy the equation*

$$(\delta F(g_{ij}))_{kl} = D_k D_l\varphi. \tag{3.4.46}$$

Proof Let $\{g_{ij}\}$ be an extremum point of F under the divergence-free constraint:

$$\int_{\mathscr{M}} (\delta F(g_{ij}))_{kl} X^{kl}\sqrt{-g}dx = 0, \qquad \forall X = \{X_{kl}\} \text{ with } D_k X^{kl} = 0.$$

By Theorem 3.26 and $A = 0$ in (3.4.42), $\delta F(g_{ij})$ is in the form

$$(\delta F(g_{ij}))_{kl} = D_k\Phi_l, \tag{3.4.47}$$

for some $\{\Phi_l\}\in H^1(T^*\mathscr{M})$. By Theorem 3.24, $\delta F(g_{ij})$ is symmetric. Hence we have

$$D_k\Phi_l = D_l\Phi_k. \tag{3.4.48}$$

In addition, by

$$D_k\Phi_l = \frac{\partial \Phi_l}{\partial x^k} - \Gamma^j_{kl}\Phi_j,$$

and $\Gamma^j_{kl} = \Gamma^j_{lk}$, it follows from (3.4.48) that

$$\frac{\partial \Phi_l}{\partial x^k} = \frac{\partial \Phi_k}{\partial x^l}. \tag{3.4.49}$$

By assumption, the first Betti number of $\mathscr{M}$ is zero, i.e. the first homology of M is zero: $H_1(\mathscr{M}) = 0$. It follows from the de Rham theorem that if

$$d(\Phi_k dx^k) = \left(\frac{\partial \Phi_k}{\partial x^l} - \frac{\partial \Phi_l}{\partial x^k}\right) dx^l \wedge dx^k = 0,$$

then there exists a scalar function φ such that

$$d\varphi = \frac{\partial \varphi}{\partial x^k} dx^k = \Phi_k dx^k.$$

Thus, we infer from (3.4.49) that

$$\Phi_k = \frac{\partial \varphi}{\partial x^k} \qquad \text{for some } \varphi \in H^2(\mathscr{M}).$$

Therefore, we derive (3.4.46) from (3.4.47), and the proof is complete. □

If the first Betti number $\beta_1(\mathscr{M}) \neq 0$, then we have the following theorem.

Theorem 3.29 *Let the first Betti number of $\mathscr{M}$ is $\beta_1(\mathscr{M}) = N$ with $N \neq 0$. Then there are a scalar field $\varphi \in H^2(\mathscr{M})$ and N vector fields ψ^j $(1 \leqslant j \leqslant N)$ in $H^1(T^*\mathscr{M})$ such that the extremum point $\{g_{ij}\}$ of F with divergence-free constraint satisfies the equation*

$$(\delta F(g_{ij}))_{kl} = D_k D_l \varphi + \sum_{j=1}^{N} \alpha_j D_k \psi^j_l, \tag{3.4.50}$$

$$D^k D_k \psi^j_l = -R^k_l \psi^j_k \qquad \text{for } 1 \leqslant j \leqslant N,\ 1 \leqslant l \leqslant n, \tag{3.4.51}$$

where α_j $(1 \leqslant j \leqslant N)$ are constants, $R^k_l = g^{kj}R_{jl}$ and R_{jl} are the Ricci curvature tensors.

Proof Since the first Betti number $\beta_1(\mathscr{M}) = N(\neq 0)$, by the de Rham theorem, there are N closed 1-forms

$$\omega_j = \psi^j_k dx^k \in H^1_d(\mathscr{M}) \qquad \text{for } 1 \leqslant j \leqslant N, \tag{3.4.52}$$

they constitute a basis for the 1-dimensional de Rham homology $H^1_d(\mathscr{M})$. Hence ω_j $(1 \leqslant j \leqslant N)$ are not exact, and satisfy that

$$d\omega_j = \left(\frac{\partial \psi^j_k}{\partial x^l} - \frac{\partial \psi^j_l}{\partial x^k}\right) dx^l \wedge dx^k = 0,$$

which imply that

$$\frac{\partial \psi_k^j}{\partial x^l} = \frac{\partial \psi_l^j}{\partial x^k} \qquad \text{for } 1 \leqslant j \leqslant N.$$

or equivalently

$$D_l \psi_k^j = D_k \psi_l^j \qquad \text{for } 1 \leqslant j \leqslant N.$$

Namely, $\nabla \psi^j \in L^2(T^*\mathscr{M} \otimes T^*\mathscr{M})$ are symmetric second-order covariant tensor fields. Hence any covector field $\Phi \in L^2(T^*\mathscr{M})$ satisfying

$$D_l \Phi_k = D_k \Phi_l \tag{3.4.53}$$

must be in the following form

$$\Phi_k = D_k \varphi + \sum_{j=1}^{N} \alpha_j \psi_k^j, \tag{3.4.54}$$

where α_j $(1 \leqslant j \leqslant N)$ are constants, φ is some scalar field. Hence, by $A = 0$ in D_A, the equation (3.4.42) in Theorem 3.26 can be expressed as

$$(\delta F(g_{ij}))_{lk} = D_l \Phi_k,$$

where Φ_k satisfy (3.4.53) and (3.4.54), which are the equations given by (3.4.50).

On the other hand, by the Hodge decomposition theorem, the 1-forms in (3.4.52) are harmonic, i.e.

$$d\omega_j = 0, \quad \delta \omega_j = 0 \qquad \text{for } 1 \leqslant j \leqslant N.$$

It follows that the covector fields ψ^j $(1 \leqslant j \leqslant N)$ in (3.4.52) satisfy

$$\Delta \psi^j = 0 \qquad \text{for } 1 \leqslant j \leqslant N, \tag{3.4.55}$$

where $\Delta = d\delta + \delta d$ is the Laplace-Beltrami operator as defined in (3.2.31). Hence, the equations in (3.4.55) can be equivalently rewritten in the form

$$D^k D_k \psi_l^j = -R_l^k \psi_k^j \qquad \text{for } 1 \leqslant j \leqslant N, \quad 1 \leqslant l \leqslant n,$$

which are exactly the equations in (3.4.51). The proof is complete. □

3.5 $SU(N)$ Representation Invariance

3.5.1 $SU(N)$ gauge representation

We briefly recapitulate the $SU(N)$ gauge theory. In the general case, a set of $SU(N)$ gauge fields consists of $K = N^2 - 1$ vector fields A_μ^a and N spinor fields ψ^j:

$$A_\mu^1, \cdots, A_\mu^K, \quad \Psi = (\psi^1, \cdots, \psi^N)^{\mathrm{T}}. \tag{3.5.1}$$

For the fields (3.5.1), a gauge invariant functional is the Yang-Mills action:

$$L_{YM} = \int [\mathscr{L}_G + \mathscr{L}_D]\,dx, \tag{3.5.2}$$

where $\mathscr{L}_G$ and $\mathscr{L}_D$ are the gauge field section and Dirac spinor section, and are written as

$$\begin{aligned} &\mathscr{L}_G = -\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu a}, \\ &F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\lambda^a_{bc}A^b_\mu A^c_\mu, \end{aligned} \tag{3.5.3}$$

and

$$\begin{aligned} &\mathscr{L}_D = \overline{\Psi}(i\gamma^\mu D_\mu - m)\Psi, \\ &D_\mu = \partial_\mu + igA^a_\mu \tau_a, \end{aligned} \tag{3.5.4}$$

where τ_a $(1 \leqslant a \leqslant K = N^2 - 1)$ are the generators of $SU(N)$.

The actions (3.5.2)-(3.5.4) are invariant under both the Lorentz transformation and the $SU(N)$ gauge transformation as follows

$$\begin{aligned} &\widetilde{\Psi} = \Omega\Psi, \\ &\widetilde{A}^a_\mu \tau_a = \Omega A^a_\mu \tau_a \Omega^{-1} + \frac{i}{g}(\partial_\mu \Omega)\Omega^{-1}, \end{aligned} \tag{3.5.5}$$

and $\Omega \subset SU(N)$ can be expressed as

$$\Omega = e^{i\theta^a \tau_a}, \quad \tau_a \text{ is in } (3.5.4).$$

For the gauge theory, a vary basic and important problem is that in the gauge transformation (3.5.5) the generators of $SU(N)$ given by

$$\{\tau_a |\ 1 \leqslant a \leqslant K\}, \tag{3.5.6}$$

have infinite numbers of families, and each family of (3.5.6) corresponds to a set of gauge fields:

$$\{\tau_a |\ 1 \leqslant a \leqslant K\} \qquad \leftrightarrow \qquad \{A^a_\mu |\ 1 \leqslant a \leqslant K\}. \tag{3.5.7}$$

Now, we assume that the generators of (3.5.6) undergo a linear transformation as follows

$$\widetilde{\tau}_b = x^a_b \tau_a, \tag{3.5.8}$$

where (x^a_b) is a K-th order complex matrix. Then the corresponding gauge field A^a_μ in (3.5.7) has to change. Namely, under the transformation (3.5.8) A^a_μ will transform as

$$\widetilde{A}^a_\mu = y^a_b A^b_\mu, \tag{3.5.9}$$

and (y^a_b) is a K-order matrix depending on (x^a_b).

Intuitively, any gauge theory should be independent of the choice of $\{\tau^a\}$, otherwise the basically logical rationality will be broken. In other words, the Yang-Mills density (3.5.3) should be invariant under the transformation (3.5.9). However, in view of (3.5.3), if

$$A^a_\mu \to y^a_b A^b_\mu \ \Rightarrow\ F^a_{\mu\nu} \to y^a_b F^b_\mu,$$

then $\mathscr{L}_G$ will be changed as

$$\mathscr{L}_G = F^a_{\mu\nu}F^{\mu\nu a} \to y^a_b y^a_c F^b_{\mu\nu}F^{\mu\nu c}.$$

Hence the Yang-Mills action (3.5.2) violates the invariance.

To solve this problem, the authors have developed in (Ma and Wang, 2014h) a mathematical theory of $SU(N)$ representation invariance, where the $SU(N)$ tensors and the Riemannian metric on $SU(N)$ are defined. Furthermore the Yang-Mills action is revised. In this subsection, we shall introduce this theory.

3.5.2 Manifold structure of $SU(N)$

To establish the representation invariance theory for $SU(N)$ gauge fields, we need to introduce the $SU(N)$ tensors and the Riemannian metric defined on $SU(N)$. The main objective of this subsection is to introduce some basic concepts on $SU(N)$, including manifold structure, tangent space, coordinate systems and coordinate transformations.

1. *Manifold structure on* $SU(N)$. In mathematics, a space $\mathscr{M}$ is an n-dimensional manifold means that for each point $p \in \mathscr{M}$ there is a neighborhood $U \subset \mathscr{M}$ of p, such that U is homeomorphic to $\mathbb{R}^n$, i.e. there exists an one to one mapping

$$\psi : U \to \mathbb{R}^n$$

and ψ has a continuous inverse $\psi^{-1} : \mathbb{R}^n \to U$.

The group $SU(N)$ consists of all N-th order unitary matrices with unit determinant:

$$SU(N) = \{A \mid A \text{ is an N-th order matrix } , A^\dagger A = I, \ \det A = 1\}.$$

Each matrix $A \subset SU(N)$ can be written as

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix},$$

and $a_{kl} = x_{kl} + iy_{kl} \in \mathbb{C}$ $(1 \leqslant k,l \leqslant N)$. Thus the matrix A can be regarded as a point p_A in $\mathbb{R}^{2N^2}$:

$$p_A = (x_{11}, y_{11}, \cdots, x_{1N}, y_{1N}, \cdots, x_{N1}, y_{N1}, \cdots, x_{NN}, y_{NN}) \in \mathbb{R}^{2N^2}. \tag{3.5.10}$$

Therefore, we have that $SU(N)$ can be regarded as a subspace of $\mathbb{R}^{2N^2}$.

By $A^\dagger A = I$ and $\det A = 1$, the entries a_{kl} $(1 \leqslant k,l \leqslant N)$ of A satisfy

$$\begin{aligned} & a_{ij}a^*_{jk} = \delta_{ik} \qquad \text{for } 1 \leqslant i,k \leqslant N, \\ & \det\,(a_{ij}) = 1, \end{aligned} \tag{3.5.11}$$

which are N^2+1 equations, as constraints for the point p_A in (3.5.10). Hence $SU(N)$ can be regarded as subspace of $\mathbb{R}^{2N^2}$ has dimension N^2-1.

Mathematically $SU(N)$ is a manifold. In fact, at any point p_A of (3.5.10), each equation of (3.5.11) represents a hypersurface near p_A in $\mathbb{R}^{2N^2}$:

$$\Sigma_{ik}: \; a_{ij}a^*_{jk} = \delta_{ik} \qquad \text{for } 1 \leqslant i,k \leqslant N,$$
$$\Sigma_1: \; \det\,(a_{ij}) = 1,$$

and the N^2+1 hypersurface Σ_{ik} and Σ_1 transversally interact in $\mathbb{R}^{2N^2}$ to constitute an N^2-1 dimensional surface $\Gamma(p_A)$ near each $p_A \in \mathbb{R}^{2N^2}$, and the sum of all $\Sigma(p_A)$ is the $SU(N)$ space:

$$SU(N) = \bigcup_{p_A} \Gamma(p_A).$$

Hence, $SU(N)$ possesses the manifold structure.

2. *Tangent space* $T_ASU(N)$. Since $SU(N)$ is an N^2-1 dimensional manifold, at each point $A \in SU(N)$ there is a tangent space, denoted by $T_ASU(N)$, which is an N^2-1 dimensional linear space, as shown in Figure 3.2.

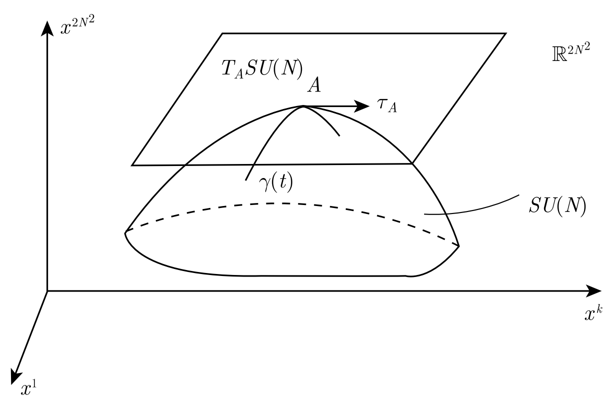


Figure 3.2 Tangent space $T_ASU(N)$ at $A \in SU(N)$

Now we derive some properties of tangent vectors τ on $T_ASU(N)$. To this end, let $\gamma(t) \subset SU(N)$ be a curve passing through the point $A \in SU(N)$ with $\tau_A \in T_ASU(N)$ as its tangent vector at A, as shown in Figure 3.2. Let $\gamma(0) = A$. Then the curve $\gamma(t)$ satisfies the following equation

$$\begin{aligned} &\frac{d\gamma(t)}{dt} = \tau_t, \qquad \tau_t \in TSU(N) \text{ with } \tau_A = \tau|_{t=0}, \\ &\gamma(0) = A. \end{aligned} \tag{3.5.12}$$

For infinitesimal t, the solution of (3.5.12) is

$$\gamma(t) = A + t\tau_A \subset SU(N).$$

It follows that

$$(A + t\tau_A)^\dagger (A + t\tau_A) = I. \tag{3.5.13}$$

As $A^\dagger A = I$ and t is infinitesimal, we deduce from (3.5.13) that

$$A^\dagger \tau_A + \tau_A^\dagger A = 0. \tag{3.5.14}$$

Hence, (3.5.14) is the condition for an N-th order complex matrix $\tau \in T_A SU(N)$: $A^\dagger \tau$ is anti-Hermitian. Namely,

$$T_A SU(N) = \{\tau \mid \tau \text{ satisfies } (3.5.14)\}. \tag{3.5.15}$$

Note that

$$A + t\tau_A = A(I + tA^\dagger \tau_A) = Ae^{tA^\dagger \tau_A}, \tag{3.5.16}$$

for infinitesimal t. If we replace τ by $i\tau$, then (3.5.15) can be expressed as

$$T_A SU(N) = \{i\tau \mid A^\dagger \tau = \tau^\dagger A,\ \text{Tr}(A^\dagger \tau) = 0\}. \tag{3.5.17}$$

By (3.5.17), $\forall A \in SU(N)$ there is a neighborhood $U_A \subset SU(N)$ such that for any $\Omega \in U_A, \Omega$ can be written as

$$\Omega = Ae^{iA^\dagger \tau} \qquad \text{for some } i\tau \in T_A SU(N). \tag{3.5.18}$$

Here the traceless condition in (3.5.17)

$$\text{Tr}(A^\dagger \tau) = 0$$

is derived by (2.2.51) and $\det \Omega = 1$ for Ω as in (3.5.18).

In particular, at the unit matrix $e = I$, we have

$$\begin{aligned} &T_e SU(N) = \{i\tau \mid \tau^\dagger = \tau,\ \text{Tr}\tau = 0\}, \\ &\Omega = e^{i\tau}, \quad \tau \in T_e SU(N), \end{aligned} \tag{3.5.19}$$

where $\Omega \subset SU(N)$ is in a neighborhood of $e = I$.

3. *Coordinate systems on* $TSU(N)$. Since the representation (3.5.17)-(3.5.18) of $SU(N)$ is essentially the same for all $A \in SU(N)$, it suffices to only consider the representation (3.5.19) of $SU(N)$ at the unit matrix $e = I$.

It is known that $T_e SU(N)$ is an $N^2 - 1$ dimensional linear space. Therefore we can take a coordinate basis, called the generator basis of $SU(N)$, denoted by

$$\tau_1, \cdots, \tau_K, \qquad K = N^2 - 1, \tag{3.5.20}$$

such that for any $\tau \in T_eSU(N)$, we have

$$\tau = \theta^a \tau_a, \tag{3.5.21}$$

and θ^a $(1 \leqslant a \leqslant K)$ are complex numbers, called the coordinate system on $T_eSU(N)$. In this case, Ω in (3.5.19) can be expressed as

$$\Omega = e^{i\theta^a \tau_a}, \qquad \Omega \subset SU(N).$$

3.5.3 $SU(N)$ tensors

Let $(\tau_1, \cdots, \tau_K) \subset T_eSU(N)$ be a generator basis of $SU(N)$. If the basis undergoes a linear transformation:

$$\begin{aligned} &\widetilde{\tau}_a = x_a^b \tau_b, \\ &X = (x_a^b) \text{ is a complex } K\text{-th order matrix,} \end{aligned} \tag{3.5.22}$$

then the coordinate system $(\theta^1, \cdots, \theta^K)$ of $T_eSU(N)$ also undergoes a corresponding transformation as follows

$$\begin{aligned} &\widetilde{\theta}^a = y_b^a \theta^b, \\ &Y = (y_b^a) \text{ is a complex } K\text{-th order matrix.} \end{aligned} \tag{3.5.23}$$

Since the expression (3.5.21) is independent of the choice of the generator bases of $SU(N)$, we have

$$\widetilde{\theta}^a \widetilde{\tau}_a = (\theta^1, \cdots, \theta^K) Y^{\mathrm{T}} X \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_K \end{pmatrix} = \theta^a \tau_a,$$

which requires that

$$Y = (X^{-1})^{\mathrm{T}}. \tag{3.5.24}$$

Thus, we see that $(\theta^1, \cdots, \theta^K)$ is a first order contra-variant tensor defined on $T_eSU(N)$. We are now ready to define more general $SU(N)$ tensors.

Definition 3.30 ($SU(N)$ Tensors) *Let T be given as*

$$T = \{T^{a_1 \cdots a_i}_{b_1 \cdots b_j} \mid 1 \leqslant a_k, b_l \leqslant K = N^2 - 1\}.$$

We say that T is a (i, j)-type of $SU(N)$ tensor, under the generator basis transformation as (3.5.22), we have

$$\widetilde{T}^{a_1 \cdots a_i}_{b_1 \cdots b_j} = y^{a_1}_{c_1} \cdots y^{a_i}_{c_i} x^{d_1}_{b_1} \cdots x^{d_j}_{b_j} T^{c_1 \cdots c_i}_{d_1 \cdots d_j},$$

where (x_b^a) and (y_b^a) are as in (3.5.22) and (3.5.23).

Based on Definition 3.30, it is easy to see that the $SU(N)$ gauge fields $(A^1_\mu, \cdots, A^a_\mu)$ is a contra-variant $SU(N)$ tensor. In other words, under (3.5.22), $(A^1_\mu, \cdots, A^a_\mu)$ transforms as

$$\widetilde{A}^a_\mu = y^a_b A^b_\mu, \quad Y = (y^a_b) \text{ as in (3.5.24)}. \tag{3.5.25}$$

This can be seen from the fact that the operator $A^a_\mu \tau_a$ in the differential operator D_μ in (3.5.4) is independent of generator bases τ_a of $SU(N)$.

We now verify that the structure constants λ^a_{bc} of $SU(N)$ constitute a (1,2)-type of $SU(N)$ tensor. By the definition of λ^a_{bc},

$$[\tau_b, \tau_c] = \tau_b \tau_c^\dagger - \tau_c \tau_b^\dagger = i\lambda^a_{bc}\tau_a.$$

By (3.5.22),

$$[\widetilde{\tau}_b, \widetilde{\tau}_c] = x^a_b x^d_c [\tau_a, \tau_d] = i x^a_b x^d_c \lambda^f_{ad} \tau_f,$$

and by definition

$$[\widetilde{\tau}_b, \widetilde{\tau}_c] = i\widetilde{\lambda}^a_{bc}\widetilde{\tau}_a = i\widetilde{\lambda}^a_{bc} x^d_a \tau_d.$$

Then it follows that

$$\widetilde{\lambda}^a_{bc} = x^f_b x^g_c y^a_d \lambda^d_{fg}, \tag{3.5.26}$$

which means that $\{\lambda^a_{bc}\}$ is a (1,2)-type $SU(N)$ tensor.

Next, we introduce two second-order covariant $SU(N)$ tensors $\mathscr{G}_{ab}$ and g_{ab}, and later we shall prove that they are equivalent.

Let $A \in SU(N)$. Then the tangent space $T_A SU(N)$ is given by (3.5.17). Let

$$\omega_1, \cdots, \omega_K \in T_A SU(N), \tag{3.5.27}$$

be a generator basis of $SU(N)$ at A.

1) *$SU(N)$ tensor $\mathscr{G}_{ab}(A)$.* By the basis (3.5.27) we can get a 2-covariant $SU(N)$ tensor defined by

$$\mathscr{G}_{ab}(A) = \frac{1}{2}\mathrm{tr}(\omega_a \omega_b^\dagger), \quad A \in SU(N), \tag{3.5.28}$$

where ω_a $(1 \leqslant a \leqslant K)$ are as in (3.5.27).

2) *$SU(N)$ tensor g_{ab}.* The structure constants λ^a_{bc} generated by the generators ω_a in (3.5.27) satisfy

$$[\omega_b, \omega_c] = i\lambda^a_{bc}\omega_a A^\dagger, \qquad \forall A \in SU(N) \tag{3.5.29}$$

then we can define another 2-covariant tensor g_{ab} on $T_A SU(N)$ by the structure constants λ^a_{bc} as follows

$$g_{ab} = \frac{1}{4N}\lambda^c_{ad}\lambda^d_{cb}. \tag{3.5.30}$$

The following theorem shows that both $\mathscr{G}_{ab}$ and g_{ab} are symmetric second-order $SU(N)$ tensors.

Theorem 3.31 *The fields $\mathscr{G}_{ab}(A)$ and g_{ab} given by (3.5.28) and (3.5.30) are 2-order symmetric $SU(N)$ tensors, and for any $A \in SU(N)$, the generator basis (3.5.27) satisfies (3.5.29), where λ^a_{bc} are the structure constants of $SU(N)$ independent of A.*

Proof We first consider g_{ab}. By the anti-symmetry of the structure constants:

$$\lambda_{ab}^{c} = -\lambda_{ba}^{c},$$

we deduce the symmetry of g_{ab}:

$$g_{ab} = \lambda_{ad}^{c}\lambda_{cb}^{d} = \lambda_{da}^{c}\lambda_{bc}^{d} = g_{ba}.$$

Now we verify the symmetry of $\mathscr{G}_{ab}$. By (3.5.17) the basis $(\omega_1, \cdots, \omega_K)$ of $T_A SU(N)$ satisfy that

$$A^{\dagger}\omega_a = (A^{\dagger}\omega_a)^{\dagger}, \quad \mathrm{tr}(A^{\dagger}\omega_a) = 0, \quad A^{\dagger} = A^{-1}, \tag{3.5.31}$$

Namely $A^{\dagger}\omega_a$ are Hermitian. Let $A^{\dagger}\omega_a = \tau_a$, then

$$\tau_a = \tau_a^{\dagger}, \quad \mathrm{tr}\tau_a = 0. \tag{3.5.32}$$

Thus we have

$$\mathscr{G}_{ab} = \frac{1}{2}\mathrm{tr}(\omega_a\omega_b^{\dagger}) = \frac{1}{2}\mathrm{tr}(A\tau_a\tau_b^{\dagger}A^{\dagger}) = \frac{1}{2}\mathrm{tr}(\tau_a\tau_b^{\dagger}). \tag{3.5.33}$$

Thanks to (3.5.32), $(\tau_1, \cdots, \tau_K) \in T_e SU(N)$. Hence we have

$$\tau_a\tau_b^{\dagger} = \tau_b\tau_a^{\dagger} + i\lambda_{ab}^{c}\tau_c, \quad \mathrm{tr}\tau_c = 0. \tag{3.5.34}$$

It follows from (3.5.33) and (3.5.34) that

$$\mathscr{G}_{ab} = \frac{1}{2}\mathrm{tr}(\tau_a\tau_b^{\dagger}) = \frac{1}{2}\mathrm{tr}(\tau_b\tau_a^{\dagger}) = \mathscr{G}_{ba}.$$

Hence $\mathscr{G}_{ab}$ is symmetry.

Finally we prove that for any $A \in SU(N)$, each generator basis $(\omega_1, \cdots, \omega_K)$ of $T_A SU(N)$ satisfy (3.5.29). In fact, by (3.5.31), $\tau_a = A^{\dagger}\omega_a$ $(1 \leqslant a \leqslant K)$ constitute a basis of $T_e SU(N)$. Therefore τ_a satisfies (3.5.34), i.e.

$$A^{\dagger}[\omega_a, \omega_b]A = i\lambda_{ab}^{c}A^{\dagger}\omega_c.$$

Hence we obtain (3.5.29). The proof is complete. □

3.5.4 Intrinsic Riemannian metric on $SU(N)$

By Theorem 3.31, $\mathscr{G}_{ab}(A)$ and g_{ab} are symmetric. We now show that both $\mathscr{G}_{ab}$ and g_{ab} are positive definite, and consequently $\mathscr{G}_{ab}(A)$ is an intrinsic Riemannian metric on $SU(N)$.

Theorem 3.32 *For the 2-order $SU(N)$ tensors g_{ab} and $\mathscr{G}_{ab}$, the following assertions hold true*:

1) *The tensor $\{g_{ab}\}$ given by (3.5.30) is positive definite*; *and*

2) *For each point $A \in SU(N)$, there is a coordinate system in where $\mathscr{G}_{ab}(A) = g_{ab}$. Therefore $\mathscr{G}_{ab}(A)$ is a Riemannian metric on $SU(N)$.*

Proof First, we prove Assertion 1). As a 2-order covariant tensor, g_{ab} transforms as

$$(\widetilde{g}_{ab}) = X(g_{ab})X^{\mathrm{T}}, \quad X = (x_b^a) \text{ as in (3.5.22)}. \tag{3.5.35}$$

Hence, if for a given basis $\{\tau_a |\ 1 \leqslant a \leqslant K\}$ of $T_eSU(N)$ we can verify that $g_{ab} = \dfrac{1}{4N}\lambda_{ad}^c\lambda_{cb}^d$ is positive definition, then by (3.5.35) we derive Assertion 1). In the following, we proceed first for $SU(2)$ and $SU(3)$, then for the general $SU(N)$.

For $SU(2)$, we take the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{3.5.36}$$

as the generator basis of $SU(2)$. The structure constants λ_{ab}^c of (3.5.36) are as follows

$$\lambda_{ab}^c = 2\varepsilon_{abc}, \quad \varepsilon_{abc} = \begin{cases} 1 & \text{if } (abc) \text{ is even}, \\ -1 & \text{if } (abc) \text{ is odd}, \\ 0 & \text{if otherwise}. \end{cases} \tag{3.5.37}$$

Based on (3.5.37), direct calculation shows that

$$g_{ab} = \frac{1}{8}\lambda_{ad}^c\lambda_{cb}^d = \delta_{ab}.$$

Namely $(g_{ab}) = I$ is identity.

For $SU(3)$, we can take the following Gell-Mann matrices as the generator basis of $SU(3)$:

$$\begin{aligned}
&\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
&\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \\
&\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.
\end{aligned} \tag{3.5.38}$$

The structure constants are

$$\lambda_{ab}^c = 2f_{abc}, \quad 1 \leqslant a, b, c \leqslant 8,$$

and f_{abc} are anti-symmetric, given by

$$\begin{aligned} &f_{123} = 1, && f_{147} = f_{246} = f_{257} = f_{345} = \frac{1}{2}, \\ &f_{156} = f_{367} = -\frac{1}{2}, && f_{458} = f_{678} = \sqrt{3}/2, \\ &f_{abc} = 0, && \text{for others.} \end{aligned} \tag{3.5.39}$$

By (3.5.39) we can deduce that

$$\lambda_{ad}^{c}\lambda_{cb}^{d} = \begin{cases} 0, & \text{if } a \neq b, \\ 12, & \text{if } a = b. \end{cases}$$

Hence we get

$$g_{ab} = \frac{1}{12}\lambda_{ad}^{c}\lambda_{cb}^{d} = \delta_{ab}.$$

Namely, under the Gell-Mann representation (3.5.38), $(g_{ab}) = I$.

We are now in position to consider general $SU(N)$. In fact, for all $N \geqslant 2$, there exists a generator basis $\{\tau_a |\ 1 \leqslant a \leqslant N^2 - 1\}$ of $SU(N)$ such that $(g_{ab}) = \frac{1}{4N}(\lambda_{ad}^{c}\lambda_{cb}^{d}) = I$. The generators τ_a are given by the following traceless Hermitian matrices:

$$\begin{aligned}
&\tau_1^{(1)} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tau_2^{(1)} = \begin{pmatrix} \sigma_2 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tau_3^{(1)} = \begin{pmatrix} \sigma_3 & 0 \\ 0 & 0 \end{pmatrix}, \\
&\tau_1^{(2)} = \begin{pmatrix} \lambda_4 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tau_2^{(2)} = \begin{pmatrix} \lambda_5 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tau_2^{(2)} = \begin{pmatrix} \lambda_6 & 0 \\ 0 & 0 \end{pmatrix}, \\
&\tau_4^{(2)} = \begin{pmatrix} \lambda_7 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tau_5^{(2)} = \begin{pmatrix} \lambda_8 & 0 \\ 0 & 0 \end{pmatrix}, \\
&\quad \vdots \\
&\tau_1^{(N-1)} = \begin{pmatrix} 0 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 0 \end{pmatrix}, \quad \tau_2^{(N-1)} = \begin{pmatrix} 0 & \cdots & -i \\ \vdots & & \vdots \\ i & \cdots & 0 \end{pmatrix}, \\
&\tau_3^{(N-1)} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \end{pmatrix}, \quad \tau_4^{(N-1)} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -i \\ \vdots & \vdots & & \vdots \\ 0 & i & \cdots & 0 \end{pmatrix}, \\
&\quad \vdots \\
&\tau_{2N-1}^{(N-1)} = \frac{\sqrt{2}}{\sqrt{N(N-1)}} \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & -(N-1) \end{pmatrix},
\end{aligned} \tag{3.5.40}$$

where σ_k $(1 \leqslant k \leqslant 3)$ are as in (3.5.36) and λ_j $(3 \leqslant j \leqslant 8)$ as in (3.5.38). Corresponding to (3.5.40), g_{ab} are as follows

$$g_{ab} = \frac{1}{4N}\lambda^c_{ad}\lambda^d_{cb} = \delta_{ab}. \tag{3.5.41}$$

Hence the 2-order $SU(N)$ tensor $\{g_{ab}\}$ is positive definition.

Now, we prove Assertion 2). For each $A \in SU(N)$ we take the matrices

$$\omega_a = A\tau_a \qquad \text{for } 1 \leqslant a \leqslant K, \tag{3.5.42}$$

where τ_a $(1 \leqslant a \leqslant K)$ form a basis of $T_eSU(N)$. It is clear that ω_a satisfy the properties (3.5.31). Hence the matrices $\{\omega_a |\ 1 \leqslant a \leqslant K\}$ constitute a basis of $T_ASU(N)$. On the other hand, we see that

$$\begin{aligned}\mathscr{G}_{ab}(A) =& \frac{1}{2}\text{tr}(\omega_a\omega_b^\dagger)\\ =& \frac{1}{2}\text{tr}(A\tau_a\tau_b^\dagger A^\dagger) \qquad \text{(by (3.5.41))}\\ =& \frac{1}{2}\text{tr}(\tau_a\tau_b^\dagger) \qquad \text{(by (2.3.16))}.\end{aligned} \tag{3.5.43}$$

If we take (3.5.40) as the basis τ_a $(1 \leqslant a \leqslant K)$, then we have

$$\frac{1}{2}\text{tr}(\tau_a\tau_b^\dagger) = \delta_{ab}.$$

It follows from (3.5.43) that with the basis (3.5.40) of $T_ASU(N)$,

$$\mathscr{G}_{ab}(A) = \mathscr{G}_{ab}(I) = \delta_{ab} = g_{ab} \quad \text{(by (3.5.42))}.$$

Assertion 2) and the theorem are proved. □

3.5.5 Representation invariance of gauge theory

In this subsection, we consider the representation invariance for the $SU(N)$ gauge theory. In Subsection 3.5.1 we see that the classical Yang-Mills action (3.5.2)-(3.5.4) will change under the transformation of generator bases of $SU(N)$. The modified version of the Yang-Mills action obeying the representation invariance is given by

$$\begin{aligned}&\mathscr{L}_{YM} = \mathscr{L}_G + \mathscr{L}_D,\\ &\mathscr{L}_G = -\frac{1}{4}\mathscr{G}_{ab}F^a_{\mu\nu}F^{\mu\nu b},\\ &\mathscr{L}_D = \overline{\Psi}\left[i\gamma^\mu(\partial_\mu + igA^a_\mu\tau_a) - m\right]\Psi,\end{aligned} \tag{3.5.44}$$

where $\mathscr{G}_{ab}$ is as in (3.5.28).

It is clear that the action density (3.5.44) is invariant under the transformation of $T_A SU(N)$:

$$\widetilde{\tau}_a = x_a^b \tau_b. \tag{3.5.45}$$

In fact, the following are three terms in (3.5.44), which involve contractions of $SU(N)$ tensors:

$$\mathscr{G}_{ab} F_{\mu\nu}^a F^{\mu\nu b}, \quad A_\mu^a \tau_a, \quad \lambda_{bc}^a A_\mu^b A_\nu^c.$$

Obviously, these terms are also Lorentz invariant.

Remark 3.33 The purely mathematical logic requires the introduction of the modified Yang-Mills action (3.5.44) and the SU(N) tensors. There is a profound physical significance. This invariance dictates that mixing different gauge potentials from different gauge groups will often lead to the violation this simple principle. As we shall see, the new invariance theory is very important and crucial in the unified field model presented in the next chapter, where this principle is called the Principle of Representation Invariance (PRI).

3.6 Spectral Theory of Differential Operators

3.6.1 Physical background

Based on the Bohr atomic model, an atom consists of a proton and its orbital electron, bounded by electromagnetic energy. Due to the quantum effect, the orbital electron is in proper discrete energy levels:

$$0 < E_1 < \cdots < E_N, \tag{3.6.1}$$

which can be expressed as

$$E_n = E_0 + \lambda_n \qquad (\lambda_n < 0), \tag{3.6.2}$$

where λ_n $(1 \leqslant n \leqslant N)$ are the negative eigenvalues of a symmetric elliptic operator. Here E_0 stands for the intrinsic energy, and λ_n stands for the bound energy of the atom, holding the orbital electrons, due to the electromagnetism. Hence there are only N energy levels E_n for the atom, which are certainly discrete.

To see this, let Z be the atomic number of an atom. Then the potential energy for electrons is given by

$$V(r) = -\frac{Ze^2}{r}.$$

With this potential, the wave function ψ of an orbital electron satisfies the Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t} + \frac{\hbar^2}{2m_0}\nabla^2\psi + \frac{Ze^2}{r}\psi = 0. \tag{3.6.3}$$

Let ψ take the form

$$\psi = e^{-i\lambda t/\hbar}\varphi(x),$$

where λ is the bound energy. Putting ψ into (3.6.3) leads to

$$-\frac{\hbar^2}{2m_0}\nabla^2\varphi - \frac{Ze^2}{r}\varphi = \lambda\varphi.$$

Since the orbital electrons are bound in the interior of the atom, the following condition holds true:

$$\varphi = 0 \qquad \text{for } |x| > r_0,$$

where r_0 is the radius of an atom. Thus, if ignoring the electromagnetic interactions between orbital electrons, then the bound energy of an electron is a negative eigenvalue of the following elliptic boundary problem

$$\begin{aligned} & -\frac{\hbar^2}{2m_0}\nabla^2\varphi - \frac{Ze^2}{r}\varphi = \lambda\varphi && \text{for } x \in B_{r_0}, \\ & \varphi = 0 && \text{for } x \in \partial B_{r_0}, \end{aligned} \tag{3.6.4}$$

where B_{r_0} is a ball with the atom radius r_0.

According to the spectral theory for elliptic operators, the number of negative eigenvalues of (3.6.4) is finite. Hence, it is natural that the energy levels in (3.6.1) and (3.6.2) are finite and discrete.

3.6.2 Classical spectral theory

Consider the eigenvalue problem of linear elliptic operators as follows

$$\begin{aligned} & -D^2\psi + A\psi = \lambda\psi && \text{for } x \in \Omega, \\ & \psi = 0 && \text{for } x \in \partial\Omega, \end{aligned} \tag{3.6.5}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded domain, $\psi = (\psi_1, \cdots, \psi_m)^{\mathrm{T}} : \ \Omega \to \mathbb{C}^m$ is a complex-valued function with m components,

$$D = \nabla + i\vec{B}, \quad \vec{B} = (B_1, \cdots, B_n), \tag{3.6.6}$$

and A, B_k $(1 \leqslant k \leqslant n)$ are m-th order Hermitian matrices:

$$A = (A_{ij}(x)), \quad B_k = (B^k_{ij}(x)). \tag{3.6.7}$$

Let λ_0 be an eigenvalue of (3.6.5). The corresponding eigenspace at λ_0 is

$$E_{\lambda_0} = \{\psi \in L^2(\Omega, \mathbb{C}^m) |\ \psi \text{ satisfy (3.6.5) with } \lambda = \lambda_0\}$$

is finite dimensional, and its dimension

$$N = \dim E_{\lambda_0}$$

is called the multiplicity of λ_0. Physically, N is also called the degeneracy provided $N > 1$. Usually, we count the multiplicity N of λ_0 as N eigenvalues, i.e., we denote

$$\lambda_1 = \cdots = \lambda_N = \lambda_0.$$

Based on the physical background, we mainly concern the negative eigenvalues. However, for our purpose the following classical spectral theorem is very important.

Theorem 3.34 (Spectral Theorem of Elliptic Operators) *Let the matrices in (3.6.7) are Hermitian, and the functions $A_{ij}, B^k_{ij} \in L^\infty(\Omega)$. The the following assertions hold true*:

1) *All eigenvalues of (3.6.5) are real with finite multiplicities, and form an infinite consequence as follows*:

$$-\infty < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_k \leqslant \cdots, \quad \lambda_k \to \infty \text{ as } k \to \infty.$$

where λ_k is counting the multiplicity.

2) *The eigenfunctions ψ_k corresponding to λ_k are orthogonal to each other, i.e.*

$$\int_\Omega \psi_k^\dagger \psi_j dx = 0, \qquad \forall k \neq j.$$

In particular, $\{\psi_k\}$ is an orthogonal basis of $L^2(\Omega, \mathbb{C}^m)$.

3) *There are only finite number of negative eigenvalues in $\{\lambda_k\}$,*

$$-\infty < \lambda_1 \leqslant \cdots \leqslant \lambda_N < 0, \tag{3.6.8}$$

and the number N of negative eigenvalues depends on the matrices A, B_j in (3.6.7) and the domain Ω.

Remark 3.35 For the energy levels of subatomic particles introduced in Chapter 5, we are mainly interested in the negative eigenvalues of (3.6.5) and in the estimates of the number N in (3.6.8). □

Theorem 3.34 is a corollary of the classical Lagrange multiplier theorem. We recall the variational principle with constraint. Let H be a linear normed space, and F and G are two functionals on H:

$$F, G: \ H \to \mathbb{R}.$$

Let $\Gamma \subset H$ be the set

$$\Gamma = \{u \in H | \ G(u) = 1\}.$$

If $u_0 \in \Gamma$ is a minimum point of F with constraint on Γ:

$$F(u_0) = \min_{u \in \Gamma} F(u),$$

then u_0 satisfies the equation

$$\delta F(u_0) = \lambda \delta G(u_0), \tag{3.6.9}$$

where λ is a real number.

For the eigenvalue equation of (3.6.5), the corresponding functional is

$$F(\psi) = \int_\Omega \left[|D\psi|^2 + \psi^\dagger A\psi\right] dx, \tag{3.6.10}$$

and the constraint functional G is given by

$$G(\psi) = \int_\Omega |\psi|^2 dx. \tag{3.6.11}$$

It is easy to see that the equation of (3.6.5) is of the form:

$$\delta F(\psi) = \lambda \delta G(\psi),$$

and F, G are as in (3.6.10) and (3.6.11), which is as the variational equation (3.6.9) with the constraint on Γ.

Hence, the eigenvalues λ_k $(k = 1, 2, \cdots)$ of (3.6.5) can be expressed in the following forms

$$\begin{aligned} \lambda_1 &= \min_{\psi\in\Gamma} F(\psi), \\ \lambda_k &= \min_{\psi\in\Gamma, \psi\in H_{k-1}^\perp} F(\psi), \end{aligned} \tag{3.6.12}$$

where F is as in (3.6.10), Γ and $H_{k-1}^\perp$ are the sets:

$$\begin{aligned} &\Gamma = \{\psi \in H_0^1(\Omega, \mathbb{C}^m) |\ ||\psi||_{L^2} = 1\}, \\ &H_{k-1}^\perp = \left\{\psi \in H_0^1(\Omega, \mathbb{C}^m) \ \Big|\ \int_\Omega \psi^\dagger \psi_j dx = 0,\ 1 \leqslant j \leqslant k-1\right\}, \end{aligned} \tag{3.6.13}$$

and ψ_j $(1 \leqslant j \leqslant k-1)$ are the eigenfunctions corresponding to the first $(k-1)$ eigenvalues $\lambda_1, \cdots, \lambda_{k-1}$. Namely $H_{k-1}^\perp$ is the orthogonal complement of $H_{k-1} = \text{span}\{\psi_1, \cdots, \psi_{k-1}\}$ in $H_0^1(\Omega, \mathbb{C}^m)$.

Based on (3.6.12)-(3.6.13), we readily deduce the spectral theorem, Theorem 3.34, for the elliptic eigenvalue problem (3.6.5).

3.6.3 Negative eigenvalues of elliptic operators

The following theorem provides a necessary and sufficient condition for the existence of negative eigenvalues of (3.6.5), and a criterion to estimate the number of negative eigenvalues.

Theorem 3.36 *For the eigenvalue problem (3.6.5), the following assertions hold true:*

1) *Equations (3.6.5) have negative eigenvalues if and only if there is a function* $\psi \in H_0^1(\Omega, \mathbb{C}^m)$*, such that*

$$\int_\Omega [(D\psi)^\dagger (D\psi) + \psi^\dagger A\psi] dx < 0, \tag{3.6.14}$$

where D is as in (3.6.6).

2) *If there are K linear independent functions* $\psi_1, \cdots, \psi_K \in H_0^1(\Omega, \mathbb{C}^m)$*, such that*

$$\psi \text{ satisfies (3.6.14)} \quad \text{for any } \psi \in E^K = \text{span}\{\psi_1, \cdots, \psi_K\}, \tag{3.6.15}$$

then the number N of negative eigenvalues is larger than K, i.e., $N \geqslant K$*.*

Proof Assertion 1) follows directly from the following classical formula for the first eigenvalue λ_1 of (3.6.5):

$$\lambda_1 = \min_{\psi \in H_0^1(\Omega, \mathbb{C}^m)} \frac{1}{||\psi||_{L^2}} \int_\Omega \left[(D\psi)^\dagger (D\psi) + \psi^\dagger A\psi\right] dx.$$

We now prove Assertion 2) by contradiction. Assume that it is not true, then $K > N$. By Theorem 3.34, the K functions ψ_j in (3.6.15) can be expended as

$$\psi_j = \sum_{i=1}^N \alpha_{ji} e_i + \sum_{l=1}^\infty \beta_{jl} \varphi_l \qquad \text{for } 1 \leqslant j \leqslant K, \tag{3.6.16}$$

where e_i $(1 \leqslant i \leqslant N)$ and φ_l are eigenfunctions corresponding to negative and nonnegative eigenvalues. Since $K > N$, there exists a K-th order matrix P such that

$$P\alpha = \begin{pmatrix} 0 & \cdots & 0 \\ & * & \end{pmatrix}, \tag{3.6.17}$$

where

$$\alpha = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1N} \\ \vdots & & \vdots \\ \alpha_{K1} & \cdots & \alpha_{KN} \end{pmatrix} \qquad \text{with } \alpha_{ij} \text{ as in (3.6.16)}.$$

Thus, under the transformation P,

$$\widetilde{\psi} = P \begin{pmatrix} \psi \\ 0 \end{pmatrix} \in E^K, \qquad \psi = (\psi_1, \cdots, \psi_N)^{\mathrm{T}}, \tag{3.6.18}$$

where E^K is as in (3.6.15).

However, by (3.6.14) and (3.6.15), the first term $\widetilde{\psi}_1$ in (3.6.18) can be expressed in the form

$$\widetilde{\psi}_1 = \sum_{l=1}^\infty \theta_l \varphi_l \in E^K. \tag{3.6.19}$$

Note that φ_l are the eigenfunctions corresponding to the nonnegative eigenvalues of (3.6.5). Hence we have

$$\int_\Omega [(D\widetilde{\psi}_1)^\dagger (D\widetilde{\psi}_1) + \widetilde{\psi}_1^\dagger A\widetilde{\psi}_1]dx = \int_\Omega \widetilde{\psi}_1^\dagger(-D^2\widetilde{\psi}_1 + A\widetilde{\psi}_1)dx = \sum_{l=1}^\infty |\theta_l|^2\lambda_l > 0. \qquad (3.6.20)$$

Here $\lambda_l \geqslant 0$ are the nonnegative eigenvalues of (3.6.5). Hence we derive, from (3.6.19) and (3.6.20), a contradiction with the assumption in Assertion (2). The proof of the theorem is complete. □

3.6.4 Estimates for number of negative eigenvalues

For simplicity, it is physically sufficient for us to consider the eigenvalue problem of the Laplace operators, given by

$$\begin{aligned} & -\nabla^2\psi + V(x)\psi = \lambda\psi && \text{for } x \in B_r, \\ & \psi = 0 && \text{for } x \in \partial B_r, \end{aligned} \qquad (3.6.21)$$

where $B_r \subset \mathbb{R}^n$ is a ball with radius r.

In physics, V represents a potential function and takes negative value in a bound state, ensuring by Theorem 3.36 that (3.6.21) possesses negative eigenvalues.

Here, for the potential function $V(x)$, we assume that

$$V(\rho x) \simeq \rho^\alpha V_0(x) \qquad (\alpha > -2), \qquad (3.6.22)$$

where $V_0(x)$ is defined in the unit ball B_1, and

$$\Omega = \{x \in B_1 \mid V_0(x) < 0\} \neq \varnothing. \qquad (3.6.23)$$

Let $\theta > 0$ be defined by

$$\theta = \inf_{\psi \in L^2(\Omega,\mathbb{C}^m)} \frac{1}{||\psi||_{L^2}} \int_\Omega |V(x)|\, |\psi|^2 dx. \qquad (3.6.24)$$

The main result in this section is the following theorem, which provides a relation between N, θ and r, where N is the number of negative eigenvalues of (3.6.21). Let λ_1 be the first eigenvalue of the equation

$$\begin{aligned} & -\Delta e = \lambda e && \text{for } x \in \Omega, \\ & e = 0 && \text{for } x \in \partial\Omega, \end{aligned} \qquad (3.6.25)$$

where $\Omega \subset B_1$ is as defined by (3.6.23).

To state the main theorem, we need to introduce a lemma, leading to the Weyl asymptotic relation (Weyl, 1912).

Lemma 3.37(H. Weyl) *Let λ_N be the N-th eigenvalue of the m-th order elliptic operator*

$$\begin{aligned} &(-1)^m \Delta^m e = \lambda e && \text{for } x \in \Omega \subset \mathbb{R}^n, \\ &D^k e|_{\partial\Omega} = 0 && \text{for } 0 \leqslant k \leqslant m-1, \end{aligned} \tag{3.6.26}$$

then λ_N has the asymptotical relation

$$\lambda_N \sim \lambda_1 N^{2m/n}, \tag{3.6.27}$$

where λ_1 is the first eigenvalue of (3.6.26).

We are now ready for the main theorem.

Theorem 3.38 *Under the assumptions of (3.6.22) and (3.6.23), the number N of the negative eigenvalues of (3.6.21) satisfies the following approximative relation*

$$N \simeq \left(\frac{\theta r^{2+\alpha}}{\lambda_1} \right)^{n/2}, \tag{3.6.28}$$

provided that $\theta r^{2+\alpha}/\lambda_1 \gg 1$ is sufficiently large, where r and θ are as in (3.6.21) and (3.6.24), and λ_1 is the first eigenvalue of (3.6.25).

Proof The ball B_r can be written as

$$B_r = \{y = rx \mid x \in B_1\}.$$

Note that $\partial/\partial y = r^{-1}\partial/\partial x$, (3.6.21) can be equivalently expressed as

$$\begin{aligned} &-\Delta\varphi + r^2 V(rx)\varphi = \beta\varphi && \text{for } x \in B_1, \\ &\varphi = 0 && \text{for } x \in \partial B_1, \end{aligned} \tag{3.6.29}$$

and the eigenvalue λ of (3.6.21) is

$$\lambda = \frac{1}{r^2}\beta, \qquad \text{where } \beta \text{ is the eigenvalue of (3.6.29).}$$

Hence the number of negative eigenvalues of (3.6.21) is the same as that of (3.6.29), and we only need to prove (3.6.28) for (3.6.29).

By (3.6.22), the equation (3.6.29) is approximatively in the form

$$\begin{aligned} &-\Delta\varphi + r^{2+\alpha} V_0(x)\varphi = \beta\varphi && \text{for } x \in B_1, \\ &\varphi = 0 && \text{for } x \in \partial B_1. \end{aligned} \tag{3.6.30}$$

Based on Assertion (2) in Theorem 3.36, we need to find N linear independent functions $\varphi_n \in H_0^1(B_1)$ $(1 \leqslant n \leqslant N)$ satisfying

$$\int_{B_1} \left[|\nabla\varphi|^2 + r^{2+\alpha} V_0(x)\varphi^2 \right] dx < 0, \tag{3.6.31}$$

for any $\varphi \in \text{span}\,\{\varphi_1, \cdots, \varphi_N\}$ with $||\varphi||_{L^2} = 1$.

To this end, we take the eigenvalues $\{\lambda_n\}$ and eigenfunctions $\{e_n\}$ of (3.6.25) such that

$$0 < \lambda_1 \leqslant \cdots \leqslant \lambda_N < \lambda_{N+1},$$

and

$$\lambda_N < \theta r^{2+\alpha} \leqslant \lambda_{N+1}. \tag{3.6.32}$$

For the eigenfunctions e_n, we make the extension

$$\varphi_n = \begin{cases} e_n & \text{for } x \in \Omega, \\ 0 & \text{for } x \in B_1/\Omega. \end{cases}$$

It is known that φ_n is weakly differentiable, and $\varphi_n \in H_0^1(\Omega)$. These functions φ_n $(1 \leqslant n \leqslant N)$ are what we need. Let

$$\varphi = \sum_{n=1}^{N} \alpha_n \varphi_n, \quad ||\varphi||_{L^2} = 1.$$

By Assertion (2) in Theorem 3.36, φ_n $(1 \leqslant n \leqslant N)$ are orthonormal:

$$\int_{B_1} \varphi_i \varphi_j dx = \int_{\Omega} e_i e_j dx = \delta_{ij}.$$

Therefore we have

$$||\varphi||_{L^2} = \sum_{n=1}^{N} \alpha_n^2 = 1. \tag{3.6.33}$$

Thus the integral in (3.6.31) is

$$\begin{aligned}
&\int_{B_1} [|\nabla \varphi|^2 + r^{2+\alpha} V_0(x) \varphi^2] dx \\
=& \int_{\Omega} -\left(\sum_{n=1}^{N} \alpha_n e_n\right)\left(\sum_{n=1}^{N} \alpha_n \Delta e_n\right) dx + r^{2+\alpha} \int_{\Omega} V_0(x) \varphi^2 dx \\
=& \sum_{n=1}^{N} \alpha_n^2 \lambda_n + r^{2+\alpha} \int_{\Omega} V_0(x) \varphi^2 dx \\
\leqslant& \sum_{n=1}^{N} \alpha_n^2 \lambda_n - \theta r^{2+\alpha} \quad \text{(by (3.6.24))} \\
<& 0 \quad \text{(by (3.6.32) and (3.6.33))}.
\end{aligned}$$

It follows from Theorem 3.36 that there are at least N negative eigenvalues for (3.6.30). When $\theta r^{2+\alpha} \gg 1$ is sufficiently large, the relation (3.6.32) implies that

$$\lambda_N \simeq \theta r^{2+\alpha}. \tag{3.6.34}$$

On the other hand, by (3.6.27) in Lemma 3.37,

$$\lambda_N \sim \lambda_1 N^{2/n} \qquad (m=1). \tag{3.6.35}$$

Hence the relation (3.6.28) follows from (3.6.34) and (3.6.35). The proof is complete. □

Remark 3.39 In Section 6.4.6, we shall see that for particles with mass m, the parameters in (3.6.28) are

$$\alpha = 0, \quad r = 1, \quad n = 3, \quad \theta = 4m\rho_1^2 A g^2/\hbar^2 \rho,$$

where $g = g_w$ or g_s is the weak or strong interaction charge, ρ is the particle radius, ρ_1 is the weak or strong attracting radius, and A is the weak or strong interaction constant. Hence the number of energy levels of massive particles is given by

$$N = \left[\frac{4}{\lambda_1}\frac{\rho_1^2 A}{\rho}\frac{mc}{\hbar}\frac{g^2}{\hbar c}\right]^{3/2}.$$

where λ_1 is the first eigenvalue of $-\Delta$ in the unit ball B_1. □

Example 3.40 (Number of Atomic Energy Levels) As an application of Theorem 3.38, we consider here the estimates for the number of atomic energy levels. Let the atom number be Z. If ignoring interactions between orbital electrons, then the spectral equation is as follows

$$\begin{aligned} -\Delta\psi + V(x)\psi &= \lambda\psi && \text{for } x \in B_{r_0}, \\ \psi &= 0 && \text{for } x \in \partial B_{r_0}, \end{aligned}$$

and r_0 is the atom radius, V is the potential energy, given by

$$V(x) = -\frac{2Zme^2}{\hbar^2}\frac{1}{r}, \quad m \text{ the mass of electron.}$$

The parameters in (3.6.28) for this system are

$$\alpha = -1, \quad n = 3, \quad r = r_0 = 10^{-8}\text{cm}. \tag{3.6.36}$$

In addition, V_0 and Ω in (3.6.23) and (3.6.24) are as

$$V_0 = -\frac{2Zme^2}{\hbar^2}, \quad \Omega = B_1.$$

Therefore, the parameter θ is given by

$$\theta = \frac{2Zme^2}{\hbar^2}. \tag{3.6.37}$$

According to physical parameters, it is known that

$$r_0 \times \frac{mc}{\hbar} = \frac{1}{4}\times 10^3, \quad \frac{e^2}{\hbar c} = \frac{1}{137}. \tag{3.6.38}$$

Hence, by (3.6.36)-(3.6.38) the formulas (3.6.28) becomes

$$N=\left(\frac{2Z}{\lambda_1}\frac{e^2}{\hbar c}\frac{mcr_0}{\hbar}\right)^{3/2}=\left(\frac{10^3Z}{274\lambda_1}\right)^{3/2}, \tag{3.6.39}$$

where λ_1 is the first eigenvalue of $-\delta$ on B_1.

The formulas is derived in the ideal situation ignoring interactions between orbital electrons, and only holds for a bigger atom number Z.

3.6.5 Spectrum of Weyl operators

In Section 6.4.2, we shall deduce from Basic Postulates of Quantum Mechanics that the spectral equations for the massless subatomic particles are in the following form

$$\begin{aligned} &-\hbar c(\vec{\sigma}\cdot\vec{D})^2\varphi+\frac{\mathrm{i}g}{2}\{(\vec{\sigma}\cdot\vec{D}),A_0\}\varphi=\mathrm{i}\lambda(\vec{\sigma}\cdot\vec{D})\varphi,\\ &\varphi|_{\partial\Omega}=0,\end{aligned} \tag{3.6.40}$$

where $\varphi=(\varphi_1,\varphi_2)^{\mathrm{T}}:\ \Omega\to\mathbb{C}^2$ is a complex-valued function with two components, called the Weyl spinor, $\vec{\sigma}=(\sigma_1,\sigma_2,\sigma_2)$ is the Pauli matrix operator as given by (3.5.36), $\vec{D}=(D_1,D_2,D_3)$ is the derivative operator given by

$$D_k=\partial_k+igA_k, \qquad \text{for } k=1,2,3, \tag{3.6.41}$$

and $\{(\vec{\sigma}\cdot\vec{D}),A_0\}$ is the anti-commutator defined by

$$\{(\vec{\sigma}\cdot\vec{D}),A_0\}=(\vec{\sigma}\cdot\vec{D})A_0+A_0(\vec{\sigma}\cdot\vec{D}).$$

Remark 3.41 The equation (3.6.40) is essentially an eigenvalue problem of the first order differential operator:

$$i\hbar c(\vec{\sigma}\cdot\vec{D})+gA_0,$$

which is called the Weyl operator. In addition, the operator $(\vec{\sigma}\cdot\vec{D})^2$ in (3.6.40) is elliptic and can be written as

$$(\vec{\sigma}\cdot\vec{D})^2=D^2-\frac{g}{\hbar c}\vec{\sigma}\cdot\mathrm{curl}\vec{A}, \tag{3.6.42}$$

and $\vec{A}=(A_1,A_2,A_3)$ as in (3.6.41). The ellipticity of (3.6.42) g in (3.6.41)-(3.6.42) represents the weak or strong interaction charge, and $A_\mu=(A_0,A_1,A_2,A_3)$ the weak or strong interaction potential.

Note also that $\vec{A}=(A_1,A_2,A_3)$ stands for the magnetic component of the weak or strong interaction. Hence, in (3.6.42), the term

$$g\vec{\sigma}\cdot\mathrm{curl}\vec{A}$$

represents magnetic energy generated by the weak or strong interactions, which is an important byproduct of the unified field theory based on PID and PRI. □

Since (3.6.40) is essentially an eigenvalue problem of first-order differential equations, its negative and positive eigenvalues are infinite. However, if we only consider the physically meaningful eigenstates, then the number of negative eigenvalues of (3.6.40) is finite. We now give introduce these physical meaningful eigenvalues and eigenfunctions for (3.6.40).

Definition 3.42 *A real number λ and a two-component wave function $\varphi \in H_0^1(\Omega, \mathbb{C}^2)$ are called the eigenvalue and eigenfunction of (3.6.40), if (λ, φ) satisfies (3.6.40) and*

$$\int_\Omega \varphi^\dagger \left[i(\vec{\sigma} \cdot \vec{D})\varphi \right] dx > 0. \tag{3.6.43}$$

The physical significance of (3.6.43) is that the kinetic energy E of the eigenstate φ is positive: $E > 0$.

The following theorem ensures the mathematical rationality of the eigenvalue problem of the Weyl operators.

Theorem 3.43 (Spectral Theorem of Weyl Operators) *For the eigenvalue problem (3.6.40), the following assertions hold true*:

1) *The eigenvalues of (3.6.40) are real and discrete, with finite multiplicities, and satisfy*

$$-\infty < \lambda_1 \leqslant \cdots \leqslant \lambda_k \leqslant \cdots, \quad \lambda_k \to \infty \text{ as } k \to \infty.$$

2) *The eigenfunctions are orthogonal in the sense that*

$$\int_\Omega \psi_k^\dagger \left[i(\vec{\sigma} \cdot \vec{D})\psi_j \right] dx = 0, \qquad \forall k \neq j. \tag{3.6.44}$$

3) *The number of negative eigenvalues is finite*

$$-\infty < \lambda_1 \leqslant \cdots \leqslant \lambda_N < 0.$$

4) *Equations (3.6.40) have negative eigenvalues if and only if there exists a function $\varphi \in H_0^1(\Omega, \mathbb{C}^2)$ satisfying (3.6.43) such that*

$$\int_\Omega \left[\hbar c |(\vec{\sigma} \cdot \vec{D})\varphi|^2 + \frac{ig}{2} \varphi^\dagger \{(\vec{\sigma} \cdot \vec{D}), A_0\}\varphi \right] < 0.$$

Proof It is easy to see that the operator

$$L = i(\vec{\sigma} \cdot \vec{D}) : \ H_0^1(\Omega, \mathbb{C}^2) \to L^2(\Omega, \mathbb{C}^2)$$

is a Hermitian operator. Consider a functional $F : \ H^1(\Omega, \mathbb{C}) \to \mathbb{R}$:

$$F(\psi) = \int_\Omega \left[\hbar c |L\psi|^2 + \frac{g}{2} \psi^\dagger \{L, A_0\}\psi \right] dx.$$

By (3.6.42), the operator $L^2 = -(\vec{\sigma}\cdot\vec{D})^2$ is elliptic. Hence F has the following lower bound on S:

$$S = \left\{\psi \in H_0^1(\Omega, \mathbb{C}^2) \,\middle|\, \int_\Omega \psi^\dagger L\psi dx = 1\right\},$$

namely

$$\min_{\psi\in S} F(\psi) > -\infty.$$

Based on the Lagrange multiplier theorem of constraint minimization, the first eigenvalue λ_1 and the first eigenfunction $\varphi_1 \in S$ satisfy

$$\lambda_1 = F(\psi_1) = \min_{\psi\in S} F(\psi). \tag{3.6.45}$$

In addition, if

$$\lambda_1 \leqslant \cdots \leqslant \lambda_m$$

are the first m eigenvalues with eigenfunctions $\psi_k, 1 \leqslant k \leqslant m$, then the $(m+1)$-th eigenvalue λ_{m+1} and eigenfunction ψ_{m+1} satisfy

$$\lambda_{m+1} = F(\psi_{m+1}) = \min_{\psi\in S, \psi\in H_m^\perp} F(\psi), \tag{3.6.46}$$

where $H_m = \text{span}\{\psi_1, \cdots, \psi_m\}$, and $H_m^\perp$ is the orthogonal complement of H_m in the sense of (3.6.44).

It is clear that Assertions (1)-(4) of the theorem follow from (3.4.45) and (3.4.26). The proof is complete. □

In the following, we consider the estimates of the number of negative eigenvalues for the Weyl operators. If the interaction potential A_μ takes approximatively the following

$$A_\mu = (K, 0, 0, 0) \qquad \text{with } K > 0 \text{ being a constant.}$$

Then (3.6.40) becomes

$$\begin{aligned} &-\Delta\varphi = i(\lambda + K)(\vec{\sigma}\cdot\vec{\partial})\varphi && \text{for } x \in \Omega \subset \mathbb{R}^n, \\ &\varphi = 0 && \text{for } x \in \partial\Omega. \end{aligned} \tag{3.6.47}$$

Obviously, the number N of negative eigenvalues of (3.6.47) satisfies

$$\beta_N < K < \beta_{N+1}, \tag{3.6.48}$$

where β_k is the k-th eigenvalue of the equation

$$\begin{aligned} &-\Delta\varphi_k = i\beta_k(\vec{\sigma}\cdot\vec{\partial})\varphi_k, \\ &\varphi_k|_{\partial\Omega} = 0. \end{aligned} \tag{3.6.49}$$

For (3.6.49) we have the Weyl asymptotical relation

$$\beta_N \sim \beta_1 N^{1/n}. \tag{3.6.50}$$

Here the exponent is $1/n$, since (3.6.47) is an $2m$-th order elliptic equation with $m = 1/2$. Hence we deduce, from (3.6.48) and (3.6.50), the estimates of the number N of negative eigenvalues of (3.6.47) as

$$N \simeq \left(\frac{K}{\beta_1}\right)^n, \tag{3.6.51}$$

where β_1 is the first eigenvalue of (3.6.49).

Remark 3.44 For the mediators such as the photon and gluons, the number of energy levels is given by N in (3.6.51), which can be estimated as

$$N = \left(\frac{A}{\beta_1}\frac{\rho_1}{\rho}\frac{g_w^2}{\hbar c}\right)^3,$$

where ρ_1, ρ, A, g_w are as in Remark 3.39.

Chapter 4

Unified Field Theory of Four Fundamental Interactions

Once again, the goal of this book is to derive experimentally verifiable laws of Nature based on a few fundamental mathematical principles. The aims of this chapter are are as follows:

- to address the basic principles for the unified field theory coupling the four fundamental interactions,
- to derive the unified model based on these principles, and
- to study the mechanism and nature of individual interactions.

This chapter is based entirely on the recent work of the authors (Ma and Wang, 2015a, 2014h,c, 2013a, 2014e). The key ingredients of the unified field theory include the following.

First, we have established two new principles, the principle of interaction dynamics (PID) and the principle of representation invariance (PRI). PID was first discovered by (Ma and Wang, 2014e, 2015a). It requires that for the four fundamental interactions, the variation be taken under the energy-momentum conservation constraints. The validity of PID for the four fundamental interactions of Nature has been demonstrated through strong experimental and observation supports. For gravity, PID is induced by the presence of the dark matter and dark energy phenomena. PID is also required by the Higgs field and the quark confinement, as we explained in Chapter 1.

PRI, originally discovered by (Ma and Wang, 2014h), states that the $SU(N)$ gauge theory should be invariant, under the representation transformations of the generator bases. PRI is simply a logic requirement for the $SU(N)$ theory.

Second, with PRI, the unification through a large symmetry group appears to be not feasible. Then we have demonstrated that the two first principles, PID and PRI, together with the principle of symmetry-breaking, offer an entirely different route for the unification:

1) *the general relativity and the gauge symmetries dictate the Lagrangian*; *and*

2) *the coupling of the four interactions is achieved through PID and PRI in the unified field equations, which obey the PGR and PRI, but break spontaneously the gauge symmetry.*

In this chapter, there are four sections. Section 4.1 presents the general view of the unified field theory, the geometry of unified fields, gauge-symmetry breaking, PID, and PRI. Section 4.2 presents the experimental and observational physical supports for PID. The mathematical reason from the well-posedness point of view is also given. Section 4.3 introduces the unified field equations coupling the four fundamental interactions based on PID. Section 4.4 presents the natural duality between the gauge fields and their dual fields, as well as the decoupling of the unified field equations to the field equations for individual interaction when the other interactions are negligible.

4.1 Principles of Unified Field Theory

4.1.1 Four interactions and their interaction mechanism

The four fundamental interactions/forces of Nature include

1) the gravitational force, generated by the mass charge M, which is responsible to all macroscopic motions;
2) the electromagnetic force, generated by the electric charge e, which holds the atoms and molecules together;
3) the strong force, generated by the strong charge g_s, which mainly acts at three levels: quarks and gluons, hadrons, and nucleons;
4) the weak force, generated by the weak charge g_w, which provides the binding energy to hold the mediators, the leptons and quarks together.

The most crucial ingredient of each interaction is its corresponding interaction potential Φ and charge g. The relation between the corresponding force F, its associated potential Φ and its charge g is as follows:

$$F = -g\nabla\Phi, \tag{4.1.1}$$

where ∇ is the three-dimensional spatial gradient operator. The charge g for each interaction is given as follows:

$$\begin{aligned}
&m && \text{the mass charge for gravity,}\\
&e && \text{the electric charge for electromagnetism,}\\
&g_s && \text{the strong charge for strong interaction, and}\\
&g_w && \text{the weak charge for the weak interaction.}
\end{aligned} \tag{4.1.2}$$

We know now that the four interactions are dictated respectively by the following symmetry principles:

$$\begin{aligned} &\text{gravity:} && \text{principle of general relativity,} \\ &\text{electromagnetism:} && U(1) \text{ gauge invariance,} \\ &\text{weak interaction:} && SU(2) \text{ gauge invariance,} \\ &\text{strong interaction:} && SU(3) \text{ gauge invariance.} \end{aligned} \tag{4.1.3}$$

The last three interactions also obey the Lorentz invariance. As a natural consequence, the three charges e, g_w, g_s in (4.1.2) are the coupling constants of the $U(1), SU(2), SU(3)$ gauge fields.

Following the simplicity principle of laws of Nature as stated in Principle 2.2, the three basic symmetries—the Einstein general relativity, the Lorentz invariance and the gauge invariance—uniquely determine the interaction fields and their Lagrangian actions for the four interactions:

1. *Gravity.* The gravitational fields are the Riemannian metric defined on the space-time manifold $\mathscr{M}$:

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \tag{4.1.4}$$

and then second-order tensor $\{g_{\mu\nu}\}$ stands for the gravitational potential. The Lagrangian action for the metric (4.1.4) is the Einstein-Hilbert functional

$$\mathscr{L}_{EH} = R + \frac{8\pi G}{c^4} S, \tag{4.1.5}$$

where R stands for the scalar curvature of the tangent bundle $T\mathscr{M}$ of $\mathscr{M}$.

2. *Electromagnetism.* The field describing electromagnetic interaction is the $U(1)$ gauge field

$$A_\mu = (A_0, A_1, A_2, A_3),$$

representing the electromagnetic potential, and the Lagrangian action is

$$\mathscr{L}_{EM} = -\frac{1}{4} A_{\mu\nu} A^{\mu\nu}, \tag{4.1.6}$$

which stands for the scalar curvature of the vector bundle $\mathscr{M} \otimes_p \mathbb{C}^4$, with

$$A_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu.$$

being the curvature tensor.

3. *Weak interaction.* The weak fields are the $SU(2)$ gauge fields

$$W_\mu^a = (W_0^a, W_1^a, W_2^a, W_3^a) \qquad \text{for } 1 \leqslant a \leqslant 3,$$

and their action is

$$\mathscr{L}_W = -\frac{1}{4}\mathscr{G}^w_{ab} W^a_{\mu\nu} W^{\mu\nu b}, \tag{4.1.7}$$

which also stands for the scalar curvature of the spinor bundle: $\mathscr{M} \otimes_p (\mathbb{C}^4)^2$. Here

$$W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g_w \lambda^a_{bc} W^b_\mu W^c_\nu \qquad \text{for } 1 \leqslant a \leqslant 3.$$

4. *Strong interaction.* The strong fields are the $SU(3)$ gauge fields

$$S^k_\mu = (S^k_0, S^k_1, S^k_2, S^k_3) \qquad \text{for } 1 \leqslant k \leqslant 8,$$

and the action is

$$\mathscr{L}_S = -\frac{1}{4}\mathscr{G}^s_{kl} S^k_{\mu\nu} S^{\mu\nu l}, \tag{4.1.8}$$

which corresponds to the scalar curvature of $\mathscr{M} \otimes_p (\mathbb{C}^4)^3$. Here

$$S^k_{\mu\nu} = \partial_\mu S^k_\nu - \partial_\nu S^k_\mu + g_s \Lambda^k_{rl} S^r_\mu S^l_\nu \qquad \text{for } 1 \leqslant k \leqslant 8.$$

The Yukawa Interaction Mechanism, briefly mentioned in Section 2.1.6 and restated below, is the main reason why the weak interaction is described by an $SU(2)$ gauge theory and the strong interaction is described by an $SU(3)$ gauge theory.

One great vision of Albert Einstein is his principle of equivalence, which says that gravity is manifested as the curved effect of the space-time manifold $\{\mathscr{M}, g_{\mu\nu}\}$. Based on the recent work by the authors (Ma and Wang, 2015a, 2014h,d), the Einstein vision leads us to postulate the Geometric Interaction Mechanism 2.13, which is restated here for convenience:

Geometric Interaction Mechanism 4.1 *The gravitational force is the curved effect of the time-space, and the electromagnetic, weak, strong interactions are the twisted effects of the underlying complex vector bundles $\mathscr{M} \otimes_p \mathbb{C}^n$.*

Yukawa's viewpoint, entirely different from Einstein's, is that the other three fundamental forces—the electromagnetism, the weak and the strong interactions——take place through exchanging intermediate bosons:

Yukawa Interaction Mechanism 4.2 *The four fundamental interactions of Nature are mediated by exchanging interaction field particles, called the mediators. The gravitational force is mediated by the graviton, the electromagnetic force is mediated by the photon, the strong interaction is mediated by the gluons, and the weak interaction is mediated by the intermediate vector bosons $W^\pm$ and Z.*

It is the Yukawa mechanism that leads to the $SU(2)$ and $SU(3)$ gauge theories respectively for the weak and the strong interactions. In fact, the three mediators $W^{\pm}$ and Z for the weak interaction are regarded as the $SU(2)$ gauge fields W^a_μ $(1 \leqslant a \leqslant 3)$, and the eight gluons for the strong interaction are considered as the $SU(3)$ gauge fields S^k_μ $(1 \leqslant k \leqslant 8)$. Of course, the three color quantum numbers for the quarks are an important evidence for choosing the $SU(3)$ gauge theory to describe the strong interaction.

The two interaction mechanisms lead to two entirely different directions to develop the unified field theory. The need for quantization for all current theories for the four interactions is based on the Yukawa Interaction Mechanism. The new unified field theory in this article is based on the Geometric Interaction Mechanism, which focuses directly on the four interaction forces as in (4.1.1), and does not involve a quantization process.

A radical difference for these two mechanisms is that the Yukawa Mechanism is oriented toward to computing the transition probability for the particle decays and scatterings, and the Geometric Interaction Mechanism is oriented toward to fundamental laws, such as interaction potentials, of the four interactions.

4.1.2 General introduction to unified field theory

Einstein's unification

The aim of a unified field theory is to establish a set of field equations coupling the four fundamental interactions. Albert Einstein was the first person who attempted to establish a unified field theory. The basic philosophy of his unification is that all fundamental forces of Nature should be dictated by one large symmetry group, which can degenerate into a sub-symmetry for each interaction:

$$\boxed{\text{Unification through an action under a large symmetry}} \tag{4.1.9}$$

In essence, with the Einstein unification, under the large symmetry, the four fundamental forces can be regarded as one fundamental force.

Recall that there are four fundamental interactions of Nature: gravitational, electromagnetic, strong, and weak, whose fields and actions:

1) fields $g_{\mu\nu}, A_\mu, \{W^a_\mu \mid 1 \leqslant a \leqslant 3\}$ and $\{S^k_\mu \mid 1 \leqslant k \leqslant 8\}$, and

2) their actions $\mathscr{L}_{EH}, \mathscr{L}_{EM}, \mathscr{L}_W$ and $\mathscr{L}_S$,

are dictated by the symmetries in (4.1.3), as described in (4.1.5)-(4.1.8).

Basically, the Einstein unification is to search for a large symmetry, which dictates an N component field G:

$$G = (G_1, \cdots, G_N), \tag{4.1.10}$$

and an action

$$\mathscr{L} = \mathscr{L}(G_1, \cdots, G_N). \tag{4.1.11}$$

The basic requirements for such a unification are as follows: Under certain conditions,

1) the large symmetry degenerates into sub-symmetries: the general invariance, the Lorentz invariance, and the $U(1), SU(2), SU(3)$ invariance;

2) the field G of (4.1.10) is then decomposed into the fields of the four fundamental interactions:

$$(G_1, \cdots, G_N) \xrightarrow{\text{degenerate}} g_{\mu\nu}, A_\mu, W^a_\mu, S^k_\mu, \tag{4.1.12}$$

3) the action (4.1.11) also becomes the simple sum of these actions given in (4.1.5)-(4.1.8):

$$\mathscr{L}(G_1, \cdots, G_N) \to \mathscr{L}_{EH} + \mathscr{L}_{EM} + \mathscr{L}_W + \mathscr{L}_S. \tag{4.1.13}$$

For almost a century, a great deal of effort has been made to find the unified field model based on the above mentioned approach. However, all efforts in this aspect are not successful. In fact, among other reasons, this route of unification violates the principle of representation invariance (PRI), discovered in (Ma and Wang, 2014h); see also remaining part of this chapter for details.

Moreover, the basic principles and the field equations from all attempted unified field theories based on (4.1.9) are often too complex, and violate the simplicity principle of physics. Most importantly, despite of many attempts, the current theories following (4.1.9) fail to provide solutions to the following longstanding problems and challenges:

1) interaction force formulas
2) quark confinement,
3) asymptotical freedom,
4) strong interaction potentials of nucleus,
5) dark matter and dark energy phenomena,
6) decoupling to the individual interactions as required by (4.1.13),
7) spontaneous symmetry breaking from first principles,
8) mechanism of subatomic particle decay and scattering,
9) violation of PRI, and
10) reasons why leptons do not participate in strong interactions.

Unified Field Theory based on PID, PRI and PSB

The unified field theory introduced in this chapter is based on the three new principles, PID, PRI and PSB, postulated recently by the authors. In this theory, the Lagrangian is a simple sum of the known four actions in in (4.1.5)-(4.1.8):

$$\mathscr{L} = \mathscr{L}_{EH} + \mathscr{L}_{EM} + \mathscr{L}_{W} + \mathscr{L}_{S}, \tag{4.1.14}$$

and with which the unification is achieved as follows:

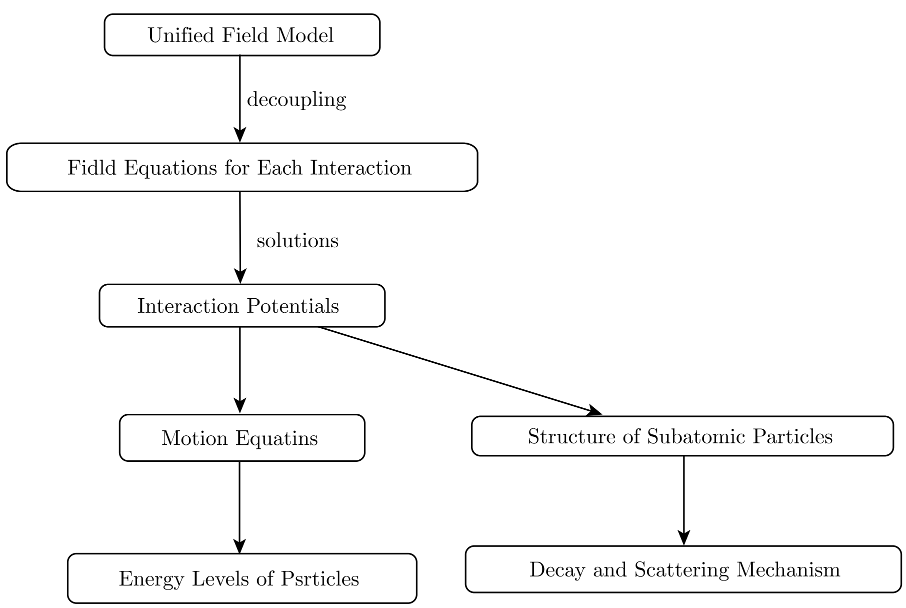
$$\boxed{\text{Unification through the Field Equations Based on PID, PRI and PSB}} \tag{4.1.15}$$

A few remarks are now in order.

First, the unified field model of (4.1.14) and (4.1.15) not only solves the most basic problems 1)-10) mentioned above, but also is the simplest with respect to the underlying physical principles and to the explicit form of the field equations coupling four forces.

Second, the new unified field theory based on PID, PRI and PSB addressed in this chapter offers an answer to dark energy and dark matter problem; see also Section 7.6.3.

Third, thanks to PRI, we have shown that the classical $SU(3)$ Yang-Mills theory will only provide a repulsive force. The attractive bounding force between quarks are due to the dual fields in the PID-induced $SU(3)$ gauge theory. In other words, the quark confinement problem is solved in (Ma and Wang, 2014c), and will be addressed in detail in Section 4.5.3.



Also, the route of unification (4.1.14) and (4.1.15) is readily applied to multi-particle interacting systems, and gives rise to a first dynamic interacting model for multi-particle systems; see Chapter 6 for details.

Finally, we present a diagram to illustrate the framework of the unified field theory, based on (4.1.14) and (4.1.15). We note that quantization is used mainly for deriving transition probability from the field equations for each interaction.

4.1.3 Geometry of unified fields

Hereafter we always assume that the space-time manifold $\mathscr{M}$ of our Universe is a 4-dimensional Riemannian manifold. We adopt the view that symmetry principles determine the geometric structure of $\mathscr{M}$, and the geometries of $\mathscr{M}$ associated with the fundamental interactions of Nature dictate all motion laws defined on $\mathscr{M}$. The process that symmetries determine the geometries of $\mathscr{M}$ and the associated vector bundles is achieved in the following three steps:

1) The symmetric principles, such as the Einstein general relativity, the Lorentz invariance, and the gauge invariance, determine that the fields reflecting geometries of $\mathscr{M}$ are the Riemannian metric $\{g_{\mu\nu}\}$ and the gauge fields $\{G^a_\mu\}$. In addition, the symmetric principles also determine the Lagrangian actions of $g_{\mu\nu}$ and G^a_μ;

2) PID determines the field equations governing $g_{\mu\nu}$ and G^a_μ; and

3) The solutions $g_{\mu\nu}$ and G^a_μ of the field equations determine the geometries of $\mathscr{M}$ and the vector bundles.

The geometry of the unified fields refers to the geometries of $\mathscr{M}$, determined by the following known physical symmetry principles:

$$\begin{aligned}
&\text{principle of general relativity,}\\
&\text{principle of Lorentz invariance,}\\
&U(1)\times SU(2)\times SU(3)\ \text{gauge invariance,}\\
&\text{principle of representation invariance (PRI).}
\end{aligned} \tag{4.1.16}$$

We shall introduce the two principles PID and PRI in Section 4.1.5.

The fields determined by the symmetries in (4.1.16) are given by

- general relativity: $g_{\mu\nu}: \mathscr{M} \to T^0_2\mathscr{M}$, the Riemannian metric,
- Lorentz invariance: $\psi: \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^N$, the Dirac spinor fields,
- $U(1)$ gauge invariance: $A_\mu: \mathscr{M} \to T^*\mathscr{M}$, the $U(1)$ gauge field,
- $SU(2)$ gauge invariance: $W^a_\mu: \mathscr{M} \to (T^*\mathscr{M})^3$, the $SU(2)$ gauge fields,

- $SU(3)$ gauge invariance: $S^k_\mu : \mathscr{M} \to (T^*\mathscr{M})^8$, the $SU(3)$ gauge fields.

The Lagrange action for the geometry of the unified fields is given by

$$L = \int_{\mathscr{M}} [\mathscr{L}_{EH} + \mathscr{L}_{EM} + \mathscr{L}_W + \mathscr{L}_S + \mathscr{L}_D] \sqrt{-g} dx \tag{4.1.17}$$

where $\mathscr{L}_{EH}$, $\mathscr{L}_{EM}$, $\mathscr{L}_W$ and $\mathscr{L}_S$ are the Lagrangian actions for the four interactions defined by (4.1.5)–(4.1.8), and the action for the Dirac spinor fields is given by

$$\mathscr{L}_D = \overline{\Psi}(i\gamma^\mu D_\mu - m)\Psi. \tag{4.1.18}$$

Here

$$\Psi = (\psi^e, \psi^w, \psi^s),$$
$$m = (m_e, m_w, m_s),$$

and

$$\begin{aligned}
&\psi^e : \mathscr{M} \to \mathscr{M} \otimes_p \mathbb{C}^4 && \text{1-component Dirac spinor,}\\
&\psi^w : \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^2 && \text{2-component Dirac spinors,}\\
&\psi^s : \mathscr{M} \to \mathscr{M} \otimes_p (\mathbb{C}^4)^3 && \text{3-component Dirac spinors.}
\end{aligned} \tag{4.1.19}$$

The derivative operators D_μ are given by

$$\begin{aligned}
&D_\mu \psi^e = (\partial_\mu + ieA_\mu)\psi^e,\\
&D_\mu \psi^w = (\partial_\mu + ig_w W^a_\mu \sigma_a)\psi^w,\\
&D_\mu \psi^s = (\partial_\mu + ig_s S^k_\mu \tau_k)\psi^s.
\end{aligned} \tag{4.1.20}$$

The geometry of unified fields consists of 1) the field functions and 2) the Lagrangian action (4.1.17), which are invariant under the following seven transformations:

1) the general linear transformation $\mathscr{Q}_p = (a^\mu_\nu) : T_p\mathscr{M} \to T_p\mathscr{M}$ with $\mathscr{Q}_p^{-1} = (b^\mu_\nu)^T$, for any $p \in \mathscr{M}$:

$$\begin{aligned}
&(\widetilde{g}_{\mu\nu}) = \mathscr{Q}_p (g_{\mu\nu}) \mathscr{Q}_p^{\mathrm{T}},\\
&\widetilde{A}_\mu = a^\nu_\mu A_\nu,\\
&\widetilde{W}^a_\mu = a^\nu_\mu W^a_\nu && \text{for} 1 \leqslant a \leqslant 3,\\
&\widetilde{S}^k_\mu = a^\nu_\mu S^k_\nu && \text{for} 1 \leqslant k \leqslant 8,\\
&\widetilde{\gamma}^\mu = b^\mu_\nu \gamma^\nu, && \widetilde{\partial}_\mu = a^\nu_\mu \partial_\nu,
\end{aligned} \tag{4.1.21}$$

with no change on other fields, where γ^μ are the Dirac matrices;

2) the Lorentz transformation on $T_p\mathscr{M}$:

$$\begin{aligned}
&L=(l_\mu^\nu):\ T_p\mathscr{M}\to T_p\mathscr{M}, && L \text{ is independent of } p\in M,\\
&(\widetilde{g}_{\mu\nu})=L(g_{\mu\nu})L^{\mathrm{T}}, && \widetilde{A}_\mu=l_\mu^\nu A_\nu,\\
&\widetilde{W}_\mu^a=l_\mu^\nu W_\nu^a && \text{for } 1\leqslant a\leqslant 3,\\
&\widetilde{S}_\mu^k=l_\mu^\nu S_\nu^k && \text{for } 1\leqslant k\leqslant 8,\\
&\widetilde{\Psi}=R_L\Psi, && R_L \text{ is the spinor transformation matrix},\\
&\widetilde{\partial}_\mu=l_\mu^\nu\partial_\nu,
\end{aligned}\tag{4.1.22}$$

with no change on other fields;

3) the $U(1)$ gauge transformation on $\mathscr{M}\otimes_p\mathbb{C}^4$:

$$\begin{aligned}
&\Omega=e^{i\theta}:\ \mathbb{C}^4\to\mathbb{C}^4\in U(1),\\
&\left(\widetilde{\psi}^e,\widetilde{A}_\mu\right)=\left(e^{i\theta}\psi^e,A_\mu-\frac{1}{e}\partial_\mu\theta\right),
\end{aligned}\tag{4.1.23}$$

4) $SU(2)$ gauge transformation on $\mathscr{M}\otimes_p(\mathbb{C}^4)^2$:

$$\begin{aligned}
&\Omega=e^{i\theta^a\sigma_a}:\ (\mathbb{C}^4)^2\to(\mathbb{C}^4)^2\in SU(2),\\
&\left(\widetilde{\psi}^w,\widetilde{W}_\mu^a\sigma_a,\widetilde{m}_w\right)=\left(\Omega\psi^w,W_\mu^a\Omega\sigma_a\Omega^{-1}+\frac{i}{g_w}\partial_\mu\Omega\Omega^{-1},\Omega m_w\Omega^{-1}\right).
\end{aligned}\tag{4.1.24}$$

5) $SU(3)$ gauge transformation on $\mathscr{M}\otimes_p(\mathbb{C}^4)^3$:

$$\begin{aligned}
&\Omega=e^{i\theta^k\tau_k}:\ (\mathbb{C}^4)^3\to(\mathbb{C}^4)^3\in SU(3),\\
&\left(\widetilde{\psi}^s,\widetilde{S}_\mu^k\tau_k,\widetilde{m}_s\right)=\left(\Omega\psi^s,S_\mu^k\Omega\tau_k\Omega^{-1}+\frac{i}{g_s}\partial_\mu\Omega\Omega^{-1},\Omega m_s\Omega^{-1}\right).
\end{aligned}\tag{4.1.25}$$

6) $SU(2)$ representation transformation on $T_eSU(2)$:

$$\begin{aligned}
&X=(x_a^b):T_eSU(2)\to T_eSU(2), && (y_b^a)^{\mathrm{T}}=X^{-1},\\
&\widetilde{\sigma}_s=x_a^b\sigma_b, && (\widetilde{G}_{ab}^w)=X(G_{ab}^w)X^{\mathrm{T}},\\
&\widetilde{W}_\mu^a=y_b^aW_\mu^b.
\end{aligned}\tag{4.1.26}$$

7) $SU(3)$ representation transformation on $T_eSU(3)$:

$$\begin{aligned}
&X=(x_k^l):\ T_eSU(3)\to T_eSU(3), && (y_l^k)^{\mathrm{T}}=X^{-1},\\
&\widetilde{\tau}_k=x_k^l\tau_l, && (\widetilde{G}_{kl}^s)=X(G_{kl}^s)X^{\mathrm{T}},\\
&\widetilde{S}_\mu^k=y_l^kS_\mu^l.
\end{aligned}\tag{4.1.27}$$

Remark 4.3 Here we adopt the linear transformations of the bundle spaces instead of the coordinate transformations in the base manifold $\mathscr{M}$. In this case, the two transformations (4.1.21) and (4.1.22) are compatible. Otherwise, we have to introduce the Vierbein tensors to overcome the incompatibility between the Lorentz transformation and the general coordinate transformation.

4.1.4 Gauge symmetry-breaking

In physics, symmetries are displayed in two levels in the laws of Nature:

$$\text{the invariance of Lagrangian actions } L, \tag{4.1.28}$$

$$\text{the covariance of variation equations of } L. \tag{4.1.29}$$

The following three symmetries:

$$\begin{aligned} &\text{the Einstein principle of general relativity (PGR),} \\ &\text{the Lorentz invariance,} \\ &\text{the principle of representation invariance (PRI),} \end{aligned} \tag{4.1.30}$$

represent the universality of physical laws— the validity of laws of Nature is independent of the coordinate systems expressing them. Consequently, the symmetries in (4.1.30) cannot be broken at both levels of (4.1.28) and (4.1.29).

The physical implication of the gauge symmetry, however, is different at the two levels:

(1) the gauge invariance of the Lagrangian action, (4.1.28), says that the energy contributions of particles in a physical system are indistinguishable; and

(2) the gauge invariance of the variational equations, (4.1.29), means that the particles involved in the interaction are indistinguishable.

It is clear that the first aspect (1) above is universally true, while the second aspect (2) is not universally true. In other words, the Lagrange actions obey the gauge invariance, but the corresponding variational equations break the gauge symmetry. This suggests us to postulate the following principle of gauge symmetry breaking for interactions described by the gauge theory.

Principle 4.4(Gauge Symmetry Breaking)

1) *The gauge symmetry holds true only for the Lagrangian actions for the electromagnetic, week and strong interactions*; *and*

2) *the field equations of these interactions spontaneously break the gauge symmetry.*

The principle of gauge symmetry breaking can be regarded as part of the spontaneous symmetry breaking, which is a phenomenon appearing in various physical fields. In 2008, the Nobel Prize in Physics was awarded to Y. Nambu for the discovery of the mechanism of spontaneous symmetry breaking in subatomic physics. In 2013, F. Englert and P. Higgs were awarded the Nobel Prize for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles.

This phenomenon was discovered in superconductivity by Ginzburg-Landau in 1951, and the mechanism of spontaneous symmetry breaking in particle physics was first proposed by Y. Nambu in 1960; see (Nambu, 1960; Nambu and Jona-Lasinio, 1961a,b). The Higgs mechanism, introduced in (Higgs, 1964; Englert and Brout, 1964; Guralnik, Hagen and Kibble, 1964), is an artificial method based on the Nambo-Jona-Lasinio spontaneous symmetry breaking, leading to the mass generation for the vector bosons of the weak interaction.

PID discovered by the authors, to be stated in detail in the next section, provides a new mechanism for gauge symmetry breaking and mass generation. The difference between the PID and the Higgs mechanisms is that the first one is a natural sequence of the first principle, and the second is to add artificially a Higgs field in the Lagrangian action. Also, the PID mechanism obeys PRI, and the Higgs mechanism violates PRI. symmetry-breaking!PID-induced

4.1.5 PID and PRI

The main objective in this subsection is to postulate two fundamental principles of physics, the principle of interaction dynamics (PID) and the principle of representation invariance (PRI), which are based on rigorous mathematical foundations established in Sections 3.3-3.5.

Let $\{\mathscr{M}, g_{\mu\nu}\}$ be the 4-dimensional space-time Riemannian manifold with $\{g_{\mu\nu}\}$ the Minkowski type Riemannian metric. For an (r,s)-tensor u we define the A-gradient and A-divergence operators ∇_A and div_A as

$$
\begin{aligned}
\nabla_A u &= \nabla u + u \otimes A, \\
\text{div}_A u &= \text{ div } u - A \cdot u,
\end{aligned}
$$

where A is a vector field and here stands for a gauge field, ∇ and div are the usual gradient and divergent covariant differential operators. Let $F = F(u)$ be a functional of a tensor field u. A tensor u_0 is called an extremum point of F with the div_A-free constraint, if u_0 satisfies the equation

$$
\frac{d}{d\lambda}\Big|_{\lambda=0} F(u_0 + \lambda X) = \int_{\mathscr{M}} \delta F(u_0) \cdot X \sqrt{-g} dx = 0, \qquad \forall X \text{ with } \text{div}_A X = 0. \tag{4.1.31}
$$

Principle 4.5(Principle of Interaction Dynamics)

1) *For all physical interactions there are Lagrangian actions*

$$L(g,A,\psi)=\int_{\mathscr{M}}\mathscr{L}(g_{\mu\nu},A,\psi)\sqrt{-g}dx, \tag{4.1.32}$$

where $g=\{g_{\mu\nu}\}$ *is the Riemannian metric representing the gravitational potential, A is a set of vector fields representing the gauge potentials, and* ψ *are the wave functions of particles*;

2) *The action (4.1.32) satisfy the invariance of general relativity, Lorentz invariance, gauge invariance and the gauge representation invariance;*

3) *The states* (g,A,ψ) *are the extremum points of (4.1.32) with the* div_A*-free constraint (4.1.31).*

Based on PID and Theorems 3.26 and 3.27, the field equations with respect to the action (4.1.32) are given in the form

$$\frac{\delta}{\delta g_{\mu\nu}}L(g,A,\psi)=(\nabla_\mu+\alpha_b A^b_\mu)\Phi_\nu, \tag{4.1.33}$$

$$\frac{\delta}{\delta A^a_\mu}L(g,A,\psi)=(\nabla_\mu+\beta^a_b A^b_\mu)\varphi^a, \tag{4.1.34}$$

$$\frac{\delta}{\delta\psi}L(g,A,\psi)=0 \tag{4.1.35}$$

where $A^a_\mu=(A^a_0,A^a_1,A^a_2,A^a_3)$ are the gauge vector fields for the electromagnetic, the weak and strong interactions, $\Phi_\nu=(\Phi_0,\Phi_1,\Phi_2,\Phi_3)$ in (4.1.33) is a vector field induced by gravitational interaction, φ^a is the scalar fields generated from the gauge fields A^a_μ, and α_b,β^a_b are coupling parameters.

Consider the action (4.1.32) as the natural combination of the actions for all four interactions, as given in (4.1.17)-(4.1.20):

$$\mathscr{L}=\mathscr{L}_{EH}+\mathscr{L}_{EM}+\mathscr{L}_W+\mathscr{L}_S+\mathscr{L}_D.$$

Then (4.1.33)-(4.1.35) provide the unified field equations coupling all interactions. Moreover, we see from (4.1.33)-(4.1.35) that there are too many coupling parameters which need to be determined. Fortunately, this problem can be satisfactorily resolved, leading also to the discovery of PRI (Ma and Wang, 2014h). Meanwhile, we remark that it is the gauge fields A^a_μ appearing on the right-hand sides of (4.1.33) and (4.1.34) that break the gauge symmetry, leading to the mass generation of the vector bosons for the weak interaction sector.

We are now in position to introduce the principle of representation invariance (PRI). We end this section by recalling the principle of representation invariance (PRI) first postulated in (Ma and Wang, 2014h). We proceed with the $SU(N)$ representation. In a neighborhood $U \subset SU(N)$ of the unit matrix, a matrix $\Omega \in U$ can be written as

$$\Omega = e^{i\theta^a \tau_a},$$

where

$$\tau_a = \{\tau_1, \cdots, \tau_K\} \subset T_e SU(N), \quad K = N^2 - 1, \tag{4.1.36}$$

is a basis of generators of the tangent space $T_e SU(N)$; see Section 3.5 for the mathematical theory. An $SU(N)$ representation transformation is a linear transformation of the basis in (4.1.36) as

$$\widetilde{\tau}_a = x_a^b \tau_b, \tag{4.1.37}$$

where $X = (x_a^b)$ is a nondegenerate complex matrix.

Mathematical logic dictates that a physically sound gauge theory should be invariant under the $SU(N)$ representation transformation (4.1.37). Consequently, the following principle of representation invariance (PRI) must be universally valid and was first postulated in (Ma and Wang, 2014h).

Principle 4.6(Principle of Representation Invariance) *All $SU(N)$ gauge theories are invariant under the transformation (4.1.37). Namely, the actions of the gauge fields are invariant and the corresponding gauge field equations as given by (4.1.33)-(4.1.35) are covariant under the transformation (4.1.37).*

Both PID and PRI are very important. As far as we know, it appears that the only unified field model, which obeys not only PRI but also the principle of gauge symmetry breaking, Principle 4.4, is the unified field model based on PID introduced in this chapter. From this model, we can derive not only the same physical conclusions as those from the standard model, but also many new results and predictions, leading to the solution of a number of longstanding open questions in physics, including the 10 problems mentioned in Section 4.1.2.

A few further remarks on PID and PRI are now in order.

First, there are strong theoretical, experimental and observational evidence for PID; see the next section for details.

Second, PID is based on variations with div_A-free constraint defined by (4.1.31). Physically, the div_A-free condition: $\text{div}_A X = 0$ stands for the energy-momentum conservation constraints.

Third, PRI means that the gauge theory is universally valid, and therefore should be independent of the choice of the generators τ_a of $SU(N)$. In other words, PRI is basic a logic requirement.

Fourth, the electroweak interactions is a $U(1) \times SU(2)$ gauge theory coupled with the Higgs mechanism. An unavoidable feature for the Higgs mechanism is that the gauge fields with different symmetry groups are combined linearly into terms in the corresponding gauge field equations. For example, in the Weinberg-Salam electroweak gauge equations with $U(1) \times SU(2)$ symmetry breaking, there are such linearly combinations as

$$\begin{aligned} Z_\mu &= \cos\theta_W W_\mu^3 + \sin\theta_W B_\mu, \\ A_\mu &= -\sin\theta_W W_\mu^3 + \cos\theta_W B_\mu, \\ W_\mu^\pm &= \frac{1}{\sqrt{2}}(W_\mu^1 \pm i W_\mu^2), \end{aligned} \tag{4.1.38}$$

where W_μ^a $(1 \leqslant a \leqslant 3)$ are the $SU(2)$ gauge fields, and B_μ is the $U(1)$ gauge field. It is clear that these terms in (4.1.38) are not covariant under the general $SU(2)$ representation transformations. Hence, the Higgs mechanism violates the PRI. Since the standard model is based on the Higgs mechanism, it violates PRI as well.

4.2 Physical Supports to PID

The original motivation for PID was to explain the dark matter and dark energy phenomena. We have demonstrated that the presence of dark matter and dark energy leads directly to PID for gravity.

There are strong theoretical, experimental and observational evidence for PID. The need of spontaneous symmetry-breaking for generating mass of the vector bosons for the weak interaction is a physical evidence for PID for the weak interaction. The quark confinement requires the introduction of the dual gluon fields demonstrates the necessity of PID; see Section 4.5. In this section, we address the physical evidence from the following viewpoints:

1) the discovery of dark matter and dark energy,
2) the non-existence of solutions for the classical Einstein gravitational field equations in general situations,
3) the principle of spontaneous gauge-symmetry breaking,
4) the Ginzburg-Landau superconductivity theory, and
5) the gauge-fixing problem.

4.2.1 Dark matter and dark energy

The presence of dark matter and dark energy provides a strong support for PID. We recall the Einstein gravitational equations, which are expressed as

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{4\pi G}{c^4} T_{\mu\nu}, \tag{4.2.1}$$

where $T_{\mu\nu}$ is the usual energy-momentum tensor of visible matter. By the Bianchi identity, the left-hand side of (4.2.1) is divergence-free, i.e.

$$\nabla^{\mu}(R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R) = 0. \tag{4.2.2}$$

It implies that the usual energy-momentum tensor satisfies

$$\nabla^{\mu}T_{\mu\nu} = 0. \tag{4.2.3}$$

However, due to the presence of dark matter and dark energy, the energy-momentum tensor of visible matter $T_{\mu\nu}$ may no longer be conserved, i.e. (4.2.3) is not true. Hence we have

$$\nabla^{\mu}T_{\mu\nu} \neq 0,$$

which is a contradiction to (4.2.1) and (4.2.2).

On the other hand, by the Orthogonal Decomposition Theorem 3.17, $T_{\mu\nu}$ can be orthogonally decomposed into

$$T_{\mu\nu} = \widetilde{T}_{\mu\nu} - \frac{c^4}{8\pi G}\nabla_{\mu}\Phi_{\nu}, \tag{4.2.4}$$

and $\widetilde{T}_{\mu\nu}$ is divergence-free:

$$\nabla^{\mu}\widetilde{T}_{\mu\nu} = 0. \tag{4.2.5}$$

Hence, by (4.2.2) and (4.2.5) the gravitational field equations (4.2.1) should be in the form

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}\widetilde{T}_{\mu\nu}. \tag{4.2.6}$$

By (4.2.4) we have

$$\widetilde{T}_{\mu\nu} = T_{\mu\nu} + \frac{c^4}{8\pi G}\nabla_{\mu}\Phi_{\nu},$$

which stands for all energy and momentum including the visible and the invisible matter and energy, and which, by (4.2.5), is conserved. Thus, the equations (4.2.6) are rewritten as

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} - \nabla_{\mu}\Phi_{\nu}, \tag{4.2.7}$$

The equations (4.2.7) are just the variational equations of L_{EH} with the div-free constraint as (4.1.33). Namely, (4.2.7) are the gravitational field equations obeying PID.

We remark that the term $\nabla_{\mu}\Phi_{\nu}$ in (4.2.7) has no variational structure, and cannot be derived by modifying the Einstein-Hilbert functional. Hence, (4.2.7) are just the variational equations due to PID.

4.2.2 Non well-posedness of Einstein field equations

The second strong theoretical evidence to PID is that the classical Einstein gravitational field equations are an over-determined system.

The Einstein field equations (4.2.1) possess 10 unknown functions $g_{\mu\nu}$ and 10 independent equations due to symmetry for the indices μ and ν. However, since the general coordinate system can be arbitrarily chosen, under proper coordinate transformation

$$\widetilde{x}_\mu = a^\nu_\mu x_\nu, \qquad 0 \leqslant \nu \leqslant 3,$$

the 10 unknown functions become

$$\begin{pmatrix} -1 & 0 \\ 0 & g_{ij} \end{pmatrix}, \qquad g_{ij} = g_{ji} \qquad \text{for } 1 \leqslant i, j \leqslant 3.$$

This observation implies that the number of independent unknown functions for the Einstein field equations (4.2.1) is six. Namely,

$$\begin{aligned} &N_{EQ} = 10, \\ &N_{UF} = 6, \end{aligned} \tag{4.2.8}$$

where N_{EQ} is the number of independent equations in (4.2.1), and N_{UF} is the number of independent unknown functions.

Consequently, the Einstein field equations (4.2.1) have no solutions in the general case. Some readers may think that the Bianchi identity

$$\nabla^\mu \left(R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + \frac{8\pi G}{c^4} T_{\mu\nu} \right) = 0, \tag{4.2.9}$$

reduce the number N_{EQ} to six: $N_{EQ} = 6$. But we note that (4.2.9) generates also four new equations

$$\nabla^\mu T_{\mu\nu} = 0 \qquad \text{for } 0 \leqslant \nu \leqslant 3,$$

because there are unknown functions $g_{\mu\nu}$ in the covariant derivative operators ∇_μ; see (3.1.66) or (2.3.26). Hence, the Einstein field equations (4.2.1) should be in the form

$$\begin{aligned} &R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu}, \\ &\nabla^\mu T_{\mu\nu} = 0. \end{aligned} \tag{4.2.10}$$

Thus, the fact (4.2.8) still holds for (4.2.10).

Now we note the gravitational field equations (4.2.7) derived from PID, where there are four additional new unknown functions Φ_ν $(0 \leqslant \nu \leqslant 3)$. In this case, the numbers of independent unknown functions and equations are the same.

In the following, we give an example to show the non well-posedness of the classical Einstein field equations.

It is known that the metric of central gravitational field takes the form

$$ds^2 = c^2 g_{00} dt^2 + g_{11} dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \tag{4.2.11}$$

where (ct, r, θ, φ) is the spherical coordinate system. The metric $g_{\mu\nu}$ in (4.2.11) can be expressed in the form

$$\begin{aligned} &g_{00} = -e^u && (u = u(r)), \\ &g_{11} = e^v && (v = v(r)), \\ &g_{22} = r^2, && \\ &g_{33} = r^2 \sin^2\theta, && \\ &g_{\mu\nu} = 0 && \text{for } \mu \neq \nu. \end{aligned} \tag{4.2.12}$$

Consider the influence of cosmic microwave background (CMB) radiation, the energy-momentum tensor can be approximatively written as

$$T_{\mu\nu} = \begin{pmatrix} -g_{00}\rho & 0 \\ 0 & 0 \end{pmatrix}, \tag{4.2.13}$$

where ρ is the energy density, a constant.

For the metric (4.2.12), the nonzero components of the Ricci tensor are

$$\begin{aligned} &R_{00} = -e^{\mu-\nu}\left[\frac{u''}{2} + \frac{u'}{r} + \frac{u'}{4}(u' - v')\right], \\ &R_{11} = \frac{u''}{2} - \frac{v'}{r} + \frac{u'}{4}(u' - v'), \\ &R_{22} = e^{-v}\left[1 - e^v + \frac{r}{2}(u' - v')\right], \\ &R_{33} = \sin^2\theta R_{22}. \end{aligned} \tag{4.2.14}$$

On the other hand, equations (4.2.10) can be equivalently written as

$$\begin{aligned} &R_{\mu\nu} = -\frac{8\pi G}{c^4}(T_{\mu\nu} - \frac{1}{2} g_{\mu\nu} T), \\ &\nabla^\mu T_{\mu\nu} = 0, \end{aligned} \tag{4.2.15}$$

and by (4.2.13),

$$T = g^{\mu\nu} T_{\mu\nu} = g^{00} T_{00} = -\rho.$$

Thus, the Einstein field equations for the spherically symmetric gravitation fields (4.2.15) are in the form

$$\begin{aligned} R_{00} &= \frac{4\pi G}{c^4} g_{00}\rho, \\ R_{11} &= -\frac{4\pi G}{c^4} g_{11}\rho, \\ R_{22} &= -\frac{4\pi G}{c^4} g_{22}\rho, \\ \nabla^\mu T_{\mu\nu} &= 0. \end{aligned} \tag{4.2.16}$$

Now, we deduce that the equations (4.2.16) have no solutions. By $D^\mu T_{\mu\nu} = 0$, we have

$$\Gamma^0_{10} T_{00} = \frac{1}{2} u' \rho = 0,$$

which implies that $u' = 0$. Hence, by (4.2.14) we have

$$R_{00} = 0,$$

which is a contradiction to the first equation of (4.2.16). Therefore the equations (4.2.16) have no solutions.

However, if we consider this example by using the field equations derived from PID, then the problem must have a solutions; see the theory of dark matter and dark energy in Chapter 7.

4.2.3 Higgs mechanism and mass generation

Principle 4.4 of gauge symmetry breaking is also a main motivation to postulate PID in our program for a unified field theory. In fact, the Higgs mechanism is one way to achieve the spontaneous gauge-symmetry breaking. In the Glashow-Weinberg-Salam (GWS) electroweak theory, the three intermediate vector bosons $W^\pm$ and Z for the weak interaction retain their masses by the Higgs mechanism. We now show that the masses of the intermediate vector bosons can be also obtained by PID. Furthermore, we shall show in Section 4.6 that all conclusions of the GWS electroweak theory confirmed by experiments can be derived by the unified field theory based on PID.

For convenience, we first introduce some related basic knowledge of quantum physics. In quantum field theory, a field ψ is called a fermion with mass m, if it satisfies the Dirac equation

$$(i\gamma^\mu D_\mu - m)\psi = 0, \tag{4.2.17}$$

and the action of (4.2.17) is

$$L_F = \int \mathscr{L}_F dx, \quad \mathscr{L}_F = \overline{\psi}(i\gamma^\mu D_\mu - m)\psi, \tag{4.2.18}$$

where $\overline{\psi} = \psi^\dagger \gamma^0$.

A field Φ is called a boson with mass m, if Φ satisfies a Klein-Gordon type of wave equation:

$$\Box\Phi + \left(\frac{mc}{\hbar}\right)^2 \Phi = g(\Phi), \tag{4.2.19}$$

where $g(\Phi)$ is the terms of Φ other than $k\Phi$ (k a constant), and $\Box$ is the wave operator given by

$$\Box = -\partial^\mu \partial_\mu = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta.$$

The bosonic field Φ is massless if it satisfies

$$\Box\Phi = g(\Phi). \tag{4.2.20}$$

The physical significances of the fermionic and bosonic fields ψ and Φ are as follows:

1) Macro-scale: ψ and Φ represent field energy density. In particular, if Φ is a gauge field then it stands for the interaction potential corresponding to the gauge theory.

2) Micro-scale (Quantization): ψ represents a spin-$\frac{1}{2}$ fermionic particle, and Φ represents a bosonic particle with an integer spin k if Φ is a k-th order tensor field.

In particular, in the classical Yang-Mills theory, the $SU(N)$ gauge fields $A^a_\mu = (A^a_0, A^a_1, A^a_2, A^a_3)$ $(1 \leqslant a \leqslant N^2-1)$ satisfy the following field equations:

$$\partial^\mu F^a_{\mu\nu} = o(A_\mu), \quad F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\lambda^a_{bc} A^b_\mu A^c_\nu, \tag{4.2.21}$$

which are the variational equations of the Yang-Mills action

$$L_{YM} = \int -\frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} + \mathscr{L}_F dx, \tag{4.2.22}$$

where $\mathscr{L}_F$ is as in (4.2.18) with $D_\mu = \partial_\mu + igA^a_\mu \tau_a$, and

$$\partial^\mu F^a_{\mu\nu} = -\Box A^a_\nu - \partial_\nu(\partial^\mu A^a_\mu) + o(A).$$

Thus, the gauge field equations (4.2.21) are reduced to the bosonic field equations (4.2.20). In other words, the gauge fields A^a_μ $(1 \leqslant a \leqslant N^2-1)$ satisfying (4.2.21) represent N^2-1 massless bosons with spin-1 because each A^a_μ is a vector field.

We are now in position to introduce the Higgs mechanism. Physical experiments show that the weak interacting fields should be $SU(2)$ gauge fields with masses, representing 3 massive bosonic particles. However, as mentioned in (4.2.21), the gauge fields satisfying the $SU(2)$ Yang-Mills theory are $3(=N^2-1)$ massless bosons. To overcome this difficulty, (Higgs, 1964; Englert and Brout, 1964; Guralnik, Hagen and Kibble, 1964) suggested to add a scalar field ϕ into the Yang-Mills action (4.2.22) to create masses. In fact, we cannot add a massive term $mA^a_\mu A^{\mu a}$ into the Yang-Mills action (4.2.22); otherwise this action will violate the gauge symmetry. But the Higgs mechanism can ensure the gauge invariance for the Yang-Mills action, and spontaneously break the gauge symmetry in field equations at a ground state of the Higgs field ϕ.

For clearly revealing the essence of the Higgs mechanism, we only take one gauge field (there are four gauge fields in the GWS theory). In this case, the Yang-Mills action density is in the form

$$\mathscr{L}_{YM} = -\frac{1}{4} g^{\mu\alpha} g^{\nu\beta} (\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial_\alpha A_\beta - \partial_\beta A_\alpha) + \overline{\psi}(i\gamma^\mu D_\mu - m)\psi, \tag{4.2.23}$$

where $g^{\mu\nu}$ is the Minkowski metric, and

$$D_\mu \psi = (\partial_\mu + igA_\mu)\psi. \tag{4.2.24}$$

It is clear that the action (4.2.23) is invariant under the following $U(1)$ gauge transformation

$$\psi \to e^{i\theta}\psi, \quad A_\mu \to A_\mu - \frac{1}{g}\partial_\mu \theta. \tag{4.2.25}$$

The variation equations of (4.2.23) are

$$\begin{aligned} &\Box A_\mu + \partial_\mu(\partial^\nu A_\nu) + gJ_\mu = 0, \\ &(i\gamma^\mu D_\mu - m)\psi = 0, \\ &J_\mu = \overline{\psi}\gamma^\mu \psi, \end{aligned} \tag{4.2.26}$$

which are invariant under the gauge transformation (4.2.25). It is clear that the bosonic particle A_μ in (4.2.26) is massless.

To generate mass for A_μ, we add a Higgs sector $\mathscr{L}_H$ to the Yang-Mills action (4.2.23):

$$\begin{aligned} &\mathscr{L}_H = -\frac{1}{2} g^{\mu\nu}(D_\mu \phi)^\dagger (D_\mu \phi) + \frac{1}{4}(\phi^\dagger \phi - \rho^2)^2, \\ &D_\mu \phi = (\partial_\mu + igA_\mu)\phi, \\ &(D_\mu \phi)^\dagger = (\partial_\mu - igA_\mu)\phi^\dagger, \end{aligned} \tag{4.2.27}$$

where $\rho \neq 0$ is a constant. Obviously, the following action

$$L = \int (\mathscr{L}_{YM} + \mathscr{L}_H)dx, \tag{4.2.28}$$

and its variational equations

$$\begin{aligned} &\frac{\delta L}{\delta A^\mu} = \partial^\nu(\partial_\nu A_\mu - \partial_\mu A_\nu) - gJ_\mu - \frac{ig}{2}(\phi(D_\mu \phi)^\dagger - \phi^\dagger D_\mu \phi) = 0, \\ &\frac{\delta L}{\delta \psi} = (i\gamma^\mu D_\mu - m)\psi = 0, \\ &\frac{\delta L}{\delta \phi} = D^\mu D_\mu \phi + (\phi^2 - \rho^2)\phi = 0, \end{aligned} \tag{4.2.29}$$

are invariant under the gauge transformation

$$(\psi,\phi)\to(e^{i\theta}\psi,e^{i\theta}\phi),\quad A_\mu\to A_\mu-\frac{1}{g}\partial_\mu\theta. \tag{4.2.30}$$

Equations (4.2.29) are still massless. However, we note that $(0,0,\rho)$ is a solution of (4.2.29), which represents a ground state in physics, i.e. a vacuum state. Consider a translation for $\Phi=(A,\psi,\phi)$ at $\Phi_0=(0,0,\rho)$ as

$$\Phi=\widetilde{\Phi}+\Phi_0,\quad \widetilde{\Phi}=(\widetilde{A},\widetilde{\psi},\widetilde{\phi}),$$

then the equations (4.2.29) become

$$\begin{aligned}
&\partial^\nu(\partial_\nu\widetilde{A}_\mu-\partial_\mu\widetilde{A}_\nu)-g^2\rho^2\widetilde{A}_\mu-g\widetilde{J}_\mu+g\widetilde{J}_\mu(\widetilde{\phi})=0,\\
&(i\gamma^\mu D_\mu-m)\widetilde{\psi}=0,\\
&D^\mu D_\mu(\widetilde{\phi}+\rho)-2\rho^2\widetilde{\phi}+(\widetilde{\phi}+\rho)\widetilde{\phi}^2=0,
\end{aligned} \tag{4.2.31}$$

where

$$\widetilde{J}_\mu(\phi)=\frac{i}{2}(\widetilde{\phi}(D_\mu\widetilde{\phi})^\dagger-\widetilde{\phi}^\dagger D_\mu\widetilde{\phi}).$$

We see that $\widetilde{A}_\mu$ attains its mass $m=g\rho$ in (4.2.31), but equations (4.2.31) break the invariance for the gauge transformation (4.2.30). The process that masses are created by the spontaneous gauge-symmetry breaking is called the Higgs mechanism. Meanwhile, the field $\widetilde{\phi}$ in (4.2.31), called the Higgs boson, is also obtain its mass $m=\sqrt{2}\rho$.

In the following, we show that PID provides a new mechanism for generating masses, drastly different from the Higgs mechanism.

In view of (4.1.34) and (4.1.35), based on PID, the variational equations of the Yang-Mills action (4.2.23) with the div_A-free constraint are in the form

$$\begin{aligned}
&\partial^\nu(\partial_\nu A_\mu-\partial_\mu A_\nu)-gJ_\mu=\left[\partial_\mu-\frac{1}{4}\left(\frac{mc}{\hbar}\right)^2x_\mu+\lambda A_\mu\right]\phi,\\
&(i\gamma^\mu D_\mu-m_f)\psi=0,
\end{aligned} \tag{4.2.32}$$

where ϕ is a scalar field. The term $-\frac{1}{4}\left(\frac{mc}{\hbar}\right)^2x_\mu$ is the mass potential of ϕ, and is also regarded as the interacting length of ϕ. If ϕ has a nonzero ground state $\phi_0=\rho$, then for the translation

$$\phi=\widetilde{\phi}+\rho,\quad A_\mu=\widetilde{A}_\mu,\quad \psi=\widetilde{\psi},$$

the first equation of (4.2.32) becomes

$$\partial^\nu(\partial_\nu\widetilde{A}_\mu-\partial_\mu\widetilde{A}_\nu)-\left(\frac{m_0c}{\hbar}\right)^2\widetilde{A}_\mu-g\widetilde{J}_\mu=\left[\partial_\mu-\frac{1}{4}\left(\frac{mc}{\hbar}\right)^2x_\mu+\lambda\widetilde{A}_\mu\right]\widetilde{\phi}, \tag{4.2.33}$$

where $\left(\frac{m_0 c}{\hbar}\right)^2 = \lambda\rho$. Thus the mass $m_0 = \frac{\hbar}{c}\sqrt{\lambda\rho}$ is generated in (4.2.33) as the Yang-Mills action takes the div_A-free constraint variation. Moreover, when we take divergence on both sides of (4.2.33), and by

$$\partial^\mu\partial^\nu(\partial_\nu\widetilde{A}_\mu - \partial_\mu\widetilde{A}_\nu) = 0, \quad \partial^\mu\widetilde{J}_\mu = 0,$$

we derive the field equation of $\widetilde{\phi}$ as follows

$$\partial^\mu\partial_\mu\widetilde{\phi} - \left(\frac{mc}{\hbar}\right)^2\widetilde{\phi} = -\lambda\partial^\mu(\widetilde{A}_\mu\widetilde{\phi}) + \frac{1}{4}\left(\frac{mc}{\hbar}\right)^2 x_\mu\partial^\mu\widetilde{\phi}. \tag{4.2.34}$$

This equation (4.2.34) is the field equation with mass m for the Higgs bosonic particle $\widetilde{\phi}$.

Remark 4.7 In (4.2.28) we see that the essence of the Higgs mechanism is to add artificially a Higgs sector $\mathscr{L}_H$ into the Yang-Mills action. However, for the PID model, the masses of A_μ and the Higgs field ϕ are generated naturally for the first principle, PID, taking the variation with energy-momentum conservation constraint.

4.2.4 Ginzburg-Landau superconductivity

Superconductivity studies the behavior of the Bose-Einstein condensation and electromagnetic interactions. The Ginzburg-Landau theory provides a support for PID.

The Ginzburg-Landau free energy for superconductivity is

$$G = \int_\Omega \left[\frac{1}{2m_s}\left|\left(i\hbar\nabla + \frac{e_s}{c}A\right)\psi\right|^2 + a|\psi|^2 + \frac{b}{2}|\psi|^4 + \frac{1}{8\pi}|\text{ curl }A|^2\right]dx, \tag{4.2.35}$$

where A is the electromagnetic potential, ψ is the wave function of superconducting electrons, Ω is the superconductor, e_s and m_s are charge and mass of a Cooper pair.

The superconducting current equations determined by the Ginzburg-Landau free energy (4.2.35) are:

$$\frac{\delta G}{\delta A} = 0, \tag{4.2.36}$$

which implies that

$$\frac{c}{4\pi}\text{curl}^2 A = -\frac{e_s^2}{m_s c}|\psi|^2 A - i\frac{\hbar e_s}{m_s}(\psi^*\nabla\psi - \psi\nabla\psi^*). \tag{4.2.37}$$

Let

$$J = \frac{c}{4\pi}\text{curl}^2 A, \qquad J_s = \frac{e_s^2}{m_s c}|\psi|^2 A - i\frac{\hbar e_s}{m_s}(\psi^*\nabla\psi - \psi\nabla\psi^*).$$

Physically, J is the total current in Ω, and J_s is the superconducting current. Since Ω is a medium conductor, J contains two types of currents as

$$J = J_s + \sigma E,$$

where σ is dielectric constant, σE is the current generated by the electric field E,

$$E = -\frac{1}{c}\frac{\partial A}{\partial t} - \nabla\Phi = -\nabla\Phi,$$

and Φ is the electric potential. Since $A_t = 0$, the superconducting current equations should be taken as

$$\frac{1}{4\pi}\,\text{curl}^2 A = -\frac{\sigma}{c}\nabla\Phi - \frac{e_s^2}{m_s c^2}|\psi|^2 A - \frac{i\hbar e_s}{m_s c}(\psi^*\nabla\psi - \psi\nabla\psi^*). \tag{4.2.38}$$

Since (4.2.37) is the expression of (4.2.36), then the equation (4.2.38) can be written in the abstract form

$$\frac{\delta G}{\delta A} = -\frac{\sigma}{c}\nabla\Phi. \tag{4.2.39}$$

In addition, for conductivity, the gauge fixing is given by

$$\text{div}\, A = 0, \quad A\cdot n|_{\partial\Omega} = 0,$$

which imply that

$$\int_\Omega \nabla\Phi\cdot A dx = 0.$$

Hence, the term $-\frac{\sigma}{c}\nabla\Phi$ in (4.2.39) can not be added into the Ginzburg-Landau free energy (4.2.35).

However, the equation (4.2.39) are just the variational equation with divergence-free constraint as follows

$$\left\langle \frac{\delta G}{\delta A}, X\right\rangle = \frac{d}{d\lambda}G(A+\lambda X)|_{\lambda=0} = 0, \qquad \forall\ \text{div}\, X = 0.$$

Thus, we see that the Ginzburg-Landau superconductivity theory obeys PID.

4.3 Unified Field Model Based on PID and PRI

4.3.1 Unified field equations based on PID

The abstract unified field equations (4.1.33)-(4.1.34) are derived based on PID. We now present the detailed form of this model, ensuring that these field equations satisfy both the principle of gauge-symmetry breaking, Principle 4.4, and PRI.

In Section 4.1.3, we showed that the action functional obeys all the symmetric principles, including principle of general relativity, the Lorentz invariance, the $U(1)\times SU(2)\times SU(3)$ gauge invariance and PRI, is the natural combination of the Einstein-Hilbert functional, the $U(1), SU(2), SU(3)$ Yang-Mills actions for the electromagnetic, weak and strong interactions:

$$L = \int_{\mathscr{M}} [\mathscr{L}_{EH} + \mathscr{L}_{EM} + \mathscr{L}_W + \mathscr{L}_S]\sqrt{-g}dx. \tag{4.3.1}$$

Here

$$
\begin{aligned}
&\mathscr{L}_{EH}=R+\frac{8\pi G}{c^4}S,\\
&\mathscr{L}_{EM}=-\frac{1}{4}A_{\mu\nu}A^{\mu\nu}+\overline{\Psi}^e(i\gamma^\mu D_\mu-m)\psi^e,\\
&\mathscr{L}_W=-\frac{1}{4}\mathscr{G}^w_{ab}W^a_{\mu\nu}W^{b\mu\nu}+\overline{\Psi}^w(i\gamma^\mu D_\mu-m_l)\psi^w,\\
&\mathscr{L}_S=-\frac{1}{4}\mathscr{G}^s_{kl}S^k_{\mu\nu}S^{\mu\nu l}+\overline{\Psi}^s(i\gamma^\mu D_\mu-m_q)\psi^s,
\end{aligned}
\tag{4.3.2}
$$

where R is the scalar curvature of the space-time Riemannian manifold $(\mathscr{M},g_{\mu\nu})$ with Minkowski type metric, S is the energy-momentum density, $\mathscr{G}^w_{ab}$ and $\mathscr{G}^s_{kl}$ are the $SU(2)$ and $SU(3)$ metrics as defined by (3.5.28), ψ^e, ψ^w and ψ^s are the Dirac spinors for fermions participating in the electromagnetic, weak, strong interactions, and

$$
\begin{aligned}
&A_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu,\\
&W^a_{\mu\nu}=\partial_\mu W^a_\nu-\partial_\nu W^a_\mu+g_w\lambda^a_{bc}W^b_\mu W^c_\nu,\\
&S^k_{\mu\nu}=\partial_\mu S^k_\nu-\partial_\nu S^k_\mu+g_s\Lambda^k_{lr}S^l_\mu S^r_\nu.
\end{aligned}
\tag{4.3.3}
$$

Here A_μ is the electromagnetic potential, W^a_μ $(1\leqslant a\leqslant 3)$ are the $SU(2)$ gauge fields for the weak interaction, S^k_μ $(1\leqslant k\leqslant 8)$ are the $SU(3)$ gauge fields for the strong interaction, g_w and g_s are the weak and strong charges, and

$$
\begin{aligned}
&D_\mu\psi^e=(\partial_\mu+ieA_\mu)\psi^e,\\
&D_\mu\psi^w=(\partial_\mu+ig_wW^a_\mu\sigma_a)\psi^w,\\
&D_\mu\psi^s=(\partial_\mu+ig_sS^k_\mu\tau_k)\psi^s,
\end{aligned}
\tag{4.3.4}
$$

where σ_a $(1\leqslant a\leqslant 3)$ and τ_k $(1\leqslant k\leqslant 8)$ are the generators of $SU(2)$ and $SU(3)$.

Remark 4.8 For a vector field X_μ and an antisymmetric tensor field $F_{\mu\nu}$, we have

$$
\begin{aligned}
&\nabla_\mu X_\nu-\nabla_\nu X_\mu=\partial_\mu X_\nu-\partial_\nu X_\mu,\\
&\nabla^\mu F_{\mu\nu}=\partial^\mu F_{\mu\nu},
\end{aligned}
$$

where ∇_μ is the Levi-Civita covariant derivative. Hence, the tensor fields in (4.3.3) and the action (4.3.1) obey both the Einstein general relativity and the Lorentz invariance simultaneously under the transformations (4.1.21)-(4.1.22). □

Remark 4.9 In the standard model, the wave functions ψ^w and ψ^s in $\mathscr{L}_W$ and $\mathscr{L}_S$ are as follows

$$
\begin{aligned}
&\psi^w=\begin{pmatrix}l_1\\ l_2\end{pmatrix}_L \quad \text{the left-hand lepton pairs},\\
&\psi^s=\begin{pmatrix}q_1\\ q_2\\ q_3\end{pmatrix} \quad q_i\ \text{the quark with ith color.}
\end{aligned}
\tag{4.3.5}
$$

The reason why ψ^w and ψ^s are taken in the form (4.3.5) is that the standard model, in particular the GWS electroweak theory, is oriented toward to computing the transition probability for the decay and scattering. In the PID field theory, it is unnecessary to take ψ^w and ψ^s as in (4.3.5), because this model is oriented toward to interaction potentials and the basic mechanism. □

Remark 4.10 According to the standard model, the field particles corresponding to electromagnetic, weak, and strong interactions are described by $U(1)$, $SU(2)$, $SU(3)$ gauge fields. Hence we take the $U(1)\times SU(2)\times SU(3)$ Yang-Mills action together with $\mathscr{L}_{EH}$ as the action. However, if only consider the field theory for an N-particle system with N_1 electric, N_2 weak, N_3 strong charges, then the action sectors in (4.3.1) should be taken as

$$\begin{aligned}
&\mathscr{L}_{EH} = R,\\
&\mathscr{L}_{EM} = SU(N_1) \text{ Yang-Mills action},\\
&\mathscr{L}_{W} = SU(N_2) \text{ Yang-Mills action, and}\\
&\mathscr{L}_{S} = SU(N_3) \text{ Yang-Mills action.}
\end{aligned}$$

In Section 6.5, we shall discuss the unified field theory for multi-particle systems. □

We are now in position to establish unified field equations obeying PRI and PSB. By PID, the unified field model (4.1.33)-(4.1.34) are derived as the variational equation of the action (4.1.17) under the div_A-constraint

$$\langle \delta L, X\rangle = 0 \quad \text{for any } X \text{ with } \text{div}_A X = 0.$$

Here it is required that the gradient operator ∇_A corresponding to div_A are PRI covariant. The gradient operators in different sectors are given as follows:

$$\begin{aligned}
D^g_\mu &= \nabla_\mu + \alpha^0 A_\mu + \alpha^1_b W^b_\mu + \alpha^2_k S^k_\mu,\\
D^e_\mu &= \nabla_\mu + \beta^0 A_\mu + \beta^1_b W^b_\mu + \beta^2_k S^k_\mu,\\
D^w_\mu &= \nabla_\mu + \gamma^0 A_\mu + \gamma^1_b W^b_\mu + \gamma^2_k S^k_\mu - \frac{1}{4} m^2_w x_\mu,\\
D^s_\mu &= \nabla_\mu + \delta^0 A_\mu + \delta^1_b W^b_\mu + \delta^2_k S^k_\mu - \frac{1}{4} m^2_s x_\mu,
\end{aligned} \tag{4.3.6}$$

where

$$\begin{aligned}
&m_w, m_s, \alpha^0, \beta^0, \gamma^0, \delta^0 && \text{are scalar parameters},\\
&\alpha^1_a, \beta^1_a, \gamma^1_a, \delta^1_a && \text{are first-order } SU(2) \text{ tensors},\\
&\alpha^2_k, \beta^2_k, \gamma^2_k, \delta^2_k && \text{are first-order } SU(3) \text{ tensors.}
\end{aligned} \tag{4.3.7}$$

Thus, the PID equations (4.1.33)-(4.1.34) can be expressed as

$$\begin{aligned}
&\frac{\delta L}{\delta g_{\mu\nu}} = D^g_\mu \phi^g_\nu,\\
&\frac{\delta L}{\delta A_\mu} = D^e_\mu \phi^e,\\
&\frac{\delta L}{\delta W^a_\mu} = D^w_\mu \phi^w_a,\\
&\frac{\delta L}{\delta S^k_\mu} = D^s_\mu \phi^s_k,
\end{aligned} \tag{4.3.8}$$

where ϕ^g_ν is a vector field, and ϕ^e, ϕ^w, ϕ^s are scalar fields.

With the PID equations (4.3.8), the PRI covariant unified field equations are then given as follows:[1]

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu} + D^g_\mu \phi^g_\nu, \tag{4.3.9}$$

$$\partial^\mu(\partial_\mu A_\nu - \partial_\nu A_\mu) - eJ_\nu = D^e_\nu \phi^e, \tag{4.3.10}$$

$$\mathscr{G}^w_{ab}\left[\partial^\mu W^b_{\mu\nu} - g_w \lambda^b_{cd} g^{\alpha\beta} W^c_{\alpha\nu} W^d_\beta\right] - g_w J_{\nu a} = D^w_\nu \phi^w_a, \tag{4.3.11}$$

$$\mathscr{G}^s_{kj}\left[\partial^\mu S^j_{\mu\nu} - g_s \Lambda^j_{cd} g^{\alpha\beta} S^c_{\alpha\nu} S^d_\beta\right] - g_s Q_{\nu k} = D^s_\nu \phi^s_k, \tag{4.3.12}$$

$$(i\gamma^\mu D_\mu - m)\psi^e = 0, \tag{4.3.13}$$

$$(i\gamma^\mu D_\mu - m_l)\psi^w = 0, \tag{4.3.14}$$

$$(i\gamma^\mu D_\mu - m_q)\psi^s = 0, \tag{4.3.15}$$

where $D^g_\mu, D^e_\nu, D^w_\nu, D^s_\nu$ are given by (4.3.6), and

$$\begin{aligned}
&J_\nu = \overline{\psi}^e \gamma_\nu \psi^e,\\
&J_{\nu a} = \overline{\psi}^w \gamma^\nu \sigma_a \psi^w,\\
&Q_{\nu k} = \overline{\psi}^s \gamma_\nu \tau_k \psi^s,\\
&T_{\mu\nu} = \frac{\delta S}{\delta g_{\mu\nu}} + \frac{c^4}{16\pi G} g^{\alpha\beta}(\mathscr{G}^w_{ab} W^a_{\alpha\mu} W^b_{\beta\nu} + \mathscr{G}^s_{kl} S^k_{\alpha\mu} S^l_{\beta\nu} + A_{\alpha\mu} A_{\beta\nu})\\
&\qquad - \frac{c^4}{16\pi G} g_{\mu\nu}(\mathscr{L}_{EM} + \mathscr{L}_W + \mathscr{L}_S).
\end{aligned} \tag{4.3.16}$$

Remark 4.11 It is clear that the action (4.3.1)-(4.3.4) for the unified field model is invariant under the $U(1) \times SU(2) \times SU(3)$ gauge transformation as follows

[1] We ignore the Klein-Gordon fields.

$$\begin{aligned}
&\left(\widetilde{\psi}^e,\widetilde{A}_\mu\right)=\left(e^{i\theta}\psi^e,A_\mu-\frac{1}{e}\partial_\mu\theta\right),\\
&\left(\widetilde{\psi}^w,\widetilde{W}^a_\mu\sigma_a\right)=\left(U\psi^w,W^a_\mu U\sigma_aU^{-1}+\frac{i}{g_w}\partial_\mu UU^{-1}\right),\quad U=e^{i\theta^a\sigma_a},\\
&\left(\widetilde{\psi}^s,\widetilde{S}^k\tau_k\right)=\left(e^{i\varphi^k\tau_k}\psi^s,S^k_\mu\Omega\tau_k\Omega^{-1}+\frac{i}{g_s}\partial_\mu\Omega\Omega^{-1}\right),\quad \Omega=e^{i\varphi^k\tau_k},\\
&\widetilde{m}_l=e^{i\theta^a\sigma_a}m_le^{-i\theta^a\sigma_a},\\
&\widetilde{m}_q=e^{i\varphi^k\tau_k}m_qe^{-i\varphi^k\tau_k}.
\end{aligned}\tag{4.3.17}$$

However, the equations (4.3.9)-(4.3.15) are not invariant under the transformation (4.3.17) due to the terms $D^g_\mu\phi^g_\nu, D^e_\nu\phi^e, D^w_\nu\phi^w_a, D^s_\nu\phi^s_k$ on the right-hand sides of (4.3.9)-(4.3.12) containing the gauge fields A_μ, W^a_μ and S^k_μ.

Hence, the unified field model based on PID and PRI satisfies the spontaneous gauge-symmetry breaking as stated in Principle 4.4 and PRI.

4.3.2 Coupling parameters and physical dimensions

There are a number of to-be-determined coupling parameters in the general form of the unified field equations (4.3.9)-(4.3.15), and the $SU(2)$ and $SU(3)$ generators σ_a and τ_k are taken arbitrarily. With PRI we are able to substantially reduce the number of these to-be-determined parameters in the unified model to two $SU(2)$ and $SU(3)$ tensors

$$\{\alpha^w_a\}=(\alpha^w_1,\alpha^w_2,\alpha^w_3),\quad \{\alpha^s_k\}=(\alpha^s_1,\cdots,\alpha^s_8),$$

containing 11 parameters, representing the portions distributed to the gauge potentials by the weak and strong charges.

Also, if we take σ_a $(1\leqslant a\leqslant 3)$ as the Pauli matrices (3.5.36) and $\tau_k=\lambda_k$ $(1\leqslant k\leqslant 8)$ as the Gell-Mann matrices (3.5.38), then the two metrics $\mathscr{G}^w_{ab}$ and $\mathscr{G}^s_{kl}$ are Euclidian:

$$\mathscr{G}^w_{ab}=\delta_{ab},\qquad \mathscr{G}^s_{kl}=\delta_{kl}.$$

Hence we usually take the Pauli matrices σ_a and the Gell-Mann matrices λ_k as the $SU(2)$ and $SU(3)$ generators.

For convenience, we first introduce dimensions of related physical quantities. Let E represent energy, L be the length and t be the time. Then we have

$$\begin{aligned}
&(A_\mu,W^a_\mu,S^k_\mu):\ \sqrt{E/L}, &&(e,g_w,g_s):\ \sqrt{EL},\\
&(J_\mu,J_{\mu a},Q_{\mu k}):\ 1/L^3, &&(\phi^e,\phi^w_a,\phi^s_k):\ \frac{\sqrt{E}}{\sqrt{L}L},\\
&(\hbar,c):(Et,L/t), &&mc/\hbar:\ 1/L.
\end{aligned}$$

In addition, for gravitational fields we have

$$\begin{aligned} &g_{\mu\nu}: \text{ dimensionless}, \quad R: 1/L^2, \qquad T_{\mu\nu}: E/L^3, \\ &\phi^g_\mu: 1/L, \qquad\qquad\qquad G: L^5/Et^4. \end{aligned} \tag{4.3.18}$$

The dimensions of the parameters in (4.3.9)–(4.3.15) are as follows

$$\begin{aligned} &(m_w, m_s): 1/L, &&(\alpha^0,\beta^0,\gamma^0,\delta^0): 1/\sqrt{EL}, \\ &(\alpha^1_a,\beta^1_a,\gamma^1_a,\delta^1_a): 1/\sqrt{EL}, &&(\alpha^2_k,\beta^2_k,\gamma^2_k,\delta^2_k): 1/\sqrt{EL}. \end{aligned} \tag{4.3.19}$$

Thus the parameters in (4.3.7) can be rewritten as

$$\begin{aligned} &(m_w, m_s) = \left(\frac{m_H c}{\hbar}, \frac{m_\pi c}{\hbar}\right), \\ &(\alpha^0,\beta^0,\gamma^0,\delta^0) = \frac{e}{\hbar c}(\alpha^e,\beta^e,\gamma^e,\delta^e), \\ &(\alpha^1_a,\beta^1_a,\gamma^1_a,\delta^1_a) = \frac{g_w}{\hbar c}(\alpha^w_a,\beta^w_a,\gamma^w_a,\delta^w_a), \\ &(\alpha^2_k,\beta^2_k,\gamma^2_k,\delta^2_k) = \frac{g_s}{\hbar c}(\alpha^s_k,\beta^s_k,\gamma^s_k,\delta^s_k), \end{aligned} \tag{4.3.20}$$

where m_H and m_π represent the masses of ϕ^w and ϕ^s, and all the parameters $(\alpha,\beta,\gamma,\delta)$ on the right hand side of (4.3.20) with different super and sub indices are dimensionless constants.

4.3.3 Standard form of unified field equations

With (4.3.20) at our disposal, the unified field equations (4.3.9)-(4.3.15) can be simplified in the form

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} + \left[\nabla_\mu + \frac{e\alpha^e}{\hbar c}A_\mu + \frac{g_w\alpha^w_a}{\hbar c}W^a_\mu + \frac{g_s\alpha^s_k}{\hbar c}S^k_\mu\right]\phi^g_\nu, \tag{4.3.21}$$

$$\partial^\nu A_{\nu\mu} - eJ_\nu = \left[\partial_\mu + \frac{e}{\hbar c}\beta^e A_\mu + \frac{g_w}{\hbar c}\beta^w_a W^a_\mu + \frac{g_s}{\hbar c}\beta^s_k S^k_\mu\right]\phi^e, \tag{4.3.22}$$

$$\begin{aligned} &\partial^\nu W^a_{\nu\mu} - \frac{g_w}{\hbar c}\varepsilon^a_{bc}g^{\alpha\beta}W^b_{\alpha\mu}W^c_\beta - g_w J^a_\mu \\ &\quad = \left[\partial_\mu + \frac{e}{\hbar c}\gamma^e A_\mu + \frac{g_w}{\hbar c}\gamma^w_b W^b_\mu + \frac{g_s}{\hbar c}\gamma^s_k S^k_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu\right]\phi^a_w, \end{aligned} \tag{4.3.23}$$

$$\begin{aligned} &\partial^\nu S^k_{\nu\mu} - \frac{g_s}{\hbar c}f^k_{ij}g^{\alpha\beta}S^i_{\alpha\mu}S^j_\beta - g_s Q^k_\mu \\ &\quad = \left[\partial_\mu + \frac{e}{\hbar c}\delta^e A_\mu + \frac{g_w}{\hbar c}\delta^w_b W^b_\mu + \frac{g_s}{\hbar c}\delta^s_l S^l_\mu - \frac{1}{4}\left(\frac{m_\pi c}{\hbar}\right)^2 x_\mu\right]\phi^k_s, \end{aligned} \tag{4.3.24}$$

$$(i\gamma^\mu D_\mu - m)\Psi = 0, \tag{4.3.25}$$

where $\Psi = (\psi^e, \psi^w, \psi^s)$, and

$$\begin{aligned} A_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu, \\ W^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + \frac{g_w}{\hbar c}\varepsilon^a_{bc} W^b_\mu W^c_\nu, \\ S^k_{\mu\nu} &= \partial_\mu S^k_\nu - \partial_\nu S^k_\mu + \frac{g_s}{\hbar c} f^k_{ij} S^i_\mu S^j_\nu. \end{aligned} \tag{4.3.26}$$

Equations (4.3.21)-(4.3.25) need to be supplemented with coupled gauge equations to compensate the new dual fields $(\phi^e, \phi^a_w, \phi^k_s)$. In different physical situations, the coupled gauge equations may be different.

From the field theoretical point of view instead of the field particle point of view, the coefficients in (4.3.21)-(4.3.24) should be

$$\begin{aligned} (\alpha^w_1, \alpha^w_2, \alpha^w_3) &= \alpha^w(\omega_1, \omega_2, \omega_3), \\ (\beta^w_1, \beta^w_2, \beta^w_3) &= \beta^w(\omega_1, \omega_2, \omega_3), \\ (\gamma^w_1, \gamma^w_2, \gamma^w_3) &= \gamma^w(\omega_1, \omega_2, \omega_3), \\ (\delta^w_1, \delta^w_2, \delta^w_3) &= \delta^w(\omega_1, \omega_2, \omega_3), \end{aligned} \tag{4.3.27}$$

and

$$\begin{aligned} (\alpha^s_1, \cdots, \alpha^s_8) &= \alpha^s(\rho_1, \cdots, \rho_8), \\ (\beta^s_1, \cdots, \beta^s_8) &= \beta^s(\rho_1, \cdots, \rho_8), \\ (\gamma^s_1, \cdots, \gamma^s_8) &= \gamma^s(\rho_1, \cdots, \rho_8), \\ (\delta^s_1, \cdots, \delta^s_8) &= \gamma^s(\rho_1, \cdots, \rho_8), \end{aligned} \tag{4.3.28}$$

with the unit modules:

$$\begin{aligned} |\omega| &= \sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2} = 1, \\ |\rho| &= \sqrt{\rho_1^2 + \cdots + \rho_8^2} = 1, \end{aligned}$$

using the Pauli matrices σ_a and the Gell-Mann matrices λ_k as the generators for $SU(2)$ and $SU(3)$ respectively.

The two $SU(2)$ and $SU(3)$ tensors in (4.3.27) and (4.3.28),

$$\omega_a = (\omega_1, \omega_2, \omega_3), \quad \rho_k = (\rho_1, \cdots, \rho_8), \tag{4.3.29}$$

are very important, by which we can obtain $SU(2)$ and $SU(3)$ representation invariant gauge fields:

$$W_\mu = \omega_a W^a_\mu, \quad S_\mu = \rho_k S^k_\mu. \tag{4.3.30}$$

which represent respectively the weak and the strong interaction potentials.

In view of (4.3.27)-(4.3.30), the unified field equations for the four fundamental forces are written as

$$R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R+\frac{8\pi G}{c^4}T_{\mu\nu}=\left[\nabla_\mu+\frac{e\alpha^e}{\hbar c}A_\mu+\frac{g_w\alpha^w}{\hbar c}W_\mu+\frac{g_s\alpha^s}{\hbar c}S_\mu\right]\phi^g_\nu, \tag{4.3.31}$$

$$\partial^\nu A_{\nu\mu}-eJ_\mu=\left[\partial_\mu+\frac{e\beta^e}{\hbar c}A_\mu+\frac{g_w\beta^w}{\hbar c}W_\mu+\frac{g_s\beta^s}{\hbar c}S_\mu\right]\phi^e, \tag{4.3.32}$$

$$\begin{aligned}&\partial^\nu W^a_{\nu\mu}-\frac{g_w}{\hbar c}\varepsilon^a_{bc}g^{\alpha\beta}W^b_{\alpha\mu}W^c_\beta-g_wJ^a_\mu \qquad (4.3.33)\\ &\quad=\left[\partial_\mu-\frac{1}{4}k_w^2x_\mu+\frac{e\gamma^e}{\hbar c}A_\mu+\frac{g_w\gamma^w}{\hbar c}W_\mu+\frac{g_s\gamma^s}{\hbar c}S_\mu\right]\phi^a_w,\\ &\partial^\nu S^k_{\nu\mu}-\frac{g_s}{\hbar c}f^k_{ij}g^{\alpha\beta}S^i_{\alpha\mu}S^j_\beta-g_sQ^k_\mu \qquad (4.3.34)\\ &\quad=\left[\partial_\mu-\frac{1}{4}k_s^2x_\mu+\frac{e\delta^e}{\hbar c}A_\mu+\frac{g_w\delta^w}{\hbar c}W_\mu+\frac{g_s\delta^s}{\hbar c}S_\mu\right]\phi^k_s,\end{aligned}$$

$$(i\gamma^\mu D_\mu-m)\Psi=0. \tag{4.3.35}$$

4.3.4 Potentials of the weak and strong forces

It is known that the $U(1)$ gauge fields

$$A_\mu=(A_0,A_1,A_2,A_3) \tag{4.3.36}$$

represent the electromagnetic potentials, with

$$\begin{aligned}&A_0=\text{ the Coulomb potential},\\ &\vec{A}=\text{ magnetic potential},\qquad \vec{A}=(A_1,A_2,A_3),\end{aligned} \tag{4.3.37}$$

and the electric charge e is

$$e=\text{ the } U(1) \text{ gauge coupling constant}. \tag{4.3.38}$$

The electromagnetic forces are given by

$$\begin{aligned}&F_e=-e\nabla A_0 &&\text{the Coulomb force},\\ &F_m=\frac{e}{c}\vec{v}\times\text{ curl}\vec{A} &&\text{the Lorentz force}.\end{aligned} \tag{4.3.39}$$

Now, we consider the $SU(2)$ and $SU(3)$ gauge fields:

$$\begin{aligned}&SU(2)\text{ gauge fields:}\quad W^a_\mu=(W^a_0,W^a_1,W^a_2,W^a_3),\quad 1\leqslant a\leqslant 3,\\ &SU(3)\text{ gauge fields:}\quad S^k_\mu=(S^k_0,S^k_1,S^k_2,S^k_3),\quad 1\leqslant k\leqslant 8.\end{aligned} \tag{4.3.40}$$

They are $SU(N)$ tensors with $N=2,3$, and have N^2-1 components. These components will change under the transformation of $SU(N)$ generators. Thanks to PRI, the N^2-1 $(N=2,3)$

gauge fields in (4.3.40) can be combined into two vector fields as in (4.3.30):

$$\begin{aligned} W_\mu &= \omega_a W_\mu^a = (W_0, W_1, W_2, W_3), \\ S_\mu &= \rho_k S_\mu^k = (S_0, S_1, S_2, S_3), \end{aligned} \tag{4.3.41}$$

which have the same role as (4.3.36)-(4.3.39) for the electro-magnetic $U(1)$ gauge fields.

In the same spirit as the electromagnetic fields, for the two fields given by (4.3.41), we have

$$\begin{aligned} W_0 &= \text{ the weak force potential}, \\ \vec{W} &= \text{ the weak magnetic potential}, \quad \vec{W} = (W_1, W_2, W_3), \end{aligned} \tag{4.3.42}$$

and

$$\begin{aligned} S_0 &= \text{ the strong force potential}, \\ \vec{S} &= \text{ the strong magnetic potential}, \quad \vec{S} = (S_1, S_2, S_3). \end{aligned} \tag{4.3.43}$$

In addition, the weak and strong charges g_w and g_s are

$$\begin{aligned} &\text{weak charge } g_w = SU(2) \text{ gauge coupling constant}, \\ &\text{strong charge } g_s = SU(3) \text{ gauge coupling constant}. \end{aligned} \tag{4.3.44}$$

The weak and strong forces are given by

$$\begin{aligned} &\text{weak force:} && F_w = -g_w \nabla W_0, \\ &\text{weak magnetic force:} && F_{wm} = \frac{g_w}{c} \vec{v} \times \text{ curl } \vec{W}, \end{aligned} \tag{4.3.45}$$

and

$$\begin{aligned} &\text{strong force:} && F_s = -g_s \nabla S_0, \\ &\text{strong magnetic force:} && F_{sm} = \frac{g_s}{c} \vec{v} \times \text{ curl } \vec{S}. \end{aligned} \tag{4.3.46}$$

Remark 4.12 It is the PRI that provides a physical approach to combine the $N^2 - 1$ components of the $SU(N)$ gauge fields into the forms (4.3.41)-(4.3.46) for the interacting forces. With this physical interpretation, the electromagnetic, weak and strong interactions can be regarded as a unified force, separated by three different interaction charges: the electric charge e, weak charge g_w, strong charge g_s.

4.3.5 Gauge-fixing problem

For a gauge theory, one has to supplement gauge-fixing equations to ensure a unique physical solution. In the $U(1)$ gauge theory the gauge-fixing problem is well-posed. However for the $SU(N)$ gauge theory, the problem is generally not well-posed under the Principle of Lagrange Dynamics (PLD).

We first recall the classical $U(1)$ gauge theory describing electromagnetism. The field equations by PLD are given by

$$\partial^\nu (\partial_\nu A_\mu - \partial_\mu A_\nu) = e \overline{\psi} \gamma_\mu \psi \qquad (\gamma_\mu = g_{\mu\nu} \gamma^\nu), \tag{4.3.47}$$

$$i\gamma^\mu (\partial_\mu + ieA_\mu)\psi - m\psi = 0, \tag{4.3.48}$$

which are invariant under the $U(1)$ gauge transformation

$$\widetilde{\psi}=e^{i\theta}\psi, \qquad \widetilde{A}_\mu=A_\mu-\frac{1}{e}\partial_\mu\theta.$$

It implies that $\widetilde{A}_\mu, \widetilde{\psi}$ are also solutions. Hence, the equations (4.3.47)-(4.3.48) have infinitely many solutions. However, in these solutions only one is physical. To find the physical solution, one has to provide a supplementary equation, called gauge-fixing equation,

$$F(A_\mu)=0, \tag{4.3.49}$$

such that the system (4.3.47)-(4.3.49) has a unique physical solution. Observe that

$$\begin{aligned}&\partial^\mu\partial^\nu(\partial_\nu A_\mu-\partial_\mu A_\nu)=0,\\&\partial^\mu\overline{\psi}\gamma_\mu\psi=0 \qquad (\text{by } (4.3.48)).\end{aligned}$$

Only three equations in (4.3.47) are independent. Therefore the number of independent equations in (4.3.47)-(4.3.49) is $N_{EQ}=8$, the same as the number of unknowns. Hence, (4.3.47)-(4.3.49) are well-posed. Usually physical gauge-fixing equation (4.3.47) takes one of the forms:

$$\begin{aligned}&\text{Coulomb gauge}: && \frac{\partial A_1}{\partial x^1}+\frac{\partial A_2}{\partial x^2}+\frac{\partial A_3}{\partial x^3}=0,\\&\text{Lorentz gauge}: && \partial^\mu A_\mu=0,\\&\text{Axial gauge}: && A_3=0,\\&\text{Temporal gauge}: && A_0=0.\end{aligned}$$

However, for the $SU(N)$ gauge theory by PLD, the gauge-fixing problem is in general not well-posed. In fact, the $SU(N)$ gauge field equations are given by

$$\partial^\nu F^a_{\nu\mu}=g\overline{\psi}\gamma_\mu\tau_a\psi, \qquad \text{for } 1\leqslant a\leqslant N^2-1, \tag{4.3.50}$$

$$i\gamma^\mu(\partial_\mu+igG^a_\mu\tau_a)\psi-m\psi=0, \tag{4.3.51}$$

where

$$F^a_{\nu\mu}=\partial_\nu G^a_\mu-\partial_\mu G^a_\nu+g\lambda^a_{bc}G^b_\mu G^c_\nu.$$

The equations (4.3.50)-(4.3.51) are invariant under the $SU(N)$ gauge transformations

$$\widetilde{\psi}=e^{i\theta^a}\tau_a\psi, \qquad \widetilde{G}^a_\mu\tau_a=G^a_\mu e^{i\theta^b\tau_b}\tau_a e^{-i\theta^b\tau_b}-\frac{1}{g}\partial_\mu\theta^a\tau_a. \tag{4.3.52}$$

Hence, if there is a solution for (4.3.50)-(4.3.51), then there are infinitely many solutions. Thus, we have to supplement N^2-1 gauge-fixing equations in order to get a unique physical solution:

$$F_a(G_\mu)=0 \qquad \text{for } 1\leqslant a\leqslant N^2-1. \tag{4.3.53}$$

The reason why take N^2-1 equations in (4.3.53) are that there are N^2-1 free functions θ^a in (4.3.52).

Now, the gauge-fixing problem (4.3.50)-(4.3.51) with (4.3.53) is not well-posed either, because the number of independent equations of (4.3.50) are $4(N^2-1)$ due to

$$\partial^\mu(\overline{\psi}\gamma_\mu\tau_a\psi)\neq 0.$$

Namely, the number of independent equations in the gauge-fixing problem (4.3.50)-(4.3.51) with (4.3.53) is $N_{EQ}=5(N^2-1)+4N$, larger than the number of unknowns $N_{UF}=4(N^2-1)+4N$.

The non well-posedness of $SU(N)$ gauge-fixing problem implies the the PLD is not applicable for the $SU(N)$ gauge field theory. However, based on PID, the $SU(N)$ gauge-fixing problem is well-posed.

4.4 Duality and Decoupling of Interaction Fields

The natural duality of four fundamental interactions to be addressed in this section is a direct consequence of PID. It is with this duality, together with the PRI invariant potentials S_μ and W_μ given by (4.5.1) and (4.6.1), that we establish a clear explanation for many longstanding challenging problems in physics, including for example the dark matter and dark energy phenomena, the formulas of the weak and strong forces, the quark confinement, the asymptotic freedom, and the strong interacting potentials of nucleons. Also, this duality lay a solid foundation for the weakton model of elementary particles and the energy level theory of subatomic particles, and gives rise to a new mechanism for sub-atomic decay and scattering.

The unified field model can be easily decoupled to study each individual interaction when other interactions are negligible. In other words, PID is certainly applicable to each individual interaction. For gravity, for example, PID offers to a new gravitational field model, leading to a unified model for dark energy and dark matter (Ma and Wang, 2014e).

4.4.1 Duality

In the unified field equations (4.3.21)-(4.3.24), there exists a natural duality between the interaction fields $(g_{\mu\nu},A_\mu,W^a_\mu,S^k_\mu)$ and their corresponding dual fields $(\phi^g_\mu,\phi^e,\phi^w_a,\phi^s_k)$:

$$\begin{aligned}
g_{\mu\nu} &\quad\leftrightarrow\quad \phi^g_\mu, \\
A_\mu &\quad\leftrightarrow\quad \phi^e, \\
W^a_\mu &\quad\leftrightarrow\quad \phi^a_w \quad \text{for } 1\leqslant a\leqslant 3, \\
S^k_\mu &\quad\leftrightarrow\quad \phi^k_s \quad \text{for } 1\leqslant k\leqslant 8.
\end{aligned}\tag{4.4.1}$$

Thanks to PRI, the $SU(2)$ gauge fields W^a_μ $(1\leqslant a\leqslant 3)$ and the $SU(3)$ gauge fields S^k_μ $(1\leqslant k\leqslant 8)$ are symmetric in their indices $a=1,2,3$ and $k=1,\cdots,8$ respectively. Therefore, the corresponding relation (4.4.1) can be also considered as the following dual relation

$$\begin{aligned} g_{\mu\nu} &\quad\leftrightarrow\quad \phi^g_\mu, \\ A_\mu &\quad\leftrightarrow\quad \phi^e, \\ \{W^a_\mu\} &\quad\leftrightarrow\quad \{\phi^a_w\}, \\ \{S^k_\mu\} &\quad\leftrightarrow\quad \{\phi^k_s\}. \end{aligned} \tag{4.4.2}$$

The duality relation (4.4.1) can be regarded as the correspondence between field particles for each interaction, and the relation (4.4.2) is the duality of interacting forces. We now address these two different dualities.

1. *Duality of field particles.* In the duality relation (4.4.1), if the tensor fields on the left-hand side are of k-th order, then their dual tensor fields on the right-hand side are of $(k-1)$-th order. Physically, this amounts to saying that if a mediator for an interaction has spin-k, then the dual mediator for the dual field has spin-$(k-1)$. Hence, (4.4.1) leads to the following important physical conclusion:

Duality of Interaction Mediators 4.13 *Each interaction mediator possesses a dual field particle, called the dual mediator, and if the mediator has spin-k, then its dual mediator has spin-$(k-1)$.*

The duality between interaction mediators is a direct consequence of PID used for deriving the unified field equations. Based on this duality, if there exist a graviton with spin $J=2$, then there must exist a dual graviton with spin $J=1$. In fact, for all interaction mediators, we have the following duality correspondence:

$$\begin{aligned} &\text{graviton } (J=2) &&\leftrightarrow\quad \text{dual vector graviton } (J=1), \\ &\text{photon } (J=1) &&\leftrightarrow\quad \text{dual scalar photon } (J=0), \\ &W^\pm \text{ bosons } (J=1) &&\leftrightarrow\quad \text{charged Higgs } H^\pm \ (J=0), \\ &Z \text{ boson } (J=1) &&\leftrightarrow\quad \text{neutral Higgs } H^0 \ (J=0), \\ &\text{gluons } g^k \ (J=1) &&\leftrightarrow\quad \text{dual scalar gluons } \phi^k_g \ (J=0). \end{aligned} \tag{4.4.3}$$

The neutral Higgs H^0 (the adjoint particle of Z) had been discovered experimentally. We remark that the duality (4.4.3) can also be derived using the weakton model (Ma and Wang, 2015b), which is also presented in the next chapter.

2. *Duality of interacting forces.* The correspondence (4.4.2) provides a dual relation between the attracting and repelling forces. In fact, from the interaction potentials we find that the even-spin fields yield attracting forces, and the odd-spin fields yield repelling forces.

Duality of Interaction Forces 4.14 *Each interaction generates both attracting and repelling forces. Moreover, for each pair of dual fields, the even-spin field generates an attracting force, and the odd-spin field generates a repelling force.*

This duality of interaction forces is illustrated as follows:

$$\begin{aligned}
&\text{Gravitation force} && = && \text{attraction due to } g_{\mu\nu} + \text{ repelling due to } \phi^g_\mu, \\
&\text{Electromagnetism} && = && \text{attraction due to } \phi^e + \text{ repelling due to } A_\mu, \\
&\text{Weak force} && = && \text{attraction due to } \phi_w + \text{ repelling due to } W_\mu, \\
&\text{Strong force} && = && \text{attraction due to } \phi_s + \text{ repelling due to } S_\mu.
\end{aligned} \tag{4.4.4}$$

Here, we remark that the electromagnetic force in (4.4.4) is between the charged particles with the same sign, and the force generated by ϕ^e is attractive.

4.4.2 Gravitational field equations derived by PID

If we only consider the gravitational interaction, then the gravitational field equations can be decoupled from the unified field model (4.3.21)-(4.3.25), and are given by

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu} + \left[\nabla_\mu + \frac{e}{\hbar c} A_\mu\right] \Phi_\nu, \tag{4.4.5}$$

where the term $\frac{e}{\hbar c} A_\mu \Phi_\nu$ represents the coupling between the gravitation and the cosmic microwave background (CMB) radiation.

By the Bianchi identity (4.2.2), taking divergence on both sides of (4.4.5) yields

$$\nabla^\mu \nabla_\mu \Phi_\nu + \frac{e}{\hbar c} \nabla^\mu (A_\mu \Phi_\nu) = \frac{8\pi G}{c^4} \nabla^\mu T_{\mu\nu}. \tag{4.4.6}$$

The duality of gravity is based on the field equations (4.4.5) and (4.4.6).

1. *Gravitons and dual gravitons.* It is known that as the equations describing field particles, (4.4.5) characterize the graviton as a massless, neutral bosonic particle with spin $J = 2$, and (4.4.6) indicate that the dual vector graviton is a massless, neutral bosonic particle with $J = 1$. Hence, the gravitational field equations induced by PID and PRI provide a pair of field particles:

$$\begin{aligned}
&\text{tensor graviton:} && J = 2,\ m = 0,\ Q_e = 0, \\
&\text{vector graviton:} && J = 1,\ m = 0,\ Q_e = 0,
\end{aligned} \tag{4.4.7}$$

where Q_e is the electric charge.

It is the nonlinear interaction of these two field particles in (4.4.7) that lead to the dark matter and dark energy phenomena.

2. *Gravitational force.* If we consider the gravitational force only from the Einstein field equations

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu}, \tag{4.4.8}$$

then by the Schwarzschild solution of (4.4.8), we can derive the classical Newton's gravitational force as

$$F = -\frac{mMG}{r^2}, \tag{4.4.9}$$

which is an attracting force generated by $g_{\mu\nu}$.

However, with the field equations (4.4.5), we can deduce a revised formula to (4.4.9). Actually, ignoring the microwave background radiation, the equations (4.4.5) become (Ma and Wang, 2014e):

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} - \nabla_\mu\nabla_\nu\varphi, \tag{4.4.10}$$

where $\Phi_\nu = -\nabla_\nu\varphi$, and φ is a scalar field. In Chapter 7 (see also (Ma and Wang, 2014e)), we are able to derive from (4.4.10) that the gravitational force should be in the form

$$F = mMG\left[-\frac{1}{r^2} + \frac{c^2}{2MG}\Phi r - \left(\frac{c^2}{MG} + \frac{1}{r}\right)\frac{d\varphi}{dr}\right], \tag{4.4.11}$$

where φ is the dual scalar field, representing the scalar potential, and

$$\Phi = g^{\mu\nu}\nabla_\mu\nabla_\nu\varphi. \tag{4.4.12}$$

The first term in the right-hand side of (4.4.11) is the Newton's gravitational force, and the second term (4.4.12) represents the repelling force generated by the dual field φ, and the third term

$$-\left(\frac{c^2}{MG} + \frac{1}{r}\right)\frac{d\varphi}{dr}$$

represents the force due to the nonlinear coupling of $g_{\mu\nu}$ and its dual φ. Formula (4.4.11) can be approximatively written as

$$\begin{aligned} &F = mMG\left(-\frac{1}{r^2} - \frac{k_0}{r} + k_1 r\right), \\ &k_0 = 4\times 10^{-18}\ \text{km}^{-1}, \quad k_1 = 10^{-57}\ \text{km}^{-3}. \end{aligned} \tag{4.4.13}$$

The formula (4.4.13) shows that a central gravitational field with mass M has an attracting force $-k_0/r$ in addition to the Newtonian gravitational force. This explains the dark matter phenomenon. Also there is a repelling force $k_1 r$, which explains the dark energy phenomenon; see (Ma and Wang, 2014e) for details.

4.4.3 Modified QED model

For the electromagnetic interaction only, the decoupled QED field equations from (4.3.22) and (4.3.25) are given by

$$\partial^\nu(\partial_\nu A_\mu - \partial_\mu A_\nu) - eJ_\mu = \left(\partial_\mu + \frac{\beta e}{\hbar c}A_\mu\right)\phi^e, \tag{4.4.14}$$

$$i\gamma^\mu\left(\partial_\mu + i\frac{e}{\hbar c}A_\mu\right)\psi - \frac{mc}{\hbar}\psi = 0, \tag{4.4.15}$$

where β is a dimensionless constant, and $J_\mu = \overline{\psi}\gamma^\mu\psi$ is the current density satisfying

$$\partial^\mu J_\mu = 0. \tag{4.4.16}$$

Equations (4.4.14) and (4.4.15) are the modified QED model. Taking divergence on both sides of (4.4.14), by (4.4.16) and

$$\partial^\mu \partial^\nu (\partial_\nu A_\mu - \partial_\mu A_\nu) = 0,$$

the equations (4.4.14)-(4.4.15) can be equivalently written as

$$\partial^\nu (\partial_\nu A_\mu - \partial_\mu A_\nu) - eJ_\mu = \left(\partial_\mu + \frac{\beta e}{\hbar c} A_\mu\right) \phi^e, \tag{4.4.17}$$

$$\partial^\mu \partial_\mu \phi^e + \frac{\beta e}{\hbar c} \partial^\mu (A_\mu \phi^e) = 0, \tag{4.4.18}$$

$$i\gamma^\mu \left(\partial_\mu + ieA_\mu\right) \psi - \frac{mc}{\hbar} \psi = 0. \tag{4.4.19}$$

If we take

$$H = \text{ curl} \vec{A}, \quad E = -\frac{1}{c}\frac{\partial \vec{A}}{\partial t} - \nabla \varphi, \tag{4.4.20}$$

where $A_\mu = (\varphi, \vec{A})$, $\vec{A} = (A_1, A_2, A_3)$, then (4.4.17)-(4.4.18) and (4.4.20) are a modified version of the classical Maxwell equations, which are expressed as

$$\begin{aligned}
&\frac{1}{c}\frac{\partial H}{\partial t} = - \text{ curl } E, \\
&H = \text{ curl} \vec{A}, \\
&\frac{1}{c}\frac{\partial E}{\partial t} = \text{ curl } H + \vec{J} + \nabla \phi^e + \frac{\beta e}{\hbar c} A_\mu \phi^e, \\
&\text{div } E = \rho + \frac{1}{c}\frac{\partial \phi^e}{\partial t} + \frac{\beta e}{\hbar c} \phi^e \varphi, \\
&\left(\frac{\partial^2}{\partial t^2} - \Delta\right) \phi^e + \frac{\beta e}{\hbar c}\left(\frac{1}{c}\frac{\partial}{\partial t}(\varphi \phi^e) - \text{ div } (\vec{A}\phi^e)\right) = 0,
\end{aligned} \tag{4.4.21}$$

where $\vec{J}$ is the electric current density and ρ is the electric charge density.

The equations (4.4.14)-(4.4.15) or (4.4.17)-(4.4.19) need to be supplemented with a coupled equation to compensate the gauge-symmetry breaking and the induced dual field ϕ^e:

$$F(A_\mu, \phi^e, \psi) = 0. \tag{4.4.22}$$

Remark 4.15 Usually, the compensating equation (4.4.22) is called the gauge-fixing equation. The compensating equation (4.4.22) should be determined based on first principles. However, we don't know whether there are such physical laws. In general, according to different situations physicists take the gauge-fixing equation in the following forms:

$$\begin{aligned}
&\text{Lorentz gauge:} && \partial^\mu A_\mu = 0, \\
&\text{Coulomb gauge:} && \text{div } \vec{A} = 0, \\
&\text{Axial gauge:} && A_3 = 0, \\
&\text{Temporal gauge:} && A_0 = 0 \quad (A_0 = \varphi).
\end{aligned} \tag{4.4.23}$$

The following are the two perspectives of the duality for electromagnetism.

1. *Photon and dual scalar photon.* If we view the field equations (4.4.17) and (4.4.18) as describing the field particles, then we have

$$J_\mu = 0, \quad \beta = 0, \qquad \text{in (4.4.17)-(4.4.18).}$$

Thus, the usual photon equation is given by

$$\Box A_\mu + \partial_\mu(\partial^\nu A_\nu) = 0, \tag{4.4.24}$$

and the dual photon, also called the scalar photon, is described by the following equation

$$\Box \phi^e = 0. \tag{4.4.25}$$

By (4.4.24) and (4.4.25) we deduce the following basic properties for photons and scalar photons:

$$\begin{aligned} &\text{photon:} && J = 1,\ m = 0,\ Q_e = 0, \\ &\text{scalar photon:} && J = 0,\ m = 0,\ Q_e = 0. \end{aligned} \tag{4.4.26}$$

2. *Electromagnetic force.* If we consider the electromagnetic force, then the constant $\beta \neq 0$ in (4.4.17). It is known that in the classical theory, the Coulomb potential φ satisfies the equation

$$\Delta \varphi = -4\pi\rho, \tag{4.4.27}$$

which is the stationary equation of (4.4.17) with $\mu = 0$ and $\Phi^e = 0$, and with the Coulomb gauge in (4.4.23). For the case where $\rho = e\delta(r)$, the solution of (4.4.27) is the well-known Coulomb potential:

$$\varphi = \frac{e}{r}. \tag{4.4.28}$$

For the modified equations (4.4.17) and (4.4.18), the stationary time-component equations with the Coulomb gauge are given by

$$\Delta \varphi - \frac{\beta e}{\hbar c}\phi^e \varphi = 4\pi e \delta(r), \tag{4.4.29}$$

$$\Delta \phi^e = 0. \tag{4.4.30}$$

Only the radially symmetric solutions of (4.4.29) and (4.4.30) are physical. The radial solutions of (4.4.30) are constants

$$\phi^e = \phi_0 \qquad \text{the constants.}$$

Thus the equations (4.4.29) and (4.4.30), in the spherical coordinate system, are reduced as follows

$$\frac{1}{r^2}\frac{d}{dr}(r^2\frac{d}{dr})\varphi - k\varphi = -4\pi e \delta(r), \tag{4.4.31}$$

where $k = \frac{e}{\hbar c}\beta\phi_0$. The solutions of (4.4.31) are expressed as

$$\varphi = \frac{e}{r}e^{-\sqrt{k}r}. \tag{4.4.32}$$

The parameter k is to be determined by experiments.

We discuss the solutions (4.4.32) in the following three cases.

1) *Case* $k = 0$. The solution (4.4.32) in this case is reduced to the Coulomb potential (4.4.28);

2) *Case* $k > 0$. This solution is similar to the Yukawa potential for the strong interactions of nucleons.

3) *Case* $k < 0$. In this case, the solution can be written as

$$\varphi = \frac{e}{r}\cos\sqrt{-k}r + \frac{\alpha e}{r}\sin\sqrt{-k}r, \tag{4.4.33}$$

where α is an arbitrary constant. The function

$$\varphi_0 = \frac{1}{r}\sin\sqrt{-k}r$$

in (4.4.33) satisfies that

$$\Delta\varphi_0 - k\varphi_0 = 0.$$

4.4.4 Strong interaction field equations

The decoupled strong interaction field model from (4.3.21)-(4.3.25) is given by

$$\partial^\nu S^k_{\nu\mu} - \frac{g_s}{\hbar c}f^k_{ij}g^{\alpha\beta}S^i_{\alpha\mu}S^j_\beta - g_s Q^k_\mu = \left[\partial_\mu + \frac{g_s}{\hbar c}\delta^s_l S^l_\mu - \frac{1}{4}\left(\frac{m_\pi c}{\hbar}\right)^2 x_\mu\right]\phi^k_s, \tag{4.4.34}$$

$$i\gamma^\mu\left[\partial_\mu + i\frac{g_s}{\hbar c}S^b_\mu\tau_b\right]\psi - \frac{mc}{\hbar}\psi = 0, \tag{4.4.35}$$

for $1 \leqslant k \leqslant 8$, where $\tau_k = \tau^k$ are the Gell-Mann matrices as in (3.5.38), and

$$\begin{aligned} S^k_{\mu\nu} &= \partial_\mu S^k_\nu - \partial_\nu S^k_\mu + \frac{g_s}{\hbar c}S^i_\mu S^j_\nu, \\ Q^k_\mu &= \overline{\psi}\gamma^\mu\tau^k\psi. \end{aligned} \tag{4.4.36}$$

Taking divergence on both sides of (4.4.34) and by

$$\partial^\mu\partial^\nu S^k_{\mu\nu} = 0 \qquad \text{for } 1 \leqslant k \leqslant 8,$$

we deduce the following dual field equations for the strong interaction:

$$\partial^\mu\partial_\mu\phi^k_s + \partial^\mu\left[\left(\frac{g_s}{\hbar c}\delta^k_l S^l_\mu - \frac{1}{4}\frac{m^2_\pi c^2}{\hbar^2}x_\mu\right)\phi^k_s\right] = -g_s\partial^\mu Q^k_\mu - \frac{g_s}{\hbar c}f^k_{ij}g^{\alpha\beta}\partial^\mu(S^i_{\alpha\mu}S^j_\beta). \tag{4.4.37}$$

The equations (4.4.34)-(4.4.35) also need 8 additional gauge equations to compensate the induced dual fields ϕ_s^k:

$$F_s^k(S_\mu, \phi_s, \psi) = 0, \quad 1 \leqslant k \leqslant 8. \tag{4.4.38}$$

We have the following duality for the strong interaction.

1. *Gluons and dual scalar gluons.* Based on quantum chromodynamics (QCD), the field particles for the strong interaction are the eight massless gluons with spin $J = 1$, which are described by the $SU(3)$ gauge fields S_μ^k $(1 \leqslant k \leqslant 8)$. By the duality (4.4.1), for the strong interactions we have the field particle correspondence

$$S_\mu^k \leftrightarrow \phi_s^k \qquad \text{for } 1 \leqslant k \leqslant 8.$$

It implies that corresponding to the 8 gluons S_μ^k $(1 \leqslant k \leqslant 8)$ there should be 8 dual gluons represented by ϕ_s^k, called the scalar gluons due to ϕ_s^k being scalar fields. Namely we have the following gluon correspondence

$$\text{gluons } g_k \leftrightarrow \text{ scalar gluons } g_0^k \qquad \text{for } 1 \leqslant k \leqslant 8.$$

Gluons and scalar gluons are described by equations (4.4.34) and (4.4.37) respectively, which are nonlinear. In fact, g_k and g_0^k are confined in hadrons.

2. *Strong force.* The strong interaction forces are governed by the field equations (4.3.31)-(4.3.35). The decoupled field equations are given by

$$\partial^\nu S_{\nu\mu}^k - \frac{g_s}{\hbar c} f_{ij}^k g^{\alpha\beta} S_{\alpha\mu}^i S_\beta^j - g_s Q_\mu^k = \left[\partial_\mu - \frac{1}{4} k_s^2 x_\mu + \frac{g_s \delta}{\hbar c} S_\mu\right] \phi_s^k, \tag{4.4.39}$$

$$\begin{aligned} \partial^\mu \partial_\mu \phi_s^k - k^2 \phi_s^k + \frac{1}{4} k_s^2 x_\mu \partial^\mu \phi_s^k + \frac{g_s \delta}{\hbar c} \partial^\mu (S_\mu \phi_s^k) & \\ = -g_s \partial^\mu Q_\mu^k - \frac{g_s}{\hbar c} f_{ij}^k g^{\alpha\beta} \partial^\mu (S_{\alpha\mu}^i S_\beta^j), & \end{aligned} \tag{4.4.40}$$

$$i\gamma^\mu \left[\partial_\mu + i \frac{g_s}{\hbar c} S_\mu^l \tau_l\right] \psi - \frac{mc}{\hbar} \psi = 0, \tag{4.4.41}$$

for $1 \leqslant k \leqslant 8$, where δ is a parameter, and S_μ is as in (4.3.41).

In the next section, we shall deduce the layeredformulas of strong interaction potentials from the equations (4.4.39)-(4.4.41) with gauge equations (4.4.38).

Remark 4.16 We need to explain the physical significance of the parameters k_s and δ. Usually, k_s and δ are regarded as masses of the field particles. However, since (4.4.39)-(4.4.41) are the field equations for the interaction forces, the parameters k_s and δ are no longer viewed as masses. In fact, k^{-1} represents the range of attracting force for the strong interaction, and $\left(\frac{g_s \phi_s^0}{\hbar c} \delta\right)^{-1}$ is the range of the repelling force, where ϕ_s^0 is a ground state of ϕ_s. □

4.4.5 Weak interaction field equations

The unified field model (4.3.21)-(4.3.25) can be decoupled to study the weak interaction only, leading to the following weak interaction field equations:

$$\partial^\nu W^a_{\nu\mu} - \frac{g_w}{\hbar c}\varepsilon^a_{bc}g^{\alpha\beta}W^b_{\alpha\mu}W^c_\beta - g_w J^a_\mu = \left[\partial_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu + \frac{g_w}{\hbar c}\gamma^w_b W^b_\mu\right]\phi^a_w, \tag{4.4.42}$$

$$i\gamma^\mu\left[\partial_\mu + i\frac{g_w}{\hbar c}W^a_\mu\sigma_a\right]\psi - \frac{mc}{\hbar}\psi = 0, \tag{4.4.43}$$

where m_H represents the mass of the Higgs particle, $\sigma_a = \sigma^a$ $(1 \leqslant a \leqslant 3)$ are the Pauli matrices as in (3.5.36), and

$$\begin{aligned} W^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + \frac{g_w}{\hbar c}\varepsilon^a_{bc}W^b_\mu W^c_\nu, \\ J^a_\mu &= \overline{\psi}\gamma_\mu\sigma^a\psi, \qquad \gamma_\mu = g_{\mu\nu}\gamma^\nu. \end{aligned} \tag{4.4.44}$$

Taking divergence on both sides of (4.4.42) we get

$$\begin{aligned} &\partial^\mu\partial_\mu\phi^a_w - \left(\frac{m_H c}{\hbar}\right)^2\phi^a_w + \frac{g_w}{\hbar c}\gamma^w_b\partial^\mu(W^b_\mu\phi^a_w) - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu\partial^\mu\phi^a_w \\ &= -\frac{g_w}{\hbar c}\varepsilon^a_{bc}g^{\alpha\beta}\partial^\mu(W^b_{\alpha\mu}W^c_\beta) - g_w\partial^\mu J^a_\mu. \end{aligned} \tag{4.4.45}$$

Also, we need to supplement (4.4.42)-(4.4.43) with three additional 3 gauge equations to compensate the induced dual fields ϕ^a_w:

$$F^a_w(W_\mu, \phi_w, \psi) = 0 \qquad \text{for } 1 \leqslant a \leqslant 3. \tag{4.4.46}$$

1. *Duality between $W^\pm, Z$ Bosons and Higgs Bosons $H^\pm, H^0$.* The three massive vector bosons, denoted by $W^\pm, Z^0$, has been discovered experimentally. The field equations (4.4.42) give rise to a natural duality:

$$Z^0 \leftrightarrow H^0, \quad W^\pm \leftrightarrow H^\pm, \tag{4.4.47}$$

where $H^0, H^\pm$ are three dual scalar bosons, called the Higgs particles. The neutral Higgs H^0, discovered by LHC in 2012, and the charged Higgs $H^\pm$, to be discovered experimentally.

In Section 4.6, we shall introduce the dual bosons (4.4.47) by applying the field equations (4.4.42)-(4.4.45) with gauge equations (4.4.46).

2. *Weak force.* If we consider the weak interaction force, we have to use the equations decoupled from (4.3.31)-(4.3.35):

$$\partial^\nu W^a_{\nu\mu} - \frac{g_w}{\hbar c}\varepsilon^a_{bc}g^{\alpha\beta}W^b_{\alpha\mu}W^c_\beta - g_w J^a_\mu = \left[\partial_\mu - \frac{1}{4}k^2_w x_\mu + \frac{g_w}{\hbar c}\gamma W_\mu\right]\phi^a_w, \tag{4.4.48}$$

$$\partial^\mu\partial_\mu\phi^a_w - k^2\phi^2_w + \frac{g_w}{\hbar c}\gamma\partial^\mu(W_\mu\phi^a_w) - \frac{1}{4}k^2 x_\mu\partial^\mu\phi^a_w \tag{4.4.49}$$

$$
\begin{aligned}
&= -g_w \partial^\mu J^a_\mu - \frac{g_w}{\hbar c}\varepsilon^a_{bc} g^{\alpha\beta}\partial^\mu (W^b_{\alpha\mu} W^c_\beta),\\
&i\gamma^\mu \left(\partial_\mu + i\frac{g_w}{\hbar c} W^a_\mu \sigma_a\right)\psi - \frac{mc}{\hbar}\psi = 0,
\end{aligned}
\tag{4.4.50}
$$

where γ, k_w are constants, W_μ is as in (4.3.41).

In Section 4.6, we shall deduce the layered formulas of weak interaction potentials by applying the equations (4.4.48)-(4.4.50).

Remark 4.17 The duality of four fundamental interactions is very important, and is a direct consequence of PID. With this duality, and with the PRI invariant potentials W_μ and S_μ given by (4.3.41), we obtain explanations for a number of physical problems such as the dark matter and dark energy phenomena, the weak and strong forces and potential formulas, the quark confinement, the asymptotic freedom, the strong potentials of nucleons etc. Also this study leads to the needed foundation for the weakton model of elementary particles and the energy level theory of subatomic particles, and gives rise to a mechanism for subatomic decays and scatterings.

4.5 Strong Interaction Potentials

4.5.1 Strong interaction potential of elementary particles

Thanks to PRI, the strong interaction potential takes the linear combination of the eight $SU(3)$ gauge potentials as follows

$$S_\mu = \rho_k S^k_\mu, \tag{4.5.1}$$

where $\rho_k = (\rho_1, \cdots, \rho_8)$ is the $SU(3)$ tensor as given in (4.3.29).

Let g_s be the strong charge of an elementary particle, i.e. the w^* weakton introduced in Chapter 5, and

$$\Phi_0 = S_0 \qquad \text{the time-component of (4.5.1)}$$

be the strong charge potential of this particle. Then the strong force between two elementary particles carrying strong charges is

$$F = -g_s \nabla \Phi_0.$$

However, the strong interactions are layered, i.e. the strong forces only act on the same level of particles, such as the quark level, the hadron level. Hence, the strong interaction potentials are also layered. In fact, in the next subsection we shall show that for a particle with N strong charges g_s of the elementary particles, its strong interaction potential is given by

$$
\begin{aligned}
&\Phi_s = N g_s(\rho)\left[\frac{1}{r} - \frac{A}{\rho}(1+kr)e^{-kr}\right],\\
&g_s(\rho) = \left(\frac{\rho_w}{\rho}\right)^3 g_s,
\end{aligned}
\tag{4.5.2}
$$

where ρ_w is the radius of the elementary particle (i.e. the w^* weakton), ρ is the particle radius, $k > 0$ is a constant with k^{-1} being the strong interaction attraction radius of this particle, and A is the strong interaction constant, which depends on the type of particles. Thus, the strong force between such two particles is

$$F = -Ng_s(\rho)\nabla\Phi_s, \tag{4.5.3}$$

where $g_s(\rho)$ and Φ_s are as in (4.5.2).

In particular for the w^*-weakton, which possesses one strong charge g_s, the formula (4.5.2) becomes

$$\Phi_0 = g_s\left[\frac{1}{r} - \frac{A_0}{\rho_w}(1+k_0r)e^{-k_0r}\right], \tag{4.5.4}$$

where $1/k_0$ is the attraction radius of the strong interaction for the elementary particles, i.e. the w^*-weakton. According to the physical observation, we take the quantitative order

$$k_0 = 10^{18}\ \text{cm}^{-1}. \tag{4.5.5}$$

In this subsection, we shall deduce formula (4.5.4) for the w^*-weaktons from the strong interaction field equations (4.4.39)-(4.4.41). Taking inner product of the field equation (4.4.39) and (4.4.40) with $\rho_k = (\rho_1, \cdots, \rho_8)$, we derive that

$$\partial^\nu S_{\nu\mu} - \frac{g_s}{\hbar c}\lambda_{ij}g^{\alpha\beta}S^i_{\alpha\mu}S^j_\beta - g_sQ_\mu = \left[\partial_\mu - \frac{1}{4}k_0^2x_\mu + \frac{g_s\delta}{\hbar c}S_\mu\right]\phi_s, \tag{4.5.6}$$

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta\right)\phi_s + k_0^2\phi_s + \frac{1}{4}k_0^2x_\mu\partial^\mu\phi_s = g_s\partial^\mu Q_\mu + \frac{g_s}{\hbar c}\partial^\mu(\lambda_{ij}g^{\alpha\beta}S^i_{\alpha\mu}S^j_\beta - \delta S_\mu\phi_s), \tag{4.5.7}$$

where

$$\phi_s = \rho_k\phi_s^k, \quad Q_\mu = \rho_kQ^k_\mu, \qquad \lambda_{ij} = \rho_k\lambda^k_{ij}, \qquad S_{\mu\nu} = \partial_\mu S_\nu - \partial_\nu S_\mu + \frac{g_s}{\hbar c}\lambda_{ij}S^i_\mu S^j_\nu.$$

Based on the superposition property of the strong potential for strong charges, $\Phi_s = S_0$ and ϕ_s obey a linear relationship. Namely, the time-component $\mu = 0$ equation of (4.5.6) and the equation (4.5.7) should be linear. In other words, we have to take eight gauge fixing equations as in (4.4.38) such that they contain the following two equations:

$$\begin{aligned} &\lambda_{ij}\left[\partial^\nu(S^i_\nu S^j_0) - g^{\alpha\beta}S^i_{\alpha o}S^j_\beta\right] + \delta S_0\phi_s = 0, \\ &\partial^\mu\left[\lambda_{ij}g^{\alpha\beta}S^i_{\alpha\beta}S^j_\beta - \delta S_\mu\phi_s\right] = 0, \end{aligned} \tag{4.5.8}$$

Also, in the eight supplement equations we take

$$x_\mu\partial^\mu\phi_s = 0, \qquad \partial^\mu S_\mu = 0, \tag{4.5.9}$$

together with the following static assumption:

$$\frac{\partial S_0}{\partial t} = 0, \qquad \frac{\partial \phi_s}{\partial t} = 0. \tag{4.5.10}$$

With the gauge fixing equations (4.5.8)-(4.5.9) and the static assumption (4.5.10), we derive from (4.5.6) and (4.5.7) that

$$-\Delta\Phi_s = g_s Q - \frac{1}{4}k_0^2 c\tau\phi_s, \tag{4.5.11}$$

$$-\Delta\phi_s + k_0^2\phi_s = g_s\partial^\mu Q_\mu, \tag{4.5.12}$$

where $c\tau$ is the wave length of $\phi_s, Q = -Q_0$.

In the following, we deduce the solution Φ_s and ϕ_s of (4.5.11)-(4.5.12) in a few steps.

Step 1. Solution of (4.5.12). By definition of Q_μ, we have

$$\partial^\mu Q_\mu = \rho_k \partial_\mu \overline{\Psi}\gamma^\mu \tau_k \Psi + \rho_k \overline{\Psi}\gamma^\mu \tau_k \partial_\mu \Psi.$$

In view of the Dirac equation (4.4.41),

$$\partial_\mu \overline{\Psi}\gamma^\mu \tau_k \Psi = i\frac{g_s}{\hbar c} S_\mu^j \overline{\Psi}\gamma^\mu \tau_j \tau_k \Psi + i\frac{mc}{\hbar}\overline{\Psi}\tau_k \Psi,$$

$$\overline{\Psi}\gamma^\mu \tau_k \partial_\mu \Psi = -i\frac{g_s}{\hbar c} S_\mu^j \overline{\Psi}\gamma^\mu \tau_k \tau_j \Psi - i\frac{mc}{\hbar}\overline{\Psi}\tau_k \Psi.$$

Hence we arrive at

$$\partial^\mu Q_\mu = \frac{ig_s}{\hbar c}\rho_k S_\mu^j \overline{\Psi}\gamma^\mu [\tau_j, \tau_k]\Psi = -\frac{2g_s}{\hbar c}\rho_k S_\mu^j \lambda_{jk}^i Q_i^\mu \tag{4.5.13}$$

Since $Q_i^\mu = Q_\mu^i$ is the current density, we have

$$\rho_k \lambda_{jk}^i Q_i^\mu = \theta_j^\mu \delta(r),$$

where $\delta(r)$ is the Dirac delta function, θ_j^μ is a constant tensor, inversely proportional to the volume of the particle. Hence

$$\rho_k S_\mu^j \lambda_{jk}^i \theta_i^\mu = \overline{S}_\mu^j \theta_j^\mu \delta(r),$$

where $\overline{S}_\mu^j \sim \overline{S}_\mu^j(0)$ takes the following average value

$$\overline{S}_\mu^j = \frac{1}{|B_{\rho_w}|}\int_{B_{\rho_w}} S_\mu^j dv. \tag{4.5.14}$$

Here ρ_w is the radius of a w^*-weakton. Later, we shall see that

$$S_\mu^j \sim \frac{1}{r} \quad \text{as } r \to 0.$$

Hence we deduce from (4.5.14) that

$$\overline{S}^j_\mu = \xi^j_\mu \rho_w^{-1} \quad (\xi^j_\mu \text{ is a constant tensor}).$$

Thus, (4.5.13) becomes

$$\partial^\mu Q_\mu = -\frac{\kappa}{\rho_w}\delta(r) \quad (\rho_w \text{ is the radius of a } w^*\text{-weakton}), \tag{4.5.15}$$

where κ is a parameter given by

$$\kappa = \frac{2g_s}{\hbar c}\xi^j_\mu \theta^\mu_j, \tag{4.5.16}$$

and κ is inversely proportional to the volume of w^*-weakton.

Therefore, equation (4.5.12) is rewritten as

$$-\Delta\phi_s + k_0^2\phi_s = -\frac{g_s\kappa}{\rho_w}\delta(r), \tag{4.5.17}$$

whose solution is given by

$$\phi_s = -\frac{g_s\kappa}{\rho_w}\frac{1}{r}e^{-k_0 r}. \tag{4.5.18}$$

Step 2. Solution of (4.5.11). The quantity $g_s Q = -g_s Q_0$ is the strong charge density of a w^*-weakton, and without loss of generality, we assume that

$$Q = \beta\delta(r), \tag{4.5.19}$$

and $\beta > 0$ is a constant, inversely proportional to the volume of the w^*-weakton. Hence (4.5.11) can be rewritten as

$$-\Delta\Phi_s = g_s\beta\delta(r) + \frac{g_s A}{\rho_w}\frac{1}{r}e^{-k_0 r}, \tag{4.5.20}$$

where A is a constant given by

$$A = \frac{k_0^2 c\tau\kappa}{4} \text{ with physical dimension } \frac{1}{L}.$$

Assume $\Phi_s = \Phi_s(r)$ is radially symmetric, then (4.5.20) becomes

$$-\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)\Phi_s = g_s\beta\delta(r) + \frac{g_s A}{\rho_w}\frac{1}{r}e^{-k_0 r},$$

whose solution takes the form

$$\Phi_s = g_s\left[\frac{\beta}{r} - \frac{A}{\rho_w}\varphi(r)e^{-k_0 r}\right], \tag{4.5.21}$$

where φ solves

$$\varphi'' + 2\left(\frac{1}{r} - k_0\right)\varphi' - \left(\frac{2k_0}{r} - k_0^2\right)\varphi = \frac{1}{r}. \tag{4.5.22}$$

Assume that the solution φ of (4.5.22) is given by

$$\varphi = \sum_{k=0}^{\infty} \alpha_k r^k. \tag{4.5.23}$$

Inserting φ in (4.5.22) and comparing the coefficients of r^k, then we obtain the following relations

$$\begin{aligned}
\alpha_1 &= k_0\alpha_0 + \frac{1}{2}, \\
\alpha_2 &= \frac{1}{2}k_0^2\alpha_0 + \frac{1}{3}k_0, \\
&\vdots \\
\alpha_k &= \frac{2k_0}{k+1}\alpha_{k-1} - k_0^2\alpha_{k-2}, \quad \forall k \geqslant 2.
\end{aligned} \tag{4.5.24}$$

where α_0 is a free parameter with dimension L. Hence

$$\varphi(r) = \alpha_0(1 + k_0 r + \frac{r}{2\alpha_0} + o(r)),$$

and often it is sufficient to take a first-order approximation.

Step 3. Strong interaction potential of w^*-weakton. The formula (4.5.21) with (4.5.23)-(4.5.24) provides an accurate strong interaction potential for the w^*-weaktons:

$$\Phi_0 = g_s\beta\left[\frac{1}{r} - \frac{A_0}{\rho_w}\widetilde{\varphi}(r)e^{-k_0 r}\right], \tag{4.5.25}$$

where $A_0 = A\alpha_0/\beta$ is a dimensionless parameter, and $\widetilde{\varphi}(r)$ is

$$\widetilde{\varphi} = 1 + k_0 r + \frac{r}{2\alpha_0} + o(r),$$

which is a dimensionless function, and

$$\beta \text{ is inversely proportional to the particle volume.} \tag{4.5.26}$$

Since $g_s\beta$ is to be determined for the elementary particle, we can take $g_s\beta$ as the strong charge of the w^*-weakton, still denoted by g_s. In addition, it is sufficient to take approximatively $\widetilde{\varphi} = 1 + k_0 r$. Then the formula (4.5.25) is rewritten as

$$\Phi_0 = g_s\left[\frac{1}{r} - \frac{A_0}{\rho_w}(1 + k_0 r)e^{-k_0 r}\right].$$

It is the strong interaction potential given in (4.5.4).

4.5.2 Layered formulas of strong interaction potentials

Different from gravity and electromagnetic force, strong interaction is of short-ranged with different strengths in different levels. For example, in the quark level, strong interaction confines quarks inside hadrons, in the nucleon level, strong interaction bounds nucleons inside atoms, and in the atom and molecule level, strong interaction almost diminishes. This layered phenomena can be well-explained using the unified field theory based on PID and PRI. We derive in this subsection strong interaction potentials in different levels.

Without loss of generation, we shall discuss strong interaction nucleon potential. For strong interaction of nucleons, we still use the $SU(3)$ gauge action

$$\mathscr{L} = -\frac{1}{4}S^k_{\mu\nu}S^{\mu\nu k} + \hbar c\overline{q}\left(i\gamma^\mu D_\mu - \frac{mc}{\hbar}\right)q, \tag{4.5.27}$$

where S^k_μ are the strong interaction gauge fields,

$$q = (q_1, q_2, q_3) \tag{4.5.28}$$

where q_1, q_2, q_3 are the wave functions of the three quarks constituting a nucleon, and

$$D_\mu q = \left(\hbar c\partial_\mu + \frac{ig_s}{\hbar c}S^k_\mu \tau_k\right)q. \tag{4.5.29}$$

The action $\mathscr{L}$ defined by (4.5.27) is $SU(3)$ gauge invariant. Physically this means that the contribution of each quark to the strong interaction potential energy is indistinguishable. With PID, the corresponding field equations for the action (4.5.27) are

$$\partial^\nu S^k_{\nu\mu} + \frac{g_s}{\hbar c}\lambda^k_{ij}g^{\alpha\beta}S^i_{\alpha\mu}S^j_\beta - g_sQ^k_\mu = \left(\partial_\mu - \frac{k_1^2}{4}x_\mu\right)\phi^k_n, \tag{4.5.30}$$

$$i\gamma^\mu\left(\partial_\mu + i\frac{g_s}{\hbar c}S^k_\mu\tau_k\right)q - \frac{mc}{\hbar}q = 0, \tag{4.5.31}$$

where the parameter k_1 is defined by

$$r_1 = \frac{1}{k_1} = 10^{-13}\text{ cm}, \tag{4.5.32}$$

which is the strong attraction radius of nucleons, and

$$Q^k_\mu = \overline{q}\gamma_\mu\tau^k q \quad (\tau^k = \tau_k).$$

Let the strong interaction nucleon potential Φ^k_n and its dual potential ϕ_n be defined by

$$\Phi_n = \rho_k S^k_0, \quad \phi_n = \rho_k\phi^k_n.$$

In the same spirit as for deriving the weakton potential equations (4.5.11) and (4.5.12), for Φ_n and ϕ_n we deduce that

$$\begin{aligned} &-\Delta\Phi_n = g_sQ_n - \frac{1}{4}k_1^2c\tau_1\phi_n, \\ &-\Delta\phi_n + k_1^2\phi_n = g_s\partial^\mu Q_\mu, \end{aligned} \tag{4.5.33}$$

where $c\tau_1$ is the wave length of ϕ_n, and

$$Q_\mu = \rho_k Q_\mu^k = (Q_0, Q_1, Q_2, Q_3)$$

represents the quark current density inside a nucleon. Similar to (4.5.15) and (4.5.19), for Q_n and $\partial^\mu Q_\mu$ we have

$$\partial^\mu Q_\mu = -\frac{\kappa_n}{\rho_n}\delta(r), \quad Q_n = \beta_n \delta(r), \tag{4.5.34}$$

where ρ_n is the radius of a nucleon, β_n and κ_n are constants, inversely proportional to the volume of nucleons. Hence, in the same fashion as in deducing (4.5.25), from (4.5.33) and (4.5.34) we derive the following strong nucleon potential as

$$\Phi_n = \beta_n g_s \left[\frac{1}{r} - \frac{A_n}{\rho_n}\varphi(r)e^{-k_1 r}\right], \tag{4.5.35}$$

where $\varphi(r)$ is as

$$\varphi(r) = 1 + k_1 r + \frac{r}{2\alpha_0} + o(r), \qquad \alpha_0 \text{ as in (4.5.24)}, \tag{4.5.36}$$

and A_n is a dimensionless parameter:

$$A_n = \frac{\kappa_n k_1^2 c\tau_1}{4\beta_n}.$$

Note that β_n is inversely proportional to the volume V_n of a nucleon. Hence we have

$$\frac{\beta_n}{\beta} = \frac{NV_0}{V_n} = N\left(\frac{\rho_w}{\rho_n}\right)^3 \quad (N=3), \tag{4.5.37}$$

where N is the number of strong charges in a nucleon, β is the parameter as in (4.5.25) and (4.5.26), and V_0 the volume of a w^*-weakton. Since $g_s\beta$ in (4.5.25) is taken as g_s, by (4.5.37) the formula (4.5.35) is expressed as

$$\begin{aligned} &\Phi_n = g_s(\rho_n)\left[\frac{1}{r} - \frac{A_n}{\rho_n}\varphi(r)e^{-k_n r}\right], \\ &g_s(\rho_n) = 3\left(\frac{\rho_w}{\rho_n}\right)^3 g_s, \end{aligned} \tag{4.5.38}$$

and $\varphi(r)$ is as in (4.5.36).

In summary, for a particle with N strong charges and radius ρ, its strong interaction potential can be written as

$$\begin{aligned} &\Phi = g_s(\rho)\left[\frac{1}{r} - \frac{A}{\rho}\varphi(r)e^{-kr}\right], \\ &g_s(\rho) = N\left(\frac{\rho_w}{\rho}\right)^3 g_s, \end{aligned} \tag{4.5.39}$$

where ρ_w is the radius of w^*-weakton, A is a dimensionless constant depending on the particle, $\varphi(r) = 1 + kr$, and $1/k$ is the radius of strong attraction of particles. Phenomenologically, we take

$$\frac{1}{k} = \begin{cases} 10^{-18}\text{ cm} & \text{for } w^*\text{-weaktons},\\ 10^{-16}\text{ cm} & \text{for quarks},\\ 10^{-13}\text{ cm} & \text{for neutreons},\\ 10^{-7}\text{ cm} & \text{for atom/molecule}. \end{cases} \tag{4.5.40}$$

More specifically, we give the following layered formulas of strong interaction potentials for various level of particles as the w^*-weakton potential Φ_0, the quark potential Φ_q, the nucleon/hadron potential Φ_n and the atom/molecule potential Φ_a are given as follows (Ma and Wang, 2014c):

$$\begin{aligned} \Phi_0 &= g_s \left[\frac{1}{r} - \frac{A_0}{\rho_w}(1+k_0 r)e^{-k_0 r}\right],\\ \Phi_q &= \left(\frac{\rho_w}{\rho_q}\right)^3 g_s \left[\frac{1}{r} - \frac{A_q}{\rho_q}(1+k_1 r)e^{-k_1 r}\right],\\ \Phi_n &= 3\left(\frac{\rho_w}{\rho_n}\right)^3 g_s \left[\frac{1}{r} - \frac{A_n}{\rho_n}(1+k_n r)e^{-k_n r}\right],\\ \Phi_a &= N\left(\frac{\rho_w}{\rho_a}\right)^3 g_s \left[\frac{1}{r} - \frac{A_a}{\rho_a}(1+k_a r)e^{-k_a r}\right]. \end{aligned} \tag{4.5.41}$$

Here, k_0, k_1, k_n, k_a are taken as in (4.5.40):

$$\begin{aligned} &\frac{1}{k_0} = 10^{-18}\text{ cm}, && \frac{1}{k_1} = 10^{-16}\text{ cm},\\ &\frac{1}{k_n} = 10^{-13}\text{ cm}, && \frac{1}{k_a} = 10^{-10} \sim 10^{-7}\text{ cm}. \end{aligned} \tag{4.5.42}$$

Remark 4.18 For two particles with N_1, N_2 strong charges and radii ρ_1, ρ_2, their strong potential energy is

$$V = g_s(\rho_1) g_s(\rho_2)\left[\frac{1}{r} - \frac{B}{\overline{\rho}}(1+kr)e^{-kr}\right], \tag{4.5.43}$$

where B and k depend on the two types of particles, and $\overline{\rho}$ depends on ρ_1 and ρ_2 with $\overline{\rho} = \rho$ if $\rho = \rho_1 = \rho_2$. The strong force between the two particles is

$$F = -\nabla V, \quad V \text{ as in (4.5.43)}. \tag{4.5.44}$$

4.5.3 Quark confinement

Quark model was confirmed by lots of experiments. However, no any single quark is found ever. This fact suggests that the quarks were permanently bound inside a hadron, which is called the quark confinement. Up to now, no other theories can successfully describe the quark confinement phenomena. The direct reason is that all current theories for interactions fail to provide a successful strong interaction potential to explain the various level strong interactions.

Now, we can derive the quark confinement by using the layered strong interaction potentials (4.5.41).

The strong interaction bound energy E for two particles is given by (4.5.43). In particular, for two same particles we have

$$\begin{aligned} E &= g_s^2(\rho)\left[\frac{1}{r} - \frac{A}{\rho}(1+kr)e^{-kr}\right], \\ g_s(\rho) &= \left(\frac{\rho_w}{\rho}\right)^3 g_s. \end{aligned} \tag{4.5.45}$$

The quark confinement can be well explained from the viewpoint of the strong quark bound energy E_q and the nucleon bound energy E_n. In fact, by (4.5.45) we have

$$\frac{E_q}{E_n} \simeq \frac{A_q}{A_n}\left(\frac{\rho_n}{\rho_q}\right)^7. \tag{4.5.46}$$

According to the physical observation fact,

$$\rho_n \simeq 10^{-16}\text{ cm}, \quad \rho_q \simeq 10^{-19}\text{ cm}. \tag{4.5.47}$$

Then, by (4.5.46) we derive that

$$\frac{E_q}{E_n} \simeq 10^{21}\frac{A_q}{A_n}.$$

Physically A_q/A_n is no small, and we assume that $A_q/A_n \simeq 10^{-1}$. Then we have

$$E_q = 10^{20}E_n. \tag{4.5.48}$$

It is known that the bound energy of nucleons is about

$$E_n \sim 10^{-2}GeV.$$

Thus, we obtain by (4.5.48) that

$$E_q \sim 10^{18}GeV,$$

which is at the Planck level. This clearly shows that the quarks is confined in hadrons, and no free quarks can be found.

Remark 4.19 The magnitude order (4.5.42) and (4.5.47) are not accurate, but they are enough precisely to explain the important physical problems as the quark confinement and asymptotic freedom etc.

4.5.4 Asymptotic freedom

The strong interaction potentials provide also a natural explanation for the asymptotic freedom phenomena. To this end, we need to introduce the asymptotic freedom in two persptives: the deep inelastic scattering experiments, and the QCD theory for the coupling constant of quark potentials.

1. *Deep inelastic scattering experiments.* In physical experiments, the interior of a proton is probed by using the accelerating electrons to hit this proton. The collision is called the elastic scattering if there is no momentum exchange as in the *e*-*p* elastic scattering:

$$e^- + p \to e^- + p. \tag{4.5.49}$$

The collision is an inelastic scattering if the particles are changed after the collision. For example the usual *e*-*p* inelastic scatterings are as follows

$$\begin{aligned} &e^- + p \to e^- + \pi^+ + n, \\ &e^- + p \to e^- + \pi^0 + p. \end{aligned} \tag{4.5.50}$$

In 1967, three physicists J. L. Friedman, H. W. Kendall and R. E. Taylor performed a series of deep inelastic experiments, which not only provided sufficient evidence for the existence of quarks, but also exhibited the asymptotic freedom phenomena. Due to their pioneering investigations concerning deep inelastic scattering, the three physicists were awarded the Nobel Prize in 1990.

In the *e*-*p* scattering experiments, if an electron at lower energy collides with a proton, then the proton looks as a point particle, and the collision is an elastic scattering. However, if the accelerating electron is at higher energy, this electron will hit deeply into the interior of a proton, and collides with a quark in the proton, as shown in Figure 4.1. The experiments show that the the collided quark acts as if it is a free particle.

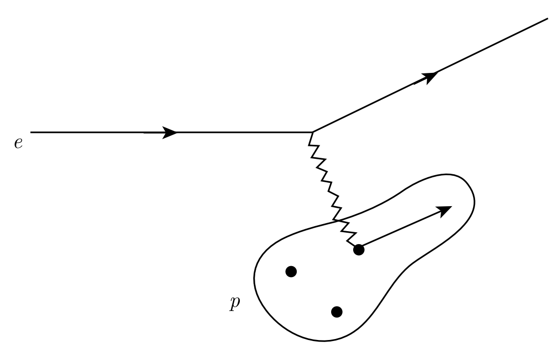


Figure 4.1 Black dots represent three quarks in a proton

2. *QCD theory for asymptotic freedom.* The notion of asymptotic freedom was first introduced by (Gross and Wilczek, 1973; Politzer, 1973), who were awarded the Nobel Prize in 2004. By using the renormalization group, they derived the strong interaction coupling constant of quarks as follows

$$\alpha_s = \frac{4\pi}{\left(11 - \frac{2}{3}n_f\right)\ln(q^2/\lambda^2)}, \tag{4.5.51}$$

where λ is a parameter, $n_f = 6$ is the flavor number of quarks, and q^2 is the transfer momentum by the incidence electron.

In QCD, the coupling constant α_s stands for the strength of the strong interactions between the quarks in a proton. It is clear that as the transfer momentum q^2 tends to infinitely large, α_s tends to zero:

$$\alpha_s \to 0 \quad \text{as} \quad q^2 \to \infty. \tag{4.5.52}$$

It means that the greater the incidence electron kinetics is, the more freedom the quarks exhibit. The asymptotic freedom is based on the fact (4.5.51)-(4.5.52).

3. *Strong interaction potential for asymptotic freedom.* The strong interaction potentials provide also a natural explanation for the asymptotic freedom phenomena. By (4.5.41) and (4.5.44), the strong force between quarks is as

$$F = -g_s(\rho_q)\frac{d\Phi_q}{dr} = g_s^2(\rho_q)\left[\frac{1}{r^2} - \frac{A_q}{\rho_q}k_1^2 r e^{-k_1 r}\right],$$

which implies that there are two radii $\overline{r}_1$ and $\overline{r}_2$ such that

$$\rho_q < \overline{r}_1 < r_1 < \overline{r}_2 \quad (r_1 = 1/k_1),$$

and the strong force F satisfies

$$F\begin{cases} >0 & \text{for } 0 < r < \overline{r}_1, \\ <0 & \text{for } \overline{r}_1 < r < \overline{r}_2. \end{cases} \tag{4.5.53}$$

By (4.5.53), we see that

$$F \sim 0 \quad \text{near } r = \overline{r}_1.$$

This indicates that there is a free shell region inside a proton with radius $R \simeq \overline{r}_1$, such that the three quarks are free in this shell region.

When a lower energy electron collides with the proton, the electromagnetic force cause the electron moving away, leading to the elastic scattering as given by (4.5.49). However, when a high speed electron collides with a proton, it can run into the inside of this proton, interacting with one of the quarks. Since the quark was in a free shell region with no force acting upon it, this particular quark will behavior as a free quark. As it moves into the attracting region of the proton, the quark confinement will hold this quark, which, at the same time, will collide with gluons, exchanging quarks, leading to the inelastic scatterings as given in (4.5.50). This explains the asymptotic freedom.

Remark 4.20 The two phenomena of the quark confinement and the asymptotic freedom are consistent. On the one hand, the quarks in a proton are held tightly together by the strong interaction to cause the quark confinement, and on the other hand, quarks look as if they are free. The consistency is the direct consequence of the quark potential formula in (4.5.41).

4.5.5 Modified Yukawa potential

One of the mysteries of the strong interaction is the conflict characteristics exhibited in the quark level and in the nucleon level. By the strong interaction potentials (4.5.41), we see that

$$\text{quark strong force is infinitely attractive at } r = 10^{-16} \text{ cm;} \tag{4.5.54}$$

while in the nucleon level,

$$\text{nucleon strong force is repulsive in } 0 < r < 10^{-13} \text{ cm.} \tag{4.5.55}$$

The conflicting characteristics of the strong interactions demonstrated in (4.5.54) and (4.5.55) can hardly be explained by any existing theory. However, the layered strong interaction potentials in (4.5.41) derived based on PID and PRI lead to a natural explanation of these characteristics, as well as explanation of the quark confinement and asymptotic freedom in previous subsections. In this subsection, we shall show that the strong interaction nucleon potential (4.5.38), which can be regarded as a modified Yukawa potential, fits experiments. Meanwhile, we also determine some parameters for the strong interaction potentials, including the strong charge g_s.

1. *Experimental results.* Experiments showed that the strong nucleon force has the following properties:

- Nucleon force is of short-ranged, with the radius of the force range: $r \sim 2 \times 10^{-13}$ cm.
- Nucleon force has a repelling center and an attracting region: it is repulsive for $r < \frac{1}{2} \times 10^{-13}$ cm, is attractive for $\frac{1}{2} \times 10^{-13} < r < 2 \times 10^{-13}$ cm, and diminishes for $r > 2 \times 10^{-13}$. Namely, experimentally, the nucleon force behaves as

$$F_{\text{exp}} \begin{cases} > 0 & \text{for } 0 < r < \overline{r} = \frac{1}{2} \times 10^{-13} \text{ cm}, \\ < 0 & \text{for } \overline{r} < r < 2 \times 10^{-13} \text{ cm}, \\ \simeq 0 & \text{for } r > 2 \times 10^{-13} \text{ cm}. \end{cases} \tag{4.5.56}$$

More precisely, the experimental data for the strong interaction nucleon potential can be schematically shown in Figure 4.2; see (Weisskopf, 1972).

2. *Yukawa potential*. Based on the classical strong interaction theory, the potential holding nucleons to form an atomic nucleus is the Yukawa potential

$$\Phi_Y = -\frac{g}{r}e^{-k_n r}, \quad r_1 = \frac{1}{k_n} = 10^{-13} \text{ cm}, \tag{4.5.57}$$

where g is the Yukawa strong charge, and

$$g^2 = 1 \sim 10\,\hbar c. \tag{4.5.58}$$

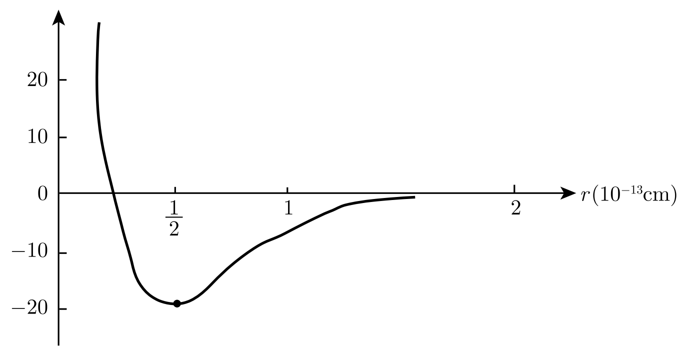


Figure 4.2 Experimental curve of nucleon potential energy

By (4.5.57) we can deduce the classical nucleon force as

$$F_Y = -g\frac{d\Phi_Y}{dr} = -g^2\left(\frac{1}{r^2} + \frac{1}{r_1 r}\right)e^{-k_n r}. \tag{4.5.59}$$

It is clear that the nucleon force (4.5.59) is alway attractive, i.e.

$$\begin{aligned} &F_Y < 0 && \text{for any } r > 0,\\ &F_Y \to -\infty && \text{for } r \to 0,\\ &F_Y/F_1 \simeq 0 && \text{for } r > 10 \times 10^{-13} \text{ cm}, \end{aligned} \tag{4.5.60}$$

where $F_1 = 2g^2/er_1^2$, and e is the base of the natural logarithm.

Comparing (4.5.60) with (4.5.56), we find that the Yukawa theory has a large error in $0 < r < \frac{1}{2} \times 10^{-13}$ cm. In particular, by (4.5.57) the Yukawa potential can be shown in Figure 4.3.

3. *Modified Yukawa potential.* The nucleon potential Φ_n derived by the unified field model based on PID and PRI is given by

$$\Phi_n = 3\left(\frac{\rho_w}{\rho_n}\right)^3 g_s\left[\frac{1}{r} - \frac{A_n}{\rho_n}(1 + k_n r)e^{-k_n r}\right],$$

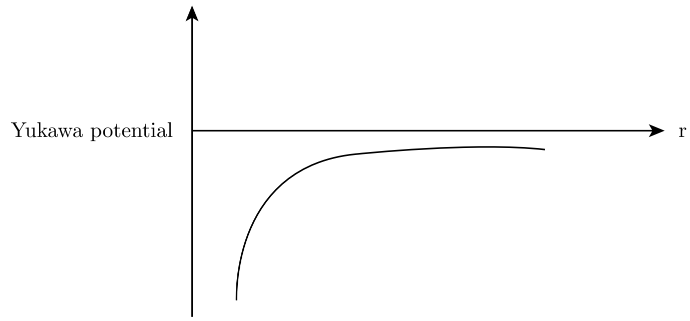


Figure 4.3 Theoretic curve of Yukawa potential energy

and the potential energy V_n of two nucleons is

$$V_n = 3\left(\frac{\rho_w}{\rho_n}\right)^3 g_s\Phi_n = 9\left(\frac{\rho_w}{\rho_n}\right)^6 g_s^2\left[\frac{1}{r} - \frac{A_n}{\rho_n}(1+k_nr)e^{-k_nr}\right]. \tag{4.5.61}$$

The nucleon force is given by

$$F_n = -\frac{dV_n}{dr} = 9\left(\frac{\rho_w}{\rho_n}\right)^6 g_s^2\left[\frac{1}{r^2} - \frac{A_n}{\rho_n}k_n^2 re^{-k_nr}\right], \tag{4.5.62}$$

where

$$\rho_n = 10^{-16}\text{ cm}, \quad k_n = 10^{13}\text{ cm}^{-1}. \tag{4.5.63}$$

4. *Parameters A_n and g_s^2.* First, we use the experimental data in (4.5.56) to determine the parameter A_n in (4.5.62). By (4.5.56) we know that

$$F_{\exp} = 0, \qquad \text{at } \overline{r} = \frac{1}{2}\times 10^{-13}\text{ cm}.$$

Hence, let

$$F_n = 0, \qquad \text{at } \overline{r} = \frac{1}{2}\times 10^{-13}\text{ cm}.$$

Then, it follows from (4.5.62) and (4.5.63) that

$$A_n = \rho_n e^{k_n\overline{r}}/k_n^2\overline{r}^3 = 8e^{1/2}\times 10^{-3}. \tag{4.5.64}$$

Next, we assume that

$$F_n = F_Y, \qquad \text{at } r_1 = 10^{-13}\text{ cm}. \tag{4.5.65}$$

Then, by (4.5.59) and (4.5.62), we deduce from (4.5.65) that

$$9\left(\frac{\rho_w}{\rho_n}\right)^6 g_s^2\left[\frac{1}{r_1^2} - \frac{A_n}{\rho_n}k_n^2 r_1 e^{-1}\right] = -g^2\frac{2}{r_1^2}e^{-1},$$

which leads to

$$g_s^2 = \frac{2}{9}\frac{\rho_n}{(A_n k_n r_1^2 - e\rho_n)}\left(\frac{\rho_n}{\rho_w}\right)^6 g^2.$$

In view of (4.5.63)-(4.5.64) and $r_1 = 1/k_n$, we derive

$$g_s^2 = \frac{2}{9}\frac{e^{-1/2}}{8-e^{1/2}}\left(\frac{\rho_n}{\rho_w}\right)^6 g^2, \tag{4.5.66}$$

where $e = 2.718$ and g is as in (4.5.58).

Finally, from $F_n = 0$ we can also deduce that

$$F_n > 0 \quad \text{as } r > r_2 \simeq 9\times 10^{-13} \text{ cm}.$$

4.5.6 Physical conclusions for nucleon force

The discussion above leads to the following physical conclusions:

1. The modified Yukawa potential based on PID and PRI is

$$\Phi_n = \beta g\left[\frac{1}{r} - \frac{8e^{1/2}}{r_1}\left(1+\frac{r}{r_1}\right)e^{-r/r_1}\right], \tag{4.5.67}$$

where $\beta = \sqrt{2}/\sqrt{8\sqrt{e}-e}$, g is the Yukawa charge, $r_1 = 10^{-13}$ cm, and the modified Yukawa force is

$$F_n = \beta^2 g^2\left[\frac{1}{r^2} - \frac{8e^{1/2}}{r_1^2}\frac{r}{r_1}e^{-r/r_1}\right]. \tag{4.5.68}$$

2. The nucleon strong interaction constant A_n is given by

$$A_n = 8e^{1/2}\times 10^{-3}, \quad e = 2.718. \tag{4.5.69}$$

3. By (4.5.66), the strong charge g_s is given by

$$g_s^2 = \frac{2}{9}\frac{e^{-1/2}}{8-e^{1/2}}\left(\frac{\rho_n}{\rho_w}\right)^6 g^2, \tag{4.5.70}$$

where ρ_w is the w^*-weakton radius and ρ_n the nucleon radius.

4. The largest attraction force of F_n is achieved at $r = 1.5r_1$, and the force is given by

$$F_{\max} = -\frac{8e^{-1/2}}{8-e^{1/2}}\frac{g^2}{r_1^2}\left[3e^2 - \frac{1}{9}\right]. \tag{4.5.71}$$

5. By (4.5.68), the attractive and repulsive regions of the strong nucleon force F_n are as follows

$$F_n \begin{cases} >0 & \text{for } 0<r<r_1/2, \\ <0 & \text{for } r_1/2<r<9r_1, \\ >0 & \text{for } 9r_1<r. \end{cases} \tag{4.5.72}$$

6. Based on (4.5.67), the theoretical curve of the strong interaction potential energy for two nucleons is as shown in Figure 4.4, where we see that the theoretical result is in agreement with the experimental curve shown in Figure 4.2.

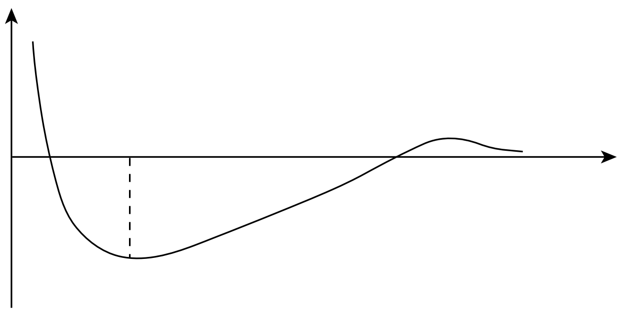


Figure 4.4 Theoretical curve of nucleon potential energy

4.5.7 Short-range nature of strong interaction

By the layered potentials (4.5.41) for the strong interaction, we see that when the nucleons form an atom, the nucleon potential is no longer valid, and the correct potential becomes the strong interaction potential for atoms given by the fourth formula in (4.5.41). The corresponding force formula is given by

$$F_a = N^2 \left(\frac{\rho_w}{\rho_a}\right)^6 g_s^2 \left[\frac{1}{r^2} - \frac{A_a}{\rho_a} k_a^2 r e^{-k_a r}\right], \tag{4.5.73}$$

where

$$\begin{aligned} &\text{for atom}: && k_a = 10^{-9} \sim 10^{-10}\text{ cm}, && \rho_a = 10^{-8}\text{ cm}, \\ &\text{for molecule}: && k_a = 10^{-7}\text{ cm}, && \rho_a = 10^{-7}\text{ cm}. \end{aligned} \tag{4.5.74}$$

It is clear that the attractive force in (4.5.73) is of short-ranged. The repulsive force in (4.5.73) looks as if it is long-ranged. However by (4.5.74) the factors $(\rho_w/\rho_a)^6$ for the atoms and molecules are very small. Hence, the strong repulsive forces for atoms and molecules almost vanishes.

In fact, the bound force between atoms and molecules is the electromagnetic force with strength given by

$$\frac{e^2}{\hbar c} = \frac{1}{137}, \quad e \text{ the electric charge.} \tag{4.5.75}$$

Hence at the atomic and molecular scale, the ratio between strong repulsive force and the electromagnetic force is

$$\frac{F_a}{F_e} = N_s^2 g_s^2 \left(\frac{\rho_w}{\rho_a}\right)^6 / N_e^2 e^2, \tag{4.5.76}$$

where N_s is the number of strong charge, and N_e is the number of the electric charge. Note that each nucleon has three strong charges, and the protons are almost the same as neutrons. Therefore, we assume that

$$N_s = 6N_e.$$

In view of (4.5.70), the ratio (4.5.76) becomes

$$\frac{F_a}{F_e} = 4\beta^2 \left(\frac{\rho_n}{\rho_a}\right)^6 g^2/e^2, \quad \beta^2 = \frac{2}{8\sqrt{e}-e} \backsimeq 0.2.$$

By (4.5.58) and (4.5.75), we have

$$\frac{g^2}{e^2} \text{ is in the range of } \frac{1}{137} \sim \frac{1}{13.7}.$$

Then, by $\rho_n = 10^{-16}$ cm and (4.5.74), we derive that

$$\frac{F_a}{F_e} \sim \begin{cases} 10^{-50} & \text{at the atom level,} \\ 10^{-56} & \text{at the molecular level.} \end{cases}$$

This clearly demonstrates the short-range nature of the strong interaction.

4.6 Weak Interaction Theory

4.6.1 Dual equations of weak interaction potentials

According to the standard model, weak interaction is described by the $SU(2)$ gauge theory. In Section 4.3.4, we have demonstrated that the weak interaction potential is given by the following PRI representation invariant

$$W_\mu = \omega_a W_\mu^a = (W_0, W_1, W_2, W_3), \tag{4.6.1}$$

where $\{\omega_a \mid 1 \leqslant a \leqslant 3\}$ is the $SU(2)$ tensor as in (4.3.29).

Also, the weak charge potential and weak force are as

$$\begin{aligned} &\Phi_w = W_0 && \text{the time component of } W_\mu, \\ &F_w = -g_w(\rho)\nabla\Phi_w, \end{aligned} \tag{4.6.2}$$

where $g_w(\rho)$ is the weak charge of a particle with radius ρ.

In this subsection, we establish the field equations for the dual potential Φ_w and ϕ_w of the weak interaction from the field equations (4.4.48)-(4.4.50).

Taking inner products of (4.4.48) and (4.4.49) with ω_a respectively, we derive the field equations of the two dual weak interaction potentials Φ_w and ϕ_w as follows

$$\partial^\nu W_{\nu\mu} - \frac{g_w}{\hbar c}\kappa_{ab} g^{\alpha\beta} W^a_{\alpha\mu} W^b_\beta - g_w J^w_\mu = (\partial_\mu - \frac{1}{4}k^2 x_\mu + \frac{g_w}{\hbar c}\gamma W_\mu)\phi_w, \tag{4.6.3}$$

$$\begin{aligned}&\left[\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta\right]\phi_w + k_0^2\phi_w - g_w\partial^\mu J^w_\mu \\ &= \frac{g_w}{\hbar c}\partial^\mu\left[\kappa_{ab} g^{\alpha\beta} W^a_{\alpha\mu} W^b_\beta - \gamma W_\mu \phi_w\right] - \frac{1}{4}k_0^2 x_\mu \partial^\mu \phi_w,\end{aligned} \tag{4.6.4}$$

where $\kappa_{ab} = \varepsilon^c_{ab}\omega_c$, W_μ is as in (4.6.1), $\phi_w = \omega_a \phi^a_w$, and

$$J^w_\mu = \omega_a \overline{\Psi}\gamma_\mu \sigma_a \Psi, \quad W_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu + \frac{g_w}{\hbar c}\kappa_{ab} W^a_\mu W^b_\nu. \tag{4.6.5}$$

Experiments showed that the $SU(2)$ gauge fields W^a_μ for weak interaction field particles possess masses. In addition, the dual Higgs fields ϕ^a_w of W^a_μ also have masses. In (4.6.3), there is no massive term of W_μ. However, we see that on the right-hand side of (4.6.3) there is a term

$$\frac{g_w}{\hbar c}\gamma \phi_w W_\mu, \tag{4.6.6}$$

which is spontaneously generated by PID, and breaks the gauge-symmetry. Namely, (4.6.6) will vary under the following $SU(2)$ gauge transformation

$$W^a_\mu \to W^a_\mu - \varepsilon^a_{bc}\theta^b W^c_\mu - \frac{1}{g_w}\partial_\mu \theta^a$$

We shall show that it is the spontaneous symmetry breaking term (4.6.6) that generates mass from the ground state of ϕ_w.

It is clear that the following state

$$(W^a_\mu, \phi^a_w, \psi) = (0, \phi^a_0, 0) \qquad \text{with } \phi^a_0 \text{ being constants,}$$

is a solution of (4.6.3) and (4.6.4), which is a ground state of ϕ^a_w. Let $a_0 = \phi^a_0 \omega_a$, which is a constant. Take the transformation

$$\phi_w \to \phi_w + a_0, \qquad W^a_\mu \to W^a_\mu, \qquad \psi \to \psi.$$

Then the equations (4.6.3) and (4.6.4) are rewritten as

$$\begin{aligned}&\partial^\nu W_{\nu\mu} - k_1^2 W_\mu - \frac{g_w}{\hbar c}\kappa_{ab} g^{\alpha\beta} W^a_{\alpha\mu} W^b_\beta - g_w J^w_\mu \\ &= \left[\partial_\mu - \frac{1}{4}k_0^2 x_\mu + \frac{g_w}{\hbar c}\gamma W_\mu\right]\phi_\mu - \frac{1}{4}a_0 k_0^2 x_\mu,\end{aligned} \tag{4.6.7}$$

$$\begin{aligned}&\left[\frac{1}{c^2}\frac{\partial}{\partial t^2} - \Delta\right]\phi_w + k_0^2\phi_w - g_w\partial^\mu J^w_\mu + k_0^2 a_0 \\ &= \frac{g_w}{\hbar c}\partial^\mu\left[\kappa_{ab} g^{\alpha\beta} W^a_{\alpha\mu} W^b_\beta - \gamma W_\mu(\phi_\mu + a_0)\right] - \frac{1}{4}k_0^2 x_\mu \partial^\mu \phi_w,\end{aligned} \tag{4.6.8}$$

where $k_1 = \sqrt{g_w \gamma a_0/\hbar c}$ represents mass.

Thus, (4.6.7) and (4.6.8) have masses as

$$k_0 = m_H c/\hbar, \quad k_1 = m_W c/\hbar, \tag{4.6.9}$$

where m_H and m_W are the masses of Higgs and $W^{\pm}$ bosons. Physical experiments measured the values of m_H and m_W as

$$m_H \backsimeq 160 \text{ GeV}/c^2, \quad m_W \backsimeq 80 \text{ GeV}/c^2. \tag{4.6.10}$$

By (4.4.46), equations (4.6.7) and (4.6.8) need to add three gauge fixing equations. Based on the superposition property of the weak charge forces, the dual potentials W_0 and ϕ_w should satisfy linear equations, i.e. the time-component $\mu = 0$ equation of (4.6.7) and the equation (4.6.8) should be linear. Therefore, we have to take the three gauge fixing equations in the following forms

$$\begin{aligned}
&\kappa_{ab}\left[g^{\alpha\beta}W^a_{\alpha 0}W^b_\beta - \partial^\mu(W^a_\mu W^b_0)\right] + \gamma W_0\phi_w - \frac{\hbar c}{g_w}\frac{a_0 k_0^2}{4}x_0 = 0,\\
&\frac{g_w}{\hbar c}\partial^\mu\left[\kappa_{ab}g^{\alpha\beta}W^a_{\alpha\mu}W^b_\beta - \gamma W_\mu\phi_w\right] - \frac{k_0^2}{4}x_\mu\partial^\mu\phi_w - k_0^2 a_0 = 0,\\
&\partial^\mu W_\mu = 0,
\end{aligned} \tag{4.6.11}$$

and with the static conditions

$$\frac{\partial}{\partial t}\Phi_w = 0, \quad \frac{\partial}{\partial t}\phi_w = 0, \quad (\Phi_w = W_0). \tag{4.6.12}$$

With the equations (4.6.11) and the static conditions (4.6.12), the time-component $\mu = 0$ equation of (4.6.7) and its dual equation (4.6.8) become

$$-\Delta\Phi_w + k_1^2\Phi_w = g_w Q_w - \frac{1}{4}k_0^2 c\tau\phi_w, \tag{4.6.13}$$

$$-\Delta\phi_w + k_0^2\phi_w = g_w\partial^\mu J^w_\mu, \tag{4.6.14}$$

where $c\tau$ is the wave length of $\phi_w, Q_w = -J^w_0$, and J^w_μ is as in (4.6.5). The two dual equations (4.6.13) and (4.6.14) for the weak interaction potentials Φ_w and ϕ_w are coupled with the Dirac equations (4.4.50), written as

$$i\gamma^\mu\left(\partial_\mu + i\frac{g_w}{\hbar c}W^a_\mu\sigma_a\right)\psi - \frac{mc}{\hbar}\psi = 0. \tag{4.6.15}$$

In addition, by (4.6.9) and (4.6.10) we can determine values of the parameters k_0 and k_1 as follows

$$k_0 = 2k_1, \quad k_1 = 10^{16} \text{ cm}^{-1}. \tag{4.6.16}$$

The parameters $1/k_0$ and $1/k_1$ represent the attracting and repulsive radii of weak interaction forces.

In the next subsection, we shall apply the equations (4.6.13)-(4.6.15) to derive the layered formulas of weak interaction potentials.

4.6.2 Layered formulas of weak forces

Now we deduce from (4.6.13)-(4.6.16) the following layered formulas for the weak interaction potential:

$$\begin{aligned}\Phi_w &= g_w(\rho)e^{-kr}\left[\frac{1}{r}-\frac{B}{\rho}(1+2kr)e^{-kr}\right],\\ g_w(\rho) &= N\left(\frac{\rho_w}{\rho}\right)^3 g_w,\end{aligned} \tag{4.6.17}$$

where Φ_w is the weak force potential of a particle with radius ρ and carrying N weak charges g_w (g_w is the unit of weak charge for each weakton, an elementary particle), ρ_w is the weakton radius, B is a parameter depending on the particles, and

$$\frac{1}{k} = 10^{-16}\text{ cm}, \tag{4.6.18}$$

represents the force-range of weak interaction.

To derive the layered formulas (4.6.17), first we shall deduce the following weak interaction potential for a weakton

$$\Phi_w^0 = g_s e^{-kr}\left[\frac{1}{r}-\frac{B_0}{\rho_w}(1+2kr)e^{-kr}\right]. \tag{4.6.19}$$

To derive the solution ϕ_w of (4.6.14), we need to compute the right-hand term of (4.6.14). By (4.6.5) we have

$$\partial^\mu J_\mu^w = \omega_a\partial_\mu\overline{\psi}\gamma^\mu\sigma_a + \omega_a\overline{\psi}\gamma^\mu\sigma_a\partial_\mu\psi.$$

Due to the dirac equation (4.6.15),

$$\begin{aligned}\partial_\mu\overline{\psi}\gamma^\mu\sigma_a\psi &= -i\frac{g_w}{\hbar c}W_\mu^b\overline{\psi}\gamma^\mu\sigma_b\sigma_a\psi + i\frac{mc}{\hbar}\overline{\psi}\sigma_a\psi,\\ \overline{\psi}\gamma^\mu\sigma_a\partial_\mu\psi &= i\frac{g_w}{\hbar c}W_\mu^b\overline{\psi}\gamma^\mu\sigma_a\sigma_b\psi - i\frac{mc}{\hbar}\overline{\psi}\sigma_a\psi.\end{aligned}$$

Thus we obtain

$$\partial^\mu J_\mu^w = i\frac{g_w}{\hbar c}\omega_a W_\mu^b\overline{\psi}\gamma^\mu[\sigma_a,\sigma_b]\psi = -2\frac{g_w}{\hbar c}\varepsilon_{ab}^c\omega_a W_\mu^b J_c^\mu. \tag{4.6.20}$$

Here we used $[\sigma_a,\sigma_b] = i2\varepsilon_{ab}^c\sigma_c$ and $J_c^\mu = \overline{\psi}\gamma^\mu\sigma_c\psi$. Note that

$$\varepsilon_{ab}^c\omega_a W_\mu^b = \begin{vmatrix}\vec{i} & \vec{j} & \vec{k}\\ \omega_1 & \omega_2 & \omega_3\\ W_\mu^1 & W_\mu^2 & W_\mu^3\end{vmatrix} = \vec{\omega}\times\vec{W}_\mu,$$

where $\vec{\omega} = (\omega_1,\omega_2,\omega_3), \vec{W}_\mu = (W_\mu^1,W_\mu^2,W_\mu^3)$. Hence, (4.6.20) can be rewritten as

$$\partial^\mu J_\mu^w = -\frac{2g_w}{\hbar c}(\vec{\omega}\times\vec{W}_\mu)\cdot\vec{J}^\mu, \tag{4.6.21}$$

and $\vec{J}^\mu = (J_1^\mu, J_2^\mu, J_3^\mu)$. The weak current density J_a^μ is as

$$J_a^\mu = \theta_a^\mu \delta(r), \qquad \theta_a^\mu \quad \text{the constant tensor},$$

and $\vec{W}_\mu$ in (4.6.21) is replaced by the average value

$$\frac{1}{|B_{\rho_w}|}\int_{B_{\rho_w}} \vec{W}_\mu dx, \tag{4.6.22}$$

where $B_{\rho_w} \subset \mathbb{R}^n$ is the ball with radius ρ_w. Similar to the case (4.5.14) for the strong interaction, the average value (4.6.22) is

$$\vec{W}_\mu = \vec{\zeta}_\mu/\rho_w, \quad \vec{\zeta}_\mu = (\zeta_\mu^1, \zeta_\mu^2, \zeta_\mu^3).$$

Thus, (4.6.21) can be expressed as

$$\partial^\mu J_\mu^w = -\kappa\delta(r)/\rho_w, \tag{4.6.23}$$

and κ is a parameter, written as

$$\kappa = \frac{2g_w}{\hbar c}\vec{\theta}^\mu \cdot (\vec{\omega} \times \vec{\zeta}_\mu). \tag{4.6.24}$$

Putting (4.6.23) in (4.6.14) we deduce that

$$-\Delta\phi_w + k_0^2\phi_w = -\frac{g_w\kappa}{\rho_w}\delta(r),$$

whose solution is given by

$$\phi_w = -\frac{g_w\kappa}{\rho_w}\frac{1}{r}e^{-k_0 r}. \tag{4.6.25}$$

Therefore we obtain the solution of (4.6.14) in the form (4.6.25).

Inserting (4.6.25) into (4.6.13) we get

$$-\Delta\Phi_w + k_1^2\Phi_w = g_w Q_w + \frac{g_w\mathscr{B}}{\rho_w}\frac{1}{r}e^{-k_0 r}, \tag{4.6.26}$$

where $\mathscr{B}$ is a parameter with dimension $1/L$, given by

$$\mathscr{B} = \frac{1}{4}\kappa k_0^2 c\tau, \quad \kappa \ \text{ is as in (4.6.24)}.$$

Since $g_w Q_w = -g_w J_0^w$ is the weak charge density, we have

$$Q_w = \beta\delta(r),$$

and β is a scaling factor. We can take proper unit for g_s such that $\beta = 1$. Thus, putting $Q_w = \delta(r)$ in (4.6.26) we derive that

$$-\Delta\Phi_w + k_1^2\Phi_w = g_w\delta(r) + \frac{g_w\mathscr{B}}{\rho_w}\frac{1}{r}e^{-k_0 r}. \tag{4.6.27}$$

Let the solution of (4.6.27) be radially symmetric, then the equation (4.6.27) can be equivalently written as

$$-\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)\Phi_w+k_1^2\Phi_w=g_w\delta(r)+\frac{g_w\mathscr{B}}{\rho_w}\frac{1}{r}e^{-k_0r}. \tag{4.6.28}$$

The solution of (4.6.28) can be expressed as

$$\Phi_w=\frac{g_w}{r}e^{-k_1r}-\frac{g_w\mathscr{B}}{\rho_w}\varphi(r)e^{-k_0r}, \tag{4.6.29}$$

where $\varphi(r)$ satisfies the equation

$$\varphi''+2\left(\frac{1}{r}-k_0\right)\varphi'-\left(\frac{2k_0}{r}+k_1^2-k_0^2\right)\varphi=\frac{1}{r}. \tag{4.6.30}$$

Let φ be in the form

$$\varphi=\sum_{n=0}^{\infty}\beta_nr^n \quad (\text{the dimension of } \varphi \text{ is } L).$$

Inserting φ into (4.6.30), comparing coefficients of r^n, we get

$$\begin{aligned}
\beta_1&=k_0\beta_0+\frac{1}{2},\\
\beta_2&=\frac{1}{2}(k_0^2-k_1^2)\beta_0+\frac{1}{3}k_0,\\
&\vdots\\
\beta_n&=\frac{2k_0}{n+1}\beta_{n-1}-(k_0^2-k_1^2)\beta_{n-2},\quad n\geqslant 2.
\end{aligned}$$

Note that the dimensions of $\mathscr{B}$ and β_0 in $\varphi(r)$ are $1/L$ and L. The parameter $B=\mathscr{B}\beta_0$ is dimensionless. Physically, we can only measure the value of B. Therefore we take φ in its second-order approximation as follows

$$\varphi=\beta_0(1+k_0r).$$

In addition, by (4.6.16) we take

$$k_1=k,\quad k_0=2k,\quad k=10^{16}\text{ cm}^{-1}.$$

Thus, the formula (4.6.29) is written as

$$\Phi_s=g_we^{-kr}\left[\frac{1}{r}-\frac{B}{\rho_w}(1+2kr)e^{-kr}\right].$$

This is the weak interaction potential of a weakton, which is as given in (4.6.19).

In the same fashion as used in the layered formula (4.5.39) for the strong interaction, for a particle with radius ρ and N weak charges g_w, we can deduce the layered formula of weak interaction potentials as in the form given by (4.6.17).

4.6.3 Physical conclusions for weak forces

As mentioned earlier, the layered weak interaction potential formula (4.6.17) plays the same fundamental role as the Newtonian potential for gravity and the Coulomb potential for electromagnetism. Hereafter we explore a few direct physical consequences of the weak interaction potentials.

1. *Short-range nature of weak interaction.* By (4.6.17) it is easy to see that for all particles, their weak interaction force-range is as

$$r=\frac{1}{k}=10^{-16}\text{ cm},$$

which is consistent with experimental data.

2. *Weak force formula.* For two particles with radii ρ_1,ρ_2, and with N_1,N_2 weak charges g_w. Their weak charges are given by

$$g_w(\rho_j)=N_j\left(\frac{\rho_w}{\rho_j}\right)^3 g_w \qquad \text{for } j=1,2. \tag{4.6.31}$$

The weak potential energy for the two particles is

$$V=g_w(\rho_1)g_w(\rho_2)e^{-kr}\left[\frac{1}{r}-\frac{\overline{B}}{\overline{\rho}}(1+2kr)e^{-kr}\right], \tag{4.6.32}$$

where $g_w(\rho_1)$ and $g_w(\rho_2)$ are as in (4.6.31), $\overline{\rho}$ is a radius depending on ρ_1 and ρ_2, and $\overline{B}$ is a constant depending on the types of these two particles.

The weak force between the two particles is given by

$$F=-\frac{d}{dr}V=g_w(\rho_1)g_w(\rho_2)e^{-kr}\left[\frac{1}{r^2}+\frac{k}{r}-\frac{4\overline{B}}{\overline{\rho}}k^2re^{-kr}\right]. \tag{4.6.33}$$

3. *Repulsive condition.* For the two particles as above, if their weak interaction constant $\overline{B}$ satisfies the inequality

$$\frac{1}{r^2}+\frac{k}{r}\geqslant\frac{4\overline{B}}{\overline{\rho}}k^2re^{-kr}, \qquad \forall 0<r\leqslant\frac{1}{k},$$

or equivalently $\overline{B}$ satisfies

$$\overline{B}\leqslant\frac{e}{2}\overline{\rho}k \quad (e=2.718,\ k=10^{16}\text{ cm}^{-1}), \tag{4.6.34}$$

then the weak force between these two particles is always repulsive.

It follows from this conclusion that for the neutrinos:

$$(\nu_1,\nu_2,\nu_3)=(\nu_e,\nu_\mu,\nu_\tau),$$
$$(\overline{\nu}_1,\overline{\nu}_2,\overline{\nu}_3)=(\overline{\nu}_e,\overline{\nu}_\mu,\overline{\nu}_\tau),$$

the weak interaction constants

$$\begin{array}{lll} \overline{B}_{ij} & \text{for } \nu_i \text{ and } \nu_j & \forall 1 \leqslant i, j \leqslant 3, \\ \overline{B}_{ij} & \text{for } \nu_i \text{ and anti-neutrinos } \overline{\nu}_j & \forall i \neq j, \end{array}$$

satisfy the exclusion condition (4.6.34).

4. *Value of weak charge g_w.* Based on the Standard Model, the coupling constant G_w of the β-decay of nucleons and the Fermi constant G_f have the following relation

$$G_w^2 = \frac{8}{\sqrt{2}} \left(\frac{m_W c}{\hbar} \right)^2 G_f, \tag{4.6.35}$$

and G_f is given by

$$G_f = 10^{-5} \hbar c / \left(\frac{m_p c}{\hbar} \right)^2, \tag{4.6.36}$$

where m_W and m_p are masses of $W^\pm$ bosons and protons. By the gauge theory, G_w is also the coupling constant of $SU(2)$ gauge fields. Therefore we can regard G_w as the weak charge of nucleons, i.e.

$$G_w = g_w(\rho_n), \quad \rho_n \text{ the nucleon radius.}$$

In addition, it is known that

$$g_w(\rho_n) = 9 \left(\frac{\rho_w}{\rho_n} \right)^3 g_s.$$

Hence, we deduce from (4.6.35) and (4.6.36) that

$$81 \left(\frac{\rho_w}{\rho_n} \right)^6 g_w^2 = \frac{8}{\sqrt{2}} \left(\frac{m_w}{m_p} \right)^2 \times 10^{-5} \hbar c.$$

Then we derive the relation

$$g_w^2 = \frac{8}{81\sqrt{2}} \left(\frac{m_w}{m_p} \right)^2 \times 10^{-5} \times \left(\frac{\rho_n}{\rho_w} \right)^6 \hbar c. \tag{4.6.37}$$

In a comparison with (4.5.66) and (4.6.37), we find that the strong charge g_s and the weak charge g_w have the same magnitude order. Direct computation shows that

$$\frac{g_w^2}{g_s^2} \simeq 0.35 \qquad \text{equivalently} \qquad \frac{g_w}{g_s} \simeq 0.6.$$

4.6.4 PID mechanism of spontaneous symmetry breaking

In the derivation of equations (4.6.7) and (4.6.8) we have used the PID mechanism of spontaneous symmetry breaking. In this subsection we shall discuss the intermediate vector bosons $W^\pm, Z$ and their dual scalar bosons $H^\pm, H^0$, called the Higgs particles by using the PID mechanism of spontaneous symmetry breaking.

We know that the interaction field equations are oriented toward two directions: i) to derive interaction forces, and ii) to describe the field particles and derive the particle transition probability. The PID-PRI weak interaction field equations describing field particles are given by (4.4.42)-(4.4.45). Here, for convenience, we write them again as follows:

$$\partial^\nu W^a_{\nu\mu} - \frac{g_w}{\hbar c}\varepsilon^a_{bc} g^{\alpha\beta} W^b_{\alpha\mu} W^c_\beta - g_w J^a_\mu = \left(\partial_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu + \frac{g_w}{\hbar c}\gamma_b W^b_\mu\right)\phi^a, \tag{4.6.38}$$

$$\begin{aligned} & -\partial^\mu\partial_\mu\phi^a + \left(\frac{m_H c}{\hbar}\right)^2\phi^a - \frac{g_w}{\hbar c}\gamma_b\partial^\mu(W^b_\mu\phi^a) \\ & \quad = \frac{g_w}{\hbar c}\varepsilon^a_{bc} g^{\alpha\beta}\partial^\mu(W^b_{\alpha\mu}W^c_\beta) + g_w\partial^\mu J^a_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu\partial^\mu\phi^a, \end{aligned} \tag{4.6.39}$$

where m_H is the Higgs particle mass, and W^a_μ $(1 \leqslant a \leqslant 3)$ describe the intermediate vector bosons as follows

$$\begin{aligned} W^\pm: &\quad W^1_\mu \pm iW^2_\mu, \\ Z: &\quad W^3_\mu, \end{aligned} \tag{4.6.40}$$

and ϕ^a describe the dual Higgs bosons as

$$\begin{aligned} H^\pm: &\quad \phi^1 \pm i\phi^2, \\ H^0: &\quad \phi^3. \end{aligned} \tag{4.6.41}$$

By (4.4.46), equations (4.6.38)-(4.6.39) need to be supplemented with three gauge fixing equations. According to physical requirement, we take these equations as

$$\partial^\mu W^a_\mu = 0 \qquad \text{for } 1 \leqslant a \leqslant 3. \tag{4.6.42}$$

To match (4.6.40) and (4.6.41), we take the $SU(2)$ generator transformation as follows

$$\begin{pmatrix}\widetilde{\sigma}_1\\ \widetilde{\sigma}_2\\ \widetilde{\sigma}_3\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & i & 0\\ 1 & -i & 0\\ 0 & 0 & \sqrt{2}\end{pmatrix}\begin{pmatrix}\sigma_1\\ \sigma_2\\ \sigma_3\end{pmatrix}.$$

Under this transformation, W^a_μ and ϕ^a are changed into

$$\begin{aligned} (\widetilde{W}^1_\mu, \widetilde{W}^2_\mu, \widetilde{W}^3_\mu) &= (W^\pm_\mu, Z_\mu) = \left(\frac{1}{\sqrt{2}}(W^1_\mu \pm iW^2_\mu), W^3_\mu\right), \\ (\widetilde{\phi}^1, \widetilde{\phi}^2, \widetilde{\phi}^3) &= (\phi^\pm, \phi^0) = \left(\frac{1}{\sqrt{2}}(\phi^1 \pm i\phi^2), \phi^3\right). \end{aligned}$$

The equations (4.6.38) and (4.6.39) obey PRI, and under the above transformation they becomes

$$\partial^\nu W^\pm_{\nu\mu} \pm \frac{ig_w}{\hbar c} g^{\alpha\beta}(W^\pm)\alpha\mu Z_\beta - Z_{\alpha\mu}W^\pm_\beta) - g_w J^\pm_\mu \tag{4.6.43}$$
$$= \left[\partial_\mu + k_W^2 W^\pm_\mu + k_Z^2 Z_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu\right]\phi^\pm,$$

$$\partial^\nu Z_{\nu\mu} - \frac{ig_w}{\hbar c} g^{\alpha\beta}(W^+_{\alpha\mu}W^-_\beta - W^-_{\alpha\mu}W^+_\beta) - g_w J^0_\mu \tag{4.6.44}$$
$$= \left[\partial_\mu + k_W^2 W^\pm_\mu + k_Z^2 Z_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu\right]\phi^0,$$

$$\Box H^\pm + \left(\frac{m_H c}{\hbar}\right) H^\pm - \frac{g_w}{\hbar c}\partial^\mu(\widetilde{\gamma}_b \widetilde{W}^b_\mu H^\pm) \tag{4.6.45}$$
$$= \frac{g_w}{\hbar c} g^{\alpha\beta}\partial^\mu(\widetilde{\varepsilon}^\pm_{bc}\widetilde{W}^b_{\alpha\mu}\widetilde{W}^c_\beta) + g_w\partial^\mu J^\pm_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu \partial^\mu H^\pm,$$

$$\Box H^0 + \left(\frac{m_H c}{\hbar}\right)^2 H^0 - \frac{g_w}{\hbar c}\partial^\mu(\widetilde{\gamma}_b \widetilde{W}^b_\mu H^0) \tag{4.6.46}$$
$$= \frac{g_w}{\hbar c} g^{\alpha\beta}\partial^\mu(\widetilde{\varepsilon}^0_{bc}\widetilde{W}^b_{\alpha\mu}\widetilde{W}^c_\beta) + g_w\partial^\mu J^0_\mu - \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu \partial^\mu H^0,$$

where $H^\pm = \frac{1}{\sqrt{2}}(\phi^1 \pm i\phi^2), H^0 = \phi^3$ in (4.6.45) and (4.6.46), and

$$\begin{aligned} &J^\pm_\mu = \frac{1}{\sqrt{2}}(J^1_\mu \pm iJ^2_\mu) && \text{the charged current,} \\ &J^0_\mu = J^3_\mu && \text{the neutral current,} \\ &W^\pm_{\nu\mu} = \partial_\nu W^\pm_\mu - \partial_\mu W^\pm_\nu \pm \frac{ig_w}{\hbar c}(Z_\mu W^\pm_\nu - Z_\nu W^\pm_\mu), \\ &Z_{\nu\mu} = \partial_\nu Z_\mu - \partial_\mu Z_\nu + \frac{ig_w}{\hbar c}(W^+_\mu W^-_\nu - W^+_\nu W^-_\mu), \end{aligned} \tag{4.6.47}$$

and

$$k_W^2 = \frac{g_w\gamma_1}{\sqrt{2}\hbar c}, \qquad k_Z^2 = \frac{g_w\gamma_3}{\hbar c}, \tag{4.6.48}$$

which are derived by the following transformation

$$\begin{pmatrix} k_W^2 \\ k_W^2 \\ k_Z^2 \end{pmatrix} = \frac{g_w}{\sqrt{2}\hbar c}\begin{pmatrix} 1 & i & 0 \\ 1 & -i & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix},$$

with $\gamma_2 = 0$ in the Pauli matrix representation.

The equation (4.6.43)-(4.6.46) are the model to govern the behaviors of the weak interaction field particles (4.6.40) and (4.6.41). Two observations are now in order.

First, we note that these equations are nonlinear, and consequently, no free weak interaction field particles appear.

Second, there are two important solutions of (4.6.43)-(4.6.44), dictating two different weak interaction procedures.

The first solution sets

$$W_\mu^\pm = 0, \quad \phi^0 = 1. \tag{4.6.49}$$

Then Z satisfies the equation

$$\Box Z_\mu + k_z^2 Z_\mu = -g_w J_\mu^0 + \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu. \tag{4.6.50}$$

The second solution takes

$$Z_\mu = 0, \quad \phi^\pm = 1. \tag{4.6.51}$$

Then $W_\mu^\pm$ satisfy the equations

$$\Box W_\mu^\pm + k_w^2 W_\mu^\pm = -g_w J_\mu^\pm + \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 x_\mu. \tag{4.6.52}$$

We are now ready to obtain the following physical conclusions for the weak interaction field particles.

1. *Duality of field particles.* The field equations (4.6.43)-(4.6.46) provide a natural duality between the field particles:

$$W^\pm \leftrightarrow H^\pm, \quad Z \leftrightarrow H^0.$$

2. *Lack of freedom for field particles.* Due to the nonlinearity of (4.6.43)-(4.6.46), the weak interaction field particles $W^\pm, Z, H^\pm, H^0$ are not free bosons.

3. *PID mechanism of spontaneous symmetry breaking.* In the field equations (4.6.50) and (4.6.52), it is natural that the masses m_W and m_Z of $W^\pm$ and Z bosons appear at the ground states (4.6.49) and (4.6.51), with

$$m_W = \frac{\hbar}{c} k_W, \quad m_Z = \frac{\hbar}{c} k_Z,$$

and k_W and k_Z are as in (4.6.48).

4. *Basic properties of field particles.* From (4.6.43)-(4.6.46) we can obtain some basic information for the bosons as follows:

$W^\pm:$	spin $J = 1$,	electric charge $= \pm e$,	mass m_W,
$Z:$	spin $J = 1$,	electric charge $= 0$,	mass m_Z,
$H^\pm:$	spin $J = 0$,	electric charge $= \pm e$,	mass m_H^+,
$H^0:$	spin $J = 0$,	electric charge $= 0$,	mass m_H^0.

Remark 4.21 Under the translation

$$Z_\mu \to Z_\mu + \frac{1}{4}\left(\frac{m_H c}{\hbar}\right)^2 \frac{1}{k_Z^2} x_\mu,$$
$$W_\mu^\pm \to W_\mu^\pm + \frac{1}{4}(\frac{m_H c}{\hbar})^2 \frac{1}{k_W^2} x_\mu,$$

the equations (4.6.50) and (4.6.52) become

$$\begin{aligned} &\Box Z_\mu + k_Z^2 Z_\mu = -g_w J_\mu^0, \\ &\Box W_\mu^\pm + k_W^2 W_\mu^\pm = -g_w J_\mu^\pm. \end{aligned} \tag{4.6.53}$$

Similarly, if we take the gauge fixing equations

$$x_\mu \partial^\mu \phi^a = 0 \qquad \text{for } 1 \leqslant a \leqslant 3,$$

instead of (4.6.42), then under the conditions

$$W_\mu^a = 0 \qquad \text{for } 1 \leqslant a \leqslant 3,$$

the equations (4.6.46) and (4.6.45) are in the form

$$\begin{aligned} &\Box H^0 + \left(\frac{m_H c}{\hbar}\right)^2 H^0 = g_w \partial^\mu J_\mu^0, \\ &\Box H^\pm + \left(\frac{m_H c}{\hbar}\right)^2 H^\pm = g_w \partial^\mu J_\mu^\pm. \end{aligned} \tag{4.6.54}$$

The equations (4.6.53) and (4.6.54) are the standard Klein-Gordon models describing the $W^\pm, Z, H^\pm, H^0$ bosons, which are derived only by the PID-PRI unified field model.

4.6.5 Introduction to the classical electroweak theory

In comparison with the PID-PRI weak interaction theory, in this subsection we briefly introduce the classical $U(1) \times SU(2)$ electroweak theory; see among many others (Kaku, 1993; Griffiths, 2008; Quigg, 2013).

1. The fields in the electroweak theory are as follows:

$SU(2)$ gauge fields:	$W_\mu^1, W_\mu^2, W_\mu^3,$
$U(1)$ gauge field:	$B_\mu,$
Dirac spinor doublets:	$L = \begin{pmatrix} \nu_L \\ l_L \end{pmatrix},$
Dirac spinor singlet:	$R = l_R,$
Higgs scalar doublet:	$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix},$

where ν_L and l_L are the left-hand neutrino and lepton, l_R is the right-hand lepton, (ϕ^+, ϕ^0) are scalar fields with electric charge (1,0).

2. The Lagrangian action of the electroweak theory is given by

$$\mathscr{L}_{WS} = \mathscr{L}_G + \mathscr{L}_D + \mathscr{L}_H, \tag{4.6.55}$$

where $\mathscr{L}_G$ is the gauge sector, $\mathscr{L}_D$ is the Dirac sector, and $\mathscr{L}_H$ is the Higgs sector:

$$\begin{aligned}
&\mathscr{L}_G = -\frac{1}{4}W^a_{\mu\nu}W^{\mu\nu a} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu},\\
&\mathscr{L}_D = i\overline{L}\gamma^\mu D_\mu L + i\overline{R}\gamma^\mu D_\mu R,\\
&\mathscr{L}_H = \frac{1}{2}(D^\mu\phi)^\dagger(D_\mu\phi) + \frac{\lambda}{4}(\phi^\dagger\phi - a^2)^2 - G_l(\overline{L}\phi R + \overline{R}\phi^\dagger L).
\end{aligned} \tag{4.6.56}$$

Here λ, a, G_l are constants,

$$\begin{aligned}
&W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g_1\varepsilon^a_{bc}W^b_\mu W^c_\nu,\\
&B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu,\\
&D_\mu R = (\partial_\mu + ig_2B_\mu)R,\\
&D_\mu L = (\partial_\mu + i\frac{g_2}{2}B_\mu - i\frac{g_1}{2}W^a_\mu\sigma_a)L,\\
&D_\mu\phi = (\partial_\mu - i\frac{g_2}{2}B_\mu - i\frac{g_1}{2}W^a_\mu\sigma_a)\phi,
\end{aligned} \tag{4.6.57}$$

g_1, g_2 are coupling constants of $SU(2)$ and $U(1)$ gauge fields, and σ_a $(1 \leqslant a \leqslant 3)$ are the Pauli matrices.

3. The action (4.6.55)-(4.6.57) is invariant under the following $SU(2)$ and $U(1)$ gauge transformations:

- *$SU(2)$ gauge transformation*:

$$\begin{aligned}
&L \to e^{\frac{i}{2}\theta^a\sigma_a}L,\\
&\phi \to e^{\frac{i}{2}\theta^a\sigma_a}\phi,\\
&R \to R,\\
&W^a_\mu \to W^a_\mu - \frac{2}{g_1}\partial_\mu\theta^a + \varepsilon^a_{bc}\theta^b W^c_\mu, \quad \text{i.e. as in (2.4.38)},\\
&B_\mu \to B_\mu.
\end{aligned}$$

- *$U(1)$ gauge transformation*:

$$\begin{aligned}
&L \to e^{\frac{i}{2}\beta}L,\\
&\phi \to e^{-\frac{i}{2}\beta}\phi,\\
&R \to e^{i\beta}R,\\
&W^a_\mu \to W^a_\mu - \frac{2}{g_2}\partial_\mu\beta,\\
&B_\mu \to B_\mu + \frac{2}{g_2}\partial_\mu\beta.
\end{aligned}$$

4. The electroweak field equations are derived by

$$\frac{\delta L_{WS}}{\delta W^a_\mu}=0,\ \frac{\delta L_{WS}}{\delta B_\mu}=0,\ \frac{\delta L_{WS}}{\delta \phi}=0,\ \frac{\delta L_{WS}}{\delta L}=0,\ \frac{\delta L}{\delta R}=0. \tag{4.6.58}$$

We remark here that the Higgs field in this setting is included in the Lagrangian action, drastically different from the mechanism based on PID developed earlier.

In addition, the particle ϕ^+ represents a massless boson with a positive electric charge. However, in reality no such particles exist. Hence we have to take the Higgs scalar field as

$$\phi=\begin{pmatrix}0\\ \varphi\end{pmatrix}. \tag{4.6.59}$$

In fact, in the WS theory, the Higgs field ϕ is essentially taken as the form (4.6.59). Under the condition (4.6.59), the variational equations (4.6.58) of the SW action (4.6.55)-(4.6.57) are expressed as follows:

Gauge field equations (massless):

$$\begin{aligned}
&\partial^\nu W^1_{\nu\mu}-g_1g^{\alpha\alpha}(W^2_{\alpha\mu}W^3_\alpha-W^3_{\alpha\mu}W^2_\alpha)+\frac{g_1}{2}J^1_\mu-\frac{g_1^2}{2}\varphi^2W^1_\mu=0,\\
&\partial^\nu W^2_{\nu\mu}-g_1g^{\alpha\alpha}(W^3_{\alpha\mu}W^1_\alpha-W^1_{\alpha\mu}W^3_\alpha)+\frac{g_1}{2}J^2_\mu-\frac{g_1^2}{2}\varphi^2W^2_\mu=0,\\
&\partial^\nu W^3_{\nu\mu}-g_1g^{\alpha\alpha}(W^1_{\alpha\mu}W^2_\alpha-W^2_{\alpha\mu}W^1_\alpha)+\frac{g_1}{2}J^3_\mu-\frac{g_1}{2}\varphi^2(g_1W^3_\mu-g_2B_\mu)=0,\\
&\partial^\nu B_{\nu\mu}-\frac{g_2}{2}J^L_\mu-g_2J^R_\mu-\frac{g_2}{2}\varphi^2(g_2B_\mu-g_1W^3_\mu)=0.
\end{aligned} \tag{4.6.60}$$

Higgs field equations:

$$\begin{aligned}
&\partial^\mu\partial_\mu\varphi-\frac{1}{4}\varphi(g_1^2W^a_\mu W^{\mu a}+g_2^2B_\mu B^\mu-2g_1g_2W^3_\mu B^\mu)\\
&-\lambda\varphi(\varphi^2-a^2)+G_l(\bar{l}_LR+\overline{R}l_L)=0.
\end{aligned} \tag{4.6.61}$$

Dirac equations:

$$\begin{aligned}
&i\gamma^\mu(\partial_\mu+ig_2B_\mu)R-G_l\varphi l_L=0,\\
&i\gamma^\mu(\partial_\mu+i\frac{g_2}{2}B_\mu-i\frac{g_1}{2}W^a_\mu\sigma_a)\begin{pmatrix}\nu_L\\ l_L\end{pmatrix}-G_lR\begin{pmatrix}0\\ \varphi\end{pmatrix}=0.
\end{aligned} \tag{4.6.62}$$

Here

$$\begin{aligned}
&J^a_\mu=\overline{L}\gamma_\mu\sigma_aL && 1\leqslant a\leqslant 3,\\
&J^L_\mu=\overline{\nu}_L\gamma_\mu\nu_L+\bar{l}_L\gamma_\mu l_L,\\
&J^R_\mu=\overline{R}\gamma_\mu R, && \gamma_\mu=g_{\mu\alpha}\gamma^\alpha.
\end{aligned}$$

5. Masses are generated at the ground states. It is clear that the following state

$$\varphi=a,\quad W^a_\mu=0,\quad B_\mu=0,\quad L=0,\quad R=0,$$

is a solution of (4.6.60)-(4.6.62), called a ground state. We take the translation transformation

$$\varphi \to \varphi + a, \quad W_\mu^a \to W_\mu^a, \quad B_\mu \to B_\mu, \quad L \to L, \quad R \to R,$$

then the massless equations (4.6.60) become massive, written as

$$\partial^\nu \begin{pmatrix} W_{\nu\mu}^1 \\ W_{\nu\mu}^2 \\ W_{\nu\mu}^3 \\ B_{\nu\mu} \end{pmatrix} - M \begin{pmatrix} W_\mu^1 \\ W_\mu^2 \\ W_\mu^3 \\ B_\mu \end{pmatrix}$$

$$= \begin{pmatrix} g_1 g^{\alpha\beta} \varepsilon_{ab}^1 W_{\alpha\mu}^a W_\beta^b - \frac{g_1}{2} J_\mu^1 + \frac{1}{2} g_1^2 (\varphi^2 + 2a\varphi) W_\mu^1 \\ g_1 g^{\alpha\beta} \varepsilon_{ab}^2 W_{\alpha\mu}^a W_\beta^b - \frac{g_1}{2} J_\mu^2 + \frac{1}{2} g_1^2 (\varphi^2 + 2a\varphi) W_\mu^2 \\ g_1 g^{\alpha\beta} \varepsilon_{ab}^3 W_{\alpha\mu}^a W_\beta^b - \frac{g_1}{2} J_\mu^3 + \frac{1}{2} g_1 (\varphi^2 + 2a\varphi)(g_1 W_\mu^3 - g_2 B_\mu) \\ \frac{1}{2} g_2 J_\mu^L + g_2 J_\mu^R + \frac{1}{2} g_2 (\varphi^2 + 2a\varphi)(g_2 B_\mu - g_1 W_\mu^3) \end{pmatrix}, \tag{4.6.63}$$

where M is the mass matrix given by

$$M = \frac{c^2}{\hbar^2} \begin{pmatrix} m_1^2 & 0 & 0 & 0 \\ 0 & m_1^2 & 0 & 0 \\ 0 & 0 & m_1^2 & -m_3^2 \\ 0 & 0 & -m_3^2 & m_2^2 \end{pmatrix}, \tag{4.6.64}$$

and

$$\frac{m_1 c}{\hbar} = \frac{g_1 a}{\sqrt{2}}, \quad \frac{m_2 c}{\hbar} = \frac{g_2 a}{\sqrt{2}}, \quad \frac{m_3 c}{\hbar} = \frac{\sqrt{g_1 g_2} a}{\sqrt{2}}.$$

6. The masses m_W and m_Z can be derived from (4.6.63) and (4.6.64) as follows. According to the IVB theory for particle transition, the $W^\pm$ particles are characterized as

$$W^\pm : \; W_\mu^1 \pm i W_\mu^2.$$

Hence we need the following transformation for W_μ^1 and W_μ^2:

$$\begin{pmatrix} W_\mu^+ \\ W_\mu^- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix} \begin{pmatrix} W_\mu^1 \\ W_\mu^2 \end{pmatrix}, \tag{4.6.65}$$

which requires the following transformation of $SU(2)$ generators from the Pauli matrices σ_a:

$$\begin{pmatrix} \widetilde{\sigma}_1 \\ \widetilde{\sigma}_2 \\ \widetilde{\sigma}_3 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i & 0 \\ 0 & -i & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{pmatrix}.$$

In addition, we also need to obtain the mass m_Z of Z boson by diagonalizing the massive matrix (4.6.64). It is easy to see that

$$UMU^\dagger = \frac{c^2}{\hbar^2}\begin{pmatrix} m_W^2 & 0 & 0 & 0 \\ 0 & m_W^2 & 0 & 0 \\ 0 & 0 & m_Z^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & \alpha & \beta \\ 0 & 0 & -\beta & \alpha \end{pmatrix}, \qquad \alpha = \frac{g_1}{|g|}, \quad \beta = \frac{g_2}{|g|}, \tag{4.6.66}$$

where $|g| = \sqrt{g_1^2 + g_2^2}$, and

$$\frac{c^2 m_W^2}{\hbar^2} = \frac{a^2}{2} g_1^2, \quad \frac{c^2 m_Z}{\hbar^2} = \frac{a^2}{2}|g|^2. \tag{4.6.67}$$

7. The field equations governing $W^\pm$ and Z bosons are obtained from the equations (4.6.63) under the following transformation

$$\begin{pmatrix} W_\mu^+ \\ W_\mu^- \\ Z_\mu \\ A_\mu \end{pmatrix} = U \begin{pmatrix} W_\mu^1 \\ W_\mu^2 \\ W_\mu^3 \\ B_\mu \end{pmatrix} \qquad \text{for } U \text{ as in (4.6.66).} \tag{4.6.68}$$

In this case, the equations (4.6.63) become

$$\begin{pmatrix} \partial^\nu W_{\nu\mu}^+ - \left(\frac{m_W c}{\hbar}\right)^2 W_\mu^+ \\ \partial^\nu W_{\nu\mu}^- - \left(\frac{m_W c}{\hbar}\right)^2 W_\mu^- \\ \partial^\nu Z_{\nu\mu} - \left(\frac{m_Z c}{\hbar}\right)^2 Z_\mu \\ \partial^\nu A_{\nu\mu} \end{pmatrix} = \begin{pmatrix} \frac{g_1}{\sqrt{2}} J_\mu^+ \\ \frac{g_1}{\sqrt{2}} J_\mu^- \\ |g| J_\mu^0 \\ -e J_\mu^{em} \end{pmatrix} + \text{higher order terms.} \tag{4.6.69}$$

where $e = g_1 g_2 / |g|$, and

$$\begin{aligned} J_\mu^\pm &= \frac{1}{2}(J_\mu^1 \pm i J_\mu^2), \\ J_\mu^0 &= \frac{1}{2|g|^2}(g_1^2 J_\mu^3 - g_2^2 J_\mu^L - 2 g_2^2 J_\mu^R), \\ J_\mu^{em} &= \frac{1}{2}(J_\mu^3 + J_\mu^L + 2 J_\mu^R). \end{aligned}$$

Under the transformation (4.6.68),

$$
\begin{aligned}
Z_\mu &= \cos\theta_w W_\mu^3 + \sin\theta_w B_\mu, \\
A_\mu &= -\sin\theta_w W_\mu^3 + \cos\theta_w B_\mu,
\end{aligned} \tag{4.6.70}
$$

where θ_w is called the Weinberg angle, defined by

$$
\cos\theta_w = \frac{g_1}{|g|}, \quad \sin\theta_w = \frac{g_2}{|g|}.
$$

8. The Higgs field equation governed H^0 boson is given by (4.6.61), and at the ground state $\varphi = a$ which can be written as

$$
\begin{aligned}
\partial^\mu\partial_\mu\varphi - \left(\frac{m_H c}{\hbar}\right)^2\varphi =& \frac{1}{4}(\varphi + a)(g_1^2 W_\mu^a W^{\mu a} + g_2^2 B_\mu B^\mu - 2g_1 g_2 W_\mu^3 B_\mu) \\
&+ \lambda\varphi(\varphi^2 + 2a\varphi) + G_l(\bar{l}_L R + \overline{R} l_L),
\end{aligned} \tag{4.6.71}
$$

and the Higgs boson mass is

$$
\frac{m_H c}{\hbar} = \sqrt{2\lambda}a. \tag{4.6.72}
$$

4.6.6 Problems in WS theory

The classical electroweak theory provides a model with some experimental supports. However, this theory faces a number of problems, which are difficult, if not impossible, to resolve.

1. *Lack of weak force formulas.* This problem is that all weak interaction theories have to face, and it is also that all existed theories cannot solve.

In fact, in the original field equations (4.6.60) there are four gauge field components:

$$
W_\mu^1,\ W_\mu^2,\ W_\mu^3,\ B_\mu, \tag{4.6.73}
$$

and we don't know which of these potentials plays the role of weak interaction potential. In fact, with the mixed fields

$$
W_\mu^\pm, Z_\mu, A_\mu,
$$

it is even more difficult to determine the weak force.

If we combine the classical electroweak theory with PRI, and take

$$
W_\mu = \omega_1 W_\mu^1 + \omega_2 W_\mu^2 + \omega_3 W_\mu^3
$$

as the weak interaction potential, then from the field equations (4.6.60), we can only deduce the weak force potential in the following form:

$$
\Phi_w = \frac{g_1}{r}e^{-kr} \quad \text{or} \quad \Phi_w = -\frac{g_1}{r}e^{-kr}, \quad k = \frac{1}{\sqrt{2}}ag_1.
$$

It implies that the weak force is only repulsive or attractive, which is not consistent with experiments.

2. *Violation of PRI.* In the classical electroweak theory, a key ingredient is the linear combinations of W_μ^3 and B_μ as given by (4.6.70). By PRI,

$$\begin{aligned}&W_\mu^3 \ \text{ is the third component of a } SU(2) \ \text{ tensor } \{W_\mu^a\},\\&B_\mu \ \text{ is the } U(1) \ \text{ gauge field.}\end{aligned}$$

Hence, for the combinations of two different types of tensors:

$$\begin{aligned}Z_\mu &= \cos\theta_w W_\mu^3 + \sin\theta_w B_\mu,\\A_\mu &= -\sin\theta_w W_\mu^3 + \cos\theta_w B_\mu,\end{aligned}$$

their field equations (4.6.69) must vary under general $SU(2)$ generator transformations as follows

$$\widetilde{\sigma}_a = x_a^b \sigma_b. \tag{4.6.74}$$

In other words, such linear combinations violates PRI.

3. *Decoupling obstacle.* The classical electroweak theory has a difficulty for decoupling the electromagnetic and the weak interactions. In reality, electromagnetism and weak interaction often are independent to each other. Hence, as a unified theory for both interactions, one should be able to decouple the model to study individual interactions. However, the classical electroweak theory manifests a radical decoupling obstacle.

In fact, it is natural to require that under the condition

$$W_\mu^\pm = 0, \quad Z_\mu = 0, \tag{4.6.75}$$

the WS field equations (4.6.69) should return to the $U(1)$ gauge invariant Maxwell equations. But we see that

$$A_\mu = \cos\theta_w B_\mu - \sin\theta_w W_\mu^3,$$

where B_μ is a $U(1)$ gauge field, and W_μ^3 is a component of $SU(2)$ gauge field. Therefore, $A_\mu$is not independent of $SU(2)$ gauge transformation. In particular, the condition (4.6.75) means

$$W_\mu^1 = 0, \quad W_\mu^2 = 0, \quad W_\mu^3 = -\operatorname{tg}\theta_w B_\mu. \tag{4.6.76}$$

Hence, as we take the transformation (4.6.74), W_μ^a becomes

$$\begin{pmatrix}\widetilde{W}_\mu^1\\ \widetilde{W}_\mu^2\\ \widetilde{W}_\mu^3\end{pmatrix} = \begin{pmatrix}y_3^1 W_\mu^3\\ y_3^2 W_\mu^3\\ y_3^3 W_\mu^3\end{pmatrix}, \quad (y_a^b)^T = (x_a^b)^{-1}.$$

It implies that under a transformation (4.6.74), a nonzero weak interaction can be generated from a zero weak interaction field of (4.6.75)-(4.6.76):

$$\widetilde{W}^{\pm}_{\mu} \neq 0, \quad \widetilde{Z}_{\mu} \neq 0 \quad \text{as } y^a_3 \neq 0 \; (1 \leqslant a \leqslant 3),$$

and the nonzero electromagnetic field $A_\mu \neq 0$ will become zero:

$$\widetilde{A}_\mu = 0 \quad \text{as} \quad y^3_3 = \cot\theta_w.$$

Obviously, it is not reality.

4. *Artificial Higgs mechanism.* In the classical electroweak action (4.6.55)-(4.6.57), the Higgs sector $\mathscr{L}_H$ is not based on a first principle, and is artificially added into the action.

5. *Presence of a massless and charged boson* ϕ^+. In the WS theory, the Higgs scalar doublet $\phi = (\phi^+, \phi^0)$ contains a massless boson ϕ^+ with positive electric charge. Obviously there are no such particles in reality. In particular, the particle ϕ^+ is ”formally suppressed” in the WS theory by transforming it to zero in the WS theory. However, from a field theoretical point of view, this particle field still represents a particle. This is one of major flaws for the electroweak theory and for the standard model.

Chapter 5
Elementary Particles

The aims of this chapter are as follows:

1) to give a brief introduction to the particle physics,
2) to introduce the weakton model of elementary particles,
3) to address the mechanism of subatomic decays,
4) to introduce color algebra, as the mathematical basis for the color quantum number, and
5) to derive the structure of mediator clouds around subatomic particles.

Searching for the main constituents of matter has a long history going back to ancient Greeks, to Robert Boyle (1600s), John Dalton (early 1800s), J.J. Thomson and Ernest Rutherford (end of the 19th century), and Niels Bohr (1913).

In the current standard model of particle physics, all forms of matter are made up of 6 leptons and 6 quarks, and their antiparticles, which are treated as elementary particles. The forces are mediated by the mediators, including the photon mediating the electromagnetism, the vector bosons $W^{\pm}$, Z and the Higgs H^0 mediating the weak interaction, the eight gluons mediating the strong interaction, and the hypothetical graviton mediating gravity.

However, there are many challenging problems related to subatomic particles. One such problem is that why leptons do not participate in strong interactions. In fact, the most difficult challenge is associated with the puzzling decay and reaction behavior of subatomic particles. For example, the electron radiations and the electron-positron annihilation into photons or quark-antiquark pair clearly shows that there must be interior structure of electrons, and the constituents of an electron contribute to the making of photon or the quark in the hadrons formed in the process. In fact, all sub-atomic decays and reactions show clearly the following conclusion:

There must be interior structure of charged leptons, quarks and mediators.

This conclusion motivates us to propose a weakton model for sub-lepton, sub-quark, and sub-mediators (Ma and Wang, 2015b), based on the new field theory and the new insights for the weak and strong interactions presented in the last chapter.

One important theoretical basic for the weakton model is the field theory developed in the last chapter. In particular, the weak and strong charges are responsible for the weak and strong interactions. The confinement property demonstrated by the weak and strong interaction potentials give rise to the needed confinement for the weakton constituents of the composite particles.

Remarkably, the weakton model offers a perfect explanation for all sub-atomic decays. In particular, all decays are achieved by 1) exchanging weaktons and consequently exchanging newly formed quarks, producing new composite particles, and 2) separating the new composite particles by weak and/or strong forces.

One aspect of this decay mechanism is that now we know the precise constituents of particles involved in all decays/reactions both before and after the reaction. It is therefore believed that the new decay mechanism provides clear new insights for both experimental and theoretical studies.

This chapter is organized as follows. Sections 5.1 and 5.2 recall the basic knowledge of particle physics and the quark model. Sections 5.3 and 5.4 are based entirely on (Ma and Wang, 2015b), with the weakton model introduced in Section 5.3, and with the mechanism of subatomic decays and electron radiations presented in Section 5.4. The last section, Section 5.5 studies the color algebra associated with the color quantum number, leading to detailed structure on mediator clouds around subatomic particles. Section 5.5 is based entirely on (Ma and Wang, 2014b).

5.1 Basic Knowledge of Particle Physics

5.1.1 Classification of particles

In particle physics, subatomic particles are classified into two basic classes, bosons and fermions:

$$\begin{aligned} &\text{bosons} &&= \text{integral spin particles,} \\ &\text{fermions} &&= \text{fractional spin particles.} \end{aligned}$$

Also, based on their properties and levels, all subatomic particles are currently classified into four types:

$$\text{leptons}, \quad \text{quarks}, \quad \text{mediators}, \quad \text{hadrons}.$$

In the following, we recall basic definitions and the quantum characterizations of these particles; see among many others (Kane, 1987; Griffiths, 2008; Halzen and Martin, 1984).

1. *Leptons.* Leptons are fermions which do not participate in the strong interaction, and consist of three generations:

$$\begin{pmatrix} e \\ \nu_e \end{pmatrix}, \quad \begin{pmatrix} \mu, \\ \nu_\mu \end{pmatrix}, \quad \begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix},$$

where e, μ, τ are the electron, the muon, the tau, and ν_e, ν_μ, ν_τ are the e-neutrino, the μ-neutrino, the τ-neutrino. Together with their antiparticles, there are total 12 leptons:

particles:	(e^-, ν_e),	(μ^-, ν_μ),	(τ^-, ν_τ),
antiparticles:	$(e^+, \overline{\nu}_e)$,	$(\mu^+, \overline{\nu}_\mu)$,	$(\tau^+, \overline{\nu}_\tau)$.

2. *Quarks.* Based on the standard model, there are three generations of quarks consisting of 6 particles and 6 antiparticles, which participate all four fundamental interactions:

quarks:	(u,d),	(c,s),	(t,b),
antiquarks:	$(\overline{u},\overline{d})$,	$(\overline{c},\overline{s})$,	$(\overline{t},\overline{b})$,

where u,d,c,s,t,b are the up quark, down quark, charm quark, strange quark, top quark, and bottom quark.

The quark model asserts that three quarks are bounded together to form a baryon, and a pair of quark and antiquark are bounded to form a meson. As mentioned in Section 4.5.3, quarks are confined in hadrons, and no free quarks have been found in Nature. This phenomena is called the quark confinement, which has been very well explained by the layered formulas of strong interactions in the last chapter.

3. *Mediators.* The standard model shows that each interaction is associated with a class of field particles, called mediators. Hence, there are four classes of mediators:

Gravitation:	graviton G,
Electromagnetism:	photon γ,
Weak interaction:	vector and Higgs bosons $W^\pm, Z, H^\pm, H^0$,
Strong interaction:	gluons g^k $(1 \leqslant k \leqslant 8)$.

By the unified field theory introduced in the last chapter, there exists a natural duality between the interacting fields $\{g_{\mu\nu}, A_\mu, W^a_\mu, S^k_\mu\}$ and their dual fields $\{\phi^g_\mu, \phi^e, \phi^a_w, \phi^k_s\}$:

$$\begin{aligned}
& g_{\mu\nu} \leftrightarrow \phi^g_\mu, \\
& A_\mu \leftrightarrow \phi^e, \\
& W^a_\mu \leftrightarrow \phi^a_w, \quad 1 \leqslant a \leqslant 3, \\
& S^k_\mu \leftrightarrow \phi^k_s, \quad 1 \leqslant k \leqslant 8.
\end{aligned}$$

This duality leads to four classes of new dual particle fields, called the dual mediators:

$$\begin{array}{lcl}
\text{tensor graviton } G & \leftrightarrow & \text{vector graviton } g, \\
\text{vector photon } \gamma & \leftrightarrow & \text{scalar photon } \gamma_0, \\
\text{charged vector bosons } W^\pm & \leftrightarrow & \text{charged Higgs bosons } H^\pm, \\
\text{neutral vector boson } Z & \leftrightarrow & \text{neutral Higgs boson } H^0, \\
\text{vector gluons } g^k & \leftrightarrow & \text{scalar gluons } g^k_0,\ 1 \leqslant k \leqslant 8,
\end{array} \tag{5.1.1}$$

where the upper index k in g^k and g^k_0 represents the color index of gluons.

The mediator duality (5.1.1) is derived from the unified field theory. However, the duality (5.1.1) can also be naturally deduced from the weakton model postulated in Section 5.3. Namely, the mediator duality (5.1.1) is derived from two very different perspectives in two different theories. This can hardly be a coincident.

4. *Hadrons.* Hadrons are classified into two types: baryons and mesons. Baryons are fermions and mesons are bosons, which are all made up of quarks:

$$\text{Baryons} = q_i q_j q_k, \quad \text{mesons} = q_i \overline{q}_j,$$

where $q_k = \{u,d,c,s,t,b\}$. The quark constituents of main hadrons are listed as follows:

- Baryons with $J = \frac{1}{2}$:

$$\begin{array}{lll} p\ (uud), & n\ (udd), & \Lambda\ (s(du-ud)/\sqrt{2}), \\ \Sigma^+\ (uus), & \Sigma^-\ (dds), & \Sigma^0\ (s(du+ud)/\sqrt{2}), \\ \Xi^0\ (uss), & \Xi^-\ (dss). & \end{array}$$

- Baryons with $J = \frac{3}{2}$:

$$\begin{array}{llll} \Delta^{++}\ (uuu), & \Delta^+\ (uud), & \Delta^-\ (ddd), & \Delta^0\ (udd), \\ \Sigma^{*+}(uus), & \Sigma^{*-}(dds), & \Sigma^0(uds), & \\ \Xi^{*0}(uss), & \Xi^{*-}(dss), & \Omega^-(sss). & \end{array}$$

- Mesons with $J = 0$:

$$\begin{array}{lll} \pi^+(u\overline{d}), & \pi^-(\overline{u}d), & \pi^0((u\overline{u}-d\overline{d})/\sqrt{2}), \\ K^+(u\overline{s}), & K^-(\overline{u}s), & K^0(d\overline{s}), \\ \overline{K}^0(d\overline{s}), & \eta((u\overline{u}+d\overline{d}-2s\overline{s})/\sqrt{6}). & \end{array}$$

- Mesons with $J = 1$:

$$\begin{array}{lll} \rho^+(u\overline{d}), & \rho^-(\overline{u}d), & \rho^0((u\overline{u}-d\overline{d})/\sqrt{2}), \\ K^{*+}(u\overline{s}), & K^{*-}(\overline{u}s), & K^{*0}(d\overline{s}), \\ \overline{K}^{*0}(d\overline{s}), & \psi(c\overline{c}), & \gamma(b\overline{b}). \end{array}$$

5.1.2 Quantum numbers

It is easy to identify a macro-body, but not easy to identify a sub-atomic particle. Usually experimental physicists distinguish these sub-atomic particles by measuring their physical parameters. There are numerous types of particles, which obey various conservation laws. Each conservation law is characterized by one parameter. These physical parameters are called quantum numbers.

The main quantum numbers of particles are:

$$\begin{array}{lll} \text{mass } m, & \text{electric charge } Q, & \text{lifetime } \tau, \\ \text{spin } J, & \text{lepton number } L, & \text{baryon number } B, \\ \text{parity } \pi, & \text{isospin } (I, I_3), & \text{strange number } S, \\ G\text{-parity}, & \text{hypercharge } Y. & \end{array} \tag{5.1.2}$$

In addition, there are some new quantum numbers, including the weak charge Q_w, strong charge Q_s, and colour index k, which play an important role in in the weakton model introduced in Section 5.3.

We now briefly introduce the quantum numbers listed in (5.1.2).

1. *Mass.* This is a quantity to characterize the inertia, and also plays the role of gravitational charge. In the macro-world, masses are continuous in distribution, however in quantum world masses are discrete. The same particles have the same masses, but the different massive particles have a gap between their masses.

In non-quantum physics, the masses are additive, i.e. if a body A consists of two bodies B and C, then the mass m_A of A is the sum of masses m_B and m_C of B and C:

$$A = B + C \;\Rightarrow\; m_A = m_B + m_C. \tag{5.1.3}$$

But, the additive relation (5.1.3) is not valid in a quantum system.

Mass is a most important quantum number, which plays a role of an identity card for all massive particles.

2. *Electric charge Q.* This is an important quantum number. It is the source of electromagnetic force. In the four interaction charges: m, e, g_w, g_s, the electric charge e is unique one possessing positive and negative values. The electric charges are discrete, they appear only at an integral multiples of the electron charge e:[1]

$$Q = \pm ne \quad (n = 0, 1, 2, \cdots).$$

Electric charge is an additive conservation quantity. For any particle reaction:

$$A_1 + \cdots + A_n \rightarrow B_1 + \cdots + B_N,$$

we have

$$\sum_{k=1}^{n} Q_{A_k} = \sum_{k=1}^{N} Q_{B_k},$$

[1] In the weakton model of elementary particles, w^* carries $2/3$ electric charge, and w_1 and w_2 carry $-1/3$ and $-2/3$ electric charges respectively.

where Q_{A_k} and Q_{B_k} are the electric charges of particles A_k and B_k

3. *Lifetime of particles.* Due to the decay property for most particles, the lifetime becomes a quantum number. Except the long lifetime particles: the electron e, proton p, neutrino ν, photon γ, gluon g^k, graviton G, all other particles undergo decays.

Usual particles with decay have very short lifetime, and in general their lifetimes τ do not exceed 10^{-5}s:

$$\tau < 10^{-5}\text{s} \quad \text{on an average.}$$

The neutron n is special, and has a longer lifetime:

$$\tau = 885.7\text{s} \quad \text{on an average.}$$

4. *Spin J.* The spin is an intrinsic property of particles. Although it exhibits angular momentum characteristics, the spin does not represent particle rotation around its axis. However, for a better understanding, we may imaginarily illustrate the concept of spin in Figure 5.1. Also, the spin of a right-handed particle is denoted by $\uparrow$, and a left-handed particle by $\downarrow$.

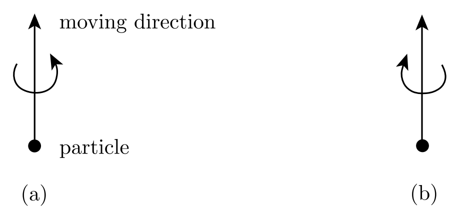


Figure 5.1 (a) right-hand spin($J>0$), and (b) left-hand spin($J<0$).

In quantum mechanics, the spin operator $\vec{S}$ is a 3-dimensional vector operator, defined as

$$\begin{aligned} &\vec{S} = s\hbar(\tau_1, \tau_2, \tau_3), \quad s \text{ the spin value,} \\ &\tau_k = \begin{pmatrix} \sigma_k & 0 \\ 0 & \sigma_k \end{pmatrix} \quad \text{for } 1 \leqslant k \leqslant 3, \end{aligned} \tag{5.1.4}$$

and σ_k are the Pauli matrices. The spin value s of the spin operator (5.1.4) represents the right-hand and left-hand spins. For example, if $\pm s$ $(s>0)$ are the eigenvalues of third component S_3,

$$S_3\psi_R = s\psi_R, \quad S_3\psi_L = -s\psi_L,$$

then the eigenfunctions ψ_R and ψ_L are the wave functions of the right-handed and left-handed particles with spin $J = s$.

5. *Lepton numbers L.* There are 3 types of lepton numbers: L_e, L_μ, L_τ, which are special parameters to characterize the three generations of leptons (e, ν_e), (μ, ν_μ), (τ, ν_μ). They

only take three values $L = \pm 1, 0$. For all non-lepton particles, their lepton number is zero, and

$$L_e = \begin{cases} 1 & \text{for } e^-, \nu_e, \\ -1 & \text{for } e^+, \overline{\nu}_e, \\ 0 & \text{others,} \end{cases}$$

$$L_\mu = \begin{cases} 1 & \text{for } \mu^-, \nu_\mu, \\ -1 & \text{for } \mu^+, \overline{\nu}_\mu, \\ 0 & \text{others,} \end{cases}$$

$$L_\tau = \begin{cases} 1 & \text{for } \tau^-, \nu_\tau, \\ -1 & \text{for } \tau^+, \overline{\nu}_\tau, \\ 0 & \text{others.} \end{cases}$$

The lepton numbers are additive and conservative quantities.

6. *Baryon number B.* The hadrons are classified into baryons and mesons. Baryons are fermions and mesons are bosons. Also, baryons have an additive and conservative quantum number: the baryon number B, defined by

$$B = \begin{cases} 1 & \text{for a baryon,} \\ -1 & \text{for an antibaryon,} \\ 0 & \text{for all other particles.} \end{cases}$$

7. *Parity π.* Due to the Noether Theorem 2.38, each symmetry is associated with a conservation law. The parity is a conservative quantum number corresponding to the spatial reflection symmetry:

$$x \to -x \quad (x \in \mathbb{R}^3).$$

Parity π only takes two values:

$$\pi = \pm 1.$$

The weak interaction violates parity; see (Lee and Yang, 1956; Wu, Ambler, Hayward, Hoppes and Hudson, 1957). The parities π of most particles are obtained by two methods: the experimental and the artificial means.

The parity is a multiplication quantum number. For example, for an N particles system:

$$A = A_1 + \cdots + A_N,$$

its tatol parity π_A is given by

$$\pi_A = (-1)^l \pi_{A_1} \cdots \pi_{A_N},$$

where l is the sum of orbital quantum numbers of all particles, and π_{A_k} is the parity of A_k particle.

The parity conservation is defined as follows. For a particle reaction:

$$A_1 + \cdots + A_N \longrightarrow B_1 + \cdots + B_K, \tag{5.1.5}$$

the total parities of both sides of (5.1.5) are the same:

$$(-1)^{l_1}\pi_{A_1}\cdots\pi_{A_N} = (-1)^{l_2}\pi_{B_1}\cdots\pi_{B_K}.$$

In Subsection 5.1.4, we shall introduce the violation of parity conservation in the weak interaction in detail.

8. *Isospin* (I, I_3). The isospin was first presented in 1932 by Heisenberg, and was used to describe that the strong interaction between protons and neutrons, independently of electric charges. Later, along with the development of particle physics, it was discovered that the isospin is a good quantum number for all hadrons.

Isospin has two components (I, I_3), where I is called the isospin and I_3 is the third component of isospin. In a many-particle system, the isospin I obeys the vectorial additive rule and I_3 obeys the usual additive rule. We shall introduce the isospin again in Subsection 5.1.4.

9. *Strange number S.* In 1947, C. D. Rochester and C. C. Butler first discovered the strange particle K^0, a neutral meson, from cosmic rays, and later many strange particles such as $K^{\pm}, \Lambda, \Sigma^{\pm}, \Xi^-, \Xi^0, \Omega^-$ were found. They are named the "strange" because these particles possess a kind quantum number called the strange number.

The strange number S is an additive quantum number, which is conserved only in the electromagnetic and strong interactions, and takes the integral values:

$$S = 0, 1, 2, 3, \cdots.$$

10. *Hypercharge Y.* Hypercharge Y is another quantum number for hadrons, defined by

$$Y = S + B,$$

where S and B are the strange number and the baryon number.

11. *G-parity.* G parity is a conservative quantity only for strong interactions under the G-transformation

$$G\psi = \hat{C}e^{i\pi I_2}\psi,$$

where $\hat{C}$ is the transformation of particles and antiparticles:

$$\hat{C}A = \overline{A}, \quad \hat{C}\overline{A} = A, \quad \overline{A} \text{ the antiparticle of } A,$$

and I_2 is the second axis of the isospin space.

5.1.3 Particle transitions

Particle transition is the main dynamic form of sub-atomic particles, which includes particle decays, scatterings and various radiations. It is called ”transition” because in these processes particles will undergo a transition from an energy level to other energy levels.

The particle transition is a crucial way for us to understand the particle structures and properties. In particular, these transition processes can reveal the mysteries of the weak and the strong interactions.

1. *Decays.* Particle decay means that a particle is spontaneously decomposed into several other particles. The most remarkable example is the β-decay, i.e., a neutron n decays to a proton p, an electron e^- and an anti-neutrino $\overline{\nu}_e$:

$$n \to p + e^- + \overline{\nu}_e. \tag{5.1.6}$$

Current list of discovered particle consists of hundreds of numbers. Except electrons and protons, all massive particles will decay therefore they are finite lifetime. The massless particles, such as the photon, the gluons and the neutrinos, do not decay, and have infinite lifetime.

All decays obey the following laws and rules:

1) Decays are caused by the three fundamental forces: the electromagnetic, weak, strong interactions, and, therefore, are mainly classified into two types: weak interaction decays and strong interaction decays;

2) Decays always take place from higher masses toward to lower masses. For example in the β-decay (5.1.6), the masses of particles on the right-hand side are smaller than the mass of the neutron n.

3) Both strong and weak interaction decays have to obey some basic conservation laws, as shown in the next subsection.

Remark 5.1 The following transition

$$p \to n + e^+ + \nu_e \tag{5.1.7}$$

is often called β-decay as well. In fact, in Nature the process (5.1.7) cannot spontaneously take place, and it always occurs under certain energetic excitation. Hence, (5.1.7) is an excited scattering. □

2. *Scattering.* If the decay is a spontaneous behavior, then scattering is a forced behavior of particles under certain force actions, such as collisions and energetic excitations. The transition (5.1.7) is a scattering, and the precise decay mechanism is written as

$$p + \gamma \to n + e^+ + \nu_e.$$

In general, a scattering reaction is written as

$$A_1 + \cdots + A_N \to B_1 + \cdots + B_K, \tag{5.1.8}$$

where A_n $(1 \leqslant n \leqslant N)$ are the initial particles, and B_k $(1 \leqslant k \leqslant K)$ are the final particles.

There are two forms of scatterings: the elastic and non-elastic scatterings, both of which are caused by the three interactions: the electromagnetic, the weak and the strong interactions. Collision is the most important experimental method to detect new particles and new phenomena. The scatterings (5.1.8) satisfy various conservation relations.

Here are a few important scatterings in the history of quantum physics:

Compton scattering:	$\gamma + e^- \to \gamma + e^-$,
Pair annihilation:	$e^+ + e^- \to \gamma + \gamma$,
Pair creation:	$\gamma + \gamma \to e^+ + e^-$,
Moller scattering:	$e^- + e^- \to e^- + e^-$,
Bhahba scattering:	$e^- + e^+ \to e^- + e^+$.
Deep inelastic scattering :	$e^- + p \to e^- + p + \pi^0$.

3. *Radiations.* Radiations include the electromagnetic radiation and the gluon radiation. The first one is that electrons emit photons:

$$e^- \to e^- + \gamma, \tag{5.1.9}$$

and the second one is that quarks emit gluons

$$q \to q + g^k, \tag{5.1.10}$$

Electromagnetic radiations are caused in two scenarios:

1) As an electron changes its velocity, it emits photons. This radiation is called the bremsstrahlung. The radiation energy in unit time can be expressed by the formula:

$$W = \frac{2}{3}\frac{e^2}{c^3}\overline{a}^2 \qquad \text{with } \overline{a} \text{ being the average acceleration.}$$

2) As an electron at a higher energy level E_1 undergoes a transition to a lower energy level $E_0 < E_1$, it emits photons. The energy ε of the emitting photons equals to the difference of energy levels:

$$\varepsilon = E_1 - E_0.$$

In summary, there are three types of particle transitions: decays, scatterings, and radiations. They are the main dynamic behavior for micro-particles, and reveal the interior structure of particles, and provide crucial information about the three interactions associated with the electromagnetic, the weak, and the strong forces.

Here we list some principal decay forms:

- Lepton decays:

$$\begin{aligned}
&\mu^- \to e^- + \overline{\nu}_e + \nu_\mu, \\
&\mu^+ \to e^+ + \nu_e + \overline{\nu}_\mu, \\
&\tau^- \to e^- + \overline{\nu}_e + \nu_\tau, \\
&\tau^- \to \mu^- + \overline{\nu}_\mu + \nu_\tau, \\
&\tau^- \to \pi^- + \nu_\tau, \\
&\tau^- \to \rho^- + \nu_\tau, \\
&\tau^- \to K^- + \nu_\tau.
\end{aligned} \tag{5.1.11}$$

- Quark decays:

$$\begin{aligned}
&d \to u + e^- + \overline{\nu}_e, \\
&s \to u + e^- + \overline{\nu}_e, \\
&s \to d + \gamma, \\
&c \to d + \overline{s} + u.
\end{aligned} \tag{5.1.12}$$

- Mediator decays:

$$\begin{aligned}
&W^+ \to e^+ + \nu_e, \quad \mu^+ + \nu_\mu, \quad \tau^+ + \nu_\tau, \\
&W^- \to e^- + \overline{\nu}_e, \quad \mu^- + \overline{\nu}_\mu, \quad \tau^- + \overline{\nu}_\tau, \\
&Z \to e^+ + e^-, \quad \mu^+ + \mu^-, \quad \tau^+ + \tau^-.
\end{aligned} \tag{5.1.13}$$

- Baryon decays:

$$\begin{aligned}
&n \to p + e^- + \overline{\nu}_e, \\
&\Lambda \to p + \pi^-, \quad n + \pi^0, \\
&\Sigma^+ \to p + \pi^0, \quad n + \pi^+, \\
&\Sigma^- \to n + \pi^-, \\
&\Sigma^0 \to \Lambda + \gamma, \\
&\Xi^- \to \Lambda + \pi^-, \\
&\Delta^{++} \to p + \pi^+, \\
&\Delta^+ \to p + \pi^0,
\end{aligned} \tag{5.1.14}$$

$$\begin{aligned}
&\Delta^- \to n + \pi^-, \\
&\Delta^0 \to n + \pi^0, \\
&\Sigma^{*\pm} \to \Sigma^{\pm} + \pi^0, \\
&\Sigma^{*0} \to \Sigma^0 + \pi^0, \\
&\Xi^{*-} \to \Xi^- + \pi^0, \\
&\Xi^{*0} \to \Xi^0 + \pi^0.
\end{aligned}$$

- Meson decays:

$$\begin{aligned}
&\pi^+ \to \mu^+ + \nu_\mu, \\
&\pi^- \to \mu^- + \overline{\nu}_\mu, \\
&\pi^0 \to 2\gamma, \\
&K^+ \to \mu^+ + \nu_\mu, \quad \pi^+ + \pi^0, \quad \pi^+ + \pi^+ + \pi^-, \\
&K^- \to \mu^- + \overline{\nu}_\mu, \quad \pi^- + \pi^0, \quad \pi^- + \pi^- + \pi^+, \\
&K^0 \to \pi^+ + e^- + \overline{\nu}_e, \quad \pi^+ + \pi^-, \quad \pi^+ + \pi^- + \pi^0, \\
&\eta \to 2\gamma, \quad \pi^+ + \pi^- + \pi^0, \\
&\rho^\pm \to \pi^\pm + \pi^0, \\
&\rho^0 \to \pi^+ + \pi^-, \\
&K^{*\pm} \to K^\pm + \pi^0, \\
&K^{*0} \to K^0 + \pi^0, \\
&\omega \to \pi^0 + \gamma, \quad \pi^+ + \pi^- + \pi^0.
\end{aligned} \tag{5.1.15}$$

5.1.4 Conservation laws

All particle transitions can be expressed in the form:

$$A_1 + \cdots + A_N \to B_1 + \cdots + B_K. \tag{5.1.16}$$

They are driven by the three fundamental forces: the electromagnetic, the weak, and the strong interactions. Particle transitions have to obey certain basic conservation laws in the sense that for certain conservative quantum number q, its total values on both sides of (5.1.16) are the same:

$$q_A = q_B. \tag{5.1.17}$$

In addition, some conservation laws may not be valid for all interactions. For example, parity is valid only in the strong and the electromagnetic interactions, and is violated in the weak interaction.

We now introduce the basic conservation laws.

1. *Energy conservation.* Energy is an additive quantum number. The energy of a particle is given by the formula

$$\varepsilon = c\sqrt{P^2 + m^2c^2},$$

where P is the momentum, and m the mass. Energy conservation is a law that all transitions must obey. Namely, for (5.1.16),

$$\sum_{n=1}^{N} \varepsilon_{A_n} = \sum_{k=1}^{K} \varepsilon_{B_k},$$

where ε_{A_n} and ε_{B_k} are the energies of particles A_n and B_k.

2. *Momentum conservation.* Momentum is vectorial additive. For two particles A_1 and A_2, their total momentum $P = (P_1, P_2, P_3)$ equals to

$$P_k = P_k^{A_1} + P_k^{A_2}, \quad 1 \leqslant k \leqslant 3,$$

where P^{A_1} and P^{A_2} are the momentums of A_1 and A_2.

Momentum conservation is a universal conservation law, and for (5.1.16) the following equality holds true:

$$\sum_{n=1}^{N} P_j^{A_n} = \sum_{k=1}^{K} P_j^{B_k}, \quad 1 \leqslant j \leqslant 3.$$

3. *Angular momentum conservation.* This quantum number is vectorial additive. The angular momentum $\vec{J}$ of a particle consists of the orbital angular momentum $\overline{L}$ and spin $\vec{S}$:

$$\vec{J} = \vec{L} + \vec{S}, \quad \vec{L} = \vec{r} \times P, \quad P \text{ is the momentum.}$$

Angular momentum conservation law is also a universal conservation law, and for (5.1.16),

$$\sum_{n=1}^{N} \vec{J}_{A_n} = \sum_{k=1}^{K} \vec{J}_{B_k}.$$

4. *Other universal conservation laws.* In particle physics, the following quantum numbers are additive and conservative in all interactions: the electric charge Q_e, the lepton numbers L_e, L_μ, L_τ, and the baryon number B.

Table 5.1 lists the conservation or non-conservation properties of quantum numbers for the three interactions.

Remark 5.2 *CP* combines charge conjugation C and parity P, and CPT combines CP and time reversal T.

Table 5.1 Conservation Laws

Conservative Quantities		Strong	Electromagnetic	Weak
Energy E		Yes	Yes	Yes
Momentum P		Yes	Yes	Yes
Angular Momentum J		Yes	Yes	Yes
Electric charge Q_e		Yes	Yes	Yes
Lepton Numbers L_e, L_μ, L_τ		Yes	Yes	Yes
Baryon Number B		Yes	Yes	Yes
Strange Number S		Yes	Yes	No
Parity π		Yes	Yes	No
Isospin	I	Yes	No	No
	I_3	Yes	Yes	No
G Parity G		Yes	No	No
Charge Conjugation C		Yes	Yes	No
Time Reversal T		Yes	Yes	No
CP		Yes	Yes	No
CPT		Yes	Yes	Yes

Combined CPT conservation was proved by G. Lüders and W. Pauli independently in 1954, called CPT theorem. The CPT conservation is an important result in quantum physics. Based on CPT theorem, we can deduce that particles and antiparticles have the same masses and lifetimes, and their magnetic moments are reversal with the same magnitudes. □

5.1.5 Basic data of particles

Here, we list the basic data for leptons, quarks, and hadrons in Tables 5.2-5.5. The units are: mass in MeV/c^2, lifetime in seconds, and electric charge in the unit of proton charge.

Note that the quantum number Q_w in Table 5.2 represents the weak charge number. Here the values of Q_w are based on the weakton model in Section 5.3.

Table 5.2 Leptons $\left(\text{spin } J = \frac{1}{2}\right)$

Leptons	m	Q_e	Q_w	L_e	L_μ	L_τ	τ
e^-	0.51	−1	3	1	0	0	∞
e^+	0.51	1	3	−1	0	0	
μ^-	105.7	−1	3	0	1	0	2.2×10^{-6}
μ^+	105.7	1	3	0	−1	0	
τ^-	1777	−1	3	0	0	1	2.9×10^{-13}
τ^+	1777	1	3	0	0	−1	
ν_+	0	0	1	1	0	0	∞
$\overline{\nu}_e$	0	0	1	−1	0	0	∞
ν_μ	0	0	1	0	1	0	∞
$\overline{\nu}_\mu$	0	0	1	0	−1	0	∞
ν_τ	0	0	1	0	0	1	∞
$\overline{\nu}_\tau$	0	0	1	0	0	−1	∞

Also, all neutrinos possess left-hand spin with $J = -\frac{1}{2}$, and antineutrinos possess right-hand spin with $J = \frac{1}{2}$. Namely, neutrinos move at the speed of light.

In addition, the data of $W^{\pm}, Z, H^0$ are given as follows.

$$\begin{aligned} &W^{\pm}: \ m = 80.4\ \text{GeV}/c^2, && Q_e = \pm 1, && \tau = 3.11 \times 10^{-25},\\ &Z: \ m = 91.2\ \text{GeV}/c^2, && Q_e = 0, && \tau = 2.67 \times 10^{-25},\\ &H^0: \ m = 126\ \text{GeV}/c^2, && Q_e = 0, && \tau = 10^{-21}. \end{aligned}$$

Table 5.3 Quarks $\left(\text{spin } J = \frac{1}{2}\right)$

Quarks	m	Q_e	Q_s	Q_w	B	I	I_3	S	Y
d	7	$-1/3$	1	3	$1/3$	$1/2$	$-1/2$	0	$1/3$
u	3	$2/3$	1	3	$1/3$	$1/2$	$1/2$	0	$1/3$
s	120	$-1/3$	1	3	$1/3$	0	0	-1	$-2/3$
c	1200	$2/3$	1	3	$1/3$	$1/2$	$1/2$	0	$1/3$
b	4300	$-1/3$	1	3	$1/3$	$1/2$	$-1/2$	0	$1/3$
t	174000	$2/3$	1	3	$1/3$	0	0	1	$4/3$

5.2 Quark Model

5.2.1 Eightfold way

We see from Tables 5.4 and 5.5 that there are many hadrons, and in fact their number arrives at hundreds. The abundance of hadrons implies that there must be some rules to classify

Table 5.4 Baryons

Baryons	J	m	Q_e	Q_w	Q_s	B	I	I_3	S	Y	τ
p	$1/2$	938.3	1	9	3	1	$1/2$	$1/2$	0	1	∞
n	$1/2$	939.6	0	9	3	1	$1/2$	$-1/2$	0	1	885.7
Λ	$1/2$	1115.7	0	9	3	1	0	0	-1	0	2.7×10^{-10}
Σ^+	$1/2$	1189.4	1	9	3	1	1	1	-1	0	8×10^{-11}
Σ^-	$1/2$	1197.5	-1	9	3	1	1	0	-1	0	1.5×10^{-10}
Σ^0	$1/2$	1192.6	0	9	3	1	1	-1	-1	0	7.4×10^{-20}
Ξ^-	$1/2$	1321.3	-1	9	3	1	$1/2$	$1/2$	-2	-1	1.6×10^{-10}
Ξ^0	$1/2$	1314.8	0	9	3	1	$1/2$	$-1/2$	-2	-1	2.9×10^{-10}
Δ^{++}	$3/2$	1230	2	9	3	1	$3/2$	$3/2$	0	1	5.6×10^{-24}
Δ^+	$3/2$	1231	1	9	3	1	$3/2$	$1/2$	0	1	5.6×10^{-24}
Δ^-	$3/2$	1234	-1	9	3	1	$3/2$	$-1/2$	0	1	5.6×10^{-24}
Δ^0	$3/2$	1232	0	9	3	1	$3/2$	$-3/2$	0	1	5.6×10^{-24}
Σ^{*+}	$3/2$	1383	1	9	3	1	1	1	-1	0	1.8×10^{-23}
Σ^{*-}	$3/2$	1387	-1	9	3	1	1	0	-1	0	1.8×10^{-23}
Σ^{*0}	$3/2$	1384	0	9	3	1	1	-1	-1	0	1.8×10^{-23}
Ξ^{*-}	$3/2$	1535	-1	9	3	1	$1/2$	$1/2$	-2	-1	6.9×10^{-23}
Ξ^{*0}	$3/2$	1532	0	9	3	1	$1/2$	$-1/2$	-2	-1	6.9×10^{-23}
Ω^-	$3/2$	1672	-1	9	3	1	0	0	-3	-2	8.2×10^{-11}

them into groups based on certain common properties. The Eightfold Way can be regarded as a successful classification for hadrons. In fact, the Eightfold Way, together with the irreducible representations of $SU(3)$, has played a navigation role for the introduction of the quark model.

Table 5.5 Mesons

Mesons	J	m	Q_e	Q_w	Q_s	B	I	I_3	S	Y	τ
π^0	0	135	0	6	2	0	1	0	0	0	8.4×10^{-17}
π^+	0	139.6	1	6	2	0	1	1	0	0	2.6×10^{-8}
π^-	0	139.6	-1	6	2	0	1	-1	0	0	2.6×10^{-8}
K^+	0	493.7	1	6	2	0	1/2	1/2	1	1	1.2×10^{-8}
K^-	0	493.7	-1	6	2	0	1/2	$-1/2$	1	1	1.2×10^{-8}
K^0	0	497.7	0	6	2	0	1/2	$-1/2$	1	1	5.1×10^{-8}
η	0	547.5	0	6	2	0	0	0	0	0	5.1×10^{-19}
η'	0	957.8	0	6	2	0	0	0	0	0	3.2×10^{-21}
D^0	0	1864.5	0	6	2	0	1/2	$-1/2$	0	0	4.1×10^{-13}
D^+	0	1869.3	1	6	2	0	1/2	1/2	0	0	10^{-12}
D^-	0	1869.3	-1	6	2	0	1/2	$-1/2$	0	0	10^{-12}
B^+	0	5279	1	6	2	0	1/2	1/2	0	0	1.6×10^{-12}
B^-	0	5279	-1	6	2	0	1/2	$-1/2$	0	0	1.6×10^{-12}
B^0	0	5279.4	0	6	2	0	1/2	$-1/2$	0	0	1.5×10^{-12}
ρ^+	1	775.5	1	6	2	0	1	1	0	0	4×10^{-24}
ρ^-	1	775.5	-1	6	2	0	1	-1	0	0	4×10^{-24}
ρ^0	1	775.5	0	6	2	0	1	0	0	0	4×10^{-24}
K^{*+}	1	894	1	6	2	0	1/2	1/2	1	1	10^{-23}
K^{*-}	1	894	-1	6	2	0	1/2	$-1/2$	1	1	10^{-23}
K^{*0}	1	894	0	6	2	0	1/2	$-1/2$	1	1	10^{-23}
ω	1	782.6	0	6	2	0	0	0	0	0	8×10^{-23}
ψ	1	3097	0	6	2	0	0	0	0	0	7×10^{-21}
D^{*+}	1	2008	1	6	2	0	1/2	1/2	0	0	3×10^{-21}
D^{*-}	1	2008	-1	6	2	0	1/2	$-1/2$	0	0	3×10^{-21}
D^{*0}	1	2008	0	6	2	0	1/2	$-1/2$	0	0	3×10^{-21}

The Eightfold Way was introduced by Gell-Mann and Ne'eman independently in 1961. This scheme arranges the baryons and mesons into certain geometric patterns, according to their charge, strangeness, hypercharge and isospin I_3. These geometric patterns include hexagons and triangles, and all particles put in each diagram are considered as a class.

The first group of hadrons consists of the eight lightest baryons:

$$n, p, \Sigma^-, \Sigma^0, \Sigma^+, \Lambda, \Xi^-, \Xi^0, \tag{5.2.1}$$

which fit into a hexagonal array with two particles Σ^0 and Λ at the center; see Figure 5.2.

The eight baryons (5.2.1) with $J = 1/2$ are known as the baryon octet. Note that particles in Figure 5.2 range with charges lie along the downward-sloping diagonal line: $Q = 1$ for p and $\Sigma^+, Q = 0$ for n, Λ, Σ^0 and Ξ^0, and $Q = -1$ for Σ^- and Ξ^-. Horizontal lines associate particles of the same strangeness $S = 0$ and hyper-charge $Y = 1$ for the neutron and proton, $S = -1$ and $Y = 0$ for $\Sigma^-, \Sigma^0, \Lambda$ and $\Sigma^+, S = -2$ and $Y = -1$ for Ξ^- and Ξ^0. The third component I_3 of isospin indicates particles in column.

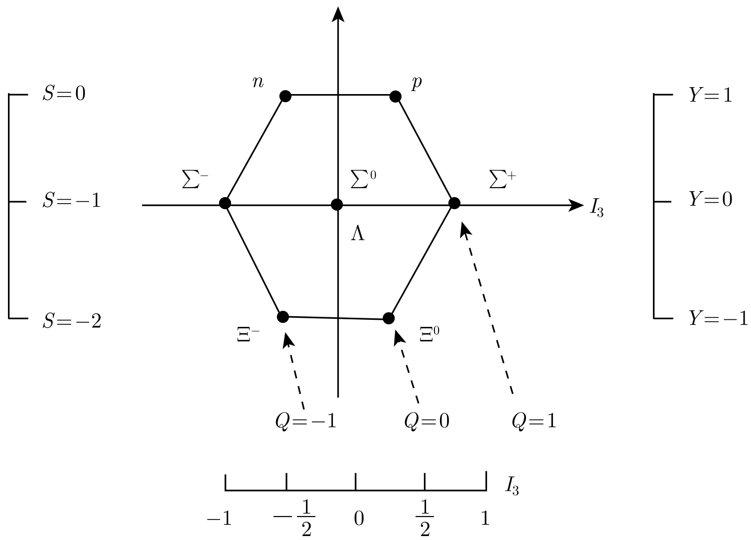


Figure 5.2 Baryon octet

The second group of hadrons consists of the eight lightest mesons:

$$\pi^+, \pi^0, \pi^-, K^+, K^0, K^-, \overline{K}^0, \eta. \tag{5.2.2}$$

In the same fashion as the baryon octet (5.2.1), this group (5.2.2) fits into a hexagonal array with π^0 and η at the center; see Figure 5.3.

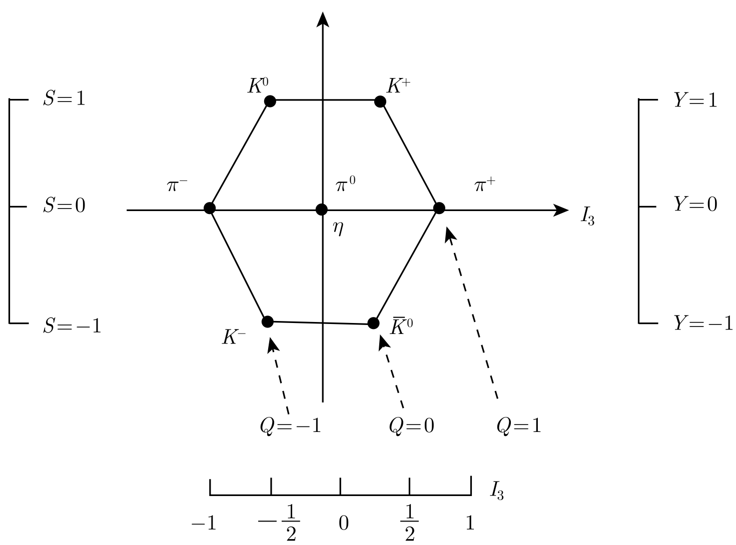


Figure 5.3

The third group of hadrons consists of ten baryons:

$$\Delta^{++}, \Delta^{+}, \Delta^{-}, \Delta^{0}, \Sigma^{*+}, \Sigma^{*-}, \Sigma^{*0}, \Xi^{*-}, \Xi^{*0}, \Omega^{-}, \tag{5.2.3}$$

which are fited into a triangular array as in Figure 5.4.

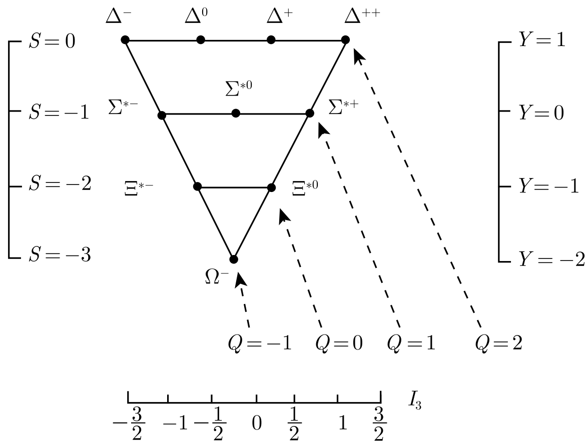


Figure 5.4

The fourth group of hadrons is given by the eight messons:

$$\rho^{+}, \rho^{-}, \rho^{0}, K^{*+}, K^{*-}, K^{*0}, \overline{K}^{*0}, \omega, \tag{5.2.4}$$

which are arranged as in Figure 5.5.

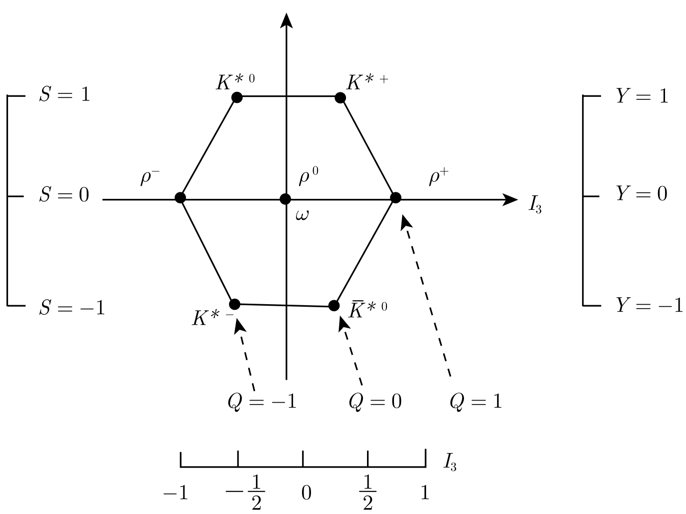


Figure 5.5

Remark 5.3 The Eightfold Way based on the patterns given by Figures 5.2-5.5 provides a beautiful classification for hadrons. However, the most important point is that the Eightfold way and the irreducible representation of $SU(N)$ also provide key clues to discover the quark model for the hadron structure.

5.2.2 Irreducible representations of $SU(N)$

To better understand the process from the Eightfold way to the quark model, it is necessary to know the irreducible representation of $SU(N)$ and its connection with particle physics. We proceed in a few steps as follows.

1. *Irreducible representation of Lie groups.* We begin with the definition of representation of abstract groups. Let G and M be two groups, and M consist of N-th order real or complex matrices satisfying certain properties. Let H be a mapping from G to M:

$$H: \ G \to M, \tag{5.2.5}$$

preserving the multiplication:

$$H(g_1 \cdot g_2) = H(g_1)H(g_2), \qquad \forall g_1, g_2 \in G. \tag{5.2.6}$$

Then, the image $H(G) \subset M$ of group G is called an N-dimensional representation of G.

Because M is a group of matrices, for any $g \in G, H(g) \in M$ is an N-th order matrix, written as

$$H(g) = \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}.$$

If there exists a matrix A (not necessary in M) such that $A^{-1}H(g)A \in M$ for all $g \in G$, and

$$AH(g)A^{-1} = \begin{pmatrix} H_1(g) & & 0 \\ & \ddots & \\ 0 & & H_n(g) \end{pmatrix}, \qquad \forall g \in G, \tag{5.2.7}$$

where $H_k(g)$ $(1 \leqslant k \leqslant n)$ are m_k-th order matrices with $\sum\limits_{k=1}^{n} m_k = N$, then the representation $H(G)$ is reducible, and each block matrix $H_k(g)$ in (5.2.7) is a smaller representation of G. In mathematics, (5.2.7) can be equivalently expressed as

$$H(G) = H_1(G) \oplus \cdots \oplus H_n(G). \tag{5.2.8}$$

If all block matrices $H_k(G)$ $(1 \leqslant k \leqslant n)$ in (5.2.8) cannot be split into smaller pieces anymore, then the direct sum of all sub-representations $H_k(G)$ as

$$H_1(G) \oplus \cdots \oplus H_n(G)$$

is called an irreducible representation of G, which can be simply written in the following form

$$H(G) = m_1 \oplus \cdots \oplus m_n, \tag{5.2.9}$$

where m_k is the order of $H_k(G)$.

2. *Fundamental representation* $SU(N)$. Let G be a linear transformation group made up of all linear norm-preserving mappings of $\mathbb{C}^N$:

$$g: \ \mathbb{C}^N \to \mathbb{C}^N. \tag{5.2.10}$$

It is known that for each linear operator $g \in G$ as defined in (5.2.10), there is a unique matrix $U \in SU(N)$ such that

$$g(\psi) = U\psi, \qquad \forall \psi \in \mathbb{C}^N. \tag{5.2.11}$$

Hence, relation (5.2.11) provides a correspondence

$$g \ \mapsto \ U \qquad \text{for } g \in G \ \text{ and } U \in SU(N),$$

which is a one to one and onto mapping

$$H: \ G \to \ SU(N), \tag{5.2.12}$$

and satisfies the multiplication relation (5.2.6).

Usually, the representation given by (5.2.12):

$$SU(N) = H(G)$$

is called an N-dimensional fundamental representation of linear norm-preserving transformation group G, which for simplicity is denoted by $SU(N)$.

3. *Conjugate representation* $SU(\overline{N})$. The conjugate group $SU(\overline{N})$ of $SU(N)$ is called the conjugate representation, expressed as

$$SU(\overline{N}) = \{\overline{U} \mid U \in SU(N)\}, \tag{5.2.13}$$

where $\overline{U}$ is the complex conjugate of U.

If $SU(N)$ and $SU(\overline{N})$ are regarded as linear norm-preserving transformation groups of N-dimensional complex space $\mathbb{C}^N$, then they represent such linear operators as follows. Let

$$\{e_1, \cdots, e_N\} \subset \mathbb{C}^N$$

constitute a complex orthogonal basis of $\mathbb{C}^N$, i.e.

$$\mathbb{C}^N = \left\{ \sum_{k=1}^{N} c_k e_k \middle| c_k \in \mathbb{C} \right\}.$$

Then the conjugate space $\overline{\mathbb{C}}^N$ of $\mathbb{C}^N$ can be written as

$$\overline{\mathbb{C}}^N = \left\{ \sum_{k=1}^{N} \beta_k \overline{e}_k \middle| \beta_k \in \mathbb{C} \right\}.$$

Thus, each matrix $U \in SU(N)$ is a linear transformation:

$$U: \mathbb{C}^N \to \mathbb{C}^N, \tag{5.2.14}$$

and each $\overline{U} \in SU(\overline{N})$ gives

$$\overline{U}^*: \overline{\mathbb{C}}^N \to \overline{\mathbb{C}}^N. \tag{5.2.15}$$

Remark 5.4 In particle physics, a complex orthogonal basis $\{e_1, \cdots, e_N\}$ of $\mathbb{C}^N$ stands for N different particles, and its conjugate basis $\{\overline{e}_1, \cdots, \overline{e}_N\}$ stands for the N antiparticles. Hence we have

$$\begin{aligned} \mathbb{C}^N &= \text{ the space of all states of particles } e_1, \cdots, e_N, \\ \overline{\mathbb{C}}^N &= \text{ the space of all states of antiparticles } \overline{e}_1, \cdots, \overline{e}_N. \end{aligned} \tag{5.2.16}$$

Thus, the mappings $U \in SU(N)$ in (5.2.14) stand for state transformations of particles $e_1, \cdots, e_N$, and $U^* \in SU(\overline{N})$ for state transformations of antiparticles $\overline{e}_1, \cdots, \overline{e}_N$. □

4. *Tensor product of matrices.* In quantum physics we often see tensor products of matrices. Here we give their definition. Let A, B be two matrices given by

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mm} \end{pmatrix}.$$

Then the tensor product $A \otimes B$ is defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}, \tag{5.2.17}$$

where $a_{ij}B$ are the block matrices

$$a_{ij}B = \begin{pmatrix} a_{ij}b_{11} & \cdots & a_{ij}b_{1m} \\ \vdots & & \vdots \\ a_{ij}b_{m1} & \cdots & a_{ij}b_{mm} \end{pmatrix}.$$

Hence $A \otimes B$ is an $(n \times m)$-th order matrix.

5. *Irreducible representation of $SU(N)$.* In the quark model, we shall meet the notations:

$$\begin{aligned} \text{meson} &= 3 \otimes \overline{3}, \\ \text{baryon} &= 3 \otimes 3 \otimes 3, \end{aligned} \tag{5.2.18}$$

representing the tensor products:

$$\begin{aligned} &\text{meson} = SU(3)\otimes SU(\overline{3}),\\ &\text{baryon} = SU(3)\otimes SU(3)\otimes SU(3). \end{aligned} \tag{5.2.19}$$

To understand the implications of (5.2.18) and (5.2.19), we have to know the irreducible representation of the tensor products

$$\underbrace{SU(N)\otimes\cdots\otimes SU(N)}_{k_1}\otimes\underbrace{SU(\overline{N})\otimes\cdots\otimes SU(\overline{N})}_{k_2}, \tag{5.2.20}$$

and its physical significance, which will be discussed in more detail in the next two subsections. Here we just give a brief introduction to the irreducible representation of (5.2.20).

A representation of (5.2.20) is a mapping defined as

$$\begin{aligned} &H: SU(N)\to\underbrace{SU(N)\otimes\cdots\otimes SU(N)}_{k_1}\otimes\underbrace{SU(\overline{N})\otimes\cdots\otimes SU(\overline{N})}_{k_2},\\ &H(U)=\underbrace{U\otimes\cdots\otimes U}_{k_1}\otimes\underbrace{\overline{U}\otimes\cdots\otimes\overline{U}}_{k_2} \qquad \text{for } U\in SU(N). \end{aligned} \tag{5.2.21}$$

We can show that mapping (5.2.21) satisfies (5.2.6), i.e.

$$H(U_1\cdot U_2)=H(U_1)H(U_2), \qquad \forall U_1,U_2\in SU(N).$$

By Definition (5.2.17) for tensor products of matrices, each representation $H(U)$ in (5.2.21) is an N^k-th order matrix, and is also a linear transformation of the complex space as

$$H(U):\ X\to X, \quad X=\underbrace{\mathbb{C}^N\otimes\cdots\otimes\mathbb{C}^N}_{k_1}\otimes\underbrace{\overline{\mathbb{C}}^N\otimes\cdots\otimes\overline{\mathbb{C}}^N}_{k_2}. \tag{5.2.22}$$

It is (5.2.16) that bestows the physical implication of the representation (5.2.21)-(5.2.22) of $SU(N)$, which will be explained in the next subsection.

Based on the irreducible representation theory of $SU(N)$, if $k=k_1+k_2\geqslant 2$ and $N\geqslant 2$, the representation (5.2.21) must be reducible, i.e. $H(U)$ can be split into smaller block diagonal form:

$$A^{-1}H(U)A=H_1(U)\oplus\cdots\oplus H_K(U), \qquad \forall U\in SU(N). \tag{5.2.23}$$

Usually, (5.2.23) is simply denoted as

$$\underbrace{N\otimes\cdots\otimes N}_{k_1}\otimes\underbrace{\overline{N}\otimes\cdots\otimes\overline{N}}_{k_2}=m_1\oplus\cdots\oplus m_K, \tag{5.2.24}$$

where m_j is the order of $H_j(U)$. In Subsection 5.2.4 we shall give the computational method of (5.2.23) (or (5.2.24)) by the Young tableau.

5.2.3 Physical explanation of irreducible representations

The physical implications of irreducible representations of $SU(N)$ were revealed first by Sakata in 1950's. In the following we give the Sakata explanation in a few steps.

1. The dimension N of $SU(N)$ represents N particles:

$$\psi_1, \cdots, \psi_N, \tag{5.2.25}$$

and their complex linear combination constitute $\mathbb{C}^N$:

$$\mathbb{C}^N = \left\{ \sum_{j=1}^N z_k \psi_k \,\middle|\, z_k \in \mathbb{C},\ 1 \leqslant k \leqslant N \right\},$$

which contains all physical states of the N particles (5.2.25).

In addition, the N fundamental particles of $SU(\overline{N})$ are the antiparticles of (5.2.25) given by

$$\overline{\psi}_1, \cdots, \overline{\psi}_N, \tag{5.2.26}$$

where $\overline{\psi}_k$ is the complex conjugate of ψ_k. The linear space

$$\overline{\mathbb{C}}^N = \left\{ \sum_{j=1}^N y_k \overline{\psi}_k \,\middle|\, y_k \in \mathbb{C},\ 1 \leqslant k \leqslant N \right\}$$

contains all physical states of the N antiparticles (5.2.26).

2. Each matrix $U \in SU(N)$ and each $\overline{U} \in SU(\overline{N})$ represent the transformations of physical states of particles (5.2.25) and antiparticles (5.2.26) as follows

$$\begin{aligned} &\sum_{j=1}^N z_k \psi_k \to \sum_{j=1}^N \widetilde{z}_k \psi_k, \\ &\sum_{j=1}^N y_k \overline{\psi}_k \to \sum_{j=1}^N \widetilde{y}_k \overline{\psi}_k, \end{aligned} \tag{5.2.27}$$

where

$$\begin{pmatrix} \widetilde{z}_1 \\ \vdots \\ \widetilde{z}_N \end{pmatrix} = U \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix}, \quad \begin{pmatrix} \widetilde{y}_1 \\ \vdots \\ \widetilde{y}_N \end{pmatrix} = \overline{U} \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}. \tag{5.2.28}$$

3. The tensor product of fundamental particles (5.2.25) and (5.2.26) are denoted by

$$\underbrace{N \otimes \cdots \otimes N}_{k_1} \otimes \underbrace{\overline{N} \otimes \cdots \otimes \overline{N}}_{k_2} = \{\psi_{i_1} \cdots \psi_{i_{k_1}} \overline{\psi}_{j_1} \cdots \overline{\psi}_{j_{k_2}}\}, \tag{5.2.29}$$

which stands for a new particle system where each particle

$$\psi_{i_1\cdots i_{k_1} j_1\cdots j_{k_2}} = \psi_{i_1}\cdots\psi_{i_{k_1}}\overline{\psi}_{j_1}\cdots\overline{\psi}_{j_{k_2}} \tag{5.2.30}$$

is a composite particle made up of $\psi_{i_1},\cdots,\psi_{i_{k_1}},\overline{\psi}_{j_1},\cdots,\overline{\psi}_{j_{k_2}}$.

For example for $N=3$, the tensor product

$$3\otimes\overline{3} = \begin{pmatrix} \psi_1\overline{\psi}_1 & \psi_1\overline{\psi}_2 & \psi_1\overline{\psi}_3 \\ \psi_2\overline{\psi}_1 & \psi_2\overline{\psi}_2 & \psi_2\overline{\psi}_3 \\ \psi_3\overline{\psi}_1 & \psi_3\overline{\psi}_2 & \psi_3\overline{\psi}_3 \end{pmatrix}$$

contains 9 new particles

$$\psi_{ij} = \psi_i\overline{\psi}_j, \qquad 1\leqslant i,j\leqslant 3,$$

composed of a fundamental particle ψ_i and an anti-particle $\overline{\psi}_j$.

4. The state space of composite particle system (5.2.29) is the tensor product of k_1 complex spaces $\mathbb{C}^N$ and k_2 complex conjugate spaces $\overline{\mathbb{C}}^N$, expressed as

$$\begin{aligned} &\underbrace{\mathbb{C}^N\otimes\cdots\otimes\mathbb{C}^N}_{k_1}\otimes\underbrace{\overline{\mathbb{C}}^N\otimes\cdots\otimes\overline{\mathbb{C}}^N}_{k_2} \\ =&\Big\{\sum_{i_1=1}^N\cdots\sum_{i_{k_1}=1}^N\sum_{j_1=1}^N\cdots\sum_{j_{k_2}=1}^N z_{i_1\cdots i_{k_1}j_1\cdots j_{k_2}}\psi_{i_1\cdots i_{k_1}j_1\cdots j_{k_2}} \mid \\ & z_{i_1\cdots i_{k_1}j_1\cdots j_{k_2}}\in\mathbb{C},\ \psi_{i_1\cdots i_{k_1}j_1\cdots j_{k_2}} \text{ are as in (5.2.30)}\Big\}. \end{aligned} \tag{5.2.31}$$

5. Denote the space (5.2.31) as

$$\mathbb{C}^{N^{k_1}}\otimes\overline{\mathbb{C}}^{N^{k_2}} = \underbrace{\mathbb{C}^N\otimes\cdots\otimes\mathbb{C}^N}_{k_1}\otimes\underbrace{\overline{\mathbb{C}}^N\otimes\cdots\otimes\overline{\mathbb{C}}^N}_{k_2}.$$

Then a representation $H(U)$ of (5.2.21) for $U\in SU(N)$ is a linear transformation of state space of composite particle system:

$$H(U):\ \mathbb{C}^{N^{k_1}}\otimes\overline{\mathbb{C}}^{N^{k_2}} \to \mathbb{C}^{N^{k_1}}\otimes\overline{\mathbb{C}}^{N^{k_2}}, \tag{5.2.32}$$

which represents state transformation of composite particles, similar to the state transformation (5.2.27)-(5.2.28) for a single particle system.

6. We recall the irreducible representation of $SU(N)$ given by (5.2.23) and (5.2.24). The irreducible representation implies that there is a decomposition in the N^k composite particles (5.2.29), which are classified into K groups as

$$G_1 = \{\Psi_1^1,\cdots,\Psi_{m_1}^1\},\ \cdots,\ G_K = \{\Psi_1^K,\cdots,\Psi_{m_K}^K\}, \tag{5.2.33}$$

where each group G_j contain m_j composite particles $\Psi_1^j, \cdots, \Psi_{m_j}^j$, as in (5.2.30), such that the space (5.2.31) can be decomposed into the direct sum of the K subspaces of span G_j as

$$\mathbb{C}^{N^{k_1}} \otimes \overline{C}^{N^{k_2}} = E_1 \oplus \cdots \oplus E_K, \tag{5.2.34}$$

where

$$E_j = \text{ span } G_j = \text{ span}\{\Psi_1^j, \cdots, \Psi_{m_j}^j\}, \quad 1 \leqslant j \leqslant K, \tag{5.2.35}$$

and then, the linear transformations $H(U)$ in (5.2.32) are also decomposed into the direct sum for all $U \in SU(N)$ as in (5.2.23). Namely, under the decomposition (5.2.33) of basis of $\mathbb{C}^{N^{k_1}} \otimes \overline{\mathbb{C}}^{N^{k_2}}$, the representations

$$H(U) \in \underbrace{SU(N) \otimes \cdots \otimes SU(N)}_{k_1} \otimes \underbrace{SU(\overline{N}) \otimes \cdots \otimes SU(\overline{N})}_{k_2}$$

can also be decomposed in the form

$$H(U) = H_1(U) \oplus \cdots \oplus H_K(U), \qquad \forall U \in SU(N), \tag{5.2.36}$$

and $H_j(U)$ $(1 \leqslant j \leqslant K)$ are as in (5.2.23), such that

$$H_j(U): \ E_j \to E_j, \quad \dim E_j = m_j, \quad 1 \leqslant j \leqslant K. \tag{5.2.37}$$

In other words, the subspace E_j of (5.2.35) is invariant for the linear transformation (5.2.36)-(5.2.37).

7. *Sakata's explanation of irreducible representation of $SU(N)$.* Now, we can deduce the following physical conclusions from the discussions in above steps 1-6.

Physical Explanation 5.5 *Let (5.2.29) be a family of composite particles as given by (5.2.30). The irreducible representation (5.2.36), which usually is expressed as*

$$\underbrace{N \otimes \cdots \otimes N}_{k_1} \otimes \underbrace{\overline{N} \otimes \cdots \otimes \overline{N}}_{k_2} = m_1 \oplus \cdots \oplus m_K,$$

means that

- *the composite particle system (5.2.29) can be classified into K groups of particles:*

$$G_j = \{\Psi_1^j, \cdots, \Psi_{m_j}^j\}, \quad 1 \leqslant j \leqslant K;$$

- *each group G_j has m_j particles, such that under the state transformation (5.2.27)-(5.2.28) of fundamental particles:*

$$U: \ \mathbb{C}^N \to \mathbb{C}^N, \quad \overline{U}: \ \overline{\mathbb{C}}^N \to \overline{\mathbb{C}}^N \quad (U \in SU(N)),$$

  the particles in G_j only transform between themselves as in (5.2.37).

We now examine the Physical Explanation 5.5 from the mathematical viewpoint. The N^k $(k=k_1+k_2)$ elements of (5.2.29)-(5.2.30) form a basis of $\mathbb{C}^{N^{k_1}}\otimes\overline{\mathbb{C}}^{N^{k_2}}$. We denote these elements as

$$E=\{\Psi_1,\cdots,\Psi_{N^k}\}, \tag{5.2.38}$$

with each Ψ_j as in (5.2.30). The irreducible representation (5.2.23):

$$AH(U)A^{-1}=H_1(U)\oplus\cdots\oplus H_K(U), \qquad \forall U\in SU(N)$$

implies that if we take the basis transformation for (5.2.38)

$$\begin{pmatrix}\widetilde{\Psi}_1\\ \vdots\\ \widetilde{\Psi}_{N^k}\end{pmatrix}=A\begin{pmatrix}\Psi_1\\ \vdots\\ \Psi_{N^k}\end{pmatrix}, \tag{5.2.39}$$

then under the new basis $\widetilde{E}=\{\widetilde{\Psi}_1,\cdots,\widetilde{\Psi}_{N^k}\}$, the linear mapping

$$H(U):\ X\to X \quad (X\ \text{ as in (5.2.22)})$$

becomes the block diagonal form

$$H(U)=\begin{pmatrix}H_1(U) & & 0\\ & \ddots & \\ 0 & & H_K(U)\end{pmatrix}, \qquad \forall U\in SU(N).$$

This means that the new basis $\widetilde{E}$ of X is divided into K sub-bases

$$G_1=\{\widetilde{E}_1,\cdots,\widetilde{E}_{m_1}\},\cdots,\ G_K=\{\widetilde{E}_{J_K+1},\cdots,\widetilde{E}_{J_k+m_K}\},$$

with $J_K+m_K=N^K$, such that each subspace X_J of X, spanned by the j-th sub-basis G_j as

$$X_j=\ \text{span}\ G_j \quad (1\leqslant j\leqslant K)$$

is an invariant subspace of the mapping $H(U)$ for all $U\in SU(N)$. In particular, the block matrix $H_j(U)$ is the restriction of $H(U)$ on X_j $(1\leqslant j\leqslant K)$:

$$H(U)|_{X_j}=H_j(U):\ X_j\to X_j. \tag{5.2.40}$$

Hence, when we take any linear transformation on $\mathbb{C}^N$ as

$$U:\ \mathbb{C}^N\to\mathbb{C}^N, \quad U\in SU(N),$$

then the subspaces X_j will undergo themselves a linear transformation in the fashion as given by (5.2.40).

Thus, from the mathematical viewpoint, the Physical Explanation 5.5 is sound, and provides the mathematical foundation for the quark model.

5.2.4 Computations for irreducible representations

We have understood the physical implications of the irreducible representations of $SU(N)$. The remaining crucial problem is how to compute the irreducible representations. Namely, for an N-dimensional representation as (5.2.21), we need to determine its irreducible decomposition (5.2.24). In other words, we need to determine $m_1, \cdots, m_K$ and K in

$$\underbrace{N\otimes\cdots\otimes N}_{k_1}\otimes\underbrace{\overline{N}\otimes\cdots\otimes\overline{N}}_{k_2}=m_1\oplus\cdots\oplus m_K. \tag{5.2.41}$$

A very effective method to compute (5.2.41) is the Young tableaux, which uses square diagrams to deduce these numbers m_j $(1\leqslant j\leqslant K)$. The method is divided in two steps.

The first step is to obtain the rule to group together square diagrams, and to obtain the irreducible representation (5.2.41) in the form

$$\underbrace{N\otimes\cdots\otimes N}_{k_1}\otimes\underbrace{\overline{N}\otimes\cdots\otimes\overline{N}}_{k_2}=\boxed{1}\oplus\boxed{2}\oplus\cdots\oplus\boxed{K}. \tag{5.2.42}$$

The second step provides the rule and method to compute the dimensional m_j from the j-th square diagrams $\boxed{j}$ on the right-hand side of (5.2.42):

$$\text{Computation of } m_j \text{ from } \boxed{j}. \tag{5.2.43}$$

1. *Rule to group the square diagram in (5.2.42).* First of all, in the Young tableau we use a square $\square$ to stand for a fundamental representation N of $SU(N)$, and use a column of $N-1$ squares to stand for the conjugate representation $\overline{N}$ in the right-hand side of (5.2.42):

$$N=\square,\qquad \overline{N}=\left.\begin{array}{|c|}\hline\ \\ \hline\ \\ \hline\ \\ \hline\ \\ \hline\end{array}\right\}N-1.$$

For example, we can write $3\otimes 3\otimes\overline{3}$ as

$$3\otimes 3\otimes\overline{3}=\square\otimes\square\otimes\begin{array}{|c|}\hline\ \\ \hline\ \\ \hline\end{array}\ .$$

By this rule, the left-hand side of (5.2.42) can be expressed as

$$\underbrace{N\otimes\cdots\otimes N}_{k_1}\otimes\underbrace{\overline{N}\otimes\cdots\otimes\overline{N}}_{k_2}=\underbrace{\square\otimes\cdots\otimes\square}_{k_1}\otimes\underbrace{\begin{array}{|c|}\hline\ \\ \hline\ \\ \hline\ \\ \hline\ \\ \hline\end{array}\otimes\cdots\otimes\begin{array}{|c|}\hline\ \\ \hline\ \\ \hline\ \\ \hline\ \\ \hline\end{array}}_{k_2} \tag{5.2.44}$$

Now, we give rules to group the square diagrams of the right-hand side of (5.2.42) for different k_1 and k_2.

(a) *Case* $k_1 = 2, k_2 = 0$: By the rule (5.2.44), the left-hand side is

$$N \otimes N = \square \otimes \square.$$

Then the right-hand side is defined as

$$\square \otimes \boxed{\alpha} = \square\boxed{\alpha} \oplus \begin{array}{c}\square\\ \boxed{\alpha}\end{array} = \square\square \oplus \begin{array}{c}\square\\ \square\end{array}. \tag{5.2.45}$$

(b) *Case* $k_1 = 1$ *and* $k_2 = 1$: We have

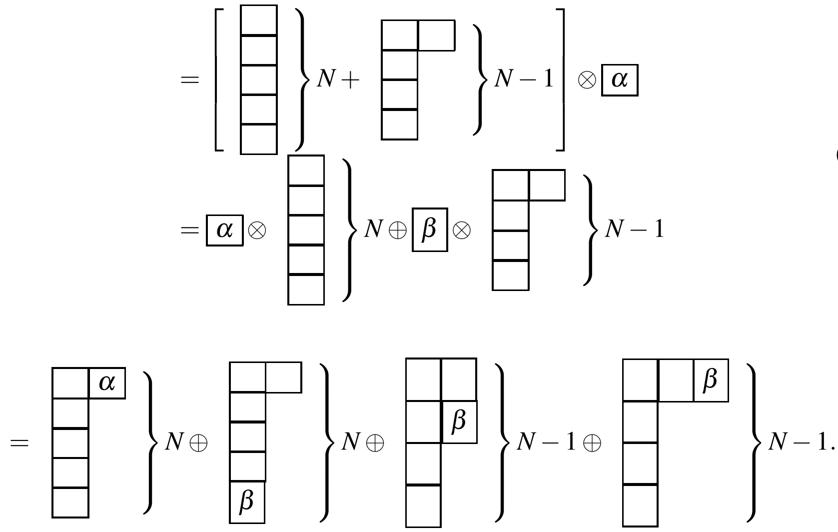
$$\boxed{\alpha} \otimes \left.\begin{array}{c}\square\\ \square\\ \square\\ \square\\ \square\end{array}\right\} N-1 = \left.\begin{array}{c}\boxed{\alpha}\\ \square\\ \square\\ \square\\ \square\\ \square\end{array}\right\} N + \left.\begin{array}{ll}\square & \boxed{\alpha}\\ \square & \\ \square & \\ \square & \\ \square & \end{array}\right\} N-1 \tag{5.2.46}$$

(c) *Case* $k_1 = 3$ *and* $k_2 = 0$: We define the right-hand side of (5.2.42) as shown

$$\begin{aligned} N \otimes N \otimes N &= \square \otimes \square \otimes \square \\ &= (\square \otimes \square) \otimes \square \\ &= \left(\square\square \oplus \begin{array}{c}\square\\ \square\end{array}\right) \otimes \square \qquad \text{(by (5.2.45))} \\ &= \square\square \otimes \square \oplus \begin{array}{c}\square\\ \square\end{array} \otimes \square \\ &= \square\square \otimes \boxed{\alpha} \oplus \square \otimes \begin{array}{c}\square\\ \square\end{array} \qquad \text{(use (5.2.46) for 2nd term)} \\ &= \square\square\boxed{\alpha} \oplus \begin{array}{ll}\square & \square\\ \boxed{\alpha} & \end{array} \oplus \begin{array}{c}\boxed{\beta}\\ \square\\ \square\end{array} \oplus \begin{array}{ll}\square & \boxed{\beta}\\ \square & \end{array}. \end{aligned} \tag{5.2.47}$$

(d) *Case* $k_1 = 2$ *and* $k_2 = 1$: We have

$$\begin{aligned} N \otimes N \otimes \overline{N} &= \square \otimes \square \otimes \left.\begin{array}{c}\square\\ \square\\ \square\\ \square\end{array}\right\} N-1 \\ &= \left(\square \otimes \left.\begin{array}{c}\square\\ \square\\ \square\\ \square\end{array}\right\} N-1\right) \otimes \square \qquad \text{(use (5.2.46))} \end{aligned}$$



We know that the representations of $SU(N)$ for particles and antiparticles are conjugate relation, i.e. if

$$\underbrace{N\otimes\cdots\otimes N}_{k_1}\otimes\underbrace{\overline{N}\otimes\cdots\otimes\overline{N}}_{k_2} \tag{5.2.49}$$

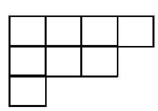
represents a particle system, then its conjugate

$$\underbrace{\overline{N}\otimes\cdots\otimes\overline{N}}_{k_1}\otimes\underbrace{N\otimes\cdots\otimes N}_{k_2} \tag{5.2.50}$$

represents the antiparticle system. Due to the symmetry of particles and antiparticles, (4.2.49) and (4.2.50) have the same irreducible representations. In addition each composite particle is composed of N particles with $N \leqslant 3$. Hence it is enough to give the four cases 1)-4) above for the physical purpose.

2. *Computation of (5.2.43).* In the right-hand side of the Young tableaux (4.2.45)-(4.2.48), the number of square diagrams is the K as in (4.2.42), and each of which represents a dimension m_j $(1 \leqslant j \leqslant K)$ of an irreducible representation. The rule to compute m_j is as follows.

For example, for the following square diagram



(5.2.51)

the dimension m is given by the formula

$$m = \frac{\alpha_N}{\beta_N}, \tag{5.2.52}$$

where α_N and β_N can be computed by the diagram as (5.2.51).

(a) *Computation of* α_N. Fill the blanks of (4.2.51) in numbers in the following fashion:

N	N+1	N+2	N+2
$N-1$	N	N+1	
$N-2$			

(5.2.53)

The α_N equals to the multiplication of all numbers in (5.2.53),

$$\alpha_N = N(N+1)(N+2)(N+3)(N-1)N(N+1)(N-2). \tag{5.2.54}$$

(b) *Computation of* β_N. Fill the blanks of (5.2.51) in the fashion:

6	4	3	1
4	2	1	
1			

(5.2.55)

where the data in a square equals to the number of all squares on its right-hand side and below it adding one. For example, for the square marked 6, there are 3 squares on its right-hand side, and 2 square below it. Hence the number k in this blank is

$$k = 3+2+1 = 6.$$

Then, the number β_N is the multiplication of all numbers in (4.2.55)

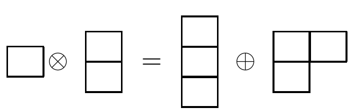
$$\beta_N = 6\times 4\times 3\times 1\times 4\times 2\times 1\times 1. \tag{5.2.56}$$

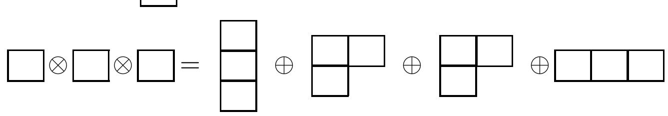
Thus, by (4.2.54) and (4.2.56) we can get the value of (4.2.52). In the following, we give a few examples to show how to use the Young tableau to compute the irreducible representations of $SU(N)$.

Example 5.6 For $N = 3$, consider the two cases given by

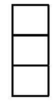
$$3\otimes\overline{3} \quad \text{and} \quad 3\otimes 3\otimes 3, \tag{5.2.57}$$

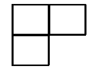
which are the most important cases in particle physics. By (5.2.46) and (5.2.47), the Young tableaux of (5.2.57) are as follows

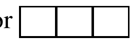
$3\otimes\overline{3} =$  (5.2.58)

$3\otimes 3\otimes 3 =$  (5.2.59)

According to (5.2.52) and the methods to compute α_N and β_N $(N=3)$, we infer from (5.2.58) and (5.2.59) that

$\alpha_N = 3\times 2\times 1,\quad \beta_N = 3\times 2\times 1,\quad m=1,$ for 

$\alpha_N = 3\times 4\times 2,\quad \beta_N = 3\times 1\times 1,\quad m=8,$ for 

$\alpha_N = 3\times 4\times 5,\quad \beta_N = 3\times 2\times 1,\quad m=10,$ for 

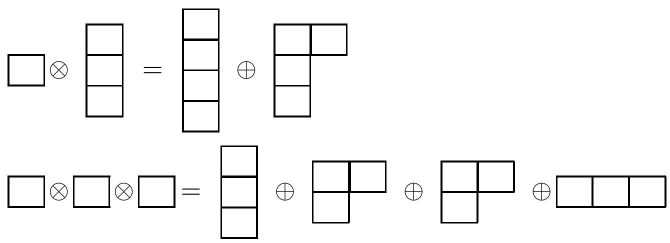
Consequently, we derive, from (5.2.58) and (5.2.59), the following irreducible representations of (5.2.57):

$$\begin{aligned} &3\otimes\overline{3} = 1\oplus 8, \\ &3\otimes 3\otimes 3 = 1\oplus 8\oplus 8\oplus 10. \end{aligned} \tag{5.2.60}$$

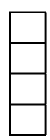
Example 5.7 The other important cases in physics are the two irreducible representations for $N=4$:

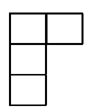
$$4\otimes\overline{4} \quad \text{and} \quad 4\otimes 4\otimes 4.$$

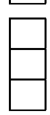
Their Young tableaux are given by

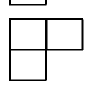
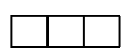


Then we can compute that

$m=1,$ for 

$m=15,$ for 

$m=4,$ for 

$m=20,$ for  and 

Hence we deduce that

$$\begin{aligned} &4\otimes\overline{4} = 1\oplus 15, \\ &4\otimes 4\otimes 4 = 4\oplus 20\oplus 20\oplus 20. \end{aligned} \tag{5.2.61}$$

5.2.5 Sakata model of hadrons

In 1950's, many hadrons were discovered, leading to many attempts to investigate the deeper hadron structure. Based on the irreducible representation (5.2.60) of $SU(3)$, i.e..

$$3\otimes\overline{3}=8\oplus 1, \tag{5.2.62}$$

Sakata presented a model for hadron structure, called the Sakata model. This was an early precursor to the quark model, and also resulted in the physical implications of irreducible representations as stated by Physical Explanation 5.5.

Sakata model proposed three particles

$$p,\quad n,\quad \Lambda \tag{5.2.63}$$

as the fundamental particles for all strong interacting particles. In his scheme, Sakata used the three particles in (5.2.63) as a basis of $SU(3)$, such that each hadron consists of a fundamental particle and an antiparticle as

$$\text{hadron}=S_i\overline{S}_j \qquad \text{for } 1\leqslant i,j\leqslant 3, \tag{5.2.64}$$

where $(S_1,S_2,S_3)=(p,n,\Lambda)$ are called the sakataons.

In (5.2.62), the left-hand side represents the pairs $S_i\overline{S}_j$, and the right-hand side represents the following eight mesons:

$$\pi^+,\pi^-,\pi^0,K^+,K^-,K^0,\overline{K}^0,\eta. \tag{5.2.65}$$

It is a coincidence that the eight particles just constitute an eight multiple state of hadrons, and can be illustrated by the Eightfold Way; see Figure 5.3 and Figure 5.6.

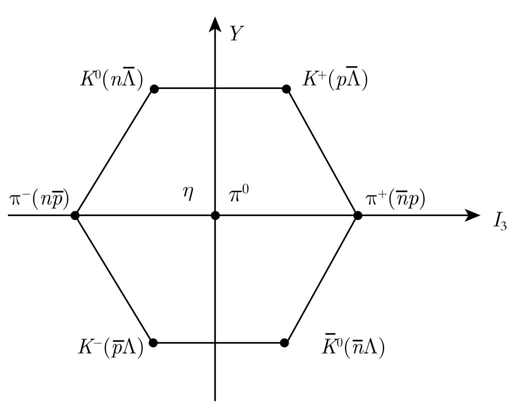


Figure 5.6

Table 5.6

Sakataons	I	I_3	Q	S	B
p	$1/2$	$+1/2$	+1	0	+1
n	$1/2$	$-1/2$	0	0	+1
Λ	0	0	0	-1	+1

In addition, from the quantum numbers of the three particles(5.2.63), we can deduce the quantum numbers for the eight mesons in(5.2.65); see Table 5.6 and 5.7.

In Table 5.7, J is the spin and π is the parity. Hence $J^{\pi} = 0^{-}$ represents that spin $J = 0$ and parity $\pi = -1$.

The Sakata model explains the mesons well. However, it encounters difficulties if we use this model to describe baryons. In fact, if we use the three sakataons to form a baryon, then the baryon number $B = 3$ for a baryon. Hence, we have to use two sakataons and one anti-sakataon to form a baryon to ensure $B = 1$. Unfortunately those combinations do not agree with experiments.

Table 5.7

	mesons	I	I_3	Q	S	J^{π}	$m(MeV)$
$p\overline{n}$	π^+	1	1	1	0	0^-	~ 140
$-\overline{p}n$	π^-	1	-1	-1	0	0^-	~ 140
$\frac{n\overline{n}-p\overline{p}}{\sqrt{2}}$	π^0	1	0	0	0	0^-	~ 140
$p\overline{\Lambda}$	K^+	$\frac{1}{2}$	$\frac{1}{2}$	1	1	0^-	~ 495
$n\overline{\Lambda}$	K^0	$\frac{1}{2}$	$-\frac{1}{2}$	0	1	0^-	~ 495
$\Lambda\overline{n}$	$\overline{K}^0$	$\frac{1}{2}$	$\frac{1}{2}$	0	-1	0^-	~ 495
$\Lambda\overline{p}$	$\overline{K}^-$	$\frac{1}{2}$	$-\frac{1}{2}$	-1	-1	0^-	~ 495
$\frac{n\overline{n}+p\overline{p}-2\Lambda\overline{\Lambda}}{\sqrt{6}}$	η	0	0	0	0	0^-	~ 548
$\frac{n\overline{n}+p\overline{p}+\Lambda\overline{\Lambda}}{\sqrt{3}}$	η'	0	0	0	0	0^-	~ 548

5.2.6 Gell-Mann-Zweig's quark model

Based on the Eightfold Way, as shown in Figures 5.2-5.5, the hadrons are classified as follows:

$$\begin{cases} \text{Mesons } (J=0): & 8 \text{ particles,} \\ \text{Mesons } (J=1): & 8 \text{ particles,} \end{cases}$$

$$\begin{cases} \text{Baryons } \left(J=\dfrac{1}{2}\right): & 8 \text{ particles,} \\ \text{Baryons } (J=3/2): & 10 \text{ particles.} \end{cases}$$

In view of the irreducible representations of $SU(3)$:

$$3\otimes\overline{3}=8\oplus 1,$$
$$3\otimes 3\otimes 3=10\oplus 8\oplus 8\oplus 1,$$

it is natural to guess the relations (5.2.18) or (5.2.19), i.e.

$$\begin{aligned}\text{mesons}&=3\otimes\overline{3}=8\oplus 1,\\ \text{baryons}&=3\otimes 3\otimes 3=10\oplus 8\oplus 8\oplus 1.\end{aligned}\tag{5.2.66}$$

According to the Physical Explanation 5.5, we infer immediately, from (5.2.66), that the hadrons in the Eightfold Way are composed of three fundamental particles, denoted by

$$q_1,\quad q_2,\quad q_3\tag{5.2.67}$$

and mesons consist of a fundamental particle and an antiparticle. In a nutshell, baryons consist of three fundamental particles:

$$\begin{aligned}\text{mesons}&=q_i\overline{q}_j &&\text{for } 1\leqslant i,j\leqslant 3,\\ \text{baryons}&=q_iq_jq_k &&\text{for } 1\leqslant i,j,k\leqslant 3.\end{aligned}\tag{5.2.68}$$

Physicists Gell-Mann and Zweig did this work independently in 1964, and presented the celebrated Quark Model. The three fundamental particles (5.2.67) were termed the up, down and strange quarks by Gell-Mann, denoted by

$$q_1=u,\quad q_2=d,\quad q_3=s.$$

Zweig called these particles the aces.

By using the three quarks (u,d,s) to replace the sakataons (p,n,Λ), we can perfectly explain all hadrons described by the Eightfold Way, i.e. the hadrons given by (5.2.1)-(5.2.4).

An important step to establish the quark model is to determine the quantum numbers of quarks, which are introduced in the following several procedures:

1) The spins of quarks have to be $J=\frac{1}{2}$, as required by

$$\begin{aligned}&\text{spins of mesons:} && J=0(\uparrow\downarrow), && J=1(\uparrow\uparrow), && J=-1(\downarrow\downarrow),\\ &\text{spins of baryons:} && J=\frac{1}{2}(\uparrow\uparrow\downarrow), && J=-\frac{1}{2}(\downarrow\downarrow\uparrow),\\ &&& J=\frac{3}{2}(\uparrow\uparrow\uparrow), && J=-\frac{3}{2}(\downarrow\downarrow\downarrow).\end{aligned}\tag{5.2.69}$$

Namely, the unique choice to ensure (5.2.69) is that $J=\frac{1}{2}$ for quarks.

2) The success of (p,n,Λ) for describing mesons suggests that for the strange number S and the isospin (I,I_3), u and p, d and n, s and Λ should be the same respectively. Hence we have

$$u:\ (S,I,I_3)=\left(0,\frac{1}{2},\frac{1}{2}\right),$$
$$d:\ (S,I,I_3)=\left(0,\frac{1}{2},-\frac{1}{2}\right),$$
$$s:\ (S,I,I_3)=(-1,0,0).$$

3) Since the baryon numbers of all baryons are $B=1$, by the constituents (5.2.68) of baryons, it is natural that

$$u:\ B=\frac{1}{3},\quad d:\ B=\frac{1}{3},\quad s:\ B=\frac{1}{3}.$$

4) For all hadrons, the following formula, well known as the Gell-Mann-Nishijima relation, holds true:

$$Q=I_3+\frac{B}{2}+\frac{S}{2}. \tag{5.2.70}$$

This relation should also be valid for quarks. Hence, we deduce from (5.2.70) that the electric charges of quarks are

$$u:\ Q=\frac{2}{3},\quad d:\ Q=-\frac{1}{3},\quad S:\ Q=-\frac{1}{3}.$$

The data derived in 1)-4) above are collected in Table 5.3. According to the quantum numbers of hadrons and quarks, we can determine the quark constituents of all hadrons.

For example, for *uud* its quantum numbers are derived from those of u and d as follows

$$uud:\ B=1,\ Q=1,\ S=0,\ J=\frac{1}{2},\ I=\frac{1}{2},\ I_3=\frac{1}{2}, \tag{5.2.71}$$

which dictate that *uud* is the proton:

$$uud=p.$$

The quark constituents of the main hadrons are listed in Subsection 5.1.1.

In 1974, Ting and Richter discovered independently J/ψ particle, which implies the existence of a new quark, named as the c quark. Thus, the quark family then was extended to four members:

$$u,d,s,c.$$

As fundamental particles of $SU(4)$, the irreducible representations (5.2.61), written as

$$\begin{aligned}&4\otimes\overline{4}=15\oplus 1,\\&4\otimes 4\otimes 4=20\oplus 20\oplus 20\oplus 4\end{aligned} \tag{5.2.72}$$

suggest that the meson groups with 8 particles should be extended to 15 particles, and the baryons be extended to 20 particles in a group. The extended group of mesons with 15 particles has been verified, which is given by

$$\pi^{\pm},\ \pi^0,\ K^{\pm},\ K^0,\ \overline{K}^0,\ \eta,\ \eta',\ D_0,\ \overline{D}_0,\ D^{\pm},\ D_s^{\pm}.$$

However, the 20 baryons corresponding to the irreducible representation (5.2.72) have not been discovered now.

Up to now, the quark family has six members:

$$u,d,s,c,b,t.$$

Their irreducible representation is given by

$$\begin{aligned} &6\otimes\overline{6}=35\oplus 1,\\ &6\otimes 6\otimes 6=70\oplus 70\oplus 56\oplus 20. \end{aligned} \tag{5.2.73}$$

The classification scheme corresponding to (5.2.73) does not seem to be realistic. The irreducible representation of $SU(N)$ is only a phenomenological theory in elementary particle physics.

5.3 Weakton Model of Elementary Particles

5.3.1 Decay means the interior structure

As we addressed in Section 5.1.3, all charged leptons, quarks and mediators can undergo decay as follows:

1) Charged lepton radiation and decay:

$$\begin{aligned} &e^- \to e^- + \gamma,\\ &\mu^- \to e^- + \overline{\nu}_e + \nu_\mu,\\ &\tau^- \to \mu^- + \overline{\nu}_\mu + \nu_\tau. \end{aligned} \tag{5.3.1}$$

2) Quark decay:

$$\begin{aligned} &d \to u + e^- + \overline{\nu}_e,\\ &s \to d + \gamma,\\ &c \to d + \overline{s} + u. \end{aligned} \tag{5.3.2}$$

3) Mediator decay:

$$\begin{aligned} &2\gamma \to e^+ + e^-,\\ &W^{\pm} \to l^{\pm} + \overline{\nu}_{l^{\pm}},\\ &Z \to l^+ + l^-, \end{aligned} \tag{5.3.3}$$

where $l^{\pm}$ are the charged leptons.

All leptons, quarks and mediators are currently regarded as elementary particles. However, the decays in (5.3.1)-(5.3.3) show that these particles must have an interior structure, and consequently they should be considered as composite particles rather than elementary particles:

Decay Means Interior Structure.

5.3.2 Theoretical foundations for the weakton model

Subatomic decays and electron radiations indicate that there must be interior structure for charged leptons, quarks and mediators. The main objective of this section is to propose an elementary particle model, which we call weakton model, based on the weak and strong interaction theories developed in the last chapter.

Angular momentum rule

It is known that the dynamic behavior of a particle is described by the Dirac equations:

$$i\hbar\frac{\partial\psi}{\partial t} = H\psi, \tag{5.3.4}$$

where $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)^{\mathrm{T}}$ is the Dirac spinor, H is the Hamiltonian:

$$H = -i\hbar c(\alpha^k\partial_k) + mc^2\alpha^0 + V(x), \tag{5.3.5}$$

V is the potential energy, α^k $(1 \leqslant k \leqslant 3)$ are the matrices as given by (2.2.48), and α^0 is the matrix as

$$\alpha^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

By the conservation laws in quantum mechanics, if an Hermitian operator L commutes with H in (5.3.5):

$$LH = HL,$$

then the physical quantity L is conserved.

Consider the total angular momentum $\vec{J}$ of a particle as

$$\vec{J} = \vec{L} + s\vec{S},$$

where $\vec{L}$ is the orbital angular momentum

$$\vec{L} = \vec{r} \times \vec{p}, \quad \vec{p} = -i\hbar\nabla,$$

s is the spin, and

$$\vec{S} = (S_1, S_2, S_3), \qquad S_k = \hbar\begin{pmatrix} \sigma_k & 0 \\ 0 & \sigma_k \end{pmatrix},$$

and σ_k $(1 \leqslant k \leqslant 3)$ are the Pauli matrices; see (2.2.47).

It is readily to check that for H in (5.3.5),

$$\begin{aligned} &\vec{J}_{1/2} = \vec{L} + \frac{1}{2}\vec{S} \qquad \text{commutes with } H, \\ &\vec{J}_s = \vec{L} + s\vec{S} \qquad \text{does not commute with } H \text{ for } s \neq \frac{1}{2} \text{ in general.} \end{aligned} \tag{5.3.6}$$

Also, we know that

$$s\vec{S} \text{ commutes with } H \text{ in straight line motion for any } s. \tag{5.3.7}$$

The properties in (5.3.6) imply that only particles with spin $s = \frac{1}{2}$can make a rotational motion in a center field with free moment of force. However, (5.3.7) implies that the particles with $s \neq \frac{1}{2}$will move in a straight line, i.e. $\vec{L} = 0$, unless they are in a field with nonzero moment of force.

In summary, we have derived the following angular momentum rule for subatomic particle motion, which is important for our weakton model established in the next subsection. The more general form of the angular momentum rule will be addressed in Section 6.2.4.

Angular Momentum Rule 5.8 *Only the fermions with spin $s = \frac{1}{2}$ can rotate around a center with zero moment of force. The fermions with $s \neq \frac{1}{2}$ will move on a straight line unless there is a nonzero moment of force present.*

For example, the particles bounded in a ball rotating around the center, as shown in Figure 5.7, must be fermions with $s = 1/2$.

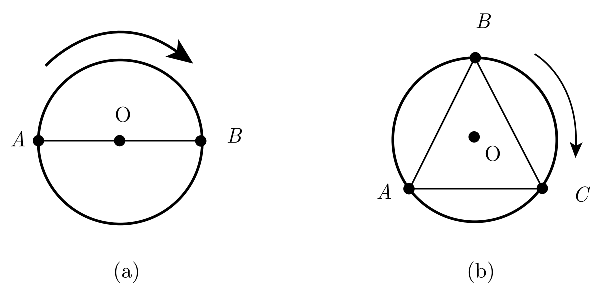


Figure 5.7 (a) Two particles A, B rotate around the center 0, amd (b) three particles A, B, C rotate around the center 0.

Mass generation mechanism

For a particle moving with velocity v, its mass m and energy E obey the Einstein relation

$$E = mc^2 \Big/ \sqrt{1 - \frac{v^2}{c^2}}. \tag{5.3.8}$$

Usually, we regard m as a static mass which is fixed, and energy is a function of velocity v.

Now, taking an opposite viewpoint, we regard energy E as fixed, mass m as a function of velocity v, and the relation (5.3.8) is rewritten as

$$m = \sqrt{1 - \frac{v^2}{c^2}} \frac{E}{c^2}. \tag{5.3.9}$$

Thus, (5.3.9) means that a particle with an intrinsic energy E has zero mass $m = 0$ if it moves at the speed of light $v = c$, and will possess nonzero mass if it moves with a velocity $v < c$.

All particles including photons can only travel at the speed sufficiently close to the speed of light. Based on this viewpoint, we can think that if a particle moving at the speed of light (approximately) is decelerated by an interaction force $\vec{F}$, obeying

$$\frac{d\vec{P}}{dt} = \sqrt{1 - \frac{v^2}{c^2}} \vec{F},$$

then this massless particle will generate mass at the instant. In particular, by this mass generation mechanism, several massless particles can yield a massive particle if they are bound in a small ball, and rotate at velocities less than the speed of light.

Interaction charges

In the unified field model introduced in the last chapter, we derived that both weak and strong interactions possess charges, as for gravity and electromagnetism; see Section 4.3.4:

$$\begin{aligned}
&\text{gravitation:} && \text{mass charge } m,\\
&\text{electromagnetism:} && \text{electric charge } e,\\
&\text{weak interaction:} && \text{weak charge } g_w,\\
&\text{strong interaction:} && \text{strong charge } g_s.
\end{aligned} \tag{5.3.10}$$

If Φ is a charge potential corresponding to an interaction, then the interacting force generated by its charge g is given by

$$F = -g \nabla \Phi,$$

where ∇ is the spatial gradient operator.

It is very crucial to introduce both weak and strong interaction charges for us to develop the weakton model. The charges in (5.3.10) possess the following physical properties:

1) Electric charge Q_e, weak charge Q_w, strong charge Q_s are conservative. The energy is a conserved quantity, but the mass M is not a conserved quantity due to the mass generation mechanism as mentioned in (5.3.9).

2) There is no interacting force between two particles without common charges. For example, if a particle A possesses no strong charge, then there is no strong interacting force between A and any other particles.

3) Only the electric charge e can take both positive and negative values, and other charges take only nonnegative values.

Layered formulas of strong interaction potentials

The layered properties of strong and weak interaction potentials are also crucial for us to establish the weakton model. We recall briefly the strong interaction potentials in the general form as:

$$\begin{aligned} &\Phi_s = g_s(\rho)\left[\frac{1}{r} - \frac{A}{\rho}(1+kr)e^{-kr}\right], \\ &g_s(\rho) = N\left(\frac{\rho_w}{\rho}\right)^3 g_s, \end{aligned} \tag{5.3.11}$$

where ρ is the particle radius, N is the number of strong charges.

There are about five levels of strong interactions between particles: w^*-weaktons, quarks, gluons, hadrons, atoms/molecules. Hence, based on (5.3.11) we can derive the layered formulas for these five level particles as follows:

$$\begin{aligned} &\Phi^s_{w^*} = g_s\left[\frac{1}{r} - \frac{A_0}{\rho_w}(1+k_0r)e^{-k_0r}\right], \\ &\Phi^s_q = \left(\frac{\rho_w}{\rho_q}\right)^3 g_s\left[\frac{1}{r} - \frac{A_q}{\rho_q}(1+k_1r)e^{-k_1r}\right], \\ &\Phi^s_g = 2\left(\frac{\rho_w}{\rho_g}\right)^3 g_s\left[\frac{1}{r} - \frac{A_g}{\rho_g}(1+k_1r)e^{-k_1r}\right], \\ &\Phi^s_n = 3\left(\frac{\rho_w}{\rho_n}\right)^3 g_s\left[\frac{1}{r} - \frac{A_n}{\rho_n}(1+k_nr)e^{-k_nr}\right], \\ &\Phi^s_a = N\left(\frac{\rho_w}{\rho_a}\right)^3 g_s\left[\frac{1}{r} - \frac{A_a}{\rho_a}(1+k_ar)e^{-k_ar}\right]. \end{aligned} \tag{5.3.12}$$

where $\rho_w, \rho_q, \rho_g, \rho_n, \rho_a$ are the radii of weakton, quark, gluon, nucleon, atom/molecule, and g_s is the strong charge of w^*-weakton. By (4.5.70), g_s can be expressed as

$$g_s^2 = \frac{1}{9}\beta^2\left(\frac{\rho_n}{\rho_w}\right)^6 g^2, \tag{5.3.13}$$

where $\beta = \sqrt{2}e^{-1/4}/(8-e^{1/2})^{1/2}$ and g is the Yukawa charge.

Layered formulas for weak interaction potentials

The weak interaction potential for a particle with radius ρ and with N weak charges g_w is given by

$$\begin{aligned}\Phi_w &= g_w(\rho)e^{-kr}\left[\frac{1}{r}-\frac{B}{\rho}(1+2kr)e^{-kr}\right],\\ g_w(\rho) &= N\left(\frac{\rho_w}{\rho}\right)^3 g_w,\end{aligned} \tag{5.3.14}$$

where $k=\dfrac{1}{r_0}=10^{16}\ \text{cm}^{-1}$, and $r_0=10^{-16}$ cm.

For the weak interaction, there are about four levels: weaktons, mediators, quarks and charged leptons. By (5.3.14), the layered formulas for these four level particles are given by

$$\begin{aligned}\Phi_0^w &= g_w e^{-kr}\left[\frac{1}{r}-\frac{B_w}{\rho_w}(1+2kr)e^{-kr}\right],\\ \Phi_m^w &= 2\left(\frac{\rho_w}{\rho_m}\right)^3 g_w e^{-kr}\left[\frac{1}{r}-\frac{B_m}{\rho_m}(1+2kr)e^{-kr}\right],\\ \Phi_q^w &= 3\left(\frac{\rho_w}{\rho_q}\right)^3 g_w e^{-kr}\left[\frac{1}{r}-\frac{B_q}{\rho_q}(1+2kr)e^{-kr}\right],\\ \Phi_l^w &= 3\left(\frac{\rho_w}{\rho_l}\right)^3 g_w e^{-kr}\left[\frac{1}{r}-\frac{B_l}{\rho_l}(1+2kr)e^{-kr}\right].\end{aligned} \tag{5.3.15}$$

By (4.6.37), the weak charge g_w of weakton can be expressed as

$$g_w^2=\frac{1}{9}\alpha^2\left(\frac{\rho_n}{\rho_w}\right)^6\hbar c, \tag{5.3.16}$$

where $\alpha=\dfrac{2}{\sqrt{5\sqrt{2}}}\left(\dfrac{m_w}{m_p}\right)\times 10^{-2}$.

Duality of mediators

Based on the unified field theory, there exists a natural duality between the mediators:

$$\begin{aligned}&\text{tensor graviton } g_G && \leftrightarrow && \text{vector graviton } \phi_G,\\ &\text{vector photon } \gamma && \leftrightarrow && \text{scalar photon } \gamma_0\\ &\text{vector bosons } W^\pm, Z && \leftrightarrow && \text{Higgs } H^\pm, H^0,\\ &\text{gluons } \{g^k\ |1\leqslant k\leqslant 8\} && \leftrightarrow && \text{scalar gluons } \{g_0^k\ |\ 1\leqslant k\leqslant 8\}.\end{aligned} \tag{5.3.17}$$

We shall see that the prediction (5.3.17) are in perfect agreement with the consequences of the weakton model.

5.3.3 Weaktons and their quantum numbers

The observation of subatomic decays and electron radiations leads us to propose a set of elementary particles, which we call weaktons. They are massless, spin $-\frac{1}{2}$ particles with one unit of weak charge g_w.

The introduction of weaktons is based on the following theories, observational facts and considerations:

1) The interior structure of charged leptons, quarks and mediators demonstrated by the decays, scatterings and radiations, as as shown in (5.3.1)-(5.3.3);

2) The new quantum numbers of weak charge g_w and strong charge g_s introduced in (5.3.10);

3) The mass generation mechanism presented in Section 5.3.2;

4) The weakton confinement theory given by the layered formulas of weak interaction potentials (5.3.15)-(5.3.16); and

5) The duality (5.3.17) for the mediators.

The weaktons consist of the following 6 elementary particles and their antiparticles:

$$\begin{aligned} &w^*, \quad w_1, \quad w_2, \quad \nu_e, \quad \nu_\mu, \quad \nu_\tau, \\ &\overline{w}^*, \quad \overline{w}_1, \quad \overline{w}_2, \quad \overline{\nu}_e, \quad \overline{\nu}_\mu, \quad \overline{\nu}_\tau, \end{aligned} \tag{5.3.18}$$

where ν_e, ν_μ, ν_τ are the three flavor neutrinos, and w^*, w_1, w_2 are three new elementary particles, which we call w-weaktons.

These weaktons in (5.3.18) are endowed with the quantum numbers: electric charge Q_e, weak charge Q_w, strong charge Q_s, weak color charge Q_c, baryon number B, lepton numbers L_e, L_μ, L_τ, spin J, and mass m. The quantum numbers of weaktons are listed in Table 5.8.

Table 5.8 Weakton quantum numbers

Weakton	Q_e	Q_w	Q_s	Q_c	L_e	L_μ	L_τ	B	J	m
w^*	$2/3$	1	1	0	0	0	0	$1/3$	$\pm 1/2$	0
w_1	$-1/3$	1	0	1	0	0	0	0	$\pm 1/2$	0
w_2	$-2/3$	1	0	-1	0	0	0	0	$\pm 1/2$	0
ν_e	0	1	0	0	1	0	0	0	$-1/2$	0
ν_μ	0	1	0	0	0	1	0	0	$-1/2$	0
ν_τ	0	1	0	0	0	0	1	0	$-1/2$	0

A few remarks are now in order.

Remark 5.9 For the weaktons and antiweaktons, the quantum numbers Q_e, Q_c, B, L_e, L_μ, L_τ have opposite signs, and Q_w, Q_s, m have the same values. The neutrinos ν_e, ν_μ, ν_τ

possess left-hand helicity with spin $J=-\dfrac{1}{2}$, and the antineutrinos $\overline{\nu}_e,\overline{\nu}_\mu,\overline{\nu}_\tau$ possess right-hand helicity with spin $J=\dfrac{1}{2}$. □

Remark 5.10 The weak color charge Q_c is a new quantum number introduced for the weaktons only, which will be used to rule out some unrealistic combinations of weaktons. □

Remark 5.11 Since the fundamental composite particles as quarks only contain one w^*-weakton, there is no strong interaction between the constituent weaktons of a composite particle, except the gluons. Therefore, for the weaktons (5.3.18), there is no need to introduce the classical strong interaction quantum numbers as strange number S, isospin (I,I_3) and parity π etc. □

Remark 5.12 It is known that the quark model is based on the irreducible representations of $SU(3)$ as

$$\begin{aligned}&\text{meson}=3\otimes\overline{3}=8\oplus 1,\\&\text{baryon}=3\otimes 3\otimes 3=10\oplus 8\oplus 8\oplus 1.\end{aligned}$$

However, the weakton model is based on the aforementioned theories and observational facts 1)-5), different from the quark model.

5.3.4 Weakton constituents and duality of mediators

In this section we introduce the weakton compositions of charged leptons, quarks and mediators. Meanwhile, in the weakton compositions of mediators there exists a natural duality in the spin arrangements, which give rise to the same conclusions as derived in (5.3.17). In addition, the neutrinos ν_l $(l=e,\mu,\tau)$ form a new particle: the ν-mediator $\nu=\sum\alpha_l\nu_l\overline{\nu}_l$, which is of special importance because it can not only explain many decays, but also provide a more reasonable explanation for the well-known solar neutrino problem in Section 6.3.

1. *Charged leptons and quarks.* The weakton constituents of charged leptons and quarks are given by

$$\begin{aligned}&e=\nu_e w_1 w_2, && \mu=\nu_\mu w_1 w_2, && \tau=\nu_\tau w_1 w_2,\\&u=w^* w_1\overline{w}_1, && c=w^* w_2\overline{w}_2, && t=w^* w_2\overline{w}_2,\\&d=w^* w_1 w_2, && s=w^* w_1 w_2, && b=w^* w_1 w_2,\end{aligned}\tag{5.3.19}$$

where c,t and d,s,b are distinguished by their spin arrangements. We suppose that

$$\begin{aligned}&u=w^* w_1\overline{w}_1(\uparrow\uparrow\downarrow,\downarrow\downarrow\uparrow,\uparrow\downarrow\uparrow,\downarrow\uparrow\downarrow,\uparrow\downarrow\downarrow,\downarrow\uparrow\uparrow),\\&c=w^* w_2\overline{w}_2(\uparrow\uparrow\downarrow,\downarrow\downarrow\uparrow,\uparrow\downarrow\downarrow,\downarrow\uparrow\uparrow),\\&t=w^* w_2\overline{w}_2(\uparrow\downarrow\uparrow,\downarrow\uparrow\downarrow),\end{aligned}\tag{5.3.20}$$

and

$$\begin{aligned}&d=w^*w_1w_2(\uparrow\uparrow\downarrow,\downarrow\downarrow\uparrow),\\&s=w^*w_1w_2(\uparrow\downarrow\downarrow,\downarrow\uparrow\uparrow),\\&b=w^*w_1w_2(\uparrow\downarrow\uparrow,\downarrow\uparrow\downarrow).\end{aligned}\tag{5.3.21}$$

These arrangements (5.3.20) and (5.3.21) are speculative, and the true results will have to determined by experiments.

2. *Mediators.* According to the matched quantum numbers, the mediators $\gamma, W^{\pm}, Z, g^k$ should have the following weakton constituents:

$$\begin{aligned}&\gamma=\alpha_1w_1\overline{w}_1+\alpha_2w_2\overline{w}_2 \quad (\alpha_1^2+\alpha_2^2=1),\\&Z=\beta_1w_1\overline{w}_1+\beta_2w_2\overline{w}_2 \quad (\beta_1^2+\beta_2^2=1),\\&W^+=\overline{w}_1\overline{w}_2,\\&W^-=w_1w_2,\\&g^k=w^*\overline{w}^* \quad (k=\text{ color index}).\end{aligned}\tag{5.3.22}$$

In view of the WS electroweak theory (Quigg, 2013):

$$\begin{aligned}&\gamma=\cos\theta_wB_\mu-\sin\theta_wW_\mu^3,\\&Z=\sin\theta_wB_\mu+\cos\theta_wW_\mu^3,\\&\sin^2\theta_w=0.23,\end{aligned}$$

we take $\alpha_1,\alpha_2,\beta_1,\beta_2$ in (5.3.22) as follows

$$\alpha_1=\cos\theta_w, \qquad \alpha_2=-\sin\theta_w, \qquad \beta_1=\sin\theta_w, \qquad \beta_2=\cos\theta_w.$$

There is a natural duality in the spin arrangements:

$$(\uparrow\uparrow,\downarrow\downarrow) \leftrightarrow (\uparrow\downarrow,\downarrow\uparrow),\tag{5.3.23}$$

which not only yields new mediators with spin $J=0$, but also gives the same conclusions as in (5.3.17).

Thus, based on (5.3.23), the weakton model also leads to the dual mediators as follows:

$$\begin{aligned}&\gamma=\cos\theta_ww_1\overline{w}_1-\sin\theta_ww_2\overline{w}_2\ (\uparrow\uparrow,\downarrow\downarrow),\\&Z=\sin\theta_ww_1\overline{w}_1+\cos\theta_ww_2\overline{w}_2\ (\uparrow\uparrow,\downarrow\downarrow),\\&W^-=w_1w_2\ (\uparrow\uparrow,\downarrow\downarrow),\\&W^+=\overline{w}_1\overline{w}_2(\uparrow\uparrow,\downarrow\downarrow),\\&g^k=w^*\overline{w}^*(\uparrow\uparrow,\downarrow\downarrow),\end{aligned}\tag{5.3.24}$$

and their dual particles

$$\begin{aligned}
&\gamma_0 = \cos\theta_w w_1\overline{w}_1 - \sin\theta_w w_2\overline{w}_2\ (\uparrow\downarrow,\downarrow\uparrow),\\
&H^0 = \sin\theta_w w_1\overline{w}_1 + \cos\theta_w w_2\overline{w}_2\ (\uparrow\downarrow,\downarrow\uparrow),\\
&H^- = w_1w_2(\uparrow\downarrow,\downarrow\uparrow),\\
&H^+ = \overline{w}_1\overline{w}_2(\uparrow\downarrow,\downarrow\uparrow),\\
&g_0^k = w^*\overline{w}^*(\uparrow\downarrow,\downarrow\uparrow).
\end{aligned} \tag{5.3.25}$$

3. *The ν-mediator.* By the weak interaction potential formula Φ_m^w in (5.3.15), the neutrino pairs

$$\nu_e\overline{\nu}_e, \quad \nu_\mu\overline{\nu}_\mu, \quad \nu_\tau\overline{\nu}_\tau \quad (\downarrow\uparrow) \tag{5.3.26}$$

should be bounded by the weak interacting force to form a new mediator, although they have not been discovered. The three pairs in (5.3.26) may be indistinguishable. Hence, they will be regarded as a particle, i.e. their linear combination

$$\nu = \sum_l \alpha_l \nu_l\overline{\nu}_l\ (\downarrow\uparrow), \quad \sum_l \alpha_l = 1, \tag{5.3.27}$$

is an additional mediator, and we call it the ν-mediator. We believe that ν is an independent new mediator.

5.3.5 Weakton confinement and mass generation

Since the weaktons are assumed to be massless and no free w-weaktons are found, we have to explain: i) the w-weakton confinement, and ii) the mass generation mechanism for the massive composite particles, including the charged leptons e,μ,τ, the quarks u,d,s,c,t,b, and the bosons $W^\pm, Z, H^\pm, H^0$.

1. *Weakton confinement.* The weak interaction potentials (5.3.15) and the weak charge formula (5.3.16) can help us to understand why no free w^*, w_1, w_2 are found, while single neutrinos ν_e, ν_μ, ν_τ can be detected.

In fact, by (5.3.15) the weak interaction potential reads

$$\Phi_w = g_s e^{-kr}\left[\frac{1}{r} - \frac{B}{\rho_w}(1+2kr)e^{-kr}\right],$$

The bound energy to hold particles together is negative. Hence for weaktons, their weak interaction bound energy E is the negative part of $g_s\Phi_w$, i.e.

$$E = -\frac{B}{\rho_w}g_s^2(1+2kr)e^{-2kr}. \tag{5.3.28}$$

By (5.3.16),

$$g_w^2 = 0.63 \times \left(\frac{\rho_n}{\rho_w}\right)^6 \hbar c. \tag{5.3.29}$$

It is estimated that

$$\frac{\rho_n}{\rho_w} = 10^4 \sim 10^6.$$

Therefore, the bound energy E given by (5.3.28) and (5.3.29) is very large provided the weak interaction constant $B > 0$.

Thus, by the sufficiently large bound energy, the weaktons can form triplets confined in the interior of charged leptons and quarks as (5.3.19), and doublets confined in mediators as (5.3.24)-(5.3.26). They cannot be opened unless the exchange of weaktons between the composite particles.

The free neutrinos ν_e, ν_μ, ν_τ and antineutrinos $\overline{\nu}_e, \overline{\nu}_\mu, \overline{\nu}_\tau$ can be found in Nature. The reason is that in the weakton exchange process there appear pairs of different types of neutrinos such as ν_e and $\overline{\nu}_\mu$, and between which the governing weak force is repelling because the constant B in (5.3.28) is non-positive, i.e.

$$B \leqslant 0 \quad \text{for different types of neutrinos.} \tag{5.3.30}$$

2. *Massless mediators.* For the mass problem, we know that the mediators:

$$\gamma,\ g^k,\ \nu \quad \text{and their dual particles,} \tag{5.3.31}$$

have no masses. To explain this, we note that these particles in (5.3.31) consist of pairs as

$$w_1\overline{w}_1, \quad w_2\overline{w}_2, \quad w^*\overline{w}^*, \quad \nu_l\overline{\nu}_l. \tag{5.3.32}$$

The weakton pairs in (5.3.32) are bound in a circle with radius R_0 as shown in Figure 5.8. Since the interacting force on each weakton pair is in the direction of their connecting line, they rotate around the center 0 without resistance. As $\vec{F} = 0$ in the moving direction, by the relativistic motion law:

$$\frac{d}{dt}\vec{P} = \sqrt{1 - \frac{v^2}{c^2}}\vec{F}, \tag{5.3.33}$$

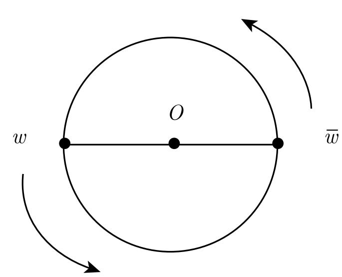


Figure 5.8

the massless weaktons rotate at the speed of light. Hence, the composite particles formed by the weakton pairs in (5.3.32) have no rest mass.

3. *Mass generation.* For the massive particles

$$e,\ \mu,\ \tau,\ u,\ d,\ s,\ c,\ t,\ b, \tag{5.3.34}$$

by (5.3.19), they are made up of weakton triplets with different electric charges. Hence the weakton triplets are not arranged in an equilateral triangle as shown in Figure 5.7 (b), and in fact are arranged in an irregular triangle as shown in Figure 5.9. Consequently, the weakton triplets rotate with nonzero interacting forces $\vec{F} \neq 0$ from the weak and electromagnetic interactions. By (5.3.33), the weaktons in the triplets at a speed less than the speed of light due to the resistance force. Thus, by the mass generating mechanism introduced in Section 5.3.2, the weaktons become massive. Hence, the particles in (5.3.34) are massive.

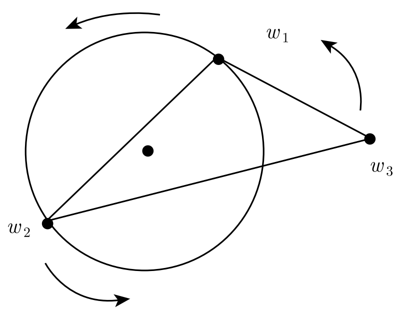


Figure 5.9

4. *Massive mediators.* Finally, we need to explain the masses for the massive mediators:

$$W^{\pm},\ Z,\ H^{\pm},\ H^0, \tag{5.3.35}$$

which consist of the weakton doublets

$$w_1w_2,\ \overline{w}_1\overline{w}_2,\ w_1\overline{w}_1,\ w_2\overline{w}_2. \tag{5.3.36}$$

Actually, in the weakton exchange theory in the next section, we can see that the particles in (5.3.36) are in some transition states in the weakton exchange procedure. At the moment of exchange, the weaktons in (5.3.36) are at a speed v with $v < c$. Hence, the particles in (5.3.35) are massive, and their life-times are very short ($\tau \simeq 10^{-25}s$).

5.3.6 Quantum rules for weaktons

By carefully examining the quantum numbers of weaktons, the composite particles in (5.3.19) and (5.3.24)-(5.3.26) are well-defined.

In the last subsection, we solved the free weakton problem and the mass problem. In this subsection, we propose a few rules to resolve the remainning problems.

1. *Weak color neutral rule*: *All composite particles by weaktons must be weak color neutral.*

Based on this rule, many combinations of weaktons are ruled out. For example, it is clear that there are no particles corresponding to the following *www* and *ww* combinations, because they all violate the weak color neutral rule:

$$\nu_e w_1 w_2,\ w^* w_2 w_2,\ w^* w_1 \overline{w}_2,\ \text{ etc.},\ \nu_e w_1,\ w^* w_1,\ w^* w_2\ \text{ etc.}$$

2. $BL = 0, L_i L_j = 0\ (i \neq j)$*, where B is the baryon number, and* $L = L_j = L_e, L_\mu, L_\tau$ *are the lepton numbers.*

The following combinations of weaktons

$$w^* \nu_i, \quad \nu_i \nu_j, \quad \nu_i \overline{\nu}_k\ (i \neq k),$$

are not observed in Nature, and are ruled out by this rule.

3. $L + Q_e = 0, |B + Q_e| \leqslant 1$ *for* $L, B \neq 0$*.*

The following combinations of weaktons

$$\nu_i w_1 \overline{w}_1,\ \nu_i w_2 \overline{w}_2,\ \overline{\nu}_i w_1 w_2,\ w^* w^*\ \text{ etc} \tag{5.3.37}$$

cannot be found in Nature, and are ruled out this rule.

4. *Spin selection.* In reality, there are no weakton composites with spin $J = \frac{3}{2}$ as

$$w^* w_1 \overline{w}_1 (\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow),\ w^* w_2 \overline{w}_2 (\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow),\ w^* w_1 w_2 (\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow), \tag{5.3.38}$$

and as

$$\nu w_1 w_2\ (\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow). \tag{5.3.39}$$

The cases (5.3.38) are excluded by the Angular Momentum Rule 5.8. The reasons for this exclusion are two-fold. First, the composite particles in (5.3.38) carry one strong charge, and consequently, will be confined in a small ball by the strong interaction potential as the quarks confined in hadrons, as shown in Figure 5.7 (b). Second, due to the uncertainty principle, the bounding particles will rotate, at high speed with almost zero moment of force, which must be excluded for composite particles with $J \neq \dfrac{1}{2}$ based on the angular momentum rule.

The exclusion for (5.3.39) is based on the observation that by the left-hand helicity of neutrinos with spin $J=-\frac{1}{2}$, one of w_1 and w_2 must be in the state with $J=+\frac{1}{2}$ to combine with ν_j $(1\leqslant j\leqslant 3)$, i.e. in the manner as

$$\nu_j w_1 w_2\ (\downarrow\upuparrows, \downarrow\uparrow\downarrow, \downdownarrows\uparrow).$$

In summary, under the above rules 1-4, only the weakton constituents in (5.3.19) and (5.3.24)-(5.3.26) are allowed.

5. *Eight quantum states of gluons.* It is known that the gluons have eight quantum states

$$g^k:\ g^1, \cdots, g^8.$$

In (5.3.24) and (5.3.25), the vector and scalar gluons have the forms

$$w^*\overline{w}^*(\upuparrows, \downdownarrows) \quad \text{and} \quad w^*\overline{w}^*(\uparrow\downarrow, \downarrow\uparrow).$$

According to *QCD*, quarks have three colors

$$\text{red}(r), \quad \text{green}(g), \quad \text{blue}(b),$$

and anticolor $\overline{r}, \overline{g}, \overline{b}$. They obey the rules

$$b\overline{b}=r\overline{r}=g\overline{g}=w \quad (\text{ white}).$$

Based on (5.3.19), w^* is endowed with three colors and anticolors:

$$w_r^*,\ w_g^*,\ w_b^* \quad \text{and} \quad \overline{w}_{\overline{r}}^*,\ \overline{w}_{\overline{g}}^*,\ \overline{w}_{\overline{b}}^*,$$

which form the bases of $SU(3)$ and $SU(\overline{3})$. By the irreducible representation:

$$3\times\overline{3}=8\oplus 1,$$

we can derive the eight-multiple states as

$$g^1=(w^*\overline{w})_w, \qquad g^2=w_b^*\overline{w}_{\overline{r}}^*, \qquad g^3=w_b^*\overline{w}_{\overline{g}}^*, \qquad g^4=w_r^*\overline{w}_{\overline{g}}^*, \tag{5.3.40}$$
$$g^5=(w^*\overline{w})_w, \qquad g^6=w_r^*\overline{w}_{\overline{b}}^*, \qquad g^7=w_g^*\overline{w}_{\overline{b}}^*, \qquad g^8=w_g^*\overline{w}_{\overline{r}}^*. \tag{5.3.41}$$

where $(w^*\overline{w}^*)_w$ is a linear combination of $w_b^*\overline{w}_{\overline{b}}^*, w_r^*\overline{w}_{\overline{r}}^*, w_g^*\overline{w}_{\overline{g}}^*$. Namely, the gluons in (5.3.41) are the antigluons of these in (5.3.40).

In summary, the quantum rules presented above can be simply expressed in the following formulas for quantum numbers Q_c, Q_e, L and B:

$$\begin{array}{ll} \text{weak color neutral:} & Q_c=0, \\ \text{mutual exclusion of leptons and baryons:} & BL=0,\ L_iL_j=0\ (i\neq j), \\ \text{relation of leptons and charges:} & L+Q_e=0\ (L\neq 0), \\ \text{relation of baryons and charges:} & |B+Q_e|\leqslant 1\ (B\neq 0). \end{array} \tag{5.3.42}$$

These relations are summed up from realistic particle data. Based on (5.3.42), together with the mass generation mechanism, the Angular Momentum Rule 5.8 and the layered formulas of weak and strong interaction potentials, all of the most basic problems in the weakton model have a reasonable explanation.

5.4 Mechanisms of Subatomic Decays and Electron Radiations

5.4.1 Weakton exchanges

We conclude that all particle decays are caused by weakton exchange. The exchanges occur between composite particles as mediators, charged leptons, and quarks.

1. *Weakton exchange in mediators.* First we consider one of the most important decay processes in particle physics, the electron-positron pair creation and annihilation:

$$\begin{aligned} &2\gamma \to e^+ + e^-, \\ &e^+ + e^- \to 2\gamma. \end{aligned} \tag{5.4.1}$$

In fact, the reaction formulas in (5.4.1) are not complete, and the correct formulas should be as follows

$$2\gamma + \nu \rightleftarrows \ e^+ + e^-, \tag{5.4.2}$$

where ν is the ν-mediator as in (5.3.27).

Note that the weakton components of γ and ν are

$$\begin{aligned} &\gamma = \cos\theta_w w_1 \overline{w}_1 - \sin\theta_w w_2 \overline{w}_2, \\ &\nu = \alpha_1 \nu_e \overline{\nu}_e + \alpha_2 \nu_\mu \overline{\nu}_\mu + \alpha_3 \nu_\tau \overline{\nu}_\tau, \end{aligned}$$

which imply that the probability of the photon γ at the state $w_1\overline{w}_1$ is $\cos^2\theta_w$ and at the state $-w_1\overline{w}_2$ is $\sin^2\theta_w$, and the probability of ν at $\nu_i\overline{\nu}_i$ is α_i^2. Namely, for photon, the densities of the $w_1 w_1(\uparrow\uparrow)$ and $-w_2\overline{w}_2(\downarrow\downarrow)$ particle states are $\cos^2\theta_w$ and $\sin^2\theta_w$, and the densities of the states $\nu_e\overline{\nu}_e, \nu_\mu\overline{\nu}_\mu,\ \nu_\tau\overline{\nu}_\tau$ are $\alpha_1^2, \alpha_2^2, \alpha_3^2$. Hence the formula (5.4.2) can be written as

$$w_1\overline{w}_1(\uparrow\uparrow) + w_2\overline{w}_2(\downarrow\downarrow) + \nu_e\overline{\nu}_e(\downarrow\uparrow) \ \rightleftarrows \ \nu_e w_1 w_2(\downarrow\uparrow\downarrow) + \overline{\nu}_e\overline{w}_1\overline{w}_2(\uparrow\uparrow\downarrow). \tag{5.4.3}$$

It is then clear to see from (5.4.3) that the weakton constituents $w_1, \overline{w}_1, w_2, \overline{w}_2, \nu_e, \overline{\nu}_e$ can regroup due to the weak interaction, and we call this process weakton exchange. The mechanism of this exchanging process can be explained using the layered formulas in (5.3.15) of weak interaction potentials.

The layered potential formulas in (5.3.15) indicate that each composite particle has an exchange radius R, which satisfies

$$\rho < R < \rho_1, \tag{5.4.4}$$

where ρ is the radius of this particle and ρ_1 is its repelling radius of weak force. If two composite particles A and B are in a distance less than their common exchange radius, then the weaktons in A and B are attracting to each other by their weak interacting forces generated by Φ_0^w in (5.3.15), given by

$$\Phi_0^w = g_w e^{-kr}\left[\frac{1}{r} - \frac{B_w}{\rho_w}(1+2kr)e^{-kr}\right]. \tag{5.4.5}$$

There is a probability for these weaktons in A and B to recombine and form new particles. Then, after the new particles have been formed, in the exchange radius R, the weak interacting force between them is governed by the potentials of the new particles, and is repelling, driving the newly formed particles apart.

For example, in Figure 5.10 we can clearly see how the weaktons in (5.4.3) undergo the exchange process. When the randomly moving photons and ν-mediators, i.e. $w_1\overline{w}_1, w_2\overline{w}_2$ and $\nu_e\overline{\nu}_e$ come into their exchange balls, they recombine to form an electron $\nu_e w_1 w_2$ and a positron $\overline{\nu}_e\overline{w}_1\overline{w}_2$ under the weak interaction attracting forces generated by the weakton potentials (5.4.5). Then, the weak interacting force between $e^- = \nu_e w_1 w_2$ and $e^+ = \overline{\nu}_e\overline{w}_1\overline{w}_2$ governed by Φ_l^w in (5.3.15) is repelling, and pushes them apart, leading to the decay process (5.4.2)

We remark here that in this range the weak repelling force between e^+ and e^- is much stronger than the Coulomb attracting force. In fact, by (5.3.29), $g_w^2 \sim 10^{30}\hbar c$ and the electric charge square $e^2 = \frac{1}{137}\hbar c$. Hence the weak repelling force in Figure 5.10 is $(3g_s)^2/r^2$, much stronger than e^2/r^2.

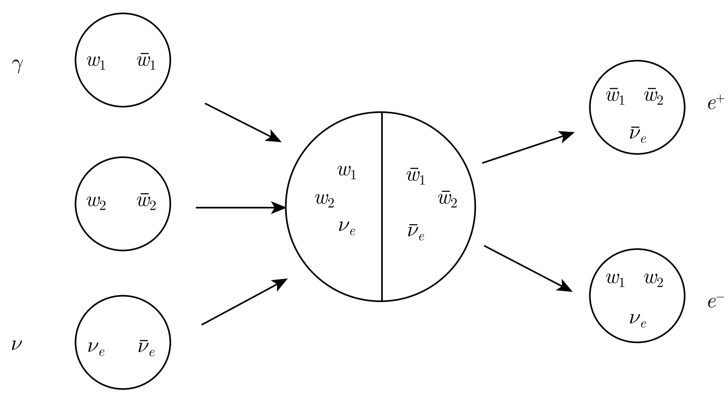


Figure 5.10

2. *Weakton exchanges between leptons and mediators.* The μ-decay reaction formula is given by

$$\mu^- \to e^- + \overline{\nu}_e + \nu_\mu. \tag{5.4.6}$$

The complete formula for (5.4.6) is

$$\mu^- + \nu \to e^- + \overline{\nu}_e + \nu_\mu,$$

which is expressed in the weakton components as

$$\nu_\mu w_1 w_2 + \nu_e \overline{\nu}_e \to \nu_e w_1 w_2 + \overline{\nu}_e + \nu_\mu. \tag{5.4.7}$$

By the exclusive rule $L_e L_\mu = 0$, the two particles ν_μ and $\overline{\nu}_e$ cannot be combined to form a particle. Hence, $\overline{\nu}_e$ and ν_μ appear as independent particles, leading to the exchange of ν_μ and ν_e as in (5.4.7).

3. *Weakton exchanges between quarks and mediators.* The d-quark decay in (5.3.2) is written as

$$d \to u + e^- + \overline{\nu}_e. \tag{5.4.8}$$

The correct formula for (5.4.8) is

$$d + \gamma + \nu \to u + e^- + \overline{\nu}_e,$$

which, in the weakton components, is given by

$$w^* w_1 w_2 + w_1 \overline{w}_1 + \nu_e \overline{\nu}_e \to w^* w_1 \overline{w}_1 + \nu_e w_1 w_2 + \overline{\nu}_e. \tag{5.4.9}$$

In (5.4.9), the weakton pair w_2 and $\overline{w}_1$ is exchanged, and ν_e is captured by the new doublet $w_1 w_2$, which is the vector boson W^-, to form an electron $\nu_e w_1 w_2$.

5.4.2 Conservation laws

The weakton exchanges must obey certain conservation laws, which are listed in the following.

1. *Conservation of weakton numbers.* The weaktons given in (5.3.18) are elementary particles, which cannot undergo any decay. Also, the w-weaktons cannot be converted between each other. Although the neutrino oscillation, which is unconfirmed, may convert one flavour of neutrino to another, for a particle decay or scattering, the neutrino number is still conserved. Namely, the lepton numbers L_e, L_μ, L_τ are conserved.

Therefore, for any particle reaction:

$$A_1 + \cdots + A_N \to B_1 + \cdots + B_K, \tag{5.4.10}$$

the number of each weakton type is invariant. Namely, for any type of weakton $\widetilde{w}$, its number is conserved in (5.4.10):

$$N^A_{\widetilde{w}} = N^B_{\widetilde{w}},$$

where $N^A_{\widetilde{w}}$ and $N^B_{\widetilde{w}}$ are the numbers of the $\widetilde{w}$ weaktons in both sides of (5.4.10).

2. *Spin conservation.* The spin of each weakton is invariant. The conservation of weakton number implies that the spin is also conserved:

$$J_{A_1} + \cdots + J_{A_N} = J_{B_1} + \cdots + J_{B_K},$$

where J_A is the spin of particle A.

In classical particle theories, the spin is not considered as a conserved quantity. The reason for the non-conservation of spin is due to the incompleteness of the reaction formulas given in Section 5.1.3. Hence, spin conservation can also be considered as an evidence for the incompleteness of those reaction formulas. The incomplete decay reaction formulas can be made complete by supplementing some massless mediators, so that the spin becomes a conserved quantum number.

3. *Energy conservation.* Energy conservation is a universal physical law for all particle systems. Hence, the energy conservation has also to be satisfied by the weakton exchanges. This law is manifested in the following prohibitions for decays:

$$\begin{aligned} e^- &\not\to \mu^- + \overline{\nu}_\mu + \nu_e, \\ e^- &\not\to \tau^- + \overline{\nu}_\tau + \nu_e, \\ \mu^- &\not\to \tau^- + \overline{\nu}_\tau + \nu_\mu. \end{aligned} \tag{5.4.11}$$

The complete reaction formulas for (5.4.11) are

$$\begin{aligned} &\nu_e w_1 w_2 + \nu_\mu \overline{\nu}_\mu \to \nu_\mu w_1 w_2 + \overline{\nu}_\mu + \nu_e, \\ &\nu_e w_1 w_2 + \nu_\tau \overline{\nu}_\tau \to \nu_\tau w_1 w_2 + \overline{\nu}_\tau + \nu_e, \\ &\nu_\mu w_1 w_2 + \nu_\tau \overline{\nu}_\tau \to \nu_\tau w_1 w_2 + \overline{\nu}_\tau + \nu_\mu. \end{aligned} \tag{5.4.12}$$

From the viewpoints of quantum rules given by (5.3.42) and weakton number conservation, the decays in (5.4.12) are allowed. However, due to the mass relations

$$m_e < m_\mu < m_\tau,$$

these weakton exchange processes (5.4.12) violate the energy conservation. Therefore these decays cannot occur. However, if ν-mediators have high energy to hit electrons, then the reactions (5.4.12) may occur.

In a weakton exchange process, the energy conservation law can be explicitly expressed as follows:

Energy Rule 5.13 *The composite particles with the lower masses cannot undergo weakton exchanges with massless mediators in lower energy to decay into the composite particles with higher masses.*

We remark that the Energy Rule 5.13 is sharper than the energy conservation law. In fact, the reaction formulas from lower masses to higher masses do not imply that the energy conservation must be violated. Hence, it is possible that there exist more basic unknown physical laws under the Energy Rule 5.13.

4. *Other conservative quantum numbers.* From the invariance of weakton numbers, we deduce immediately the following conserved quantum numbers:

electric charge Q_e,	weak charge Q_w,	strong charge Q_s,
baryon number B,	lepton numbers $L_e,\ L_\mu,\ L_\tau$.	

5.4.3 Decay types

In particle physics, the reactions listed in Section 5.1.3 are classified into two types: the weak interacting type and the strong interacting type. However, there are no clear definitions to distinguish them. Usual methods are by experiments to determine reacting intensity, i.e. the transition probability. In general, the classification is derived based on

weak type: i) presence of leptons in the reactions,

ii) change of strange numbers,

strong type: otherwise.

With the weakton model, all decays are carried out by exchanging weaktons. Hence decay types can be fully classified into three types: the weak type, the strong type, and the mixed type, based on the type of forces acting on the final particles after the weakton exchange process.

For example, the reactions

$$\begin{aligned} \nu_\mu + e^- &\to \mu^- + \nu_e, \\ n &\to p + e^- + \overline{\nu}_e, \\ \pi^0 &\to \gamma + \gamma. \end{aligned} \tag{5.4.13}$$

are weak decays,

$$\Delta^{++} \to p^+ + \pi^+ \tag{5.4.14}$$

is a strong decay, and

$$\Lambda \to p^+ + \pi^- \quad (\text{ i.e. } \Lambda + g^k + 3\gamma \to p^+ + \pi^- + \gamma) \tag{5.4.15}$$

is a mixed decay.

In view of (5.4.13)-(5.4.15), the final particles contain at most one hadron in a weak decay, contain no leptons and no mediators in a strong decay, and contain at least two hadrons and a lepton or a mediator in a mixed decay.

In summary, we give the criteria for decay types based on the particles appearing in the final step of the decay process:

$$\begin{aligned} &\text{weak decay:} && \text{at most one hadron,} \\ &\text{strong decay:} && \text{no leptons and no mediators,} \\ &\text{mixed decay:} && \text{others.} \end{aligned} \tag{5.4.16}$$

Remark 5.14 The new classification is easy to understand, and the criterion (5.4.16) provides a convenient method for us to distinguish the different types of decays and scatterings. In fact, this classification truly reflects the roles of interactions in particle reaction processes.

5.4.4 Decays and scatterings

Decays and non-elastic scatterings are caused by weakton exchanges. The massless mediators

$$\gamma,\ g^k,\ \gamma_0,\ g_0^k,\ \nu, \tag{5.4.17}$$

spread over the space or around the charged leptons and quarks at various energy levels, and most of them are at low energy states. It is these random mediators in (5.4.17) entering the exchange radius of matter particles that generate decays. In the following we shall discuss these reaction processes for various types of decays and scatterings.

Weak types

1. First we consider the e-μ scattering

$$\nu_\mu + e^- \to \mu^- + \nu_e,$$

which is rewritten in the weakton components as

$$\begin{aligned} \nu_\mu + \nu_e w_1 w_2 \quad &\to w_1 w_2 (W^-) + \nu_\mu + \nu_e \\ &\to \nu_\mu w_1 w_2 + \nu_e. \end{aligned} \tag{5.4.18}$$

In (5.4.18) we can see that the transient vector boson W^- appears, and then captures ν_μ to form the muon μ. The Feynman digram of (5.4.18) is given by Figure 5.11.

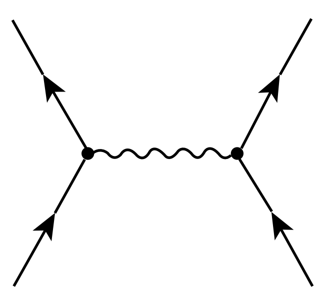


Figure 5.11

Replacing the Feynman diagram, we describe the scattering (5.4.18) by using Figure 5.12. It is clear that the scattering (5.4.18) is achieved by exchanging weaktons ν_μ and ν_e.

2. β-*decay.* Consider the classical β-decay process

$$n \to p + e^- + \overline{\nu}_e. \tag{5.4.19}$$

with the quark constituents of n and p:

$$n = udd, \quad p = uud,$$

the β-decay (5.4.19) is equivalent to the following d-quark decay:

$$d \to u + e^- + \overline{\nu}_e,$$

whose complete form should be given by

$$\begin{aligned} w^* w_1 w_2(d) + \nu_e \overline{\nu}_e(\nu) + w_1 \overline{w}_1(\gamma) &\to w^* w_1 \overline{w}_1(u) + w_1 w_2(W^-) + \nu_e \overline{\nu}_e(\nu) \\ &\to w^* w_1 \overline{w}_1(u) + \nu_e w_1 w_2(e^-) + \overline{\nu}_e. \end{aligned} \tag{5.4.20}$$

In the β-decay (5.2.20), w^* in d-quark and photon $\gamma = w_1 \overline{w}_1$ recombine to form u-quark and charged vector boson W^-, then W^- captures ν_e from ν-mediator to yield an electron e^- and an $\overline{\nu}_e$.

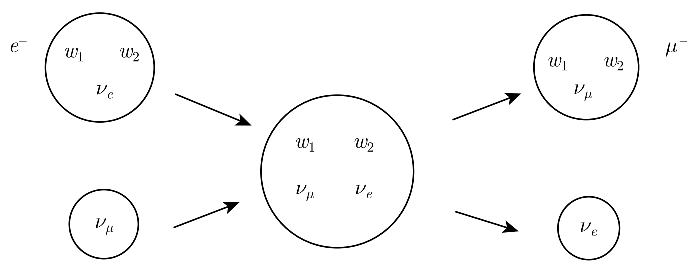


Figure 5.12 $e - \mu$ scattering

3. *Quark pair creations.* Consider

$$g^k + \gamma_0 + \gamma \to u + \overline{u},$$
$$g_0^k + \gamma_0 + \gamma_0 \to d + \overline{d}.$$

They are rewritten in the weakton constituent forms as

$$w^*\overline{w}^* \upuparrows (g^k) + w_1\overline{w}_1 \downarrow\uparrow (\gamma_0) + w_1\overline{w} \downdownarrows (\gamma) \tag{5.4.21}$$
$$\to w^*w_1\overline{w}_1 \uparrow\downarrow\uparrow (u) + \overline{w}^*w_1\overline{w}_1 \uparrow\downdownarrows (\overline{u}),$$

$$w^*\overline{w}^* \uparrow\downarrow (g_0^k) + w_1\overline{w}_1 \uparrow\downarrow (\gamma_0) + w_2\overline{w}_2 \downarrow\uparrow (\gamma_0) \tag{5.4.22}$$
$$\to w^*\overline{w}^* \uparrow\downarrow (g_0^k) + w_1w_2 \uparrow\downarrow (H^-) + \overline{w}_1\overline{w}_2 \downarrow\uparrow (H^+)$$
$$\to w^*w_1w_2 \upuparrows\downarrow (d) + \overline{w}^*\overline{w}_1\overline{w}_2 \downdownarrows\uparrow (\overline{d}).$$

In (5.4.21), w^* and $\overline{w}^*$ in a gluon are captured by a scalar photon γ_0 and a photon γ to create a pair u and $\overline{u}$. In (5.4.21), $\overline{w}_1$ and w_2 in the two scalar photons γ_0 are exchanged to form a pair of charged Higgs H^+ and H^-, then H^+ and H^- capture w^* and $\overline{w}^*$ respectively to create a pair of d and $\overline{d}$.

4. *Lepton decays.* The lepton decays

$$\mu^- + \nu \to e^- + \overline{\nu}_e + \nu_\mu,$$
$$\tau^- + \nu \to \mu^- + \overline{\nu}_\mu + \nu_\tau,$$

are rewritten in the weakton constituents as

$$\begin{aligned} \nu_\mu w_1 w_2 + \nu_e\overline{\nu}_e &\quad\to\quad \nu_e w_1 w_2 + \overline{\nu}_e + \nu_\mu, \\ \nu_\tau w_1 w_2 + \nu_\mu\overline{\nu}_\mu &\quad\to\quad \nu_\mu w_1 w_2 + \overline{\nu}_\mu + \nu_\tau. \end{aligned} \tag{5.4.23}$$

Here the neutrino exchanges form leptons in the lower energy states and a pair of neutrino and antineutrino with different lepton numbers. By the exclusive rule $L_iL_j = 0$ $(i \neq j)$, the generated neutrino and antineutino cannot be combined together, and are separated by the weak interaction repelling force.

Strong types

1. Δ^{++}*-decay.* Consider the following decay

$$\Delta^{++} \to p + \pi^+.$$

The complete decay process should be

$$\begin{aligned} &\Delta^{++} + g_0^k + 2\gamma_0 \to p + \pi^+, \\ &\Delta^{++} + g^k + 2\gamma \to p + \pi^+. \end{aligned} \tag{5.4.24}$$

It is clear that the final particles are the proton p and meson π^+. Hence by the criterion (5.4.16), the Δ^{++} decay is a strong type. Recalling the weakton constituents in (5.3.19) and the quark constituents of hadrons in Section 5.1.1, the first reaction formula in (5.4.24) is rewritten as

$$3w^*w_1\overline{w}_1(\Delta^{++}) + w^*\overline{w}^*(g_0^k) + w_1\overline{w}_1(\gamma_0) + w_2\overline{w}_2(\gamma_0) \tag{5.4.25}$$
$$\to (2w^*w_1\overline{w}_1)(w^*w_1w_2)(p) + (w^*w_1\overline{w}_1)(\overline{w}^*\overline{w}_1\overline{w}_2)(\pi^+).$$

This reaction process in (5.4.25) consists of two steps:

$$\text{weakton exchanges:} \qquad g_0^k + 2\gamma_0 \to d + \overline{d}, \qquad (5.4.26)$$

$$\text{quark exchanges:} \qquad uuu + d\overline{d} \to uud + u\overline{d}. \qquad (5.4.27)$$

The exchange mechanism of (5.4.26) was discussed in (5.4.22), which is a weak interaction, and the quark exchange (5.4.27) is both weak and strong interaction types. But, the final particles p and π^+ are driven apart by the strong hadron repelling force.

After the weakton exchange to yield d-quark pair in (5.4.26), the mechanism of the quark exchange and decay in (5.4.27) can be interpreted as follows:

1) When the quark pair $d\overline{d}$ is formed in the exchange radius R of Δ^{++}, the strong and weak interactions between the quarks $d,\overline{d}$ and u,u,u in Δ^{++} are governed by Φ_q^s in (5.3.12) and Φ_q^w in (5.3.15), which are attracting and recombine these quarks to form two new hadrons p and π^+, as shown in Figure 5.13(a); and

2) the two newly formed hadrons p and π^+ are controlled by the strong interaction potential Φ_n for hadrons, which is repulsive in the exchange radius R and pushes them apart; see Figure 5.13(b).

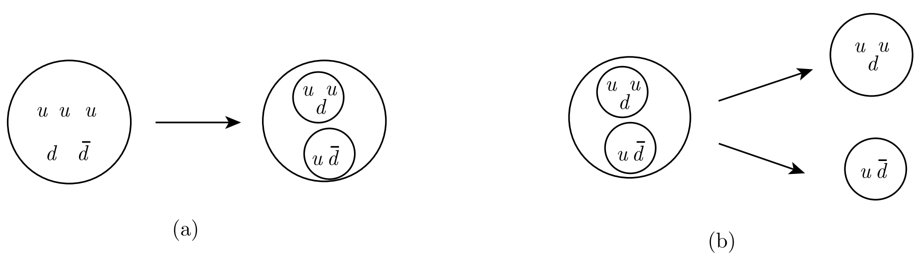


Figure 5.13

2. D^0*-decay.* Let us discuss the D^0-decay, which is considered as the weak interacting type in the classical theory, because it violates the strange number conservation. But in our classification it belongs to strong type of interactions. The D^0-decay is written as

$$D^0 \to K^- + \pi^+.$$

The complete formula is

$$D^0 + g^k + 2\gamma \to K^- + \pi^+.$$

The weakton constituents of this decay is given by

$$\begin{aligned}&(w^* w_2 \overline{w}_2)(\overline{w}^* \overline{w}_1 w_1)(c\overline{u}) + w^* \overline{w}^* (g^k) + 2 w_1 \overline{w}_1 (\gamma)\\ &\quad \to (w^* w_1 w_2)(\overline{w}^* \overline{w}_1 w_1)(c\overline{u}) + (w^* w_1 \overline{w}_1)(\overline{w}^* \overline{w}_1 \overline{w}_2)(u\overline{d}).\end{aligned}$$

This reaction is due to the c-quark decay

$$c+g^k+2\gamma \to s+u+\overline{d},$$

which is given in the weakton constituents as

$$w^*w_2\overline{w}_2(c)+w^*\overline{w}^*(g^k)+2w_1\overline{w}_1(\gamma) \to w^*w_1w_2(s)+w^*w_1\overline{w}_1(u)+w^*\overline{w}_1\overline{w}_2(d). \quad (5.4.28)$$

The reaction (5.4.28) consists of two exchange processes:

$$w^*w_2\overline{w}_2(c)+w_1\overline{w}_1(\gamma) \to w^*w_1w_2(s)+\overline{w}_1\overline{w}_2(w^-), \quad (5.4.29)$$

and

$$\overline{w}_1\overline{w}_2(W^-)+w_1\overline{w}_1(\gamma)+(w^*\overline{w}^*)(g^k) \to w^*w_1\overline{w}_1(u)+\overline{w}^*\overline{w}_1\overline{w}_2(d) \quad (5.4.30)$$

It is clear that both exchanges (5.4.29) and (5.4.30) belong to weak interactions. However, the final particles in the D^0-decay are K^- and π^+, which are separated by the strong hadron repelling force. The principle for the quark exchange and hadron separation in the D^0-decay process is the same as shown in Figure 5.13.

Remark 5.15 In the mechanism of subatomic particle decays and scatterings, the layered properties of the weak and strong interacting forces play a crucial role. Namely the attracting and repelling radii of weak and strong forces for a particle depend on its radius ρ and interaction constants A and B or equivalently the parameters

$$\frac{A}{\rho} \quad \text{in (5.3.11)} \qquad \text{and} \qquad \frac{B}{\rho} \quad \text{in (5.3.14).} \quad (5.4.31)$$

These parameters determine the attracting and repelling regions of the particle. Hence, in the exchange radius of particles, the weak and strong forces are attracting between weaktons and quarks, and are repelling between the final particles in decays and scatterings. The reason is that the parameters in (5.4.31) are different at various levels of subatomic particles. □

Mixed decays

The typical mixed decay is the Λ-decay, written as

$$\Lambda \to p+\pi^-.$$

The correct form of Λ-decay should be

$$\Lambda+g^k+2\gamma+\gamma_0 \to p+\pi^-+\gamma_0. \quad (5.4.32)$$

There are three exchange processes in (5.4.32):

$$g^k+\gamma+\gamma_0\to u+\overline{u}, \tag{5.4.33}$$

$$s+\gamma\to d+\gamma_0 \quad (\text{i.e. } uds+\gamma\to udd+\gamma_0), \tag{5.4.34}$$

$$udd(n)+u\overline{u}\to uud(p)+u\overline{d}(\pi^-). \tag{5.4.35}$$

The process (5.4.33) was described by (5.4.21), the quark exchange process (5.4.35) is clear, and (5.4.34) is the conversion from s quark to d quark, described by

$$w^*w_1w_2\uparrow\downarrow\downarrow(s)+w_1\overline{w}_1\uparrow\uparrow(\gamma)\to w^*w_1w_2\uparrow\uparrow\downarrow(d)+w_1\overline{w}_1\downarrow\uparrow(\gamma_0). \tag{5.4.36}$$

Namely, (5.4.36) is an exchange of two w_1 with reversed spins.

5.4.5 Electron structure

The weakton constituents of an electron are $\nu_e w_1 w_2$, which rotate as shown in Figure 5.9. Noting the electric charges and weak charges of ν_e, w_1, w_2 are given by

$$\begin{aligned}
&\text{electric charge:} \quad && Q_e^\nu=0, \quad && Q_e^{w_1}=-\frac{1}{3}, \quad && Q_e^{w_2}=-\frac{2}{3},\\
&\text{weak charge:} \quad && Q_w^\nu=1, \quad && Q_w^{w_1}=1, \quad && Q_w^{w_2}=1,
\end{aligned}$$

we see that the distribution of weaktons ν_e, w_1, w_2 in an electron is in an irregular triangle due to the asymmetric forces on the weaktons by the electromagnetic and weak interactions, as shown in Figure 5.14.

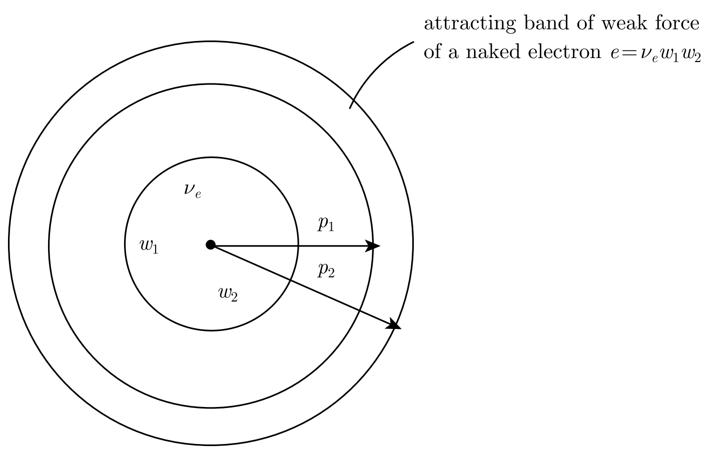


Figure 5.14 Electron structure

In addition, the weak force formula between the naked electron and massless mediators $\gamma, \gamma_0, g^k, g_0^k, \nu$, is given by

$$
\begin{aligned}
F_w = & - g_w(\rho_m) g_w(\rho_e) \frac{\partial}{\partial r}\left[\frac{1}{r} e^{-kr} - \frac{\widetilde{B}}{\widetilde{\rho}}(1+2kr)e^{-2kr}\right], \qquad (5.4.37)\\
= & g_w(\rho_m) g_w(\rho_e) e^{-kr}\left[\frac{1}{r^2} + \frac{1}{r r_0} - \frac{4\widetilde{B}}{\widetilde{\rho}}\frac{r}{r_0^2} e^{-kr}\right],
\end{aligned}
$$

where $k = 1/r_0 = 10^{16}\ \text{cm}^{-1}, g_w(\rho_m)$ and $g_w(\rho_e)$ are the weak charges of mediators and the naked electron, expressed as

$$
g_w(\rho_m) = 2\left(\frac{\rho_w}{\rho_m}\right)^3 g_w, \quad g_w(\rho_e) = 3\left(\frac{\rho_w}{\rho_e}\right)^3 g_w,
$$

and $\widetilde{B}/\widetilde{\rho}$ is a parameter determined by the naked electron and mediators.

By the weak force formula (5.4.37), there is an attracting shell region of weak interaction between naked electron and mediators

$$
\rho_1 < r < \rho_2 \qquad (5.4.38)
$$

as shown in Figure 5.13, with small weak force, where $\rho_j\ (j = 1,2)$ are the zero points of (5.4.37). Namely, they satisfy that

$$
e^{-k\rho_j}\rho_j^3 = \frac{\widetilde{\rho}}{4\widetilde{B}}\left(1 + \frac{\rho_j}{r_0}\right), \qquad \text{for } 1 \leqslant j \leqslant 2.
$$

In the region (5.4.38), the weak force is attracting, and outside this region the weak force is repelling:

$$
F_w \begin{cases} < 0 & \text{for } \rho_1 < r < \rho_2, \\ > 0 & \text{for } r < \rho_1 \ \text{ and } \rho_2 < r. \end{cases} \qquad (5.4.39)
$$

Since the mediators $\gamma, \gamma_0, g^k, g_0^k$ and ν contain two weak charges $Q_w = 2g_w$, by (5.4.39) they are attached to the electron in the attracting shell region (5.4.38), forming a cloud of mediators. The irregular triangle distribution of the weaktons ν_e, w_1, w_2 generate a small moment of force on the mediators. Meanwhile there also exist weak forces between them. Therefore the bosons will rotate at a speed less than the speed of light, and generate a small mass attached to the naked electron $\nu_e w_1 w_2$.

5.4.6 Mechanism of bremsstrahlung

It is known that an electron emits photons as its velocity changes, which is called the bremsstrahlung. The reasons why bremsstrahlung can occur is unknown in classical theories. Based on the electron structure theory established in the last subsection, we present here a mechanism of this phenomenon.

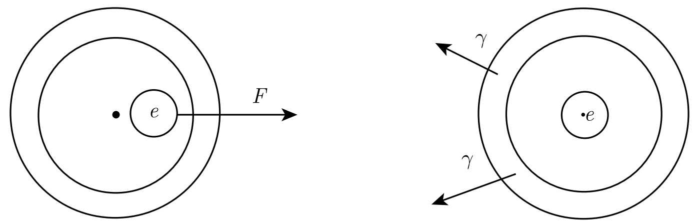


Figure 5.15 (a) The naked electron is accelerated or decelerated in electromagnetic field; amd (b) the mediators (photons) fly away from the attracting shell region under a perturbation of moment of force.

In fact, if an electron is situated in an electromagnetic field, then the electromagnetic field exerts a Coulomb force on ithe naked electron $\nu_e w_1 w_2$, but not on the attached neutral mediators. Thus, the naked electron changes its velocity, which draws the mediator cloud to move as well, causing a perturbation to moment of force on the mediators. As the attracting weak force in the shell region (5.4.38) is small, under the perturbation, the centrifugal force makes some mediators in the cloud, such as photons, flying away from the attracting shell region, and further accelerated by the weak repelling force outside this shell region to the speed of light, as shown in Figure 5.15.

5.5 Structure of Mediator Clouds Around Subatomic Particles

5.5.1 Color quantum number

In the baryon family with spin $J = \frac{3}{2}$ there are three members $\Delta^{++}, \Delta^{-}, \Omega^{-}$, whose quark constituents are as follows

$$\begin{aligned} &\Delta^{++} = uuu(\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow), \\ &\Delta^{-} = ddd(\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow), \\ &\Omega^{-} = sss(\uparrow\uparrow\uparrow, \downarrow\downarrow\downarrow). \end{aligned} \tag{5.5.1}$$

These quarks have the same spin arrangements and energy levels. The phenomenon that there are three identical fermions at one quantum state violates the Pauli exclusion principle. Thus, there exist two possibilities:

1) Pauli exclusion principle is invalid in the interior of baryons; or

2) there is a new quantum number, such that the three same flavour of quarks in the baryons of (5.5.1) possess different values for the new quantum number. Namely, they are not identical, and the Pauli exclusion principle is still valid.

To solve this problem, O. W. Greenberg presented the color quantum number in 1964. According to the constituents of (5.5.1), Greenberg thought that each flavor of quarks has three colors:

$$\text{red } r, \quad \text{green } g, \quad \text{blue } b, \tag{5.5.2}$$

and antiparticles have three anti-colors:

$$\text{anti-red } \overline{r}, \quad \text{anti-green } \overline{g}, \quad \text{anti-blue } \overline{b}. \tag{5.5.3}$$

Based on the quantum numbers (5.5.2) and (5.5.3), each flavor of quarks is endowed with three different colors

$$q_r,\ q_g,\ q_b,\ \overline{q}_{\overline{r}},\ \overline{q}_{\overline{g}},\ \overline{q}_{\overline{b}}.$$

The color indices of hadrons are given by

$$\begin{aligned} &\text{baryon} = q_{1r}q_{2g}q_{3b}, \\ &\text{meson} = q_{1r}\overline{q}_{2\overline{r}},\ q_{1g}\overline{q}_{2\overline{g}},\ q_{1b}\overline{q}_{2\overline{b}}. \end{aligned} \tag{5.5.4}$$

Since hadrons are color neutral, the color quantum number should obey the following multiplication rule:

$$\begin{aligned} &r\overline{r} = w, \quad && g\overline{g} = w, \quad && b\overline{b} = w, \\ &rgb = w, \quad && \overline{r}\,\overline{g}\,\overline{b} = w, \quad && w = \text{ white color}, \end{aligned} \tag{5.5.5}$$

and the multiplication is commutative.

The color quantum number has attained many experimental supports. Quantum Chromodynamics (QCD) for the strong interaction is based on this theory. In fact, it is natural to think that the three color quantum states of each flavour of quarks

$$q = (q_r,\ q_g,\ q_b) \tag{5.5.6}$$

are indistinguishable in the strong interaction. This gives rise to the $SU(3)$ gauge theory of QCD, i.e. the QCD action is the Yang-Milli functional of $SU(3)$ gauge fields, given by

$$\mathscr{L}_{QCD} = -\frac{1}{4}S^k_{\mu\nu}S^{\mu\nu k} + \overline{q}(i\gamma^\mu D_\mu - m)q, \tag{5.5.7}$$

where q is the quark triplet as in (5.5.6), m is the mass of the quark,

$$\begin{aligned} &S^k_{\mu\nu} = \partial_\mu S^k_\nu - \partial_\nu S^k_\mu + g_s f^k_{ij} S^i_\mu S^j_\nu, \\ &D_\mu = \partial_\mu + ig_s S^k_\mu \lambda_k, \end{aligned} \tag{5.5.8}$$

and λ_k $(1 \leqslant k \leqslant 8)$ are the generators of $SU(3)$. In QCD, λ_k are taken as the Gell-Mann matrices as defined in (3.5.38).

5.5.2 Gluons

In the last subsection we saw that the three color quantum numbers defined in (5.5.2)-(5.5.3) lead to the QCD action (5.5.7) of $SU(3)$ gauge fields. In the following, we introduce the gluons in a few steps.

Gluons derived from QCD

It was known that the $SU(3)$ gauge theory for the strong interaction is oriented toward two directions: 1) describing the field particles, i.e. the strong interaction mediators, and 2) providing strong interaction potentials. In Section 4.5, we have discussed problem 2). Here, we consider problem 1).

Experiments showed that there are eight field particles for the strong interaction, called gluons, which are massless and electric neutral, denoted by

$$g^k, \quad 1 \leqslant k \leqslant 8. \tag{5.5.9}$$

The QCD theory shows that the gluons are vector bosons with spin $J = 1$.

Based on both the unified field theory and the weakton model, corresponding to the vector gluons (5.5.9) there are eight dual field particles, which are scalar bosons with spin $J = 0$, denoted by

$$g_0^k, \quad 1 \leqslant k \leqslant 8. \tag{5.5.10}$$

In addition, by PID, it follows from (5.5.7)-(5.5.8) that the field equations describing the eight vector gluons (5.5.9) read

$$\partial^\nu S^k_{\mu\nu} - \frac{g_s}{\hbar c} f^k_{ij} g^{\alpha\beta} S^i_{\alpha\mu} S^j_\beta - g_s Q^k_\mu = (\partial_\mu + \frac{g_s}{\hbar c}\alpha_l S^l_\mu - \frac{1}{4}\alpha_0 x_\mu)\phi^k_s, \tag{5.5.11}$$

for $1 \leqslant k \leqslant 8$, and the field equations describing the eight scalar gluons (5.5.10) are

$$\partial^\mu \partial_\mu \phi^k_s + \partial^\mu \left[\left(\frac{g_s}{\hbar c}\alpha_l S^l_\mu - \frac{1}{4}\alpha_0 x_\mu\right)\phi^k_s\right] = -g_s \partial^\mu Q^k_\mu - \frac{g_s}{\hbar c} f^k_{ij} g^{\alpha\beta} \partial^\mu (S^i_{\alpha\mu} S^j_\beta), \tag{5.5.12}$$

for $1 \leqslant k \leqslant 8$, where α_l $(1 \leqslant l \leqslant 8)$ and α_0 are parameters, and

$$Q^k_\mu = \overline{q}\gamma_\mu \lambda^k q \quad (\lambda^k = \lambda_k,\ \gamma_\mu = g_{\mu\nu}\gamma^\nu).$$

The field equations (5.5.11) and (5.5.12) are as in (4.4.34) and (4.4.37).

From (5.5.11) and (5.5.12) we can deduce the following theoretical conclusions for the gluons (5.5.9) and dual gluons (5.5.10):

1) The field functions W^k_μ $(1 \leqslant k \leqslant 8)$ describing the gluons (5.5.9) are vector fields, and, therefore, g^k in (5.5.9) are vector bosons with spin $J = 1$;

2) The field function ϕ_s^k $(1\leqslant k\leqslant 8)$ describing the gluons (5.5.10) are scalar fields, and therefore g_0^k in (5.5.10) are scalar bosons with $J=0$;

3) The field equations (5.5.11) and (5.5.12) are nonlinear. Consequently, the gluons g^k and g_0^k are not in free states, and in their bound states the masses of g^k and g_0^k may appear;

4) In the bound states, the masses can be generated by the spontaneous symmetry breaking in (5.5.11) and (5.5.12), and the mass terms for g^k and g_0^k are as follows

$$g^k:\ \frac{g_s}{\hbar c}\alpha_l\widetilde{\phi}_s^k S_\mu^l \quad (\widetilde{\phi}_l^k \text{ are the ground states}),$$
$$g_0^k:\ \alpha_0\Phi_s^k;$$

5) Free gluons, if there exist, are massless.

Remark 5.16 Gluons and photons are massless only in their free states. When they are confined in mediator clouds around electrons and quarks, they may attain masses. In this case, massive photons are described by the PID electromagnetic equations as follows:

Vector photons:

$$\partial^\nu(\partial_\nu A_\mu-\partial_\mu A_\nu)-eJ_\mu=\left(\partial_\mu+\frac{e}{\hbar c}\beta A_\mu-\frac{1}{4}kx_\mu\right)\phi. \tag{5.5.13}$$

Scalar photons:

$$\partial^\mu\partial_\mu\phi-k\phi+\frac{e}{\hbar c}\beta\partial^\mu(A_\mu\phi)-\frac{1}{4}kx_\mu\partial^\mu\phi=0. \tag{5.5.14}$$

The masses of bound photons are created in (5.5.13) and (5.5.14) by the spontaneous gauge-symmetry breaking. □

Color charge

In the QCD theory, color quantum numbers are regarded as color charges. Namely, each quark carries one of the color charges:

$$\begin{array}{lll}\text{red charge } r, & \text{green charge } g, & \text{blue charge } b,\\ \text{anti-red charge } \overline{r}, & \text{anti-green charge } \overline{g}, & \text{anti-blue charge } \overline{b}.\end{array} \tag{5.5.15}$$

Each gluon is considered to carry a pair of a color charge and an anti-color charge in QCD. Thus, the three pair of color charges and anti-color charges in (5.5.15) constitute fundamental bases of $SU(3)$ and $SU(\overline{3})$, and the gluons g^k $(1\leqslant k\leqslant 8)$ form an octet of the irreducible representation

$$SU(3)\otimes SU(\overline{3})=8\oplus 1,$$

which are expressed as

$$g^1 = r\overline{g},\ g^2 = b\overline{r},\ g^3 = g\overline{b},\ g^4 = \frac{1}{\sqrt{2}}(r\overline{r} - b\overline{b}), \tag{5.5.16}$$

$$g^5 = g\overline{r},\ g^6 = r\overline{b},\ g^7 = b\overline{g},\ g^8 = \overline{g}^4. \tag{5.5.17}$$

We remark that the conclusions (5.5.16) and (5.5.17) of QCD for gluons are completely consistent with those in the weakton model in (5.3.40) and (5.3.41). However, the QCD version is based on the color charges of (5.5.15), and the version of weakton model is based on the color quantum number of the w^*-weakton:

$$w_r^*,\ w_g^*,\ w_b^*,\ \overline{w}_{\overline{r}}^*,\ \overline{w}_{\overline{g}}^*,\ \overline{w}_{\overline{b}}^*.$$

Gluon radiation

Based on the Standard Model, quarks can emit and absorb gluons in the same fashion as electrons emitting and absorbing photons. For example, a red u-quark u_r emits an $r\overline{g}$ gluon $g_{r\overline{g}}$, then u_r becomes a u_g quark, as shown in Figure 5.16.

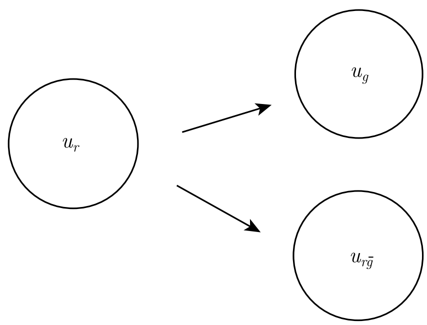


Figure 5.16

The reaction formula of the gluon radiation shown by Figure 5.16 is as

$$u_r \to u_g + g_{r\overline{g}}, \tag{5.5.18}$$

and the color index operation for (5.5.18) is given by

$$r = g(r\overline{g}) = r\overline{g}g = r. \tag{5.5.19}$$

Color index operation

However, if q_r-quark radiates a $g_{b\overline{q}}$ gluon, due to the lack of color algebra, then it is difficult to determine the type of quark to which q_r transforms, i.e.

$$q_r \to g_{b\overline{g}} + q_X, \quad X = ?.$$

In addition, the lack of color algebra also leads to another serious problem. It is known that hadrons are color neutral:

$$\text{the color of a hadron} = w.$$

But for the color index c_i defined as

$$c_1 = r,\ c_2 = g,\ c_3 = b,\ \overline{c}_1 = \overline{r},\ \overline{c}_2 = \overline{g},\ \overline{c}_3 = \overline{b}, \tag{5.5.20}$$

they is no algebraic operation, i.e.

$$c_i\overline{c}_j \neq \text{ any one of (5.5.20) for } i \neq j. \tag{5.5.21}$$

Hence the color indices of all non-neutral gluons cannot be defined. Thus, if a proton p contains three quarks and N gluons $(N \neq 0)$:

$$p = u_{c_i} + u_{c_j} + d_{c_k} + \sum_k n_k g^k, \quad \sum_k n_k = N,$$

then we cannot make sure whether p is color neutral.

In the following two subsections we shall establish the color algebra, a new mathematical theory, to solve above mentioned problems associated with the color index operation and the QCD.

5.5.3 Color algebra

The main objective of this and next subsections is to introduce a consistent color algebraic structure, and to establish a color index formula for hadrons and color transformation exchange for gluon radiation.

The color algebra for quantum chromodynamics (QCD) established here is based on color neutral principle of hadrons, and is uniquely determined. Hence it serves as the mathematical foundation of QCD.

Color algebra of color quantum numbers

First we examine the crucial problems encountered in the existing theory for color algebra. The color neutral principle of hadrons requires that the three colors must obey the following laws:

$$rgb = w, \qquad \overline{r}\,\overline{g}\,\overline{b} = w, \tag{5.5.22}$$

$$r\overline{r} = g\overline{g} = b\overline{b} = w. \tag{5.5.23}$$

Basic physical considerations imply that the product operation of color indices is commutative and associative:

$$rg = gr,\ rb = br,\ bg = gb,\ rgb = (rg)b = r(gb).$$

Hence we infer from (5.5.22) and (5.5.23) that

$$\begin{aligned} &rg=\overline{b},\ rb=\overline{g},\ bg=\overline{r},\\ &\overline{r}\,\overline{g}=b,\ \overline{r}\overline{b}=g,\ \overline{b}\overline{g}=r. \end{aligned} \tag{5.5.24}$$

Notice that the white color w is the unit element, i.e.

$$wr=r,\ wg=g,\ wb=b,\ w\overline{r}=\overline{r},\ w\overline{g}=\overline{g},\ w\overline{b}=\overline{b}.$$

Then again, we infer from (5.5.22) and (5.5.23) that

$$\begin{aligned} &rr=\overline{r},\ gg=\overline{g},\ bb=\overline{b},\\ &\overline{r}\,\overline{r}=r,\ \overline{g}\,\overline{g}=g,\ \overline{b}\,\overline{b}=b. \end{aligned} \tag{5.5.25}$$

Multiplying (5.5.22) by b and using (5.5.25), we deduce that

$$r(g\overline{b})=b, \tag{5.5.26}$$

which leads to inconsistency, no matter what color we assign for $g\overline{b}$. For example, if we assign $g\overline{b}=r$, then we derive from (5.5.26) that $rr=b$, which is inconsistent with $rr=\overline{r}$ in (5.4.25). If we assume $g\overline{b}=\overline{b}$, then $r\overline{b}=b$, which, by multiplying by b, leads to $r=bb=\overline{b}$, a contradiction again.

The above inconsistency demonstrates that in addition to the six basic color indices

$$r,\ g,\ b,\ \overline{r},\ \overline{g},\ \overline{b},$$

we need to incorporate the following color extensions:

$$r\overline{g},\ \overline{r}g,\ r\overline{b},\ \overline{r}b,\ g\overline{b},\ \overline{g}b.$$

Only two of these added color indices are independent, and in fact, we can derive from (5.5.24) that

$$r\overline{g}=b\overline{r}=g\overline{b},\quad g\overline{r}=r\overline{b}=b\overline{g}.$$

Hence we define them as yellow y and anti-yellow $\overline{y}$ as follows:

$$\begin{aligned} &y=r\overline{g}=b\overline{r}=g\overline{b},\\ &\overline{y}=g\overline{r}=r\overline{b}=b\overline{g}. \end{aligned} \tag{5.5.27}$$

In a nutshell, in order to establish a consistent color algebra, it is necessary to add two quantum numbers yellow y and anti-yellow $\overline{y}$ to the six color quantum numbers, give rise to a consistent and complete mathematical theory: color algebra.

Definition 5.17 *A color algebra is defined by the following three basic ingredients:*

1) *The generators of color algebra consists of quarks and gluons, which possesses nine color indices as*

$$r,\ g,\ b,\ y,\ \overline{r},\ \overline{g},\ \overline{b},\ \overline{y},\ w,$$

which form a finite commutative group. Here y and $\overline{y}$ are given by (5.5.27), and the group product operation is defined by

   (a) *w is the unit element, i.e.*

$$cw = c \ for\ any\ color\ index\ c;$$

   (b) $\overline{c}$ *is the inverse of* c:

$$c\overline{c} = w \ for\ c = r,\ g,\ b,\ y;$$

   (c) *in addition to the basic operations given by* (5.5.22)-(5.5.25) *and* (5.5.27), *we have*

$$\begin{aligned} &yr = b, \quad yg = r, \quad yb = g, \quad yy = \overline{y}, \\ &\overline{y}r = g, \quad \overline{y}g = b, \quad \overline{y}b = r, \quad \overline{y}\,\overline{y} = y. \end{aligned}$$

2) *Color algebra is an algebra with quarks and gluons as generators with integer coefficients, and its space is given by*

$$P = \left\{ \sum_{k=1}^{22} n_k e_k \,\middle|\, n_k \in \mathbb{Z} \right\},$$

where e_k $(1 \leqslant k \leqslant 19)$ are 18 colored quarks and 4 colored gluons, and $-e_k$ represent anti-quarks and anti-gluons.

3) *The color index of $\omega = \sum_{k=1}^{22} n_k e_k \in P$ is defined by*

$$\mathrm{Ind}_c(\omega) = \prod_{k=1}^{22} c_k^{n_k},$$

where c_k is the color index of e_k and $c_k^{n_k} = w$ if $n_k = 0$.

Two remarks are now in order.

Remark 5.18 Each element $\omega = \Sigma n_k e_k \in P$ represents a particle system, and n_k is the difference between the number of particles with color index c_k and the number of antiparticles with color index $\overline{c}_k$. In particular, particles with colors r, g, b must be quarks, particles with colors $\overline{r}, \overline{g}, \overline{b}$ must be anti-quarks, and particles with colors $y, \overline{y}$ must be gluons. □

Remark 5.19 In (5.5.27), $r\overline{g}, b\overline{r}, g\overline{b}$ are all yellow. Consequently the gluons $g_{r\overline{g}}, g_{b\overline{r}}, g_{g\overline{b}}$ have the same color. However, they represent different quantum states. In particular, in the weakton model,

$$g_{r\overline{g}} = w_r^* \overline{w}_{\overline{g}}^*, \quad g_{b\overline{r}} = w_b^* \overline{w}_{\overline{r}}^*, \quad g_{g\overline{b}} = w_g^* \overline{w}_{\overline{b}}^*,$$

which represent different quantum states. □

Color index formula of hadrons

We now study color neutral problem for hadrons and the radiation and absorption of gluons for quarks.

Let us start with color neutral problem for hadrons. Consider the constituents of a proton

$$p = u_{c_1} + u_{c_2} + u_{c_3} + \sum n_k g^k \in P, \quad \sum n_k = N,$$

whose color index is given by

$$\text{Ind}_c(p) = c_1 c_2 c_3 \prod_{k=1}^{8} (\text{ Ind}_c(g^k))^{n_k},$$

where $\text{Ind}_c(g^k)$ is the color index of the gluon g^k. The color neutral law requires that $\text{Ind}_c(p) = w$, which does not necessarily lead to $c_1 c_2 c_3 = w$. For example, for the following constituents of p:

$$p = u_r + u_r + d_g + 2g_{r\overline{g}} + g_{r\overline{b}},$$

we have

$$\begin{aligned} \text{Ind}_c(p) &= r^2 g y^2 \overline{y} = \overline{r} g y = \overline{r} r = w, \\ c_1 c_2 c_3 &= rrg = \overline{r} g = \overline{y} \neq w. \end{aligned}$$

In summary, the hadron color quantum numbers based on color algebra is very different from the classical QCD theory.

For a baryon B with its constituents given by

$$B = \sum_{i=1}^{3} q_{c_i} + \sum_{k=1}^{3} (n_k g^k + m_k \overline{g}^k) + K_1 g^4 + K_2 \overline{g}^4,$$

where g^k and $\overline{g}^k$ $(1 \leqslant k \leqslant 4)$ are as in (5.5.16) and (5.5.17), its quantum number distribution satisfies the following color index formula:

$$c_1 c_2 c_3 = \overline{y}^{N_1} y^{N_2}, \tag{5.5.28}$$

and N_1, N_2 are as

$$N_1 = \sum_{k=1}^{3} n_k, \quad N_2 = \sum_{k=1}^{3} m_k. \tag{5.5.29}$$

For a meson M with constituents:

$$M = q_{c_1} + \overline{q}_{\overline{c}_2} + \sum_{k=1}^{3}(n_k g^k + m_k \overline{g}^k) + K_1 g^4 + K_2 \overline{g}^4,$$

its color indices satisfy that

$$c_1\overline{c}_2 = \overline{y}^{N_1} y^{N_2}, \tag{5.5.30}$$

where N_1, N_2 are as in (5.5.29).

Both equalities (5.5.28) and (5.5.29) are the color index formulas for baryons and mesons, which ensure the neutral law of hadrons.

Color transformation of gluon radiation

Consider the transformation of a quark q_c with color c to another quart q_{c_3} after emitting a gluon $g_{c_1\overline{c}_2}$:

$$q_c \to g_{c_1\overline{c}_2} + q_{c_3},$$

then the color index c_3 of then induced quark q_{c_3} is given by

$$c_3 = c\overline{c}_1 c_2.$$

Also, for the transformation of a quark q_c to another quark q_{c_4} after absorbing a gluon $g_{c_1\overline{c}_2}$:

$$q_c + g_{c_1\overline{c}_2} \to q_{c_4},$$

then the color c_4 of transformed quark q_{c_4} is as follows

$$c_4 = c c_1 \overline{c}_2.$$

General color algebra

In an abstract sense, a color algebra is a triplet

$$\{G_c, P^N, \text{ Ind}_c\},$$

which consists of

1) a finite group G_c, called color group;

2) an integer modular algebra with generators $e_1, \cdots, e_N$:

$$P^N = \left\{ \sum_{k=1}^{N} n_k e_k \,\middle|\, n_k \in \mathbb{Z} \right\};$$

3) a homomorphism

$$\text{Ind}_c: \ P^N \to G_c,$$

such that for each element $w = \sum_{k=1}^N n_k e_k \in P^N$, we can define a color index for w by

$$\text{Ind}_c(w) = \prod_{k=1}^N (\ \text{Ind}_c(e_k))^{n_k},$$

where $\text{Ind}_c(e_k) \in G_c$ is the image of e_k under the homomorphism Ind_c.

The color algebra introduced earlier is such a triplet $\{G_c, P^N, \ \text{Ind}_c\}$ with $N = 22$, and

$$\begin{aligned} &G_c = \ \text{multiplication group generated by } \{r, g, b, y\}, \\ &P^N = \left\{ \sum_{k=1}^{22} n_k e_k \,\middle|\, n_k \in \mathbb{Z} \right\}, \end{aligned}$$

where e_k are colored quarks and gluons given by

$$\begin{aligned} &e_1 = u_r, \quad e_2 = u_g, \quad e_3 = u_b, \quad e_4 = d_r, \quad e_5 = d_g, \quad e_6 = d_b, \\ &e_7 = s_r, \quad e_8 = s_g, \quad e_9 = s_b, \quad e_{10} = c_r, \quad e_{11} = c_g, \quad e_{12} = c_b, \\ &e_{13} = b_r, \quad e_{14} = b_g, \quad e_{15} = b_b, \quad e_{16} = t_r, \quad e_{17} = t_g, \quad e_{18} = t_b, \\ &e_{19} = g^1, \quad e_{20} = g^2, \quad e_{21} = g^3, \quad e_{22} = g^4, \end{aligned} \tag{5.5.31}$$

and g^k $(1 \leqslant k \leqslant 4)$ are as in (5.5.16). The homomorphism Ind_c is naturally defined by the assignment in (5.5.31), i.e.

$$\begin{aligned} &\text{Ind}_c(e_1) = r, \qquad \text{Ind}_c(e_2) = g, \qquad \cdots, \qquad \text{Ind}_c(e_{18}) = b, \\ &\text{Ind}_c(e_{19}) = \ \text{Ind}_c(e_{20}) = \ \text{Ind}_c(e_{21}) = y, \qquad \text{Ind}(e_{22}) = w. \end{aligned}$$

5.5.4 w^*-color algebra

Based on the weakton model, the weakton constituents of a quark q and a gluon g^k are given by

$$\begin{aligned} &\text{quark}: \ q = w^* w w, \\ &\text{gluon: } g^k = w^* \overline{w}^*. \end{aligned}$$

The only weakton that has colors is w^*, which has three colors:

$$w_r^*, \ w_g^*, \ w_b^*, \tag{5.5.32}$$

and three anti-colors:

$$\overline{w}_{\overline{r}}^*, \ \overline{w}_{\overline{g}}^*, \ \overline{w}_{\overline{b}}^*. \tag{5.5.33}$$

With the three pairs of colored and anti-colored weaktons in (5.5.32) and (5.5.33), we can define $N = 3$ color algebra, which we call the w^*-color algebra.

Definition 5.20 *The w^*-color algebra is the triplet $\{G_c, P^3, \text{Ind}_c\}$, which is defined as follows:*

1) *the color group is*
$$G_c = \{r, g, b, y, \overline{r}, \overline{g}, \overline{b}, \overline{y}, w\},$$
with group operation given in Definition 5.17;

2) *the $\mathbb{Z}$-modular algebra P^3 given by*
$$P^3 = \left\{ \sum_{k=r,g,b} n_k w_k^* \,\middle|\, n_k \in \mathbb{Z} \right\};$$

3) *the color homomorphism* $\text{Ind}_c : P^3 \to G_c$ *defined by*
$$\text{Ind}_c(w_k^*) = k, \quad \text{Ind}_c(-w_k^*) = \overline{k}, \quad k = r, g, b.$$

The w^*-color algebra given by Definition 5.20 is based on the weakton model, which is much simpler than the QCD color algebra introduced in the last subsection, and is readily used to study the structure of subatomic particles.

Consider a particle system ω, which consists of n_k quarks $q_k, \overline{n}_k$ antiquarks $\overline{q}_k, m_1$ gluons $g^1 = g_{r\overline{g}}, m_2$ gluons $g^2 = g_{b\overline{r}}, m_3$ gluons $g^3 = g_{g\overline{b}}, m_4$ color-neutral gluons g^4, and $\overline{m}_k$ gluons $\overline{g}^k$ $(1 \leqslant k \leqslant 4)$:
$$\omega = \sum_{k=r,g,b} (n_k q_k + \overline{n}_k \overline{q}_k) + \sum_{i=1}^{4} (m_i g^i + \overline{m}_i \overline{g}^i). \tag{5.5.34}$$
Then ω corresponds to an element $X_\omega \in P^3$ expressed as
$$X_\omega = \sum_{k=r,g,b} N_k w_k^*, \tag{5.5.35}$$
where
$$\begin{aligned} N_r &= (n_r - \overline{n}_r) + (m_1 - m_2) - (\overline{m}_1 - \overline{m}_2), \\ N_g &= (n_g - \overline{n}_g) + (m_3 - m_1) - (\overline{m}_3 - \overline{m}_1), \\ N_b &= (n_b - \overline{n}_b) + (m_2 - m_3) - (\overline{m}_2 - \overline{m}_3), \end{aligned} \tag{5.5.36}$$
and, consequently, the color index for ω is defined by
$$\text{Ind}_c(\omega) = \text{Ind}_c(X_\omega) = r^{N_r} g^{N_g} b^{N_b}, \tag{5.5.37}$$
where for $N_k < 0$ we defined
$$k^{N_k} = \overline{k}^{-N_k} \quad (k = r, g, b).$$
It is then clear that
$$\text{Ind}_c(\omega_1 + \cdots + \omega_s) = \prod_{i=1}^{s} \text{Ind}_c(\omega_i).$$

The following is a basic theorem for w^*-color algebra, providing the needed foundation for the structure of charged leptons and quarks.

Theorem 5.21 *The following assertions hold true for the w^*-color algebra:*

1) *For any gluon particle system with no quarks and antiquarks*

$$\pi=\sum_{i=1}^{4}(m_i g^i+\overline{m}_i\overline{g}^i), \tag{5.5.38}$$

the color index of this system satisfies that

$$\text{Ind}_c(\pi)=\begin{cases} w & \textit{for}\sum_{i=1}^{3}(m_i-\overline{m}_i)=\pm 3n,\\ y & \textit{for}\ \sum_{i=1}^{3}(m_i-\overline{m}_i)=\pm 3n+1,\\ \overline{y} & \textit{for}\sum_{i=1}^{3}(m_i-\overline{m}_i)=\pm 3n+2,\end{cases} \tag{5.5.39}$$

for some integer $n=0,1,2,\cdots$;

2) *For any single quark system as*

$$\begin{aligned}\omega&=q+\pi\\ \overline{\omega}&=\overline{q}+\pi\end{aligned}\qquad \textit{with } \pi \textit{ as given by (5.5.38)}, \tag{5.5.40}$$

the color index of (5.5.40) satisfies that

$$\begin{aligned}\text{Ind}_c(\omega)&=r,g,b,\\ \text{Ind}_c(\overline{\omega})&=\overline{r},\overline{g},\overline{b},\end{aligned} \tag{5.5.41}$$

3) *For the hadronic systems*

$$\begin{aligned}M&=q+\overline{q}+\pi &&\textit{meson system},\\ B&=q+q+q+\pi &&\textit{baryon system},\end{aligned} \tag{5.5.42}$$

with π given by (5.5.38), the color indices of (5.5.42) must be as

$$\text{Ind}_c(M)=w,y,\overline{y},\qquad \text{Ind}_c(B)=w,y,\overline{y}. \tag{5.5.43}$$

Two remarks are now in order.

First, in particle physics, all basic and important particle systems are given by the particle systems in Assertions 1)-3) in this theorem. System (5.5.38) represents a gluon system attached to quarks and hadrons, (5.5.40) represents a cloud system of gluons around a quark or an anti-quark, and (5.5.42) represents a cloud system of gluons around a hadron (a meson or a baryon).

Second, physically, the numbers

$$\sum_{i=1}^{3}(m_i-\overline{m}_i)=\pm 3n \quad (n=0,1,2,\cdots),$$

indicate that through exchange of weaktons, a cloud system (5.5.38) of gluons can become a system consisting of white gluons and the same number of yellow and anti-yellow gluons.

Proof of Theorem 5.21 We proceed in the following three steps.

Step 1. By the basic properties (5.5.34)-(5.5.37) of color index,

$$\text{Ind}_c(\pi)=r^{M_r}g^{M_g}b^{M_b},$$

where

$$\begin{aligned}M_r&=(m_1-m_2)-(\overline{m}_1-\overline{m}_2),\\ M_g&=(m_3-m_1)-(\overline{m}_3-\overline{m}_1),\\ M_b&=(m_2-m_3)-(\overline{m}_2-\overline{m}_3).\end{aligned}$$

Consequently, using $C^{-m}=\overline{C}^m$, we have

$$\begin{aligned}\text{Ind}_c(\pi)&=(r\overline{g})^{m_1}(\overline{r}b)^{m_2}(g\overline{b})^{m_3}(\overline{r}g)^{\overline{m}_1}(r\overline{b})^{\overline{m}_2}(\overline{g}b)^{\overline{m}_3}\\ &=y^{m_1}y^{m_2}y^{m_3}\overline{y}^{\overline{m}_1}\overline{y}^{\overline{m}_2}\overline{y}^{\overline{m}_3}\\ &=y^M, \quad M=\sum_{i=1}^{3}(m_i-\overline{m}_i).\end{aligned} \tag{5.5.44}$$

Notice that $y^2=\overline{y}, \overline{y}^2=y$. Then the equality (5.5.39) follows from (5.5.44), and Assertion 1) is proved.

Step 2. For Assertion 2), with the above argument, for the particle system (5.5.40) it is easy to see that

$$\begin{aligned}\text{Ind}_c(\omega)=&\ \text{Ind}_c(q)\ \text{Ind}_c(\pi)=\ \text{Ind}_c(q)y^M,\\ \text{Ind}_c(\overline{\omega})=&\ \text{Ind}_c(\overline{q})\ \text{Ind}_c(\pi)=\ \text{Ind}_c(\overline{q})y^M.\end{aligned} \tag{5.5.45}$$

Due to the facts that

$$\text{Ind}_c(q)=r,g,b, \qquad \text{Ind}_c(\overline{q})=\overline{r},\overline{g},\overline{b},$$

and by the multiplication rule given in Definition 5.17, the conclusion (5.5.41) follows from (5.5.45).

Step 3. With the same arguments as above, for the meson and baryon system (5.5.42), we can derive that

$$\text{Ind}_c(M)=c_i\overline{c}_jy^M, \qquad \text{Ind}_c(B)=c_ic_jc_ky^M, \tag{5.5.46}$$

for $1 \leqslant i,j,k \leqslant 3$, where $c_1 = r, c_2 = g, c_3 = b$. By the basic rules for the color operation given in (5.5.24), (5.5.25) and (5.5.27), we have

$$c_i\overline{c}_j = w, y, \overline{y}, \quad c_i c_j c_k = c_i\overline{c}_l = w, y, \overline{y}.$$

Therefore, (5.5.43) follows from (5.5.46). □

5.5.5 Mediator clouds of subatomic particles

Subatomic particles include charged leptons, quarks and hadrons. As demonstrated in (5.3.12) and (5.3.15), strong and weak interactions consist of different layers, leading to different particle structures in various levels: weakton level (elementary particle level), mediator level, charged lepton and quark level, and hadron level.

In this section, we address the structure of mediator clouds for charged leptons, quarks and hadrons, based on the the color algebra and the layered properties of strong and weak interactions.

Mediator clouds for charged leptons

For simplicity and due to similarities, we consider only the case for electrons. In Section 5.4.5, we have introduced the structure of photon cloud around electrons and the mechanism of photon radiations. Here we shall again discuss the electron structure in more details.

1. *Clouds of photons and ν-mediators.* First, an electron consists of a naked electron and several layers of mediator clouds. Since electron does not contain strong charges, the mediator clouds only contain photons and ν-mediators (each gluon contains $2g_s$). Namely

$$\text{mediator cloud of electron} = \text{photon layer} + \text{the } \nu\text{-layer}. \tag{5.5.47}$$

According to the layered formulas (5.4.37) of weak forces between naked electron and mediators, the radius r of a mediator cloud is approximately given by

$$r^3 \simeq \frac{\widetilde{\rho}}{4\widetilde{B}} r_0^2, \quad r_0 = 10^{-16}\,\text{cm}, \tag{5.5.48}$$

and $\widetilde{\rho}/\widetilde{B}$ depends on the type of mediators. In fact, for gluons and naked electron, the parameter $\widetilde{B} \leqslant 0$, therefore electrons have no gluon clouds.

It is reasonable to think that the parameter $\widetilde{\rho}/\widetilde{B}$ for photon and ν-mediator is different

$$\frac{\widetilde{\rho}_\gamma}{\widetilde{B}_\gamma} \neq \frac{\widetilde{\rho}_\nu}{\widetilde{B}_\nu}.$$

Therefore, by (5.5.48) the radius r_γ of the photon layer is different from the radius r_ν of the ν-layer, i.e.

$$r_\gamma \neq r_\nu. \tag{5.5.49}$$

Due the weak repelling forces between γ and ν, only the exterior layer of cloud can emit mediators. Hence, the bremsstrahlung implies that the photon cloud should be in the exterior layer. Namely (5.5.49) should be written as

$$r_\gamma > r_\nu. \tag{5.5.50}$$

The relation (5.5.40) is only a theoretical consideration.

2. *Particle number of each layer.* Due to the weak interaction between mediators, the mediators must maintain a distance r_0 between each other, with r_0 being approximately the weak repelling radius of the mediators. Consequently, the total number N of particles in each layer of mediators satisfies that

$$N \leqslant 4\pi r^2/r_0^2, \tag{5.5.51}$$

where r is the radius of the layer as in (5.5.48). The inequality (5.5.51) means that all particles on an r-sphere are arranged with distance r_0.

Based on (5.5.51), if $r = kr_0$, then $N \leqslant 4\pi k^2$.

3. *Spin number of vector photons.* An electron refers to the system consisting of the naked electron $\nu_e w_1 w_2$ and its mediator cloud as shown in (5.5.47). The photon layer contains both vector and scalar photons. As the total spin of an electron is $J = \frac{1}{2}$, both the number $N_{J=1}$ of vector photons with spin $J = 1$ and the number $N_{J=-1}$ of photons with spin $J = -1$ should be the same, i.e. the vector photons in an electron satisfy

$$N_{J=1} = N_{J=-1}. \tag{5.5.52}$$

The condition (5.5.52) implies that the emitting and absorbing of vector photons always in pairs with one spin $J = 1$ and another spin $J = -1$. Scalar photons and ν-mediators have spin $J = 0$.

In fact, by the Angular Momentum Rule 6.14, only bosons with $J = 0$ can rotate around a center without the presence of force moment. Hence the photons and ν-mediators in the mediator cloud may all be scalar particles.

Mediator clouds of quarks

Different from the weakton constituents of charged leptons, the weakton constituents of a quark include a w^*-weakton with a strong charge, leading to different structure of mediator clouds.

1. As a quark curries three weak charges $3g_w$ and a strong charge g_s, the mediator cloud of a quark possesses three layers:

$$\begin{aligned} &\text{gluon layer:} && \text{vector and scalar gluons,} \\ &\text{photon layer:} && \text{vector and scalar photons,} \\ &\nu\text{-layer:} && \nu\text{-mediators.} \end{aligned} \tag{5.5.53}$$

2. The radii of the 3 layers in (5.5.53) are different:

$$r_g \neq r_\gamma, \quad r_\gamma \neq r_\nu, \quad r_\nu \neq r_g. \tag{5.5.54}$$

Similar to the case for an electron, the particle number in each layer of (5.5.53) satisfies a relation as (5.5.51).

3. The total spin of a quark system is also $J = \frac{1}{2}$. Therefore the vector photons and vector gluons in the mediator cloud of a quark have to obey the spin number rules as

$$N_{J=1} = N_{J=-1}, \quad \widetilde{N}_{J=1} = \widetilde{N}_{J=-1}, \tag{5.5.55}$$

where $N_{J=1}$ and $N_{J=-1}$ are the numbers of photons with spins $J=1$ and $J=-1$, and $\widetilde{N}_{J=1}$ and $\widetilde{N}_{J=-1}$ are that of gluons with spins $J=1$ and $J=-1$.

We remark that two scalar mediators can be transformed into a pair of vector particles with $J=1$ and $J=-1$ respectively.

4. By Assertion 2) of Theorem 5.21, the color indices of a quark system q and an anti-quark system $\overline{q}$, carrying gluon clouds, must take the values as follows

$$\begin{aligned} &\text{Ind}_c(q) = r, g, b, \\ &\text{Ind}_c(\overline{q}) = \overline{r}, \overline{g}, \overline{b}. \end{aligned} \tag{5.5.56}$$

5. The gluons in the cloud layers of quarks are confined in a hadron. However, gluons between quarks in a hadron can be exchanged. The gluon exchange process between quarks in a hadron is called gluon radiation. In fact, a quark can emit gluons which will be absorbed by other quarks in the hadron.

6. Consider a quark q_1, which is transformed to q_3 after emitting a gluon g_0, and consider a quark q_2, which is transformed to q_4 after absorbing the gluon g_0. The exchange process is expressed as

$$q_1 \to q_3 + g_0, \quad q_2 + g_0 \to q_4.$$

Equivalently,

$$q_1 + q_2 \to q_3 + q_4.$$

The corresponding color transformation is given by

$$\begin{aligned} \text{Ind}_c(q_1) &= \text{Ind}_c(q_3)\ \text{Ind}_c(g_0), \\ \text{Ind}_c(q_4) &= \text{Ind}_c(q_2)\ \text{Ind}_c(g_0), \end{aligned}$$

which lead to

$$\text{Ind}_c(q_1 + q_2) = \text{Ind}_c(q_3 + q_4). \tag{5.5.57}$$

The relations (5.5.53)-(5.5.57) provide the basic properties of mediator clouds of quarks. Further discussion will be given in the next chapter.

Gluon clouds of hadrons

Different from electrons and quarks, there is no photon and ν-mediator cloud layer around a hadron, due to the fact that the radius of a naked hadron ρ_H is greater than the range of the weak interaction:

$$\rho_H \geqslant 10^{-16} \text{ cm}.$$

Namely, a hadron can only have a strong attraction to vector and scalar gluons, forming a gluon cloud with radius about the same as the radius of a hadron:

$$r_H \sim 10^{-16} \sim 10^{-14} \text{ cm}.$$

Also, hadrons are colorless. For a baryon B and a meson M given by (5.5.42), we have that

$$\begin{aligned} &\text{Ind}_c(M) = c_i\overline{c}_i \text{ Ind}_c(\pi) = w \\ &\text{Ind}_c(B) = c_i c_j c_k \text{ Ind}_c(\pi) = w, \quad 1 \leqslant i, j, k \leqslant 3. \end{aligned}$$

Consequently, if the naked hadron is white, then the gluon cloud π is white as well: $\text{Ind}_c(\pi) = w$.

Free gluons

If there exist free gluons, then they must be white color:

$$\text{Ind}_c(g) = w \quad \text{for free gluons } g.$$

It is now an unknown problem that if there are free gluons.

Chapter 6
Quantum Physics

6.1 Introduction

Quantum physics is the study of the behavior of matter and energy at molecular, atomic, nuclear, and sub-atomic levels.

Quantum physics was initiated and developed in the first half of the 20th century, following the pioneering work of (Planck, 1901) on blackbody radiation, of (Einstein, 1905) on photons and energy quanta, of Niels Bohr on structure of atoms, of (de Broglie, 1924) on matter-wave duality. Quantum physics and general relativity have become the two cornerstones of modern physics. We refer interested readers to (Sokolov, Loskutov and Ternov, 1966; Greiner, 2000; Sakurai, 1994), among many others, for the basics and history of quantum physics.

Our recent work on the field theory of the four fundamental interactions and on the weakton model of elementary particles has lead to new insights to a few issues and challenges in quantum physics, including in particular 1) the basic laws for *interacting* multi-particle quantum systems, 2) energy levels of sub-atomic particles, and 3) solar neutrino problem.

Field theory for interacting multi-particle systems

Based PID and PRI, we know now that the fundamental interactions are due the corresponding charges of the particles involved, with the mass charge for gravitational effect, the electric charge for electromagnetism, the strong and the weak charges for the strong and weak interactions respectively. The geometric interaction mechanism implies that the dynamic matter distribution of the charged particles changes the geometry of the space-time manifold as well as the corresponding vector-bundles, leading to the dynamic interaction laws of the system.

Also, due to PRI, for a multi-particle system consisting of subsystems of different levels, the coupling can only be achieved through PRI and the principle of symmetry-breaking.

The above new insights from the unified field theory in Chapter 4 give rise to the following

Postulates for interacting multi-particle systems

1) *the Lagrangian action for an N-particle system satisfy the $SU(N)$ gauge invariance*;
2) gG^a_μ *represent the interaction potentials between the particles.*
3) *for an N-particle system, only the interaction field G_μ can be measured, and is the interaction field under which this system interacts with other external systems.*

With this postulate, field equations for a given multi-particle system can be naturally established using PID and PRI. In particular, one can achieve the unification so that the matter field can be geometrized as hoped by Einstein and Nambu, as stated explicitly in as stated in Nambu's Nobel lecture (Nambu, 2008).

Solar neutrino problem

Neutrino was first proposed by Wolfgang Pauli in 1930 in order to guarantee the energy and momentum conservation for β-decay. In the current standard model of particle physics, there are three flavors of neutrinos: the electron neutrino ν_e, the tau neutrino ν_τ and the mu neutrino ν_μ. The solar neutrino problem is referred to the discrepancy of the number of electron neutrinos arriving from the Sun are between one third and one half of the number predicted by the Standard Solar Model, and was first discovered by (Davis, Harmer and Hoffman, 1968).

The current dominant theory to resolve the solar neutrino problem is the neutrino oscillation theory, which are based on three basic assumptions: 1) the neutrinos are massive, and, consequently, are described by the Dirac equations, 2) the three flavors of neutrinos ν_e, ν_μ, ν_τ are not the eigenstates of the Hamiltonian, and 3), instead, the three neutrinos are some linear combinations of three distinct eigenstates of the Hamiltonian. However, the massive neutrino assumption gives rise two serious problems. First, it is in conflict with the known fact that the neutrinos violate the parity symmetry. Second, the handedness of neutrinos implies their velocity being at the speed of light.

To resolve these difficulties encountered by the classical theory, we argue that there is no physical principle that requires that neutrino must have mass to ensure oscillation. The Weyl equations were introduced by H. Weyl in 1929 to describe massless spin-$\frac{1}{2}$ free particles (Weyl, 1929), which is now considered as the basic dynamic equations of neutrino (Landau, 1957; Lee and Yang, 1957; Salam, 1957); see also (Greiner, 2000). One important property of the Weyl equations is that they violate the parity invariance. Hence by using the Weyl equations, we are able to introduce a massless neutrino oscillation model. With this massless model, we not only deduce the same oscillation mechanism, but also resolve the above two serious problems encountered in the massive neutrino oscillation model.

Despite of the success of neutrino oscillation models and certain level of experimental support, the physical principles behind the neutrino oscillation are still entirely unknown. Recently, the authors developed a phenomenological model of elementary particles, called the weakton model (Ma and Wang, 2015b). The ν mediator in the weakton model leads to an alternate explanation to the solar neutrino problem. When the solar electron neutrinos collide with anti electron neutrinos in the atmosphere, which are abundant due to the β-decay of neutrons, they can form ν mediators, causing the loss of electron neutrinos. Note that ν mediator can also have the following elastic scattering

$$\nu + e^- \longrightarrow \nu + e^-.$$

Also ν participates only the weak interaction similar to the neutrinos, and consequently possesses similar behavior as neutrinos. Consequently, the new mechanism proposed here does not violate the existing experiments (SNO and KamLAND).

Energy levels of sub-atomic particles

The classical atomic energy level theory demonstrates that there are finite number of energy levels for an atom given by $E_n = E_0 + \lambda_n$, $n = 1, \cdots, N$, where λ_n are the negative eigenvalues of the Schrödinger operator, representing the bound energies of the atom, holding the orbital electrons, due to the electromagnetism.

The concept of energy levels for atoms can certainly be generalized to subatomic particles. The key ingredients and the main results are given as follows.

1. The constituents of subatomic particles are spin-$\frac{1}{2}$ fermions, which are bound together by either weak or strong interactions. Hence the starting point of the study is the layered weak and strong potentials as presented earlier, which play the similar role as the Coulomb potential for the electromagnetic force which bounds the orbital electrons moving around the nucleons.

2. The dynamic equations of massless particles are the Weyl equations, and the dynamic equations for massive particles are the Dirac equations. The bound energies of all subatomic particles are the negative eigenvalues of the corresponding Dirac and Weyl operators, and the bound states are the corresponding eigenfunctions.

The Weyl equations were introduced by H. Weyl in 1929 to describe massless spin-$\frac{1}{2}$ free particles (Weyl, 1929), which is now considered as the basic dynamic equations of neutrino (Landau, 1957; Lee and Yang, 1957; Salam, 1957); see also (Greiner, 2000).

3. With bound state equations for both massless and massive particles, we derive the corresponding spectral equations for the bound states. We show that the energy levels of each subatomic particle are finite and discrete:

$$0 < E_1 < \cdots < E_N < \infty,$$

and each energy level E_n corresponds to a negative eigenvalue λ_n of the related eigenvalue problem. Physically, λ_n represents the bound energy of the particle, and are related to the energy level E_n with the following relation:

$$E_n = E_0 + \lambda_n, \quad \lambda_n < 0 \qquad \text{for } 1 \leqslant n \leqslant N. \tag{6.1.1}$$

Here E_0 is the intrinsic potential energy of the constituents of a subatomic particle such as the weaktons.

4. One important consequence of the above derived energy level theory is that there are both upper and lower bounds of the energy levels for all sub-atomic particles, and the largest and smallest energy levels are given by

$$0 < E_{\min} = E_0 + \lambda_1 < E_{\max} = E_0 + \lambda_N < \infty. \tag{6.1.2}$$

In particular, it follows from the energy level theory that the frequencies of mediators such as photons and gluons are also discrete and finite, and are given by $\omega_n = E_n/\hbar$ ($n = 1, \cdots, N$). In the Planck classical quantum assumption that the energy is discrete for a fixed frequency, and the frequency is continuous. Our results are different in two aspects. One is that the energy levels have an upper bound. Two is that the frequencies are also discrete and finite.

Outline of Chapter 6

Section 6.2 presents the basic postulate and facts from quantum physics from the Hamiltonian dynamics and Lagrangian dynamics points of view. We prove also a new angular momentum rule, which is useful for describing the weaktons constituents of charged leptons and quarks.

Section 6.3 presents new alternative approaches for solar neutrino problem, and is based on the recent work of authors (Ma and Wang, 2014f).

Section 6.4 introduces energy levels of subatomic particles, and is based on (Ma and Wang, 2014g).

The field theory and basic postulates for interacting multi-particle systems are established in Section 6.5, which is based entirely on (Ma and Wang, 2014d).

6.2 Foundations of Quantum Physics

6.2.1 Basic postulates

The main components of quantum physics include quantum mechanics and quantum field theory, which are based on the following basic postulates.

Postulate 6.1 *A quantum system consists of some micro-particles, which are described by a set of complex value functions $\psi = (\psi_1, \cdots, \psi_N)^T$, called wave functions. In other words, each quantum system is identified by a set of wave functions ψ:*

$$\text{a quantum system} = \psi, \tag{6.2.1}$$

which contain all quantum information of this system.

Postulate 6.2 *For a single particle system described by a wave function ψ, its modular square*

$$|\psi(x,t)|^2$$

represents the probability density of the particle being observed at point $x \in \mathbb{R}^3$ and at time t. Hence, ψ satisfies that

$$\int_{\mathbb{R}^3} |\psi|^2 dx = 1.$$

Postulate 6.3 *Each observable physical quantity L corresponds to an Hermitian operator $\hat{L}$, and the values of the physical quantity L are given by eigenvalues λ of $\hat{L}$:*

$$\hat{L}\psi_\lambda = \lambda \psi_\lambda,$$

and the eigenfunction ψ_λ is the state function in which the physical quantity L takes value λ. In particular, the Hermitian operators corresponding to position x, momentum p and energy E are given by

$$\begin{aligned}
&\text{position operator}: && \hat{x}\psi = x\psi, \\
&\text{momentum operator}: && \hat{p}\psi = -i\hbar\nabla\psi, \\
&\text{energy operator}: && \hat{E}\psi = i\hbar\frac{\partial\psi}{\partial t}.
\end{aligned} \tag{6.2.2}$$

Postulate 6.4 *For a quantum system ψ and a physical Hermitian operator $\hat{L}$, ψ can be expanded as*

$$\psi = \sum \alpha_k \psi_k + \int \alpha_\lambda \psi_\lambda d\lambda, \tag{6.2.3}$$

where ψ_k and ψ_λ are the eigenfunctions of $\hat{L}$ corresponding to discrete and continuous eigenvalues respectively. In (6.2.3) for the coefficients α_k and α_λ, their modular square $|\alpha_k|^2$ and $|\alpha_\lambda|^2$ represent the probability of the system ψ in the states ψ_k and ψ_λ. In addition, the following integral, denoted by

$$\langle\psi|\hat{L}|\psi\rangle = \int \psi^\dagger(\hat{L}\psi)dx, \tag{6.2.4}$$

represents the average value of physical quantity $\hat{L}$ of system ψ.

Postulate 6.5 *For a quantum system with observable physical quantities $l_1, \cdots, l_N$, if they satisfy a relation*

$$R(l_1, \cdots, l_N) = 0,$$

then the quantum system ψ (see (6.2.1)) satisfies the equation

$$R(\hat{L}_1, \cdots, \hat{L}_N)\psi = 0,$$

where $\hat{L}_k$ are the Hermitian operators corresponding to l_k $(1 \leqslant k \leqslant N)$, provided that $R(\hat{L}_1, \cdots, \hat{L}_N)$ is a Hermitian.

Two remarks are in order. First, in Subsection 2.2.5, Postulates 6.3and 6.5 are introduced as Basic Postulates 2.22 and 2.23.

Second, in addition to the three basic Hermitian operators given by (6.2.2), the other Hermitian operators often used in quantum physics are as follows:

$$\begin{aligned}
&\text{angular momentum:} && \hat{L} = \hat{x} \times \hat{p} = -i\hbar\vec{r} \times \nabla, \\
&\text{spin operator:} && \hat{S} = s\hbar\vec{\sigma}, \\
&\text{scalar momentum:} && \hbar p_0 = i\hbar(\vec{\sigma} \cdot \nabla) \text{ (massless fermion)}, \\
&\text{scalar momentum:} && \hat{p}_1 = -i\hbar(\vec{\alpha} \cdot \nabla) \text{ (massive fermion)}, \\
&\text{Hamiltonian energy :} && \hat{H} = \hat{K} + \hat{V} + \hat{M},
\end{aligned} \tag{6.2.5}$$

where s is the spin, $\vec{\sigma}$ and $\vec{\alpha}$ are as in (2.2.47) and (2.2.48), and $\hat{K}, \hat{V}, \hat{M}$ are the kinetic energy, potential energy, mass operators.

In addition to these Hermitians in (6.2.2) and (6.2.5), the following 5 types of particle current operators are very important in quantum field theory:

$$\begin{aligned}
&\text{scalar current operator:} && I \text{ (identity)}, \\
&\text{pseudo-scalar current operator:} && \gamma^5, \\
&\text{vector current operator:} && \gamma^\mu, \\
&\text{axial vector current operator:} && \gamma^\mu\gamma^5, \\
&\text{tensor current operator:} && \sigma^{\mu\nu} = \frac{i}{2}[\gamma^\mu, \gamma^\nu],
\end{aligned} \tag{6.2.6}$$

and the corresponding currents are

$$\begin{aligned}
&\text{scalar current:} && \rho = \psi^\dagger\psi, \\
&\text{pseudo-scalar current:} && P = \psi^\dagger\gamma^5\psi, \\
&\text{vector current:} && V^\mu = \overline{\psi}\gamma^\mu\psi, \\
&\text{axial vector current :} && A^\mu = \overline{\psi}\gamma^\mu\gamma^5\psi, \\
&\text{tensor current:} && T^{\mu\nu} = \overline{\psi}\sigma^{\mu\nu}\psi,
\end{aligned} \tag{6.2.7}$$

where $\overline{\psi} = \psi^{\dagger}\gamma^0$, and γ^{μ} $(0 \leqslant \mu \leqslant 3)$ and γ^5 are the Dirac matrices, which are expressed in the forms:

$$\begin{aligned} &\gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \qquad \gamma^k = \begin{pmatrix} 0 & \sigma^k \\ -\sigma^k & 0 \end{pmatrix} \qquad \text{for } 1 \leqslant k \leqslant 3, \\ &\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}. \end{aligned} \tag{6.2.8}$$

Remark 6.6 The particle currents defined in (6.2.7) are very important in the transition theory of particle decays and scatterings. In fact, the general form of particle currents is written as

$$J_{AB} = \overline{\psi}_A \gamma \psi_B + h \cdot c \quad (\text{ Hermitian conjugate}),$$

where γ is a current operator in (6.2.7), ψ_A and ψ_B are wave functions of particles A and B. For example, for the β-decay

$$n \to p + e^- + \overline{\nu}_e,$$

by the Fermi theory, the transition amplitude is

$$M = \frac{G_f}{\sqrt{2}}(\overline{\psi}_e \gamma^{\mu} \psi_{\nu_e})(\overline{\psi}_p \gamma_{\mu} \psi_n) + h \cdot c,$$

where G_f is the Fermi constant, and

$$\gamma_{\mu} = g_{\mu\nu}\gamma^{\nu} = (-\gamma^0, \gamma^1, \gamma^2, \gamma^3).$$

6.2.2 Quantum dynamic equations

In quantum mechanics, the following are the four basic dynamic equations:

Schrödinger equation,	governing particles at lower velocity	
Klein-Gordon equations,	describing bosons,	(6.2.9)
Weyl equations,	describing massless fermions,	
Dirac equations,	governing massive fermions.	

These four equations were initially derived in the spirit of Postulate 6.5. They can also be equivalently obtained by the Principle of Lagrangian Dynamics (PLD) or by the Principle of Hamiltonian Dynamics (PHD).

Although the three fundamental principles: Postulate 6.5, PLD, and PHD are equivalent in describing quantum mechanical systems, they offer different perspectives. In the following, we introduce the four dynamic equations based on the three principles.

Quantum dynamics based on Postulate 6.5

1. *Schrödinger equation.* In classical mechanics we have the energy-momentum relation

$$E = \frac{1}{2m}p^2 + V, \quad V \text{ is potential energy.}$$

By the Hermitian operators in (6.2.2), this relation leads to the Schrödinger equation, written as

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi, \tag{6.2.10}$$

which is clearly is non-relativistic.

2. *Klein-Gordon equation.* The relativistic energy-momentum relation is given by

$$E^2 - c^2p^2 = m^2c^4.$$

From this relation we can immediately derive the following Klein-Gordon equation:

$$\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \Delta\right)\psi - (\frac{mc}{\hbar})^2\psi = 0. \tag{6.2.11}$$

In the 4-dimensional vectorial form, the equation (6.2.11) is expressed as

$$\partial^\mu\partial_\mu\psi - \left(\frac{mc}{\hbar}\right)^2\psi = 0,$$

which is clearly Lorentz invariant.

3. *Weyl equations.* For a massless particle, the de Broglie matter-wave relation gives

$$E = \hbar\omega, \quad p_0 = \hbar/\lambda, \quad c = \omega\lambda,$$

where p_0 is a scalar momentum. It follows that

$$E = cp_0, \tag{6.2.12}$$

By (6.2.5), corresponding to p_0 the Hermitian operator reads

$$\hat{p}_0 = \mathrm{i}\hbar(\vec{\sigma}\cdot\nabla).$$

Thus, the relation (6.2.12) leads to the following Weyl equation:

$$\frac{\partial\psi}{\partial t} = c(\vec{\sigma}\cdot\nabla)\psi, \tag{6.2.13}$$

where $\psi = (\psi_1, \psi_2)^{\mathrm{T}}$ is a two-component Weyl spinor. The equation (6.2.13) describes free neutrinos.

4. *Dirac equations.* For a massive particle, the de Broglie matter-wave relation (6.2.12) should be rewritten in the form

$$E = cp_1 \pm mc^2, \tag{6.2.14}$$

where p_1 is a scalar momentum for massive fermions, and by (6.2.5) the Hermitian operators for p_1 and m are given by

$$\hat{p}_1 = -i\hbar(\vec{\alpha}\cdot\nabla), \quad \hat{m} = m\alpha_0,$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$,

$$\alpha_0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \qquad \alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix}, \quad k = 1, 2, 3,$$

and σ_k $(1 \leqslant k \leqslant 3)$ are the Pauli matrices as in (3.5.36).

Thus, the de Broglie matter-wave relation (6.2.14) for massive fermions leads to the following Dirac equations:

$$i\hbar \frac{\partial \psi}{\partial t} = -i\hbar c(\vec{\alpha} \cdot \nabla)\psi + mc^2 \alpha_0 \psi, \tag{6.2.15}$$

where $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)^{\mathrm{T}}$ is the Dirac spinor.

Multiplying both sides of (6.2.15) by the matrix α_0, then the Dirac equation is rewritten in the usual form

$$\left(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar}\right)\psi = 0, \tag{6.2.16}$$

where $\gamma^\mu = (\gamma^0, \gamma^1, \gamma^2, \gamma^3)$ is the Dirac matrices as in (6.2.8), with $\gamma^0 = \alpha_0, \gamma^k = \alpha_0 \alpha_k$ $(1 \leqslant k \leqslant 3)$.

Remark 6.7 Based on the spinor theory, the Weyl equation (6.2.13) and the Dirac equations (6.2.16) are Lorentz invariant (see Section 2.2.6) and space rotation invariant. In addition, the Dirac equations are invariant under the space reflection

$$\widetilde{x} = -x, \quad \widetilde{t} = t. \tag{6.2.17}$$

In fact, under the reflection transformation (6.2.17), the Dirac spinor ψ transforms as

$$\widetilde{\psi} = \gamma^0 \psi \quad (\text{ or } \psi = \gamma^0 \widetilde{\psi}). \tag{6.2.18}$$

Thus (6.2.16) becomes

$$i\gamma^\mu \widetilde{\partial}_\mu \widetilde{\psi} - \frac{mc}{\hbar}\widetilde{\psi} = \gamma^0 \left(i\gamma^0 \partial_0 \psi - i\gamma^0 \gamma^k \gamma^0 \partial_k \psi - \frac{mc}{\hbar}\psi\right) = \gamma^0 \left(i\gamma^\mu \partial_\mu - \frac{mc}{\hbar}\right)\psi = 0.$$

Here we used the identities

$$\gamma^0\gamma^0 = I, \quad \gamma^0\gamma^k = -\gamma^k \qquad \text{for } 1 \leqslant k \leqslant 3.$$

Remark 6.8 Because the Weyl spinor has two-components, under the reflection transformation (6.2.17), it is invariant:

$$x \to -x \;\Rightarrow\; \psi \to \psi \; (\psi \text{ the Weyl spinor}).$$

Hence the Weyl equation is not invariant under the space reflection (6.2.17), which leads to violation of parity conservation for decays and scatterings involving neutrinos. The violation of parity conservation was discovered by (Lee and Yang, 1956). □

Quantum dynamics based on PLD

Due to PLD, as the Lagrangian action of a quantum system ψ is given by

$$L = \int \mathscr{L}(\psi, D\psi) dx^\mu,$$

then its dynamic equation is derived by

$$\frac{\delta}{\delta\psi} L = 0. \tag{6.2.19}$$

Hence, we only need to give the Lagrangian action for each of (6.2.9)

1. *Schrödinger systems*. For the Schrödinger equation (6.2.10), its action takes the form

$$\mathscr{L}_s = \frac{1}{2} i\hbar \left(\frac{\partial\psi}{\partial t}\psi^* - \frac{\partial\psi^*}{\partial t}\psi \right) - \frac{1}{2}\left(\frac{\hbar^2}{2m}|\nabla\psi|^2 + V|\psi|^2 \right). \tag{6.2.20}$$

2. *Klein-Gordon systems.* The action for the Klein-Gordon equation (6.2.11) is as follows

$$\mathscr{L}_{KG} = \frac{1}{2}\partial^\mu\psi^*\partial_\mu\psi + \frac{1}{2}\left(\frac{mc}{\hbar}\right)^2 |\psi|. \tag{6.2.21}$$

3) *Weyl systems.* The action for the Weyl equation (6.2.13) reads

$$\mathscr{L}_w = \psi^\dagger \sigma^\mu \partial_\mu \psi, \tag{6.2.22}$$

where $\sigma^0 = -I, (\sigma^1, \sigma^2, \sigma^3) = \vec{\sigma}$.

4. *Dirac systems.* The action for the Dirac equation (6.2.16) is

$$\mathscr{L}_D = \overline{\psi}\left(i\gamma^\mu\partial_\mu - \frac{mc}{\hbar} \right)\psi. \tag{6.2.23}$$

Remark 6.9 Historically the four basic quantum dynamic equations were discovered all based on Postulate 5.5, and then their actions (6.2.20)-(6.2.23) were established by the known corresponding equations. However, the Lagrange actions have played crucial roles in developing the interaction field theories. □

Quantum dynamics based on PHD

In Section 2.6.4, we derived all quantum dynamic equations from the Principle of Hamiltonian Dynamics (PHD). PHD amounts to saying that each conservation system is characterized by a set of conjugate fields

$$\psi_k, \quad \varphi_k, \quad 1 \leqslant k \leqslant N, \tag{6.2.24}$$

and the conjugate fields (6.2.24) define the Hamiltonian energy

$$H=\int_{\mathbb{R}^3}\mathscr{H}(\psi,\varphi)dx. \tag{6.2.25}$$

Then the dynamic equations of this system are given by

$$\begin{aligned}\frac{\partial\psi_k}{\partial t}&=\frac{\delta}{\delta\varphi_k}H,\\ \frac{\partial\varphi_k}{\partial t}&=-\frac{\delta}{\delta\psi_k}H,\end{aligned} \tag{6.2.26}$$

where $\psi=(\psi_1,\cdots,\psi_N),\varphi=(\varphi_1,\cdots,\varphi_N)$ are as in (6.2.24), and H is the Hamiltonian energy.

In the classical quantum mechanics, if the dynamic equation of a quantum system is in the form

$$i\hbar\frac{\partial\psi}{\partial t}=\hat{H}\psi,\quad \hat{H}\ \text{ an Hermitian operator,} \tag{6.2.27}$$

then $\hat{H}$ is called the Hamiltonian operator of this system, and the physical quantity

$$H=\langle\psi|\hat{H}|\psi\rangle=\int_{\mathbb{R}^3}\psi^\dagger\hat{H}\psi dx \tag{6.2.28}$$

is the Hamilton energy of the system.

Three systems: the Schrödinger system, the Weyl system and the Dirac system, are in the form of (3.127) and (3.128). Namely the equations can be expressed in the form (3.127), with the Hamiltonian operators given by

$$\begin{aligned}&\hat{H}=-\frac{\hbar^2}{2m}\Delta+V(x) &&\text{for Schrödinger}\\ &\hat{H}=i\hbar c(\vec{\sigma}\cdot\nabla) &&\text{for Weyl spinor fields,}\\ &\hat{H}=-i\hbar c(\vec{\alpha}\cdot\nabla)+mc^2\alpha_0 &&\text{for Dirac spinor fields,}\end{aligned} \tag{6.2.29}$$

and the associated Hamiltonian energies given by

$$\begin{aligned}&H=\int_{\mathbb{R}^3}\left[\frac{\hbar^2}{2m}|\nabla\psi|^2+V|\psi|^2\right]dx &&\text{for Schrödinger system,}\\ &H=\int_{\mathbb{R}^3}\left[i\hbar c\psi^\dagger(\vec{\sigma}\cdot\nabla)\psi\right]dx &&\text{for Weyl System,}\\ &H=\int_{\mathbb{R}^3}\left[-i\hbar c\psi^\dagger(\vec{\alpha}\cdot\nabla)\psi+mc^2\psi^\dagger\alpha_0\psi\right]dx &&\text{for Dirac system.}\end{aligned} \tag{6.2.30}$$

Here the conjugate fields are the real and imaginary parts of $\psi=\psi^1+i\psi^2$.

By applying the PHD, from the Hamiltonian energies in (6.2.30) we can derive the dynamic equations of the three systems, which are equivalent to the form (6.2.27); see (2.6.49)-(2.6.55).

However the Klein-Gordon system is different. In fact, we can write the equation (6.2.11) in the form

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix}\psi\\ \varphi\end{pmatrix}=\hat{H}\begin{pmatrix}\psi\\ \varphi\end{pmatrix}, \tag{6.2.31}$$

where $\hat{H}$ is given by

$$\hat{H}=\begin{pmatrix}0 & 1\\ -\hbar^2c^2\Delta+m^2c^4 & 0\end{pmatrix}. \tag{6.2.32}$$

However, it is clear that the operator $\hat{H}$ of (6.2.32) is not Hermitian, and therefore $\hat{H}$ is not a Hamiltonian. Consequently the quantity

$$\langle\Phi|\hat{H}|\Phi\rangle=\int_{\mathbb{R}^3}\left[\psi^*\varphi+\varphi^*(-\hbar^2c^2\Delta\psi+m^2c^4\psi)\right]dx$$

is also not a physical quantity because it is not a real number in general. In other words, under the theoretic frame based on Postulate 6.5 and PLD, the Klein-Gordon equation can not be regarded as a model to describe a conservation quantum system.

With PHD, we can, however, show that the Klein-Gordon equation is a model for a conserved system. As seen in (2.6.57), we take a pair conjugate fields (ψ,φ) and the Hamiltonian energy

$$H=\frac{1}{2}\int_{\mathbb{R}^3}\left[\varphi^2+c^2|\nabla\psi|^2+\frac{m^2c^4}{\hbar^2}|\psi|^2\right]dx. \tag{6.2.33}$$

Then the Klein-Gordon equation (6.2.11) can be rewritten as

$$\frac{\partial}{\partial t}\begin{pmatrix}\psi\\ \varphi\end{pmatrix}=J\begin{pmatrix}\frac{\delta}{\delta\psi}H & 0\\ 0 & \frac{\delta}{\delta\varphi}H\end{pmatrix}\begin{pmatrix}\psi\\ \varphi\end{pmatrix}, \tag{6.2.34}$$

where H is as in (6.2.33), and

$$J=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}. \tag{6.2.35}$$

The Hamiltonian operator $\hat{H}$ for the Klein-Gordon system reads

$$\hat{H}=\delta H=\begin{pmatrix}-c^2\Delta+\frac{m^2c^4}{\hbar^2} & 0\\ 0 & 1\end{pmatrix}. \tag{6.2.36}$$

The model of (6.2.33)-(6.2.36) is in a standard form of PHD. In fact, all conservation quantum systems, including the four classical systems (6.2.9), can be expressed in the standard PHD from, which we call the Quantum Hamiltonian Dyanmics (QHD), which is stated in the following.

Principle 6.10 *A conserved quantum system can be described by a set of conjugate fields*

$$\Psi=(\psi_1,\cdots,\psi_N),\quad \Phi=(\varphi_1,\cdots,\varphi_N),$$

and an associated Hamiltonian energy $H=H(\Psi,\Phi)$, such that the dynamic equations of this system are in the form

$$\frac{\partial}{\partial t}\begin{pmatrix}\Psi\\ \Phi\end{pmatrix}=J\hat{H}(\Psi,\Phi),\quad J=\begin{pmatrix}0 & I\\ -I & 0\end{pmatrix}, \tag{6.2.37}$$

where $\hat{H}=\delta H$ is the variational derivative operator of the Hamiltonian energy H, I is the N-th order identity matrix. In particular, if the Hamiltonian H is invariant under the transformation of conjugate fields

$$\begin{pmatrix}\psi_k\\ \varphi_k\end{pmatrix}\to\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\psi_k\\ \varphi_k\end{pmatrix}\qquad \text{for } 1\leqslant k\leqslant K,$$

then the conjugate fields constitute complex valued wave functions $\psi_k+i\varphi_k$ (or $\varphi_k+i\psi_k$), $1\leqslant k\leqslant K$, for this system.

Remark 6.11 In classical quantum mechanics, the Klein-Gordon equation encounters a difficulty as mentioned in (6.2.31) that the operator $\hat{H}$ of (6.2.32) is not Hermitian, inconsistent with quantum mechanical principles for describing bosonic behaviors. But, under the QHD model (6.2.37) this difficulty is solved. In the Angular Momentum Rule in Section 6.2.4 and the spinor BEC (nonlinear quantum system) in Chapter 7, we can see this point clearly.

6.2.3 Heisenberg uncertainty relation and Pauli exclusion principle

The Heisenberg uncertainty relation and Pauli exclusion principle are two important quantum physical laws.

Uncertainty principle

We first give this relation, which is stated as follows.

Uncertainty Principle 6.12 *In a quantum system, the position x and momentum p, the time t and energy E satisfy the uncertainty relations given by*

$$\Delta x\Delta p\geqslant\frac{1}{2}\hbar,\quad \Delta t\Delta E\geqslant\frac{1}{2}\hbar, \tag{6.2.38}$$

where ΔA represents a measuring error of A value. Namely, (6.2.38) implies that x and p,t and E can not be precisely observed at the same moment.

The relations (6.2.58) can be deduced from Postulates 6.3 and 6.4. The average values of position and momentum are

$$\langle x\rangle = \int \psi^* x\psi dx,$$
$$\langle p\rangle = -\int \psi^* i\hbar\frac{\partial\psi}{\partial x}dx.$$

In statistics, the error to an average value is expressed by the squared deviation. Namely, the errors to $\langle x\rangle$ and $\langle p\rangle$ are

$$\langle(\Delta x)^2\rangle = \int \psi^*(x-\langle x\rangle)^2\psi dx = \langle x^2\rangle - \langle x\rangle^2,$$
$$\langle(\Delta p)^2\rangle = \int \psi^*(p_x-\langle p_x\rangle)^2\psi dx = \langle p_x^2\rangle - \langle p_x\rangle^2.$$

Assume that $\langle x\rangle = 0$ and $\langle p_x\rangle = 0$. Then we have

$$\begin{aligned}\langle(\Delta x)^2\rangle &= \langle x^2\rangle = \int x^2|\psi|^2dx,\\ \langle(\Delta p)^2\rangle &= \langle p_x^2\rangle = \int \psi^*\left(-i\hbar\frac{\partial}{\partial x}\right)^2\psi dx.\end{aligned} \tag{6.2.39}$$

To get the relation (6.2.38) we consider the integral

$$I(\alpha) = \int\left(\alpha x\psi^* + \frac{\partial\psi^*}{\partial x}\right)\left(\alpha x\psi + \frac{\partial\psi}{\partial x}\right)dx, \tag{6.2.40}$$

where α is a real number. The integral (6.2.40) can be written as

$$I(\alpha) = A\alpha^2 - B\alpha + C,$$

where

$$\begin{aligned}A &= \int x^2|\psi|^2dx = \langle(\Delta x)^2\rangle \quad \text{(by (6.2.39))},\\ B &= -\int x\frac{\partial}{\partial x}|\psi|^2dx = \int|\psi|^2dx = 1,\\ C &= \int\frac{\partial\psi^*}{\partial x}\frac{\partial\psi}{\partial x}dx = \frac{1}{\hbar^2}\int\psi^*\left(-i\hbar\frac{\partial}{\partial x}\right)^2\psi dx = \frac{1}{\hbar^2}\langle(\Delta p)^2\rangle \quad \text{(by (6.2.39))}.\end{aligned} \tag{6.2.41}$$

It is clear that $A,B,C>0$, and by (6.2.40),

$$I(\alpha)\geqslant 0, \qquad \forall\alpha\in\mathbb{R}^1. \tag{6.2.42}$$

Let α_0 be the minimal of $I(\alpha)$. Then α_0 satisfies

$$I'(\alpha_0) = 2A\alpha_0 - B = 0 \;\Rightarrow\; \alpha_0 = \frac{B}{2A}.$$

Inserting $\alpha_0 = B/4A$ in (6.2.42) we get that

$$AC \geqslant \frac{1}{4}B^2.$$

It follows from (6.2.41) that

$$\langle(\Delta x)^2\rangle\langle(\Delta p)^2\rangle \geqslant \frac{\hbar^2}{4},$$

which is the first relation of (6.2.38). The second relation (6.2.38) can be derived in the same fashion.

The Heisenberg uncertainty relation (6.2.38) has profound physically implications, some of which are listed as follows:

1) If a particle A consists of N more fundamental particles A_i $(1 \leqslant i \leqslant N)$, in a small ball with radius r,

$$A = A_1 + \cdots + A_N,$$

then the momentum p of each particle A_i is at least

$$p \geqslant \hbar/4r.$$

Hence, the smaller the composite particle, the greater the bounding energy is needed to hold its constituents together.

2) Most particles are unstable, and their lifetime τ is very short. In addition, the energy distribution of each particle is in a range, called the energy width Γ. By the uncertainty relation, τ and Γ satisfy that

$$\tau\Gamma \geqslant \hbar/2.$$

This relation is very important in experiments, because the width Γ is observable which determines the lifetime of a particle by the uncertainty relation $\tau \simeq \hbar/2\Gamma$.

3) Uncertainty relations (6.2.38) also imply that the energy and momentum conservations may be violated in a small scale of time and space. Both conservations are only the averaged results in larger scale ranges of time and space.

Pauli exclusion principle

We recall that particles are classified two types:

$$\begin{aligned} &\text{fermions} = \text{particles with spin } J = \frac{n}{2} \text{ for odd } n, \\ &\text{bosons} \ \ = \text{particles with spin } J = n \text{ for integer } n. \end{aligned}$$

Fermions and bosons display very different characteristics. The fermions do not like to live together with the same fermions, but bosons are sociable particles. This difference is characterized by the Pauli exclusion principle.

Pauli Exclusion Principle 6.13 *In a quantum system, there are no two or more fermions living in the same quantum states, i.e. possessing entirely the same quantum numbers.*

6.2.4 Angular momentum rule

In Section 5.3.2 we introduced the Angular Momentum Rule, which was first proved in (Ma and Wang, 2015b). It says that only the fermions with spin $J = \frac{1}{2}$ can rotate around a central force field. In fact, this rule can be generalized to scalar bosons, i.e. particles with spin $J = 0$. In this subsection we shall discuss the rule in more details. To this end, we first introduce conservation laws of quantum systems based on Principle 6.10.

Conservation laws based on quantum Hamiltonian dynamics (QHD)

Let H be the Hamiltonian energy of a conservative quantum system, which can be described by the following Hamiltonian equations:

$$\begin{aligned} \frac{\partial \Psi}{\partial t} &= \hat{H}_\Phi(\Psi, \Phi), \\ \frac{\partial \Phi}{\partial t} &= -\hat{H}_\Psi(\Psi, \Phi), \end{aligned} \tag{6.2.43}$$

where

$$\hat{H}_\Phi = \frac{\delta H}{\delta \Phi}, \qquad \hat{H}_\Psi = \frac{\delta H}{\delta \Psi}.$$

Let L be an observable physical quantity with the corresponding Hermitian operator $\hat{L}$ for the conjugate fields $(\Psi, \Phi)^{\mathrm{T}}$ of (6.2.43), and $\hat{L}$ is expressed as

$$\hat{L} = \begin{pmatrix} \hat{L}_{11} & \hat{L}_{12} \\ \hat{L}_{21} & \hat{L}_{22} \end{pmatrix}, \qquad \hat{L}_{12}^T = \hat{L}_{21}^*.$$

Then the physical quantity L of system (6.2.43) is given by

$$L = \int (\Psi^\dagger, \Phi^\dagger) \hat{L} \begin{pmatrix} \Psi \\ \Phi \end{pmatrix} dx = \int \left[\Psi^\dagger \hat{L}_{11} \Psi + \Phi^\dagger \hat{L}_{22} \Phi + 2\mathrm{Re}(\Psi^\dagger \hat{L}_{12} \Phi) \right] dx. \tag{6.2.44}$$

It is clear that the quantity L of (6.2.44) is conserved if for the solution $(\Psi, \Phi)^{\mathrm{T}}$ of (6.2.43) we have

$$\frac{dL}{dt} = 0,$$

which is equivalent to

$$\begin{aligned} \int \Big[& \hat{H}_\Phi^\dagger \hat{L}_{11} \Psi + \Psi^\dagger \hat{L}_{11} \hat{H}_\Phi - \hat{H}_\Psi^\dagger \hat{L}_{22} \Phi - \Phi^\dagger \hat{L}_{22} \hat{H}_\Psi \\ & + 2\mathrm{Re}(\hat{H}_\Phi^\dagger \hat{L}_{12} \Phi - \Psi^\dagger \hat{L}_{12} \hat{H}_\Psi) \Big] dx = 0. \end{aligned} \tag{6.2.45}$$

We remark here that if the QHD is described by a complex valued wave function:

$$\psi = \Psi + i\Phi,$$

and its dynamic equation is linear, then (6.2.43) can be written as

$$i\hbar\frac{\partial\psi}{\partial t}=\hat{H}\psi,\quad H=\int\psi^{\dagger}\hat{H}\psi dx. \tag{6.2.46}$$

In this case, the physical quantity L in (6.2.44) is in the form

$$L=\int\psi^{\dagger}\hat{L}\psi dx, \tag{6.2.47}$$

and the conservation law (6.2.45) of L is equivalent to

$$\hat{L}\hat{H}-\hat{H}\hat{L}=0. \tag{6.2.48}$$

The formulas (6.2.46)-(6.2.48) are the conservation laws of the classical quantum mechanics.

Hence, the conservation laws in (6.2.45) are the generalization to the classical quantum mechanics, and are applicable to all conservative quantum systems, including the Klein-Gordon systems and nonlinear systems.

Angular momentum rule

From the conservation laws (6.2.45) and (6.2.48) we can deduce the following Angular Momentum Rule.

Angular Momentum Rule 6.14 *Only the fermions with spin $J=\frac{1}{2}$ and the bosons with $J=0$can rotate around a center with zero moment of force. The particles with $J\neq 0,\frac{1}{2}$will move on a straight line unless there is a nonzero moment of force present.*

In the following we give a mathematical derivation of the Angular Momentum Rule.

1. *Fermions.* Consider fermions which obey the Dirac equations as (6.2.46) with the Hamiltonian

$$\hat{H}=-i\hbar c(\alpha^{k}\partial_{k})+mc^{2}\alpha^{0}+V(r), \tag{6.2.49}$$

where V is the potential energy of a central field, and α^0,α^k $(1\leqslant k\leqslant 3)$ are the Dirac matrices

$$\alpha^{0}=\begin{pmatrix}I & 0\\ 0 & -I\end{pmatrix},\quad \alpha^{k}=\begin{pmatrix}0 & \sigma^{k}\\ \sigma^{k} & 0\end{pmatrix}\qquad \text{for } 1\leqslant k\leqslant 3, \tag{6.2.50}$$

and σ^k are the Pauli matrices.

The total angular momentum $\hat{J}$ of a particle is

$$\hat{J}=\hat{L}+s\hat{S},$$

where s is the spin, $\hat{L}$ is the orbital angular momentum

$$\begin{aligned}
&\hat{L} = (\hat{L}_1, \hat{L}_2, \hat{L}_3) = \hat{r} \times \hat{p}, \quad \hat{p} = -i\hbar\nabla, \\
&\hat{L}_1 = -i\hbar(x_2\partial_3 - x_3\partial_2), \\
&\hat{L}_2 = -i\hbar(x_3\partial_1 - x_1\partial_3), \\
&\hat{L}_3 = -i\hbar(x_1\partial_2 - x_2\partial_1),
\end{aligned} \tag{6.2.51}$$

and $\hat{S}$ is the spin operator

$$\hat{S} = (\hat{S}_1, \hat{S}_2, \hat{S}_3), \quad \hat{S}_k = \hbar \begin{pmatrix} \sigma^k & 0 \\ 0 & \sigma^k \end{pmatrix} \qquad \text{for } 1 \leqslant k \leqslant 3. \tag{6.2.52}$$

By (6.2.49)-(6.2.52), we see that

$$\begin{aligned}
\hat{H}\hat{L}_1 - \hat{L}_1\hat{H} =&\hbar^2 c[(x_2\partial_3 - x_3\partial_2)(\alpha^2\partial_2 + \alpha^3\partial_3) - (\alpha^2\partial_2 + \alpha^3\partial_3)(x_2\partial_3 - x_3\partial_2)] \\
=&\hbar^2 c[\alpha^2\partial_3(x_2\partial_2 - \partial_2 x_2) - \alpha^3\partial_2(x_3\partial_3 - \partial_3 x_3)].
\end{aligned}$$

Notice that

$$x_2\partial_2 - \partial_2 x_2 = x_3\partial_3 - \partial_3 x_3 = -1.$$

Hence we get

$$\hat{H}\hat{L}_1 - \hat{L}_1\hat{H} = \hbar^2 c(\alpha^3\partial_2 - \alpha^2\partial_3). \tag{6.2.53}$$

Similarly we have

$$\begin{aligned}
\hat{H}\hat{L}_2 - \hat{L}_2\hat{H} = \hbar^2 c(\alpha^1\partial_3 - \alpha^3\partial_1), \\
\hat{H}\hat{L}_3 - \hat{L}_3\hat{H} = \hbar^2 c(\alpha^2\partial_1 - \alpha^1\partial_2).
\end{aligned} \tag{6.2.54}$$

On the other hand, we infer from (6.2.49) and (6.2.52) that

$$\hat{H}\hat{S}_j - \hat{S}_j\hat{H} = -i\hbar^2 c\gamma^5 \left[\partial_k(\sigma^k\sigma^j - \sigma^j\sigma^k)\right] = -i\hbar^2 c\gamma^5(2i\varepsilon_{kjl}\sigma^l)\partial_k = 2\hbar^2 c\varepsilon_{kjl}\alpha^l\partial_k,$$

where γ^5 is defined by

$$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}.$$

Hence we have

$$\begin{aligned}
\hat{H}\hat{S}_1 - \hat{S}_1\hat{H} = -2\hbar^2 c(\alpha^3\partial_2 - \alpha^2\partial_3), \\
\hat{H}\hat{S}_2 - \hat{S}_2\hat{H} = -2\hbar^2 c(\alpha^1\partial_3 - \alpha^3\partial_1), \\
\hat{H}\hat{S}_3 - \hat{S}_3\hat{H} = -2\hbar^2 c(\alpha^2\partial_1 - \alpha^1\partial_2).
\end{aligned} \tag{6.2.55}$$

For $\hat{J} = \hat{L} + s\hat{S}$, we derive from (6.2.53)-(6.2.55) that

$$\hat{H}\hat{J} - \hat{J}\hat{H} = 0 \Longleftrightarrow \text{ spin } s = \frac{1}{2}. \tag{6.2.56}$$

When fermions move on a straight line,

$$\hat{H} = c\alpha^3 p_3, \quad \hat{L} = 0.$$

In this case, by (6.2.53)-(6.2.54), for straight line motion,

$$\hat{H}\hat{J} - \hat{J}\hat{H} = 0 \qquad \text{for any } s. \tag{6.2.57}$$

Thus, by the conservation law (6.2.48), the assertion of Angular Momentum Rule for fermions follows from (6.2.56) and (6.2.57).

2. *Bosons.* Now, consider bosons which bey the Klein-Gordon equation in the form (6.2.43). It is known that the spins J of bosons depend on the types of Klein-Gordon fields (Ψ, Φ):

$$\begin{pmatrix} \Psi \\ \Phi \end{pmatrix} = \begin{cases} \text{a scalar field} & \Rightarrow \quad J = 0, \\ \text{a 4-vector field} & \Rightarrow \quad J = 1, \\ \text{a 2nd-order tensor field} & \Rightarrow \quad J = 2, \\ \text{a real valued field} & \Rightarrow \quad \text{neutral bosons}, \\ \text{a complex valued field} & \Rightarrow \quad \text{charged bosons}. \end{cases} \tag{6.2.58}$$

For the Klein-Gordon fields $(\Psi, \Phi)^T$, the Hamiltonian for a central force field is given by

$$H = \frac{1}{2}\int \left[|\Phi|^2 + c^2|\nabla\Psi|^2 + \frac{1}{\hbar^2}(m^2c^4 + V(r))|\Psi|^2 \right] dx \tag{6.2.59}$$

The Hamiltonian energy operator $\hat{H}$ of (6.2.59) is given by

$$\hat{H} = \begin{pmatrix} \hat{H}_\Psi & 0 \\ 0 & \hat{H}_\Phi \end{pmatrix}, \qquad \hat{H}_\Phi = \Phi, \qquad \hat{H}_\Psi = \left[-c^2\Delta + \frac{1}{\hbar^2}(m^2c^4 + V) \right] \Psi. \tag{6.2.60}$$

The angular momentum operator $\hat{J}$ is

$$\hat{J} = \begin{pmatrix} \hat{L} & 0 \\ 0 & \hat{L} \end{pmatrix} + s\hbar\hat{\sigma}, \quad \hat{\sigma} = (\sigma^1, \sigma^2, \sigma^3). \tag{6.2.61}$$

where s is the spin of bosons, and $\hat{L}$ is as in (6.2.51).

For scalar bosons, spin $s = 0$ in (6.2.61) and the Hermitian operators in the conservation law (6.2.45) are

$$\hat{L}_{11} = \hat{L}_{22} = \hat{L}, \qquad \hat{L}_{12} = \hat{L}_{21} = 0, \qquad \hat{H}_\Phi, \hat{H}_\Psi \text{ as in (6.2.60)}$$

Then by

$$\int \Phi^\dagger \hat{L}\Psi dx = -\int \Psi^* \hat{L}^\dagger \Phi dx,$$
$$\int \hat{H}_\Psi^\dagger \hat{L}\Phi dx = -\int \Phi^\dagger \hat{L}\hat{H}_\Psi dx,$$

we derive the conservation law (6.2.45), i.e.

$$\int \left[\Phi^\dagger \hat{L}\Psi + \Psi^\dagger \hat{L}\Phi - \hat{H}_\Psi^\dagger \hat{L}\Phi - \Phi^\dagger \hat{L}\hat{H}_\Psi\right] dx = 0.$$

However, it is clear that $\hat{H}$ and $\hat{J}$ in (6.2.60) and (6.2.61) don't satisfy (6.2.45) for spin $s \neq 0$. Hence the quantum rule of angular momentum for bosons holds true.

Remark 6.15 The Angular Momentum Rule is very useful in the weakton model and the theory of mediator cloud structure of charged leptons and quarks, which explain why all stable fermions with mediator clouds are at spin $J = 1/2$.

6.3 Solar Neutrino Problem

6.3.1 Discrepancy of the solar neutrinos

The solar neutrino problem is known as that the number of electron neutrinos arriving from the Sun are between one third and one half of the number predicted by the Standard Solar Model. This important discovery was made in 1968 by R. Davis, D. S. Harmer and K. C. Hoffmann (Davis, Harmer and Hoffman, 1968).

To understand this problem clearly, we begin with a brief introduction to the Standard Solar Model, following (Griffiths, 2008).

In the nineteenth century, most physicist believed that the source of the Sun's energy was gravity. However, based on this assumption, Rayleigh showed that the maximum possible age of the Sun was substantially shorter than the age of the earth estimated by geologists.

At the end of the nineteenth century, Bacquerel and Curies discovered radioactivity, and they noted that radioactive substances release a large amounts of heat. This suggested that nuclear fission, not gravity, might be the source of the Sun's energy, and it could allow for a much longer lifetime of the Sun. But, the crucial problem for this solar model was that there were no heavier radioactive elements such as uranium or radium present in the Sun, and from the atomic spectrum, it was known that the Sun is made almost entirely of hydrogen.

Up to 1920, F. W. Aston gave a series of precise measurements of atomic masses. It was found that four hydrogen atoms are more weight slight than one atom of helium-4. This implied that the fusion of four hydrogens to form a ^4He would be more favorable, and would release a substantial amounts of energy. A. Eddington proposed that the source of the Sun's energy is the nuclear fusion, and in essence he was correct.

In 1938, H. Bethe in collaboration with C. Critchfield had come up with a series of subsequent nuclear reactions, which was known as the proton-proton $p-p$ chain. The $p-p$ cycle well describes the reaction processes in the Sun, and consists of the following four steps:

Step 1: two protons yield a deuteron

$$p + p \longrightarrow {}^{2}\mathrm{H} + e^{+} + \nu_e \qquad \text{at } 99.75\%,$$
$$p + p + e^{-} \longrightarrow {}^{2}\mathrm{H} + \nu_e \qquad \text{at } 0.25\%,$$

Step 2: a deuteron and a proton produces a helium-3

$$^{2}\mathrm{H} + p \longrightarrow {}^{3}\mathrm{He} + \gamma,$$

Step 3: helium-3 makes helium-4 or beryllium

$$^{3}\mathrm{He} + p \longrightarrow {}^{4}\mathrm{He} + e^{+} + \nu_e,$$
$$^{3}\mathrm{He} + {}^{3}\mathrm{He} \longrightarrow {}^{4}\mathrm{He} + p + p \qquad \text{almost at } 86\%$$
$$^{3}\mathrm{He} + {}^{4}\mathrm{He} \longrightarrow {}^{7}\mathrm{Be} + \gamma \qquad \text{at } 14\%,$$

Step 4: beryllium makes helium -4

$$^{7}\mathrm{Be} + e^{-} \longrightarrow {}^{7}\mathrm{Li} + \nu_e \qquad \text{at } 99.89\%,$$
$$^{7}\mathrm{Li} + p \longrightarrow {}^{4}\mathrm{He} + {}^{4}\mathrm{He},$$
$$^{7}\mathrm{Be} + p \longrightarrow {}^{8}\mathrm{Be} + \gamma \qquad \text{at } 0.11\%,$$
$$^{8}\mathrm{Be} \longrightarrow {}^{8}\mathrm{Be}^{*} + e^{+} + \nu_e,$$
$$^{8}\mathrm{Be}^{*} \longrightarrow {}^{4}\mathrm{He} + {}^{4}\mathrm{He}.$$

In the $p-p$ chain, it all starts out as hydrogen (proton), and it all ends up as ^{4}He plus some electrons, positrons, photons and neutrinos. Because neutrinos interact so weakly, they are the unique products in the $p-p$ reactions reaching the earth's surface.

In the $p-p$ chain there are five reactions to yield neutrinos:

$$p + p \longrightarrow {}^{2}\mathrm{H} + e^{+} + \nu_e, \tag{6.3.1}$$
$$p + p + e^{-} \longrightarrow {}^{2}\mathrm{H} + \nu_e, \tag{6.3.2}$$
$$^{3}\mathrm{He} + p \longrightarrow {}^{4}\mathrm{He} + e^{+} + \nu_e, \quad (\simeq 0\%), \tag{6.3.3}$$
$$^{7}\mathrm{Be} + e^{-} \longrightarrow {}^{7}\mathrm{Li} + \nu_e, \tag{6.3.4}$$
$$^{8}\mathrm{Be} \longrightarrow {}^{8}\mathrm{Be}^{*} + e^{+} + \nu_e. \tag{6.3.5}$$

But the problem is that the detection of the neutrinos have an effect threshold which will lead to a nearly vanishing response to all neutrinos of lower energy. The energy spectras of neutrinos in the five reactions are

$$\begin{aligned} &E_m \simeq 0.4\ \mathrm{MeV} && \text{for } (6.3.1), \\ &E_m \simeq 1.44\ \mathrm{MeV} && \text{for } (6.3.2), \\ &E_m \simeq 18\ \mathrm{MeV} && \text{for } (6.3.3), \\ &E_m \simeq 0.9\ \mathrm{MeV} && \text{for } (6.3.4), \\ &E_m \simeq 14\ \mathrm{MeV} && \text{for } (6.3.5), \end{aligned} \tag{6.3.6}$$

where E_m is the maximum energy of neutrinos, and the energy flux are

$$\begin{aligned} F &\simeq 10^{11}/\text{cm}^2\cdot s \quad &&\text{for (6.3.1)},\\ F &\simeq 10^{8}/\ \text{cm}^2\cdot s \quad &&\text{for (6.3.2)}\\ F &\simeq 10^{2}/\ \text{cm}^2\cdot s \quad &&\text{for (6.3.3)}\\ F &\simeq 10^{10}/\ \text{cm}^2\cdot s \quad &&\text{for (6.3.4)},\\ F &\simeq 10^{6}/\ \text{cm}^2\cdot s \quad &&\text{for (6.3.5)}. \end{aligned} \tag{6.3.7}$$

Homestake experiments

The experimental search for solar neutrinos has been undertaken since 1965 by R. Davis and collaborators in the Homestake goldmine in South Dakota. Since the neutrinos cannot be directly detected by instruments, it is only by the reactions

$$\nu_e + X \longrightarrow Y + e^-$$

to detect the outgoing products that counte the neutrinos. The Homestake experiments take

$$\nu_e + {}^{37}\text{Cl} \longrightarrow {}^{37}\text{Ar} + e^-. \tag{6.3.8}$$

The effective threshold of the reaction (6.3.8) is

$$E_c = 5.8\text{MeV}.$$

Thus, by (6.3.6) only these neutrinos from both reactions (6.3.3) and (6.3.5) can be observed, which occur at a frequency of 0.015%. Theoretic computation showed that the expected counting rate of solar neutrinos is at

$$N_{Th} = (5.8 \pm 0.7)\ \text{snu}, \tag{6.3.9}$$

where *snu* stands for solar neutrino unit:

$$1\ \ \text{snu} = 10^{-36}\ \ \text{reactions}/({}^{37}\text{Cl atom}\cdot s).$$

In 1968, R. Davis et al (Davis, Harmer and Hoffman, 1968) reported the experimental results, their measuring rate is

$$N_{\text{Exp}} = (2.0 \pm 0.3)\ \text{snu}. \tag{6.3.10}$$

the experimental value (6.3.10) is only about one third of the theoretically expected value (6.3.9). It gave rise to the famous solar neutrino problem.

Super-K experiment

In 2001, the Super-Kamiokande collaboration presented its results on solar neutrinos. Unlike the Homestake experiment, Super-K uses water as the detector. The process is elastic neutrino-electron scattering:

$$\nu_x + e \longrightarrow \nu_x + e,$$

where ν_x is one of the three flavors of neutrinos. This reaction is sensitive to μ and τ neutrinos as well as e-neutrinos, but the detection efficiency is 6.5 times greater for e-neutrinos than for the other two kinds. The outgoing electron is detected by the Cherenkov radiation it emits in water. They observed the rate at

$$r = 45\% \text{ of the expected value.}$$

The Super-Kamiokande detector is located in the Mozumi Mine near Kamioka section of the city of Hida, Japan.

Sudbury Neutrino Observatory (SNO)

Meanwhile, in the summer of 2001 the SNO collaboration reported their observation results. They obtained

$$r = 35\% \text{ of the predicted value.}$$

The SNO used heavy water (^2H_2O) instead of ordinary water (H_2O), and the SNO detection method is based on the following reactions:

$$\nu_e + {}^2\mathrm{H} \longrightarrow p + p + e^-, \tag{6.3.11}$$

$$\nu_x + {}^2\mathrm{H} \longrightarrow p + n + \nu_x, \tag{6.3.12}$$

$$\nu_x + e^- \longrightarrow \nu_x + e^-. \tag{6.3.13}$$

SNO detects electrons e^-, but not τ^- and μ^-, as there is not enough energy in the solar electron-neutrino such that the transformed tau and mu neutrino can excite neutrons in ^2H to produce either τ^- or μ^-.

KamLAND

The loss of reactor electron anti-neutrino $\overline{\nu}_e$ is verified by the KamLAND experiment.

A potential alternative experiment

It is known that the following reaction

$$\nu_\mu + n \longrightarrow \mu^- + p \tag{6.3.14}$$

may occur if the energy of ν_μ satisfies

$$E_{\nu_\mu} > m_\mu c^2 = 106 \text{ MeV}.$$

By the energy spectrum (6.3.6), the maximum energy of solar neutrinos is about $14 \sim 18\,\text{MeV}$, which is much smaller than $m_\mu c^2$. Hence, assuming oscillation does occur for solar neutrinos, the reaction

$$\nu_\mu + {}^2\text{H} \longrightarrow \mu^- + p + p$$

does not occur for the transformed ν_μ from solar electron-neutrinos.

However, based on the weakton model, the complete reaction for (6.3.14) should be

$$\nu_\mu + n + \gamma \longrightarrow \mu^- + p.$$

Consequently, the following reaction

$$\nu_\mu + {}^2\text{H} + \gamma \longrightarrow \mu^- + p + p \tag{6.3.15}$$

would occur if

$$E_{\nu_\mu} + E_\gamma > 106\,\text{MeV}. \tag{6.3.16}$$

Hence one may use high energy photons to hit the heavy water to create the situation in (6.3.16), so that the reaction (6.3.15) may take place. From (6.3.15), we can detect the μ^- particle to test the neutrino oscillation.

Alternatively, by the μ-decay:

$$\mu^- \to e^- + \overline{\nu}_e + \nu_\mu,$$

we may measure the electrons to see if there are more electrons than the normal case to test the existence of mu-neutrinos.

6.3.2 Neutrino oscillations

In order to explain the solar neutrino problem, in 1968 B. Pontecorvo (Pontecorvo, 1957, 1968) introduced the neutrino oscillation mechanism, which amounts to saying that the neutrinos can change their flavors, i.e. an electron neutrino may transform into a muon or a tau neutrino. According to this theory, a large amount of electron neutrinos ν_e from the Sun have changed into the ν_μ or ν_τ, leading the discrepancy of solar electron neutrinos. This neutrino oscillation mechanism is based on the following assumptions:

- The neutrinos are massive, and, consequently, are described by the Dirac equations.
- The three types of neutrinos ν_e, ν_μ, ν_τ are not the eigenstates of the Hamiltonian (i.e. the Dirac operator)

$$\hat{H} = -i\hbar c(\vec{\alpha} \cdot \nabla) + mc^2\alpha_0. \tag{6.3.17}$$

- There are three discrete eigenvalues λ_j of (6.3.17) with eigenstates:

$$\hat{H}\nu_j = \lambda_j \nu_j \qquad \text{for } 1 \leqslant j \leqslant 3, \tag{6.3.18}$$

such that ν_e, ν_μ, ν_τ are some linear combinations of $\{\nu_j \mid 1 \leqslant j \leqslant 3\}$:

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = A \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}, \tag{6.3.19}$$

where $A \in SU(3)$ is a third-order complex matrix given by (6.3.26) below.

Remark 6.16 The formulas (6.3.17)-(6.3.19) constitute the current model of neutrino oscillation, which requires the neutrinos being massive. However, the massive neutrino assumption gives rise two serious problems. First, it is in conflict with the known fact that the neutrinos violate the parity symmetry. Second, the handedness of neutrinos implies their velocity being at the speed of light.

In fact, by using the Weyl equations as the neutrino oscillation model we can also deduce the same conclusions and solve the two mentioned problems. Moreover, the ν mediator introduced by the authors in (Ma and Wang, 2015b) leads to an alternate explanation to the solar neutrino problem. □

Under the above three hypotheses (6.3.17)-(6.3.19), the oscillation between ν_e, ν_μ and ν_τ are given in the following fashion. For simplicity we only consider two kinds neutrinos ν_e, ν_μ, i.e. $\nu_\tau = 0$. In this case, (6.3.19) becomes

$$\begin{aligned} \nu_1 &= \cos\theta \nu_\mu - \sin\theta \nu_e, \\ \nu_2 &= \sin\theta \nu_\mu + \cos\theta \nu_e. \end{aligned} \tag{6.3.20}$$

By the Dirac equations (6.2.15) and (6.3.18), ν_1 and ν_2 satisfy

$$i\hbar \frac{\partial \nu_k}{\partial t} = \lambda_k \nu_k \qquad \text{for } k = 1, 2.$$

The solutions of these equations read

$$\nu_k = \nu_k(0) e^{-i\lambda_k t/\hbar}, \quad k = 1, 2. \tag{6.3.21}$$

Assume that the initial state is at ν_e, i.e.

$$\nu_e(0) = 1, \quad \nu_\mu(0) = 0.$$

Then we derive from (6.3.20) that

$$\nu_1(0) = -\sin\theta, \quad \nu_2(0) = \cos\theta. \tag{6.3.22}$$

It follows from (6.3.21) and (6.3.22) that

$$\nu_1 = -\sin\theta e^{-i\lambda_1 t/\hbar}, \quad \nu_2 = \cos\theta e^{-i\lambda_2 t/\hbar}. \tag{6.3.23}$$

Inserting (6.3.23) into (6.3.20) we deduce that

$$\nu_\mu(t) = \cos\theta\nu_1(t) + \sin\theta\nu_2(t) = \sin\theta\cos\theta(-e^{-i\lambda_1 t/\hbar} + e^{-i\lambda_2 t/\hbar}).$$

Hence, the probability of ν_e transforming to ν_μ at time t is

$$P(\nu_e \to \nu_\mu) = |\nu_\mu(t)|^2 = \left[\sin 2\theta \sin\left(\frac{\lambda_2 - \lambda_1}{2\hbar}t\right)\right]^2. \tag{6.3.24}$$

Also, we derive in the same fashion that

$$\nu_e(t) = \cos\theta\nu_2 - \sin\theta\nu_1 = \cos^2\theta e^{-i\lambda_1 t/\hbar} + \sin^2\theta e^{-i\lambda_2 t/\hbar},$$

and the probability of ν_μ to ν_e is given by

$$P(\nu_\mu \to \nu_e) = |\nu_e(t)|^2 = \cos^2\left(\frac{\lambda_2 - \lambda_1}{2\hbar}t\right) + \cos^2 2\theta \sin^2\left(\frac{\lambda_2 - \lambda_1}{\hbar}t\right) \tag{6.3.25}$$

From formulas (6.3.24) and (6.3.25), we derive the oscillation between ν_e and ν_μ, the energy difference $\lambda_2 - \lambda_1$, and the angle θ, if the discrepancy probability $P(\nu_e \to \nu_\mu)$ is measured.

6.3.3 Mixing matrix and neutrino masses

As mentioned in Remark 6.16, the current neutrino oscillation requires mass matrix A defined in (6.3.19). In this subsection we shall discuss these two topics.

Mixing matrix

The matrix A given in (6.3.19) is called the MNS matrix, which is due to Z. Maki, M. Nakagawa and S. Sakata for their pioneering work in (Maki, Nakagawa and Sakata, 1962). This can be considered as an analog for leptons as the Cabibbo-Kobayashi-Maskawa (CKM) matrix for quarks. The MNS matrix is written as

$$A = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}c_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix}, \tag{6.3.26}$$

where δ is the phase factor, and

$$c_{ij} = \cos\theta_{ij}, \quad s_{ij} = \sin\theta_{ij},$$

with the values θ_{ij} being measured as

$$\theta_{12} \simeq 34^\circ \pm 2^\circ, \quad \theta_{23} \simeq 45^\circ \pm 8^\circ, \quad \theta_{13} \simeq 10^\circ.$$

The matrix A of (6.3.26) is a unitary matrix: $A^\dagger = A^{-1}$. Therefore, (6.3.19) can be also rewritten as

$$\begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} = A^\dagger \begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix}.$$

Neutrino masses

As masses are much less than kinetic energy $c|p|$, by the Einstein triangular relation of energy-momentum

$$E^2 = p^2c^2 + m^2c^4,$$

we obtain an approximate relation:

$$E \simeq |p|c + \frac{1}{2}\frac{m^2c^3}{|p|}.$$

The eigenvalues λ_k of (6.3.18) and E satisfy

$$\lambda_k = E_k \simeq |p|c + \frac{1}{2}\frac{m_k^2c^3}{|p|} \qquad \text{for } k = 1,2,3. \tag{6.3.27}$$

Then we have

$$\lambda_i - \lambda_j = E_i - E_j \simeq \frac{(m_i^2 - m_j^2)}{2E}c^4, \qquad E \simeq |p|c. \tag{6.3.28}$$

By (6.3.28), if we can measure the energy difference $\lambda_i - \lambda_j$, then we get the mass square difference of ν_i and ν_j:

$$\Delta_{ij} = m_i^2 - m_j^2.$$

There are three mass square differences for ν_1, ν_2, ν_3:

$$\Delta_{21} = m_2^2 - m_1^2, \quad \Delta_{32} = m_3^2 - m_2^2, \quad \Delta_{31} = m_3^2 - m_1^2, \tag{6.3.29}$$

only two of which are independent $(\Delta_{31} = \Delta_{32} + \Delta_{21})$.

Now, we consider the mass relation between ν_e, ν_μ, ν_τ and ν_1, ν_2, ν_3. Applying the Dirac operator $\hat{H}$ on both sides of (6.3.19), by (6.3.18), we have

$$\hat{H}\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = A \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}, \tag{6.3.30}$$

where A is the MNS matrix (6.3.26). By Postulate 5.4 (i.e. (6.2.4)), the energies E_e, E_μ, E_τ of ν_e, ν_μ, ν_τ are given by

$$\begin{aligned} E_e &= \int \nu_e^* \hat{H} \nu_e dx = \int A_{1k}^* A_{1j} \nu_k^* \hat{H} \nu_j dx \qquad (\hat{H}\nu_j = \lambda_j \nu_j), \\ E_\mu &= \int \nu_\mu^* \hat{H} \nu_\mu dx = \int A_{2k}^* A_{2j} \nu_k^* \hat{H} \nu_j dx, \\ E_\tau &= \int \nu_\tau^* \hat{H} \nu_\tau dx = \int A_{3k}^* A_{3j} \nu_k^* \hat{H} \nu_j dx, \end{aligned} \tag{6.3.31}$$

where A_{ij} are the matrix elements of A, A^*_{ij} are the complex conjugates of A_{ij}. The masses m_e, m_μ, m_τ of ν_e, ν_μ, ν_τ are as follows

$$E_e^2 = p^2c^2 + m_e^2c^4, \quad E_\mu^2 = p^2c^2 + m_\mu^2c^4, \quad E_\tau^2 = p^2c^2 + m_\tau^2c^4. \tag{6.3.32}$$

It is very difficult to compute E_e, E_μ, E_τ by (6.3.31). However, since $A \in SU(3)$ is norm-preserving:

$$E_e^2 + E_\mu^2 + E_\tau^2 = E_1^2 + E_2^2 + E_3^2,$$

by (6.3.32) and $E_k^2 = p^2c^2 + m_k^2c^4$, we deduce that

$$m_e^2 + m_\mu^2 + m_\tau^2 = m_1^2 + m_2^2 + m_3^2,$$

which leads to

$$m_e^2 + m_\mu^2 + m_\tau^2 = \Delta_{32} + 2\Delta_{21} + 3m_1^2, \tag{6.3.33}$$

where Δ_{32} and Δ_{21} are as in (6.3.29).

If neutrinos have masses, then only the mass square differences Δ_{ij} in (6.3.29) can be measured by current experimental methods. Hence, the only mass information of ν_e, ν_μ, ν_τ is given by the relation (6.3.33).

6.3.4 MSW effect

In 1978, L. Wolfenstein (Wolfenstein, 1978) first noted that as neutrinos pass through matter there are additional effects due to elastic scattering

$$\nu_e + e \longrightarrow \nu_e + e.$$

This phenomenon was also observed and expanded by S. Mikheyev and A. Smirnov (Mikheev and Smirnov, 1986), and is now called the MSW effect.

The MSW effect can be reflected in the neutrino oscillation model. We recall the oscillation model without MSW effect expressed as

$$\begin{aligned} &\nu_k = \varphi_k(x)e^{-i\lambda_k t/\hbar}, \\ &[-i\hbar c(\vec{\alpha}\cdot\nabla) + mc^2\alpha_0]\varphi_k = \lambda_k\varphi_k \quad k = 1,2,3, \\ &\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = A \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}, \qquad A \text{ is as in (6.3.26).} \end{aligned} \tag{6.3.34}$$

To consider the MSW effect, we have to add weak interaction potentials in the Hamiltonian operator $\hat{H}$ for neutrinos ν_e, ν_μ, ν_τ. The weak potential energy is as given in (4.6.32):

$$V_\nu = g_s(\rho_\nu)\rho_s(\rho_e)N_e e^{-kr}\left[\frac{1}{r} - \frac{B}{\rho}(1+2kr)e^{-kr}\right], \tag{6.3.35}$$

where ρ_ν, ρ_e are the radii of neutrinos and electron, g_s is the weak charge, and N_w is the weak charge density. Namely the Hamiltonian with MSW effect for $(\nu_e, \nu_\mu, \nu_\tau)$ is

$$\mathscr{H}\begin{pmatrix}\nu_e\\ \nu_\mu\\ \nu_\tau\end{pmatrix} = \begin{pmatrix}\hat{H}+V_e & 0 & 0\\ 0 & \hat{H}+V_\mu & 0\\ 0 & 0 & \hat{H}+V_\tau\end{pmatrix}\begin{pmatrix}\nu_e\\ \nu_\mu\\ \nu_\tau\end{pmatrix}, \tag{6.3.36}$$

where $\hat{H} = -i\hbar c(\vec{\alpha}\cdot\nabla) + mc^2\alpha_0$, and V is as in (6.3.35).

The equations in (6.3.34) are also in the form

$$\begin{pmatrix}e^{i\lambda_1 t/\hbar} & & \\ & e^{i\lambda_2 t/\hbar} & \\ & & e^{i\lambda_3 t/\hbar}\end{pmatrix} A^\dagger \hat{H}\begin{pmatrix}\nu_e\\ \nu_\mu\\ \nu_\tau\end{pmatrix} = \begin{pmatrix}\lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3\end{pmatrix}\begin{pmatrix}\varphi_1\\ \varphi_2\\ \varphi_3\end{pmatrix}. \tag{6.3.37}$$

Replacing $\hat{H}$ by $\mathscr{H}$ in (6.3.37), we infer from (6.3.34) that

$$A^\dagger \hat{\mathscr{H}} A\begin{pmatrix}\nu_1\\ \nu_2\\ \nu_3\end{pmatrix} = \begin{pmatrix}\beta_1 & & \\ & \beta_2 & \\ & & \beta_3\end{pmatrix}\begin{pmatrix}\nu_1\\ \nu_2\\ \nu_3\end{pmatrix}, \tag{6.3.38}$$

where $\hat{\mathscr{H}}$ is as in (6.3.36).

The equation (6.3.38) is the neutrino oscillation model with the MSW effect, where the eigenvalues β_k and eigenstates ν_k $(1 \leqslant k \leqslant 3)$ are different from that of (6.3.34). In fact, the MSW effect is just the weak interaction effect.

6.3.5 Massless neutrino oscillation model

There are several serious problems in the massive neutrino oscillation model (6.3.34), which we briefly explain as follows.

1. *Parity problem.* It is known that all weak interaction decays and scatterings involving neutrinos violate the parity symmetry, discovered by Lee and Yang in 1956 and experimentally verified by C. Wu (Lee and Yang, 1956; Wu, Ambler, Hayward, Hoppes and Hudson, 1957). It means that the neutrinos are parity non-conserved. Hence it requires that under the space reflective transformation

$$x \longrightarrow -x, \tag{6.3.39}$$

the equations governing neutrinos should violate the reflective invariance. Based on Remarks 6.7 and 6.8, the Dirac equations are invariant under the reflection (6.3.39), while the Weyl equations are not invariant. Hence, the massive neutrino oscillation model is in conflict with the violation of parity symmetry.

2. *Handedness and speed of neutrinos.* Experiments showed that all neutrinos possess only the left-handed spin $J = -\dfrac{1}{2}$, and anti-neutrinos possess the right-handed spin $J =$

$\frac{1}{2}$. It implies that the velocity of free neutrinos must be at the speed of light, which is a contradiction with massive neutrino assumption.

In fact, the handedness is allowed only for massless particles. Otherwise, there exist two coordinate systems A and B satisfying

$$v_A < v_p < v_B,$$

where v_A, v_B and v_p are the velocities of A, B and the particle. When we look at the particle ν from A and B, the spins would be reversed. Therefore, all massive particles must have both left-handed and right-handed spins.

In addition, all experiments measuring neutrino velocity had found no violation to the speed of light.

3. *Infinite number of eigenvalues and eigenstates.* The neutrino oscillation theory faces the problem of the existing of infinite number of eigenvalues. In the massive model (6.3.34), the wave functions are the Dirac spinors

$$\varphi = (\varphi^1, \varphi^2, \varphi^3, \varphi^4)^{\mathrm{T}}.$$

For free neutrinos moving on a straight line, φ depends only on z. Thus the eigenvalue equations in (6.3.34) become

$$\begin{aligned} & -i\hbar c\sigma_3 \frac{d}{dz}\begin{pmatrix}\varphi^3\\ \varphi^4\end{pmatrix} + mc^2\begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix} = \lambda\begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix}, \\ & -i\hbar c\sigma_3 \frac{d}{dz}\begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix} - mc^2\begin{pmatrix}\varphi^3\\ \varphi^4\end{pmatrix} = \lambda\begin{pmatrix}\varphi^3\\ \varphi^4\end{pmatrix}, \end{aligned} \tag{6.3.40}$$

where

$$\sigma_3 = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{6.3.41}$$

The equations (6.3.40) possess infinite number of eigenvalues

$$\lambda = \sqrt{m^2c^4 + \frac{4\pi^2 n^2 \hbar^2 c^2}{l^2}}, \qquad \forall l > 0,\ n = 0, 1, 2, \cdots, \tag{6.3.42}$$

and each eigenvalue has two eigenstates

$$\begin{aligned} \varphi_1 &= \frac{e^{i2\pi nz/l}}{\sqrt{2}l^{3/2}}\begin{pmatrix}\sqrt{1+mc^2/\lambda}\\ 0\\ \sqrt{1-mc^2/\lambda}\\ 0\end{pmatrix}, \\ \varphi_2 &= \frac{e^{i2\pi nz/l}}{\sqrt{2}l^{3/2}}\begin{pmatrix}0\\ \sqrt{1+mc^2/\lambda}\\ 0\\ -\sqrt{1-mc^2/\lambda}\end{pmatrix}. \end{aligned} \tag{6.3.43}$$

The problem is that which eigenvalues and eigenstates in (6.3.42) and (6.3.43) are the ones in the neutrino oscillation model (6.3.34), and why only three of (6.3.42)-(6.3.43) stand for the flavors of neutrinos.

The Weyl equations (6.2.13) can replace the Dirac equations to describe the neutrino oscillation, which we call massless neutrino oscillation model, expressed as follows

$$\begin{aligned} &\nu_k = \varphi_k(x)e^{-i\lambda_k t/\hbar}, \\ &i\hbar c(\vec{\sigma}\cdot\nabla)\varphi_k = \lambda_k\varphi_k \quad \text{for } k=1,2,3, \\ &\begin{pmatrix}\nu_e\\ \nu_\mu\\ \nu_\tau\end{pmatrix} = A\begin{pmatrix}\nu_1\\ \nu_2\\ \nu_3\end{pmatrix}, \quad A \text{ is as in (6.3.26),} \end{aligned} \tag{6.3.44}$$

where ν_k $(1\leqslant k\leqslant 3)$ are the two-component Weyl spinors, and $\vec{\sigma}=(\sigma_1,\sigma_2,\sigma_3)$

Based on the massless model (6.3.44), both problems of parity and handedness of neutrinos have been resolved, and we can derive in the same conclusions as given in (6.3.24) and (6.3.25). In this case, the differences $\lambda_i-\lambda_j$ of eigenvalues in the transition probabilities such as (6.3.24) and (6.3.25) stand for the differences of frequencies:

$$\lambda_i-\lambda_j=\omega_i-\omega_j, \tag{6.3.45}$$

where ω_k $(1\leqslant k\leqslant 3)$ are the frequencies of ν_k.

However, the massless model also faces the problem of infinite number of eigenvalues as mentioned above. The eigenvalue equations in (6.3.44) for the straight line motion on the y-axis is written as

$$i\hbar c\alpha^2\frac{d}{dy}\begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix} = \lambda\begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix} \qquad \text{with} \quad \alpha^2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}. \tag{6.3.46}$$

The eigenvalues of (6.3.46) are

$$\lambda_k = k\hbar c, \qquad \forall k>0, \tag{6.3.47}$$

and each eigenvalue of (6.3.47) has two eigenstates

$$\begin{pmatrix}\varphi_1^1\\ \varphi_1^2\end{pmatrix} = \begin{pmatrix}\sin ky\\ -\cos ky\end{pmatrix}, \quad \begin{pmatrix}\varphi_2^1\\ \varphi_2^2\end{pmatrix} = \begin{pmatrix}\cos ky\\ \sin ky\end{pmatrix}. \tag{6.3.48}$$

The eigenvalues of (6.3.44) at x-axis and z-axis are all the same as in (6.3.47), and the eigenstates at the x and z axes are

$$\begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix} = \begin{pmatrix}e^{-ikx}\\ e^{-ikx}\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}\varphi^1\\ \varphi^2\end{pmatrix} = \begin{pmatrix}e^{-ikz}\\ 0\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}0\\ e^{ikz}\end{pmatrix}. \tag{6.3.49}$$

6.3.6 Neutrino non-oscillation mechanism

Although the massless neutrino oscillation model can solve the parity and the handedness problems appearing in the massive neutrino oscillation mechanism, the problem of infinite numbers of eigenvalues and eigenstates still exists in the model (6.3.44).

In fact, the weakton model first introduced in (Ma and Wang, 2015b) can provide an alternative explanation to the solar neutrino problem. Based on the weakton model, there exists a ν-mediator, whose weakton constituents are given by

$$\nu = \alpha_e \nu_e \overline{\nu}_e + \alpha_\mu \nu_\mu \overline{\nu}_\mu + \alpha_\tau \nu_\tau \overline{\nu}_\tau, \tag{6.3.50}$$

where $\alpha_e^2 + \alpha_\mu^2 + \alpha_\mu^2 = 1$. The values $\alpha_e^2, \alpha_\mu^2, \alpha_\tau^2$ represent the ratio of the neutrinos ν_e, ν_μ and ν_τ in our Universe

In view of (6.3.50), we see the reaction

$$\begin{aligned} &\nu_e + \overline{\nu}_e \longrightarrow \nu \quad (\nu_e \overline{\nu}_e), \\ &\nu_\mu + \overline{\nu}_\mu \longrightarrow \nu \quad (\nu_\mu \overline{\nu}_\mu), \\ &\nu_\tau + \overline{\nu}_\tau \longrightarrow \nu \quad (\nu_\tau \overline{\nu}_\tau), \end{aligned} \tag{6.3.51}$$

which are generated by the weak interaction attracting force, as demonstrated in the weak charge potentials

$$\Phi_i = g_w e^{-r/r_0} \left[\frac{1}{r} - \frac{B_i}{\rho_\nu} \left(1 + \frac{2r}{r_0} \right) e^{-r/r_0} \right] \qquad \text{for } 1 \leqslant i \leqslant 3, \tag{6.3.52}$$

where $r_0 = 10^{-16}$ cm, $B_1, B_2, B_3 > 0$ are the weak interaction constants for ν_e, ν_μ, ν_τ respectively, and ρ_ν is the neutrino radius.

The formula (6.3.52) defines attractive radii R_i for the neutrinos and antineutrinos of the same flavors. Namely, when ν_i and $\overline{\nu}_i$ are in the radius R_i, the reaction (6.3.51) may occur:

$$\nu_i + \overline{\nu}_i \longrightarrow \nu \ (\nu_i \overline{\nu}_i) \quad \text{if } \operatorname{dist}(\nu_i, \overline{\nu}_i) < R_i \qquad \text{for } 1 \leqslant i \leqslant 3, \tag{6.3.53}$$

where $\nu_1 = \nu_e, \nu_2 = \nu_\mu, \nu_3 = \nu_\tau$, and $\operatorname{dist}(\nu_i, \overline{\nu}_i)$ is the distance between ν_i and $\overline{\nu}_i$. The condition (6.3.53) implies that the transition probability Γ_i depends on R_i:

$$\Gamma_i = \Gamma_i(R_i) \qquad \text{for } 1 \leqslant i \leqslant 3. \tag{6.3.54}$$

The attracting radius R_i satisfies that

$$\frac{d}{dr} \Phi_i(R_i) = 0. \tag{6.3.55}$$

Thus we can give a non-oscillation mechanism of neutrinos to explain the solar neutrino problem. Namely, due to the β-decay, there are large amounts of electronic anti-neutrinos

$\overline{\nu}_e$ around the earth, which generate the reaction (6.3.51) with ν_e, leading to the discrepancy of the solar neutrino.

In addition, there are three axis eigenvalues of the Weyl equations given by (6.3.48) and (6.3.49). We believe that they are the three flavors of neutrinos ν_e, ν_μ, ν_τ. Namely, the following three wave functions

$$\begin{aligned} \psi_1 &= c_1 \begin{pmatrix} \sin ky \\ -\cos ky \end{pmatrix} + c_2 \begin{pmatrix} \cos ky \\ \sin ky \end{pmatrix}, \\ \psi_2 &= \begin{pmatrix} e^{-ikx} \\ e^{-ikx} \end{pmatrix}, \\ \psi_3 &= c_3 \begin{pmatrix} e^{-ikz} \\ 0 \end{pmatrix} + c_4 \begin{pmatrix} 0 \\ e^{ikz} \end{pmatrix} \end{aligned} \tag{6.3.56}$$

represent the three flavors of neutrinos. In fact, massive particles in field equations are distinguished by different masses, and flavors of neutrinos by different axis eigenstates.

Also, neutrinos have significant scattering effect of the weak interaction. In fact, they can have weak interactions with all subatomic particles. This can certainly cause deficits of neutrinos, and may be one of the main reasons for the loss of solar neutrinos.

6.4 Energy Levels of Subatomic Particles

6.4.1 Preliminaries

For subatomic particles we can establish energy levels, and this theory is based on two basic ingredients:

1) each subatomic particle consisting of two or three weaktons or quarks bound by weak or strong interaction, and

2) the weak and strong interaction potential formulas.

For convenience, we here briefly recall them.

1. *Elementary particles.* There are six pairs of elementary particles, called weaktons:

$$\begin{aligned} &w^*,\ w_1,\ w_2,\ \nu_e,\ \nu_\mu,\ \nu_\tau, \\ &\overline{w}^*,\ \overline{w}_1,\ \overline{w}_2,\ \overline{\nu}_e,\ \overline{\nu}_\mu,\ \overline{\nu}_\tau. \end{aligned}$$

Each weakton carries one unit of weak charge g_w, and both w^* and $\overline{w}^*$ are only weaktons carrying a strong charge g_s.

2. *Six classes of subatomic particles.* There are six types of subatomic particles: charged leptons, quarks, barons, mesons, intermediate bosons, mediators, whose members are listed as follows.

1) Charged leptons:

$$e^{\pm},\ \mu^{\pm},\ \tau^{\pm}.$$

2) Quarks:

$$\begin{aligned}&u,\ d,\ s,\ c,\ b,\ t,\\ &\overline{u},\ \overline{d},\ \overline{s},\ \overline{c},\ \overline{b},\ \overline{t}.\end{aligned}$$

3) Baryons:

$$p^{\pm},\ n,\ \Lambda,\ \Sigma^{\pm},\ \Sigma^{0},\ \Delta^{++},\ \Delta^{\pm},\ \Delta^{0},\ \Xi^{\pm},\ \Xi^{0},\ \text{ etc.}$$

4) Mesons:

$$\pi^{\pm},\ \pi^{0},\ K^{\pm},\ K^{0},\ \eta,\ \rho^{\pm},\ \rho^{0},\ K^{*\pm},\ K^{*0},\ \text{ etc.}$$

5) Intermediate bosons:

$$W^{\pm},\ Z^{0},\ H^{\pm},\ H^{0}.$$

6) Mediators:

$$\gamma,\ \gamma_0,\ g^k,\ g^k_0,\ \nu.$$

3. *Constituents of subatomic particles.*

1) Weakton constituents of charged leptons:

$$\begin{aligned}&e^- = \nu_e w_1 w_2, && \mu^- = \nu_\mu w_1 w_2, && \tau^- = \nu_\tau w_1 w_2.\\ &e^+ = \overline{\nu}_e \overline{w}_1 \overline{w}_2, && \overline{\mu}^+ = \overline{\nu}_\mu \overline{w}_1 \overline{w}_2, && \tau^+ = \overline{\nu}_\tau \overline{w}_1 \overline{w}_2.\end{aligned}$$

2) Weakton constituents of quarks:

$$\begin{aligned}&u = w^* w_1 \overline{w}_1, && c = w^* w_2 \overline{w}_2, && t = w^* w_2 \overline{w}_2,\\ &d = w^* w_1 w_2, && s = w^* w_1 w_2, && b = w^* w_1 w_2.\end{aligned}$$

3) Quark constituents of baryons:

$$\text{Baron} = qqq.$$

4) Quark constituents of mesons:

$$\text{Meson} = q\overline{q}.$$

5) Weakton constituents of intermediate bosons:

$$\begin{aligned}&W^+ = \overline{w}_1 \overline{w}_2(\uparrow\uparrow,\downarrow\downarrow), && W^- = w_1 w_2(\uparrow\uparrow,\downarrow\downarrow), && Z^0 = \alpha_1 w_1 \overline{w}_1 + \alpha_2 w_2 \overline{w}_2(\uparrow\uparrow,\downarrow\downarrow),\\ &H^+ = \overline{w}_1 \overline{w}_2(\uparrow\downarrow,\downarrow\uparrow), && H^- = w_1 w_2(\uparrow\downarrow,\downarrow\uparrow), && H^0 = \alpha_1 w_1 \overline{w}_1 + \alpha_2 w_2 \overline{w}_2(\uparrow\downarrow,\downarrow\uparrow).\end{aligned}$$

6) Weakton constituents of mediators:

$$\gamma = \alpha_1 w_1 \overline{w}_1 - \alpha_2 w_2 \overline{w}_2 (\uparrow\uparrow, \downarrow\downarrow), \qquad g^k = w^* \overline{w}^* (\uparrow\uparrow, \downarrow\downarrow),$$
$$\gamma_0 = \alpha_1 w_1 \overline{w}_1 - \alpha_2 w_2 \overline{w}_2 (\uparrow\downarrow, \downarrow\uparrow), \qquad g_0^k = w^* \overline{w}^* (\uparrow\downarrow, \downarrow\uparrow),$$
$$\nu = \alpha_e \nu_e \overline{\nu}_e + \alpha_\mu \nu_\mu \overline{\nu}_\mu + \alpha_\tau \nu_\tau \overline{\nu}_\tau .$$

4. *Bound forces of subatomic particles.* The main forces holding weaktons and quarks to form subatomic particles are the weak and strong interactions, and their force sources are from the interaction charges, i.e.

$$\text{weak charge } g_w, \quad \text{strong charge } g_s.$$

The 4-dimensional interaction potentials are

$$\begin{aligned} &\text{weak potential} \quad W_\mu = \omega_a W_\mu^a = (W_0, W_1, W_2, W_3), \\ &\text{strong potential} \quad S_\mu = \rho_k S_\mu^k = (S_0, S_1, S_2, S_3), \end{aligned} \tag{6.4.1}$$

The acting forces are

$$\begin{aligned} &\text{weak force} && = -g_w \nabla W_0, \\ &\text{strong force} && = -g_s \nabla S_0, \\ &\text{weak magnetism} && = -g_w \,\text{curl}\, \vec{W}, \\ &\text{strong magnetism} && = -g_s \text{curl} \vec{S}, \end{aligned} \tag{6.4.2}$$

where $\vec{W} = (W_1, W_2, W_3), \vec{S} = (S_1, S_2, S_3)$.

Since each weakton carries one weak charge g_w, and $w^*, \overline{w}^*$ carry one strong charge g_s, by the constituents of subatomic particles, the main bound energy for various particles takes

charged leptons and quarks:	weak interaction,	
hadrons (barons and mesons) :	strong interaction,	
intermediate bosons:	weak interaction,	(6.4.3)
vector and scalar gluons:	weak and strong interactions,	
other mediators:	weak interaction.	

Remark 6.17 We need to explain that although each quark has three weak charges, due to the weak force range $r \leqslant 10^{-16}$cm and the distance $r > 10^{-16}$ cm between the quarks in a hadron, the main bound force of hadrons is strong interaction, and the weak forces between the quarks an be ignored. □

5. *Weak interaction potentials.* In Section 4.6.2, we deduced the weak interaction potential (4.6.17). For convenience, here we again write this formulas as

$$\begin{aligned} &W_0 = N g_w(\rho) e^{-r/r_0} \left[\frac{1}{r} - \frac{B}{\rho} \left(1 + \frac{2r}{r_0} \right) e^{-r/r_0} \right], \\ &g_w(\rho) = \left(\frac{\rho_w}{\rho} \right)^3 g_w, \end{aligned} \tag{6.4.4}$$

where W_0 is the weak charge potential of a composite particle with N weak charges g_w and with radius ρ, and ρ_w is the weakton radius, B the weak interaction constants of this particle, $r_0 = 10^{-16}$ cm.

Based on (6.4.4), the weak potential energy generated by two particles with N_1 and N_2 weak charges and with radii ρ_1, ρ_2 is expressed as follows

$$V_w = N_1 N_2 g_w(\rho_1) g_w(\rho_2) e^{-r/r_0} \left[\frac{1}{r} - \frac{B_{12}}{\rho_{12}} \left(1 + \frac{2r}{r_0} \right) e^{-r/r_0} \right], \tag{6.4.5}$$

where $g_w(\rho_1)$ and $g_w(\rho_2)$ are as in (6.4.4), and B_{12}/ρ_{12} depends on the types of the two particles.

The basic weak charge g_w satisfies the relation (4.6.37), i.e.

$$g_w^2 = 5 \times 10^{-3} \left(\frac{\rho_n}{\rho_w} \right)^6 \hbar c. \tag{6.4.6}$$

6. *Strong interaction potentials.* The strong interaction potential given by (4.5.39), reads as

$$\begin{aligned} S_0 &= g_s(\rho) \left[\frac{1}{r} - \frac{A}{\rho} \left(1 + \frac{r}{R} \right) e^{-r/R} \right], \\ g_s(\rho) &= \left(\frac{\rho_w}{\rho} \right)^3 g_s, \end{aligned} \tag{6.4.7}$$

where R is as

$$R = \begin{cases} 10^{-16} \text{ cm} & \text{for } w^* \text{ and quarks}, \\ 10^{-13} \text{ cm} & \text{for hadrons}. \end{cases}$$

By (6.4.7), the strong potential energy for two particles with N_1, N_2 strong charges and with radii ρ_1, ρ_2 is given by

$$V_s = N_1 N_2 g_s(\rho_1) g_s(\rho_2) \left[\frac{1}{r} - \frac{A_{12}}{\rho_{12}} \left(1 + \frac{r}{R} \right) e^{-r/R} \right], \tag{6.4.8}$$

where $g_s(\rho_1)$ and $g_s(\rho_2)$ are as in (6.4.7), and A_{12}/ρ_{12} depends on the types of the two particles.

The basic strong charge g_s satisfies the relation (4.5.66), i.e.

$$g_s^2 = 2 \times 10^{-2} \left(\frac{\rho_n}{\rho_w} \right)^6 g^2 \quad (g^2 \simeq 1\hbar c). \tag{6.4.9}$$

6.4.2 Spectral equations of bound states

In the last subsection we see that the subatomic particles have six classes, in where the mediators are massless and others are massive. By the mass generation mechanism given in Subsection 5.3.2, the weaktons in massive subatomic particles possess masses, however the weaktons in mediators are massless. The spectral equations for both massive and massless bound states are very different. in the following we shall discuss them respectively.

Massive bound states

The subatomic particles consist of two or three fermions, their wave functions are the Dirac spinors

$$\Psi^1,\cdots,\Psi^N, \qquad N=2 \text{ or } 3. \tag{6.4.10}$$

As $N=2$ or 3, for each particle its bound energy can be approximatively regarded as the superposition of the remaining $N-1$ particles. Thus the bound potential for each fermion takes the form

$$gA_\mu=\begin{cases}(N-1)g_wW_\mu & \text{for weak interaction,}\\ (N-1)g_sS_\mu & \text{for strong interaction,}\\ (N-1)(g_wW_\mu+g_sS_\mu) & \text{for weak and strong interactions}\end{cases} \tag{6.4.11}$$

where W_μ and S_μ are as in (6.4.1).

Let the masses of the N particles are

$$m=\begin{pmatrix} m_1 & & 0\\ & \ddots & \\ 0 & & m_N\end{pmatrix}.$$

Then the N wave functions of (6.4.10) satisfy the Dirac equation

$$(i\hbar c\gamma^\mu D_\mu - c^2m)\Psi=0, \tag{6.4.12}$$

where $\Psi=(\Psi^1,\cdots,\Psi^N)$, and

$$D_\mu=\partial_\mu+i\frac{g}{\hbar c}A_\mu, \qquad gA_\mu \text{ as in (6.4.11).} \tag{6.4.13}$$

It is known that each wave function is a 4-components spinor

$$\Psi^k=(\Psi^k_1,\Psi^k_2,\Psi^k_3,\Psi^k_4), \qquad 1\leqslant k\leqslant N.$$

Therefore, the equation (6.4.12) takes the equivalent form

$$\begin{aligned}(i\hbar\tfrac{\partial}{\partial t}-gA_0-c^2m_k)\begin{pmatrix}\Psi^k_1\\ \Psi^k_2\end{pmatrix}&=-i\hbar c(\vec{\sigma}\cdot\vec{D})\begin{pmatrix}\Psi^k_3\\ \Psi^k_4\end{pmatrix},\\ (i\hbar\tfrac{\partial}{\partial t}-gA_0+c^2m_k)\begin{pmatrix}\Psi^k_3\\ \Psi^k_4\end{pmatrix}&=-i\hbar c(\vec{\sigma}\cdot\vec{D})\begin{pmatrix}\Psi^k_1\\ \Psi^k_2\end{pmatrix},\end{aligned} \tag{6.4.14}$$

where $\vec{\sigma}=(\sigma^1,\sigma^2,\sigma^3)$ is the Pauli matrix operator, $\vec{D}$ is as in (6.4.13).

We now derive spectral equations for massive bound states from (6.4.14). Let the solutions of (6.4.14) be in the form

$$\Psi^k = e^{-i(\lambda+m_k c^2)t/\hbar}\psi^k.$$

Then equations (6.4.14) become

$$(\lambda - gA_0)\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix} = -ic\hbar(\vec{\sigma}\cdot\vec{D})\begin{pmatrix}\psi_3^k\\ \psi_4^k\end{pmatrix}, \tag{6.4.15}$$

$$(\lambda - gA_0 + 2m^k c^2)\begin{pmatrix}\psi_3^k\\ \psi_4^k\end{pmatrix} = -ic\hbar(\vec{\sigma}\cdot\vec{D})\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}, \tag{6.4.16}$$

for $1 \leqslant k \leqslant N$. The equation (6.4.16) can be rewritten as

$$\begin{pmatrix}\psi_3^k\\ \psi_4^k\end{pmatrix} = \frac{-i\hbar}{2m_k c}\left(1 + \frac{\lambda - gA_0}{2mc^2}\right)^{-1}(\vec{\sigma}\cdot\vec{D})\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}. \tag{6.4.17}$$

In physics, λ is the energy, and $\lambda - gA_0$ is the kinetic energy

$$\lambda - gA_0 = \frac{1}{2}m_k v^2.$$

For massive particles, $v^2/c^2 \simeq 0$. Hence, (6.4.17) can be approximatively expressed as

$$\begin{pmatrix}\psi_3^k\\ \psi_4^k\end{pmatrix} = \frac{-i\hbar}{2m_k c}(\vec{\sigma}\cdot\vec{D})\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}.$$

Inserting this equation into (6.4.15), we deduce that

$$(\lambda - gA_0)\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix} = -\frac{\hbar^2}{2m_k}(\vec{\sigma}\cdot\vec{D})^2\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}. \tag{6.4.18}$$

Now, we need to give the expression of $(\vec{\sigma}\cdot\vec{D})^2$. To this end, note that the Pauli matrices satisfy

$$(\sigma^k)^2 = 1, \quad \sigma^k\sigma^j = -\sigma^j\sigma^k = i\varepsilon_l^{jk}\sigma^l.$$

Here ε_l^{jk} is the arrange symbal:

$$\varepsilon_i^{jk} = \begin{cases} 1 & \text{as } (jkl) \text{ the even arrange,}\\ -1 & \text{as } (jkl) \text{ the odd arrange,}\\ 0 & \text{otherwise.}\end{cases}$$

Hence we obtain

$$(\vec{\sigma}\cdot\vec{D})^2 = \left(\sum_{k=1}^{3}\sigma^k D_k\right)^2 = D^2 + i\vec{\sigma}\cdot(\vec{D}\times\vec{D}). \tag{6.4.19}$$

with $\vec{D}=\nabla+i\frac{g}{\hbar c}\vec{A}$, we derive that

$$\vec{D}\times\vec{D}=i\frac{g}{\hbar c}\left[\nabla\times\vec{A}+\vec{A}\times\nabla\right].$$

Note that as an operator we have

$$\nabla\times\vec{A}=\text{ curl}\,\vec{A}-\vec{A}\times\nabla.$$

Hence we get

$$\vec{D}\times\vec{D}=i\frac{g}{\hbar c}\text{ curl}\vec{A}.$$

Thus, (6.4.19) is written as

$$(\vec{\sigma}\cdot\vec{D})^2=D^2-\frac{g}{\hbar c}\vec{\sigma}\cdot\text{curl}\vec{A}. \tag{6.4.20}$$

By (6.4.20), the spectral equation (6.4.18) is in the form

$$\left[-\frac{\hbar^2}{2m_k}D^2+gA_0\right]\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}+\vec{\mu}_k\cdot\text{curl}\vec{A}\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}=\lambda\begin{pmatrix}\psi_1^k\\ \psi_2^k\end{pmatrix}, \tag{6.4.21}$$

where $D=(D_1,D_2,D_3),(A_0,A_1,A_2,A_3)$ is as in (6.4.11), and

$$\vec{\mu}_k=\frac{\hbar g}{2m_k}\vec{\sigma},\quad D_k=\partial_k+i\frac{g}{\hbar c}A_k\quad (1\leqslant k\leqslant 3). \tag{6.4.22}$$

Since the fermions are bound in the interior Ω of subatomic particle, ψ^k $(1\leqslant k\leqslant N)$ are zero outside Ω. Therefore the equation (6.4.21) are supplemented with the Dirichlet boundary conditions:

$$(\psi_1^k,\psi_2^k)|_{\partial\Omega}=0\quad (1\leqslant k\leqslant 3), \tag{6.4.23}$$

where $\Omega\subset\mathbb{R}^3$ is a bounded domain.

The boundary value problem (6.4.21)-(6.4.23) is the model for the energy level theory of massive subatomic particles.

Massless bound states

In order to obtain the spectral equations for massless subatomic particle, we have to derive their wave equations, which are based on the basic quantum mechanics principle: the Postulate 5.5.

We first recall the Weyl equation

$$\frac{\partial\psi}{\partial t}=c(\vec{\sigma}\cdot\nabla)\psi, \tag{6.4.24}$$

which describes massless and free fermions. The Weyl equation (6.4.24) is derived from the de Broglie relation

$$E=cp\quad (\text{ see (6.2.12)}) \tag{6.4.25}$$

with

$$\hat{E} = i\hbar\frac{\partial}{\partial t}, \quad \hat{p} = i\hbar(\vec{\sigma}\cdot\nabla). \tag{6.4.26}$$

As consider the massless bound states in weak and strong interactions, the Hermitian operators in (6.4.26) are replaced by

$$\hat{E} = i\hbar\frac{\partial}{\partial t} - gA_0, \quad \hat{p} = i\hbar(\vec{\sigma}\cdot\vec{D}), \tag{6.4.27}$$

where $\vec{D} = (D_1, D_2, D_3)$ is as in (6.4.22).

In Section 6.4.1, we knew that the mediators such as photons and gluons consist of two massless weaktons, which are bound in a small ball B_r by the weak and strong interactions. Hence the Weyl spinor ψ of each weakton is restricted in a small ball B_r, i.e.

$$\psi = 0, \qquad \forall x \notin B_r,$$

which implies the boundary condition

$$\psi|_{\partial B_r} = 0. \tag{6.4.28}$$

However, in mathematics the boundary problem for the Weyl equations generated by (6.4.25) and (6.4.27) given by

$$\begin{aligned} & i\hbar\tfrac{\partial\psi}{\partial t} = i\hbar c(\vec{\sigma}\cdot\vec{D})\psi + gA_0\psi, \\ & \psi|_{\partial B_r} = 0, \end{aligned} \tag{6.4.29}$$

is in general not well-posed, i.e (6.4.29) has no solution for a given initial value $\psi(0) = \psi_0$ in general. Hence, for a massless fermion system with the boundary condition (6.4.28), we have to consider the relation

$$pE = cp^2, \tag{6.4.30}$$

which is of first order in time t. It is known that the operator $\hat{p}\hat{E}$ is Hermitian if and only if

$$\hat{p}\hat{E} = \hat{E}\hat{p}.$$

Note that $\hat{E} = i\hbar\partial/\partial t - gA_0$, and in general

$$\hat{p}A_0 \neq A_0\hat{p}.$$

Hence, in order to ensure $\hat{p}\hat{E}$ being Hermitian, we replace $\hat{p}A_0$ by $\frac{1}{2}(\hat{p}A_0 + A_0\hat{p})$, i.e. take

$$\hat{p}\hat{E} = i\hbar\hat{p}\frac{\partial}{\partial t} - \frac{g}{2}(\hat{p}A_0 + A_0\hat{p}), \quad \hat{p} = i\hbar c(\vec{\sigma}\cdot\vec{D}). \tag{6.4.31}$$

Then by Postulate 6.5, from (6.4.30) and (6.4.31) we derive the boundary problem of massless system in the following form

$$\begin{aligned} & (\vec{\sigma}\cdot\vec{D})\frac{\partial\psi}{\partial t} = c(\vec{\sigma}\cdot\vec{D})^2\psi - \frac{ig}{2\hbar}\{(\vec{\sigma}\cdot\vec{D}), A_0\}\psi, \\ & \psi|_{\partial\Omega} = 0, \end{aligned} \tag{6.4.32}$$

where $\Omega \subset R^n$ is a bounded domain, and $\{A,B\} = AB + BA$ is the anti-commutator.

Now, we derive the spectral equations from (6.4.32) for the massless bound states. Let the solutions ψ of (6.4.32) be in the form

$$\psi = e^{-i\lambda t/\hbar}\varphi, \quad \varphi = \begin{pmatrix} \varphi^1 \\ \varphi^2 \end{pmatrix},$$

Then equation (6.4.32) are reduced to the eigenvalue problem

$$-\hbar c(\vec{\sigma}\cdot\vec{D})^2\varphi + \frac{ig}{2}\{(\vec{\sigma}\cdot\vec{D}), A_0\}\varphi = i\lambda(\vec{\sigma}\cdot\vec{D})\varphi,$$

and by (6.4.20) which can be rewritten as

$$\begin{aligned}
&[-\hbar c D^2 + g\vec{\sigma}\cdot\text{curl}\vec{A}]\begin{pmatrix} \varphi^1 \\ \varphi^2 \end{pmatrix} + \frac{ig}{2}\{(\vec{\sigma}\cdot\vec{D}), A_0\}\begin{pmatrix} \varphi^1 \\ \varphi^2 \end{pmatrix} \\
=& i\lambda(\vec{\sigma}\cdot\vec{D})\begin{pmatrix} \varphi^1 \\ \varphi^2 \end{pmatrix}, \qquad (6.4.33)\\
&(\varphi^1, \varphi^2)|_{\partial\Omega} = 0,
\end{aligned}$$

where $\{(\vec{\sigma}\cdot\vec{D}), A_0\}$ is the anti-commutator defined by

$$\{(\vec{\sigma}\cdot\vec{D}), A_0\} = (\vec{\sigma}\cdot\vec{D})A_0 + A_0(\vec{\sigma}\cdot\vec{D}). \qquad (6.4.34)$$

The eigenvalue equations (6.4.33)-(6.4.34) are taken as the model for the energy levels of massless bound states. The mathematical theory (Theorem 2.42) established in Subsection 3.6.5 laid a solid foundation for the energy level theory provided by (6.4.33).

Remark 6.18 In the equations (6.4.21)-(6.4.22) and (6.4.33) for bound states, we see that there are terms

$$\vec{\mu}\cdot\text{curl}\vec{A} \quad \text{for massive particle systems,} \qquad (6.4.35)$$

$$g\vec{\sigma}\cdot\text{curl}\vec{A} \quad \text{for massless particle systems.} \qquad (6.4.36)$$

In (6.4.35), the physical quantity $\vec{\mu} = \hbar g\vec{\sigma}/2m$ represents magnetic moment, and the term in (6.4.36) is magnetic force generated by the spin coupled with ether weak or strong interaction. In other words, in the same spirit as the electric charge e producing magnetism, the weak and strong charges g_w, g_s can also produce similar effects, which we also call magnetism.

Indeed, all three interactions: electromagnetic, weak, strong interactions, enjoy a common property that moving charges yield magnetism, mainly due to the fact that they are all gauge fields.

6.4.3 Charged leptons and quarks

According to structure and interaction types, we shall discuss the energy levels for three groups of particles. charged leptons and quarks, hadrons, mediators. In this subsection we only consider the case of charged leptons and quarks.

In Section 6.4.1 we see that charged leptons and quarks are made up of three weaktons, with masses caused by the deceleration of the constituent weaktons. Let the masses of the constituent weaktons be m_1, m_2, m_3, and the wave functions of these weaktons be given by

$$\psi^k = \begin{pmatrix} \psi_1^k \\ \psi_2^k \end{pmatrix} \qquad \text{for } k = 1,2,3.$$

Here ψ_1^k and ψ_2^k represent the left-hand and right-hand states. The bound states are due to the weak interaction, and the potential in (6.4.11) takes the form

$$gA_\mu = 2g_w W_\mu = (2g_w W_0, 2g_w \vec{W}).$$

By (6.4.21)-(6.4.23), the spectral equations for charged leptons and quarks are as follows

$$\begin{aligned} & -\frac{\hbar^2}{2m_j}\left(\nabla + \mathrm{i}\frac{2g_w}{\hbar c}\vec{W}\right)^2 \psi^j + 2(g_w w_0 + \vec{\mu}_j \cdot \mathrm{curl}\vec{W})\psi^j \\ & = \lambda \psi^j \quad \text{in } \rho_w < |x| < \rho, \quad 1 \leqslant j \leqslant 3 \\ & \psi = (\psi^1, \psi^2, \psi^3) = 0 \quad \text{at } |x| = \rho_w, \rho, \end{aligned} \tag{6.4.37}$$

where ρ_w is the weakton radius, ρ the attracting radius of weak interaction, W_0 is given by (6.4.4) with $N = 1, g_w(\rho) = g_w$, and

$$\vec{\mu}_j = \frac{\hbar g_w}{2m_j}\vec{\sigma} \ \text{ is the weak magnetic moment.}$$

We are in position now to derive a few results on the energy levels for charged leptons and quarks based on (6.4.37).

1. *Bound states and energy levels.* We know that the negative eigenvalues and eigenfunctions of (6.4.37) correspond to the bound energy and bound states. Let

$$-\infty < \lambda_1 \leqslant \cdots \leqslant \lambda_N < 0$$

be all negative eigenvalues of (6.4.37) with eigenfunctions

$$\psi_1, \cdots, \psi_N \qquad \text{for } \psi_k = (\psi_k^1, \psi_k^2, \psi_k^3)^{\mathrm{T}}.$$

Each bound state ψ_k satisfies

$$\int_\Omega |\psi_k^j|^2 dx = 1 \qquad \text{for } 1 \leqslant j \leqslant 3,\ 1 \leqslant k \leqslant N, \tag{6.4.38}$$

where $\Omega = \{x \in R^3 | \ \rho_w < |x| < \rho\}$. Then, by (6.4.37) and (6.4.38) we get

$$\lambda_k = \frac{\hbar}{2m_j}\int_\Omega \left|\left(\nabla + i\frac{g_w}{\hbar c}\vec{W}\right)\psi_k^j\right|^2 dx \tag{6.4.39}$$
$$+2\int_\Omega \vec{\mu}_j \cdot \text{curl}\vec{W}|\psi_k^j|^2 dx + 2\int_\Omega g_w W_0 |\psi_k^j|^2 dx.$$

λ_k is independent of j $(1 \leqslant j \leqslant 3)$, i.e. each weakton has the same bound energy λ_k but in different bound state ψ_k^j.

In the right-hand side of (6.4.39), the first term stands for the kinetic energy, the second term for the weak magnetic energy, and the third term for the weak potential energy, the potential energy in (6.4.39) is negative. Hence, the bound energy can be written as

$$\lambda_k = \text{ kinetic energy + magnetic energy + potential energy.}$$

In addition, the energy distributions of charged leptons and quarks are discrete and finite:

$$0 < E_1 < \cdots < E_{N_0} \quad (N_0 \leqslant N), \tag{6.4.40}$$

and N is the number of negative eigenvalues. Each energy level E_k can be expressed as

$$E_k = 3(E_0 + \lambda_k), \quad E_0 = g_w^2/\rho_w \ \text{ is the intrinsic energy.}$$

2. *Masses.* At an energy level E_k of (6.4.40), the mass M_k of a lepton or a quark satisfies the relation

$$M_k = \sum_{j=1}^3 m_j + E_k/c^2 = \sum_{j=1}^3 m_j + \frac{3g_w^2}{\rho_w c^2} + \frac{3\lambda_k}{c^2}.$$

3. *Parameters of electrons.* In all charged leptons and quarks, only the electrons are long life-time and observable. Hence the physical parameters of electrons are important. By the spectral equation (6.4.37) we can derive some information for electronic parameters.

To this end, we recall the weak interaction potential for the weaktons, which is written as

$$W_0 = g_w\left[\frac{1}{r} - \frac{B_w}{\rho_w}\left(1 + \frac{2r}{r_0}\right)e^{-r/r_0}\right]e^{-r/r_0}, \tag{6.4.41}$$

where B_w is the constant for weaktons.

Assume that the masses of three weaktons are the same. We ignore the magnetism, i.e. let $\vec{W} = 0$. Then (6.4.37) is reduced in the form

$$\begin{cases} -\dfrac{\hbar^2}{2m}\Delta\psi + 2g_w W_0\psi = \lambda\psi & \text{for } \rho_w < |x| < \rho, \\ \psi = 0, & \text{for } |x| = \rho_w, \rho. \end{cases} \tag{6.4.42}$$

We shall apply (6.4.41) and (6.4.42) to derive some basic parameters and their relations for the electrons.

Let λ_e and ψ_e be the spectrum and bound state of an electron, which satisfy (6.4.42). It follows from (6.4.41) and (6.4.42) that

$$\lambda_e = \frac{1}{2}mv^2 + 2g_w^2\left(\frac{1}{r_e} - \frac{\kappa_e}{\rho_w}\right) \tag{6.4.43}$$

where $\frac{1}{2}mv^2$ is the kinetic energy of each weakton in electron r_e the radius of the naked electron, κ_e is the bound parameter of electron, and they are expressed as

$$\begin{aligned} r_e &= \int_{B_\rho} \frac{1}{r} e^{-r/r_0}|\psi_e|^2 dx, \\ \kappa_e &= B_w \int_{B_\rho} \left(1 + \frac{2r}{r_0}\right) e^{-2r/r_0}|\psi_e|^2 dx, \\ \left(\frac{m_w v}{\hbar}\right)^2 &= \int_{B_\rho} |\nabla \psi_e|^2 dx. \end{aligned}$$

These three parameters are related with the energy levels of an electron, i.e. with the λ_k and ψ_k. However, the most important case is the lowest energy level state. We shall use the spherical coordinate to disscuss the first eigenvalue λ_1 of (6.4.42). Let the first eigenfunction ψ_e be in the form

$$\psi_e = \varphi_0(r)Y(\theta, \varphi).$$

Then φ_0 and Y satisfy

$$\begin{aligned} &-\frac{\hbar^2}{2m}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)\varphi_0 + 2g_s W_0 \varphi_0 + \tfrac{\beta_k}{r^2}\varphi_0 = \lambda_e \varphi_0, \\ &\varphi_0(\rho_w) = \varphi_0(\rho) = 0, \end{aligned} \tag{6.4.44}$$

and

$$\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\right] Y_k = \beta_k Y_k, \tag{6.4.45}$$

where $\beta_k = k(k+1),\ k = 0, 1, \cdots$.

Because λ_e is the minimal eigenvalue, it implies that $\beta_k = \beta_0 = 0$ in (6.4.44). The eigenfunction Y_0 of (6.4.45) is given by

$$Y_0 = \frac{1}{\sqrt{4\pi}}.$$

Thus ψ_e is as follows

$$\psi_e = \frac{1}{\sqrt{4\pi}}\varphi_0(r),$$

and λ_e, φ_0 are the first eigenvalue and eigenfucntion of the following equation

$$\begin{aligned} &-\frac{\hbar^2}{2m}\frac{1}{r^2}\frac{d}{dr}\left(r^2\tfrac{d}{dr}\right)\varphi_0 + 2g_w W_0 \varphi_0 = \lambda_1 \varphi_0 \\ &\varphi_0(\rho_w) = \varphi_0(\rho) = 0. \end{aligned} \tag{6.4.46}$$

In this case, the parameters in (6.4.43) are simplified as

$$\begin{aligned}
r_e &= \int_{\rho_w}^{\rho} r\varphi_0^2(r)e^{-r/r_0}dr, \\
\kappa_e &= B_w\int_{\rho_w}^{\rho} r^2\left(1+\frac{2r}{r_0}\right)e^{-2r/r_0}\varphi_0^2(r)dr, \qquad (6.4.47)\\
\left(\frac{mv}{\hbar}\right)^2 &= \int_{\rho_w}^{\rho} r^2\left(\frac{d\varphi_0}{dr}\right)^2 dr,
\end{aligned}$$

where φ_0 satisfies (6.4.46).

6.4.4 Baryons and mesons

Hadrons include baryons and mesons, their spectral equations are given respectively in the following.

1. *Spectral equations of baryons.* Baryons consist of three quarks: $B = qqq$, and each quark q consists of three w-weaktons

$$q = w^*ww \quad \text{and} \quad q = w^*w\overline{w}.$$

Hence, each quark possesses one strong charge g_s and three weak charges $3g_w$. It looks as if the bound energy of baryons is provided by both weak and strong interactions. However, since the weak interaction is short-ranged, i.e.

$$\text{range of weak force} \leqslant 10^{-16}\ \text{cm},$$

and the radii of baryons are

$$r > 10^{-16}\ \text{cm}.$$

Hence the main interaction to hold three quarks together is the strong force. Let m_1, m_2, m_3 be the masses of three quarks in a baryon, and $\psi^k = (\psi_1^k, \psi_2^k)^{\mathrm{T}} (1 \leqslant k \leqslant 3)$ be the wave functions. Then the spectral equations (6.4.21)-(6.4.23) for baryons are in the form

$$\begin{aligned}
&-\frac{\hbar^2}{2m_k}\left(\nabla + i\frac{2g_s}{\hbar c}\vec{s}\right)^2\psi^k + 2g_sS_0\psi^k + 2\vec{\mu}_k\cdot \mathrm{curl}\vec{S}\psi^k \\
&= \lambda\psi^k \quad \text{in } \rho_0 < |x| < \rho_1, \quad \text{for } 1 \leqslant k \leqslant 3, \qquad (6.4.48)\\
&\psi = (\psi^1, \psi^2, \psi^3) = 0 \quad \text{at } |x| = \rho_0, \rho_1,
\end{aligned}$$

where ρ_0 is the quark radius, ρ_1 the strong attracting radius, $S_\mu = (S_0, \vec{S})$ as in (6.4.1) is 4-dimensional strong potential, and $\vec{\mu}_k = \hbar g_s\vec{\sigma}/2m_k$ the strong magnetic moment. ψ^k satisfy the nomalization

$$\int_\Omega |\psi^k|^2 dx = 1, \quad \text{in } \Omega = \{x \in \mathbb{R}^3 |\ \rho_0 < |x| < \rho_1\}.$$

2. *Spectral equations of mesons.* Mesons consist of a quark and a antiquark: $M = q\overline{q}$. Hence, the bound energy of mesons is mainly provided by strong interaction potential $S_\mu = (S_0, \vec{S})$. Let m_1, m_2 be the masses of quark and antiquark, ψ^1 and ψ^2 are their wave functions. Then the spectral equations for mesons take the form

$$\begin{aligned}
&-\frac{\hbar^2}{2m_k}\left(\nabla + i\frac{g_s}{\hbar c}\vec{S}\right)^2 \psi^k + g_s S_0 \psi^k + \vec{\mu}_k \cdot \text{curl}\vec{S}\psi^k \\
&= \lambda \psi^k \quad \text{in } \Omega = \{x \in \mathbb{R}^3 |\ \rho_0 < |x| < \rho_1\}, \quad k = 1, 2, \\
&(\psi^1, \psi^2) = 0 \quad \text{on } \partial\Omega,
\end{aligned} \tag{6.4.49}$$

and ψ^k $(k = 1, 2)$ satisfy the normalization.

3. *Physical parameters of nucleons.* In hadrons, only nucleons (protons and neutrons) are long life-time. Therefore the parameters of nucleons are very important.

By (6.4.7), the strong interaction potential for quarks is taken as

$$S_0 = \left(\frac{\rho_w}{\rho_0}\right)^3 g_s \left[\frac{1}{r} - \frac{A_q}{\rho_0}\left(1 + \frac{r}{r_0}\right) e^{-r/r_0}\right], \tag{6.4.50}$$

where $r_0 = 10^{-16}$ cm, ρ_0 is the quark radius, A_q the constant of quarks.

Similar to the relation (6.4.43), from (6.4.48) and (6.4.50) we can obtain the parameter relation of nucleons as follows

$$\lambda_n = \frac{1}{2} m_q v^2 + 2\left(\frac{\rho_w}{\rho_0}\right)^6 g_s^2 \left(\frac{1}{r_n} - \frac{\kappa_n}{\rho_0}\right) \tag{6.4.51}$$

where λ_n is the bound energy of nucleons, m_q the quark masses, v the average velocity of quarks, r_n the nucleon radius, κ_n the bound parameter of nucleons, these parameters are expressed as

$$\begin{aligned}
&r_n = \int_{\rho_0}^{\rho_1} r\varphi_n^2(r)dr, \\
&\kappa_n = A_q \int_{\rho_0}^{\rho_1} r^2 \left(1 + \frac{r}{r_0} e^{-r/r_0} \varphi_n^2(r)\right) dr, \\
&\left(\frac{m_q v}{\hbar}\right)^2 = \int_{\rho_0}^{\rho_1} r^2 \left(\frac{d\varphi_n}{dr}\right)^2 dr,
\end{aligned} \tag{6.4.52}$$

where λ_n and φ_n are the first eigenvalue and eigenfucntion of the following equation

$$\begin{aligned}
&-\frac{\hbar^2}{2m_q}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)\varphi_n + 2g_s S_0 \varphi_n = \lambda \varphi_n, \\
&\varphi_n(\rho_0) = \varphi_n(\rho_1) = 0.
\end{aligned} \tag{6.4.53}$$

Remark 6.19 The two sets of parameters (6.4.47) and (6.4.52) can characterize the physical properties of naked electrons and naked nucleons respectively. For the real electrons and nucleons, we have to consider their mediator clouds.

6.4.5 Energy spectrum of mediators

Mediators are massless bosons, which consists of a weakton and its anti-particle. From the viewpoint of bound energy, gluons are generated by both weak and strong interactions, and others are bound only by weak interaction. Hence, the spectral equations of gluons are different from those the other mediators.

1. *Gluons.* The weakton constituent of gluons is given by

$$g = w^*\overline{w}^*. \tag{6.4.54}$$

Based onthe weakton model, w^* and $\overline{w}^*$ contain a weak charge g_w and a strong charge g_s. By (6.4.6) and (6.4.9),

$$\frac{g_w^2}{g_s^2} = 0.25,$$

i.e. g_w and g_s have the same order. Therefore, the interactions for the gluon are both weak and strong forces, i.e. the 4-dimensional potential $A_\mu = (A_0, \vec{A})$ is as

$$gA_0 = g_w W_0 + g_s S_0, \quad g\vec{A} = g_w \vec{W} + g_s \vec{S}. \tag{6.4.55}$$

In (6.4.54), we only need to consider the bound state for a single weakton. Then the spectral equation is provided by (6.4.33)-(6.4.34), and for (6.4.55) which is written as

$$\begin{aligned} & -\hbar c D^2 \psi + [g_w \vec{\sigma} \cdot \text{curl}\vec{W} + g_s \vec{\sigma} \cdot \text{curl}\vec{S}]\psi \\ & \qquad + \frac{i}{2}\left\{(\vec{\sigma} \cdot \vec{D}), g_w W_0 + g_s S_0\right\}\psi = i\lambda(\vec{\sigma} \cdot \vec{D})\psi \qquad \text{for } \rho_w < |x| < \rho_g, \\ & \psi = 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{for } |x| = \rho_w,\ \rho_g, \end{aligned} \tag{6.4.56}$$

where ρ_w and ρ_g are the radii of weaktons and gluons, and

$$\vec{D} = \nabla + \frac{i}{\hbar c} g\vec{A}, \quad g\vec{A} \text{ as in (6.4.55).}$$

2. *Photons and ν-mediators.* The other mediators such as photons and ν-mediator consist of a pair of weakton and anti-weakton:

$$\begin{aligned} \gamma &= \alpha_1 w_1 \overline{w}_1 + \alpha_2 w_2 \overline{w}_2, \\ \nu &= \alpha_e \nu_e \overline{\nu}_e + \alpha_\mu \nu_\mu \overline{\nu}_\mu + \alpha_\tau \nu_\tau \overline{\nu}_\tau. \end{aligned} \tag{6.4.57}$$

The weaktons in (6.4.57) only contain a weak charge g_w, hence the bound energy of γ and ν is given by the weak force, i.e.

$$gA_\mu = g_w W_\mu = (g_w W_0,\ g_w \vec{W}).$$

In this case, the spectral equations are in the form

$$\begin{aligned} -\hbar c D^2\psi \ + g_w\vec{\sigma}\cdot\text{curl}\vec{W}\psi + \frac{ig_w}{2}\{(\vec{\sigma}\cdot\vec{D}),W_0\}\psi & \\ = i\lambda(\vec{\sigma}\cdot\vec{D})\psi, \quad \text{for } \rho_w<|x|<\rho_m, & \\ \psi=0, \quad \text{on } |x|=\rho_w,\rho_m, & \end{aligned} \tag{6.4.58}$$

where ρ_m is the radius of the mediator γ or ν, and

$$\vec{D}=\nabla+\frac{i}{\hbar c}g_w\vec{W}.$$

3. *Energy levels of mediators.* By the spectral theory of the Weyl operator established in Subsection 2.6.5, the negative eigenvalues of (6.4.56) and (6.4.57) are finite, i.e.

$$-\infty<\lambda_1\leqslant\cdots\leqslant\lambda_N<0,$$

which stand for bound energy of mediators. It shows that the energy levels of each kind mediator are finite

$$0<E_1\leqslant\cdots\leqslant E_N. \tag{6.4.59}$$

Each energy level E_k can be expressed as

$$E_k=E_0+\lambda_k \quad (1\leqslant k\leqslant N), \tag{6.4.60}$$

where E_0 is the intrinsic energy of mediators.

It follows from (6.4.59) and (6.4.60) that the frequencies of mediators are finite and discrete

$$\omega_k=\frac{1}{\hbar}E_k \quad (1\leqslant k\leqslant N), \tag{6.4.61}$$

and the difference of two adjacent frequencies is

$$\Delta\omega_k=\omega_{k+1}-\omega_k=\frac{1}{\hbar}(\lambda_{k+1}-\lambda_k). \tag{6.4.62}$$

Remark 6.20 In the classical quantum mechanics, the frequencies of particles are continuously distributed. Here we derive from the energy spectrum theory that the frequencies are finite and discrete. In fact, by the estimate (3.6.51) of the number of negative eigenvalues, we can verify that the frequency difference (6.4.62) is too small to measure in the next subsection.

6.4.6 Discreteness of energy spectrum

Based on the spectral theory developed in Section 3.6, the energy levels of all subatomic particles are finite and discrete:

$$0<E_1<\cdots<E_N, \tag{6.4.63}$$

where the number N of energy levels depends on the particle type. Each subatomic particle lies in an energy state of E_k $(1 \leqslant k \leqslant N)$ in (6.4.63). In traditional conception, the energy distribution of all particles is infinite and continuous, i.e. a particle can lie in any state of energy E with $0 < E < \infty$. Hence the energy level theory established here arrive at a very different conclution:

Physical Conclusion 6.21 *The energy distribution of subatomic particles is finite and discrete.*

Usually, we cannot observe the discreteness of energy because the number N of energy levels is so large that the average difference of adjacent energy level:

$$\Delta E_k = E_{k+1} - E_k \simeq \frac{E_N - E_1}{N} \simeq 0,$$

is very small. In the following discussion we shall show this point.

To this end, we first give the estimates of the number N of energy levels for various types of subatomic particles.

1. *Energy level number of electrons and quarks.* To consider the approximatively computation of number N of energy levels, we always ignore the magnetic effects. In this case, the spectral equations for charged leptons and quarks are given by (6.4.42), and which are written as

$$\begin{aligned} & -\frac{\hbar^2}{2m_w}\Delta\psi + 2g_w W_0 \psi = \lambda\psi, && \text{in } \rho_w < |x| < \rho, \\ & \psi = 0, && \text{on } |x| = \rho_w, \rho, \end{aligned} \tag{6.4.64}$$

where m_w is the masses of weaktons in leptons and quarks, ρ is the attracting radius of weak interaction.

For the weak interaction potential W_0, we approximatively take

$$W_0 = -\frac{g_w B_w}{\rho_w}.$$

Then take the dimensional transformation

$$x \to \rho x \quad (\rho \ \text{ as in (6.4.64)}).$$

Note that the weakton radius ρ_w for smaller than ρ,

$$\rho_w \ll \rho.$$

Hence, the problem (6.4.64) can be approximatively expressed as

$$\begin{aligned} & -\Delta\psi - \frac{4m_w B_w \rho^2}{\hbar^2 \rho_w} g_w^2 \psi = \lambda\psi && \text{for } |x| < 1, \\ & \psi = 0 && \text{on } |x| = 1. \end{aligned} \tag{6.4.65}$$

It is clear that for the equation (6.4.65), the parameters r,α,θ as in (3.6.28) of Theorem 2.37 are as follows

$$r=1,\quad \alpha=0,\quad \theta=\frac{4m_wB_w\rho^2}{\hbar^2\rho_w}g_w^2$$

Thus the number N of the energy levels of charged leptons and quarks is approximatively given by

$$N=\left[\frac{4}{\lambda_1}\frac{B_w\rho^2}{\rho_w}\frac{m_wc}{\hbar}\frac{g_w^2}{\hbar c}\right]^{\frac{3}{2}}, \tag{6.4.66}$$

where λ_1 is the first eigenvalue of $-\Delta$ in unit ball $B_1\subset R^3$.

By the de Broglie relation,

$$\frac{m_wc}{\hbar}\sim\frac{1}{\lambda\!\!\!^-},$$

where $\lambda\!\!\!^-$ is the wave length of weaktons, i.e. $\lambda\!\!\!^-=nr_m$ $(n=1,2,\cdots),r_m$ the mediator radius. We take $\lambda\!\!\!^-=r_m\simeq\rho$, and

$$\frac{4}{\lambda_1}\simeq 1,\quad \frac{B_w\rho}{\rho_w}\simeq 10^2,\quad \frac{m_wc}{\hbar}\rho\simeq 1.$$

Then, by (6.4.6) the number N in (6.4.66) has the estimate

$$N\sim\left(\frac{\rho_n}{\rho_w}\right)^9\geqslant 10^{45}. \tag{6.4.67}$$

This number is very large because $(\rho_n/\rho_w)\geqslant 10^5$.

We remark that the estimate (6.4.67) only for the naked leptons and quarks. For the natural leptons and quarks, their energy level number $\widetilde{N}$ should take as

$$\widetilde{N}=N\cdot N_1,$$

where N is as in (6.4.67) and N_1 the energy level number of mediator cloud around the charged leptons and quarks.

2. *Energy level number of hadrons.* We only consider the case of baryons, and the case of mesons is similar. For the baryons, spectral equation (6.4.48) can be approximatively reduced in the form

$$\begin{aligned} &-\Delta\psi-\frac{4m_qA_q\rho_1^2}{\hbar^2\rho_q}g_s^2\psi=\lambda\psi, && \text{for } |x|<1,\\ &\psi=0, && \text{on } |x|=1. \end{aligned} \tag{6.4.68}$$

where m_q is the quark mass, ρ_q the quark radius, ρ_1 the strong attracting radius, and A_q the strong interaction constant of quarks.

By Theorem 3.38, from (6.4.68) we can get the estimates of energy level number N of baryons as follows

$$N=\left[\frac{4}{\lambda_1}\frac{\rho_1^2 A_q}{\rho_q}\frac{m_q c}{\hbar}\frac{g_s^2}{\hbar c}\right]^{\frac{3}{2}}, \tag{6.4.69}$$

where λ_1 is as in (6.4.66). By (6.4.69) and (6.4.9) we can get the same estimate as in (6.4.67).

3. *Energy level number of mediators.* Likewise, we only consider the energy level number of photons. In this case, the spectral equation (6.4.58) can be reduced as

$$\begin{aligned} &-\Delta\varphi=\mathrm{i}\left(\lambda+\frac{B_w\rho_r}{\hbar c\rho_w}g_w^2\right)(\vec{\sigma}\cdot\nabla)\varphi, && \text{for } |x|<1,\\ &\varphi=0 && \text{on } |x|=1, \end{aligned} \tag{6.4.70}$$

where ρ_γ is the photon radius.

By (6.4.70) we can see that the parameter K as in (3.6.51) takes

$$K=\frac{B_w\rho_\gamma}{\rho_w}\frac{g_w^2}{\hbar c}.$$

Thus, by (3.6.51) the energy level number N of photons is given by

$$N=\left(\frac{K}{\beta_1}\right)^3=\left[\frac{1}{\beta_1}\frac{B_w\rho_\gamma}{\rho_w}\frac{g_w^2}{\hbar c}\right]^3, \tag{6.4.71}$$

where β_1 is as in (3.6.51), and $\beta_1\sim o(1)$.

From the physical significance, ρ_γ is approximatively the weak attracting radius of weaktons, and $\rho_w/B_w=\overline{\rho}$ satisfying

$$F_w\begin{cases} >0 & \text{for } 0<r<\overline{\rho},\\ <0 & \text{for } \overline{\rho}<r<\rho_\gamma. \end{cases}$$

Hence, it is natural to think that

$$\frac{B_w\rho_\gamma}{\rho_w}=\frac{\rho_\gamma}{\overline{\rho}}=10^2\sim 10^4.$$

Thus, by (6.4.71) and (6.4.6) we get hat

$$N\sim\left(\frac{\rho_n}{\rho_w}\right)^{18}\geqslant 10^{90}. \tag{6.4.72}$$

Remark 6.22 In the estimates (6.4.66), (6.4.69) and (6.4.71), the energy level numbers N for various subatomic particles are counting the multiplicities of eigenvalues. However, the numbers N have the same order as the real energy level numbers. It is because that the multiple eigenvalues are unstable, under the perturbation of electromagnetism together with weak and strong magnetism, most multiple eigenvalues become the simple eigenvalues. □

4. *Energy level gradient of photons.* Each energy level E_k $(1 \leqslant k \leqslant N)$ of photons can be written as

$$E_k = E_0 + \lambda_k \quad (1 \leqslant k \leqslant N).$$

It is clear that the largest and smallest energy levels are given by

$$E_{\max} = E_0 + \lambda_N, \quad E_{\min} = E_0 + \lambda_1.$$

The total energy level difference is

$$E_{\max} - E_{\min} = \lambda_N - \lambda_1.$$

Since $|\lambda_1| \gg |\lambda_N|$, the average energy level gradient (for two adjacent energy levels) is approximatively given by

$$\Delta E = \frac{E_{\max} - E_{\min}}{N} \simeq \frac{|\lambda_1|}{N}. \tag{6.4.73}$$

The first eigenvalue λ_1 of (6.4.70) is

$$\lambda_1 \simeq -K \quad \left(\text{the unit is } \frac{\hbar c}{\rho_\gamma} \right).$$

Hence, by (6.4.71) and (6.4.73)

$$\Delta E = \beta_1^3 K^{-2} \frac{\hbar c}{\rho_\gamma}, \quad K = \frac{B_w \rho_\gamma}{\rho_w} \frac{g_w^2}{\hbar c}.$$

Then we get the estimates

$$\Delta E = \left(\frac{\rho_w}{\rho_n} \right)^{12} \frac{\hbar c}{\rho_\gamma}. \tag{6.4.74}$$

As we take

$$\frac{\rho_w}{\rho_n} = 10^{-5}, \quad \rho_\gamma = 10^{-20}\text{cm},$$

then from (6.4.74) we derive that

$$\Delta E = 10^{-40} \hbar c/\text{cm} = 2 \times 10^{-45}\text{eV}.$$

This is a very small value, and it is impossible for experiments to measure.

Remark 6.23 The physical conclusion that all particles have finite and discrete energy distribution is of very important significance for the quantum field theory. It implies that the infinity appearing in the field quantization does not exist, and the renormalization theory needs to be reconsidered.

6.5 Field Theory of Multi-Particle Systems

6.5.1 Introduction

We start with the known model of multi-particle systems. Consider an N-particle system with particles

$$A_1, \cdots, A_N. \tag{6.5.1}$$

Let $x_k = (x_k^1, x_k^2, x_k^3) \in \mathbb{R}^3$ be the coordinate of A_k, and

$$\psi = \psi(t, x_1, \cdots, x_k) \tag{6.5.2}$$

be the wave function describing the N-particle system (6.5.1). Then, the classical theory for (6.5.1) is provided by the Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = -\sum_{k=1}^{N} \frac{\hbar^2}{2m_k} \Delta_k \psi + \sum_{j \neq k} V(x_j, x_k)\psi, \tag{6.5.3}$$

where $V(x_j, x_k)$ is the potential energy of interactions between A_j and A_k, m_k is the mass of A_k, and

$$\Delta_k = \frac{\partial^2}{(\partial x_k^1)^2} + \frac{\partial^2}{(\partial x_k^2)^2} + \frac{\partial^2}{(\partial x_k^3)^2}.$$

The wave function ψ satisfies the normalization condition

$$\int_{\mathbb{R}^3} \cdots \int_{\mathbb{R}^3} |\psi|^2 dx_1 \cdots dx_N = 1.$$

Namely, the physically $|\psi(t, x_1, \cdots, x_N)|^2$ represents the probability density of $A_1, \cdots,$ A_N appearing at $x_1, \cdots, x_N$ at time t.

It is clear that the Schrödinger equation (6.5.3) for an N-particle system is only an approximate model:

- It is non-relativistic model;
- The model does not involve the vector potentials $\vec{A}$ of the interactions between particles.
- By using coordinate x_k to represent the particle A_k amounts essentially to saying that the wave function ψ satisfying (6.5.3) can only describe the statistic properties of the system (6.5.1), and contains no information for each individual particle A_k $(1 \leqslant k \leqslant N)$.
- The model is decoupled with interaction fields, i.e. the interaction fields in the model are treated as given functions.

In fact, the most remarkable characteristic of interacting multi-particle systems is that both particle fields and interaction fields are closely related. Therefore, a complete field model of multi-particle systems have to couple both the particle field equations and the interaction field equations. In particular, a precise unified field theory should be based on the field model of the multi-particle system coupled with the four fundamental interactions.

6.5.2 Basic postulates for *N*-body quantum physics

As mentioned in the last subsection, the dynamic models for multi-particle quantum systems have to couple both particle and interaction fields. Therefore there should be some added quantum rules for the systems. In the following we propose the basic postulates for N-particle quantum systems.

First of all, the physical systems have to satisfy a few fundamental physical principles introduced below.

Postulate 6.24 *Any N-particle quantum system has to obey the physical fundamental principles such as:*

$$\begin{aligned} &\textit{Einstein General Relativity},\\ &\textit{Lorentz Invariance},\\ &\textit{Gauge Invariance},\\ &\textit{Gauge Representation Invariance (PRI)},\\ &\textit{Principle of Lagrange Dynamics (PLD)},\\ &\textit{Principle of Interaction Dynamics (PID)},\end{aligned} \tag{6.5.4}$$

where the gauge invariance means the invariance of the Lagrangian action under corresponding gauge transformations.

We note that in general multi-particle systems are layered, and may consist of numerous sub-systems. In particular, we know that the weak and strong interactions are also layered. Hence, here we consider the same level systems, i.e. the systems which consist of identical particles or sub-systems possessing the same level of interactions.

For multi-particle systems with N same level subsystems A_k $(1 \leqslant k \leqslant N)$, the energy contributions of A_k are indistinguishable. Hence, the Lagrangian actions for the N-particle systems satisfy $SU(N)$ gauge invariance. Thus we propose the following basic postulate:

Postulate 6.25 *An N-particle system obeys the $SU(N)$ gauge invariance, i.e. the Lagrangian action of this system is invariant under the $SU(N)$ gauge transformation*

$$\begin{pmatrix} \widetilde{\psi}_1 \\ \vdots \\ \widetilde{\psi}_N \end{pmatrix} = \Omega \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_N \end{pmatrix}, \quad \Omega \in SU(N), \tag{6.5.5}$$

where $\psi_1, \cdots, \psi_N$ are the wave functions of the N particles.

We now need to explain the physical significance of the $SU(N)$ gauge fields induced by Postulate 6.25.

Let each particle of the N-particle system carry an interaction charge g (for example a weak charge $g = g_w$). Then, there are interactions present between the N particles. By the $SU(N)$ gauge theory, the gauge invariant 4-dimensional energy-momentum operator is given by

$$D_\mu = \partial_\mu + igG^a_\mu \tau_a \qquad \text{for } 1 \leqslant a \leqslant N^2 - 1, \tag{6.5.6}$$

and the interaction energy generated by the N particles is

$$E = \begin{cases} \overline{\Psi}(i\gamma^\mu D_\mu \Psi) & \text{for fermions,} \\ |D_\mu \Psi|^2 & \text{for bosons,} \end{cases} \tag{6.5.7}$$

where $\Psi = (\psi_1, \cdots, \psi_N)^{\mathrm{T}}$, and D_μ is as in (6.5.6). From (6.5.6) and (6.5.7) we obtain the physical explanation to the $SU(N)$ gauge fields G^a_μ, stated in the following postulate:

Postulate 6.26 *For a system of N-particles in the same level with each particle carrying an interaction charge g, the N particles induce dynamic interactions between them, and the $SU(N)$ gauge fields*

$$gG^a_\mu \qquad \text{for } 1 \leqslant a \leqslant N^2 - 1 \tag{6.5.8}$$

stand for the interaction potentials between the N particles.

The N particles induce dynamic interactions between them in terms of the $SU(N)$ gauge fields (6.5.8). These interaction fields cannot be measured experimentally because they depend on the choice of generator representation τ_a of $SU(N)$. By the $SU(N)$ geometric theory in Section 3.5, there is a constant $SU(N)$ tensor

$$\alpha^N_a = (\alpha^N_1, \cdots, \alpha^N_N), \tag{6.5.9}$$

such that the contraction field using PRI

$$G_\mu = \alpha^N_a G^a \tag{6.5.10}$$

is independent of the $SU(N)$ representation τ_a. The field (6.5.10) is the interaction field which can be experimentally observed. Thus we propose the following basic postulate.

Postulate 6.27 *For an N-particle system, only the interaction field given by (6.5.10) can be measured, and is the interaction field under which this system interacts with other external systems.*

Remark 6.28 Postulates 6.24-6.27, together with the Principle of Symmetry-Breaking 2.14 and the Postulates 6.1-6.5, form a complete foundation for quantum physics. In fact, without Postulates 6.24-6.27, we cannot establish the quantum physics of multi-particle systems. □

The main motivation to introduce Postulates 6.25 and 6.26 are as follows. Consider an N-particle system with each particle carrying an interaction charge g. Let this be a fermionic system, and the Dirac spinors be given by

$$\Psi = (\psi_1, \cdots, \psi_N)^{\mathrm{T}}.$$

By Postulates 6.3 and 6.5, the Dirac equations for this system can be expressed in the general form

$$i\gamma^\mu D_\mu \Psi + M\Psi = 0, \tag{6.5.11}$$

where M is the mass matrix, and

$$D_\mu \Psi = \partial_\mu \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_N \end{pmatrix} + ig \begin{pmatrix} G_\mu^{11} & \cdots & G_\mu^{1N} \\ \vdots & & \vdots \\ G_\mu^{N1} & \cdots & G_\mu^{NN} \end{pmatrix} \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_N \end{pmatrix}, \tag{6.5.12}$$

where $G = (G_\mu^{ij})$ is an Hermitian matrix, representing the interaction potentials between the N particles generated by the interaction charge g.

Notice that the space consisting of all Hermitian matrices

$$H(N) = \{G|\ G \text{ is an N-th order Hermitian matrix}\}$$

is an N^2-dimensional linear space with basis

$$\tau_0,\ \tau_1,\ \cdots,\ \tau_K \qquad \text{with } K = N^2 - 1, \tag{6.5.13}$$

where $\tau_0 = I$ is the identity, and τ_a $(1 \leqslant a \leqslant N^2 - 1)$ are the traceless Hermitian matrices. Hence, the Hermitian matrix $G = (G_\mu^{ij}) \in H(N)$ in (6.5.12) can be expressed as

$$G = G_\mu^0 I + G_\mu^a \tau_a \qquad \text{with } \tau_a \text{ as in (6.5.13).}$$

Thus, the differential operator in (6.5.12) is in the form

$$D_\mu = \partial_\mu + igG_\mu^0 + igG_\mu^a \tau_a. \tag{6.5.14}$$

The equations (6.5.11) with (6.5.14) are just the Dirac equations in the form of $SU(N)$ gauge fields $\{G_\mu^a \mid 1 \leqslant a \leqslant N^2 - 1\}$ with a given external interaction field G_μ^0. Thus, based on Postulate 6.24, the gauge invariance of an N-particle system and the expressions (6.5.11) and (6.5.14) of the N fermionic particle field equations dictate Postulates 6.25 and 6.26.

The derivation here indicates that Postulates 6.25 and 6.26 can be considered as the consequence of 1) the gauge invariance stated in Postulate 6.24, and 2) the existence of interactions between particles as stated in (6.5.12), which can be considered as an axiom.

6.5.3 Field equations of multi-particle systems

Based on the basic axioms given by Postulates 6.24-6.27, we can establish field equations for various levels of N-particle systems. We proceed in several different cases.

Fermionic systems

Consider N fermions at the same level with interaction charge g, the wave functions (Dirac spinors) are given by

$$\Psi=(\psi_1,\cdots,\psi_N)^{\mathrm{T}},\quad \psi_k=(\psi_k^1,\psi_k^2,\psi_k^3,\psi_k^4)^{\mathrm{T}}\quad \text{for } 1\leqslant k\leqslant N, \tag{6.5.15}$$

with the mass matrix

$$M=\begin{pmatrix} m_1 & & 0\\ & \ddots & \\ 0 & & m_N\end{pmatrix}. \tag{6.5.16}$$

By Postulates 6.24 and 6.25, the Lagrangian action for the N-particle system (6.5.15)-(6.5.16) must be in the form

$$L=\int(\mathscr{L}_G+\mathscr{L}_D)dx, \tag{6.5.17}$$

where $\mathscr{L}_G$ is the sector of the $SU(N)$ gauge fields, and $\mathscr{L}_D$ is the Dirac sector of particle fields:

$$\begin{aligned}&\mathscr{L}_G=-\frac{1}{4\hbar c}\mathscr{G}_{ab}g^{\mu\alpha}g^{\nu\beta}G^a_{\nu\mu}G^b_{\alpha\beta},\\ &\mathscr{L}_D=\overline{\Psi}\left[i\gamma^\mu\left(\partial_\mu+\frac{ig}{\hbar c}G^0_\mu+\frac{ig}{\hbar c}G^a_\mu\tau_a\right)-\frac{c}{\hbar}M\right]\Psi,\end{aligned} \tag{6.5.18}$$

where G^a_μ $(1\leqslant a\leqslant N^2-1)$ are the $SU(N)$ gauge fields representing the interactions between the N particles, τ_a $(1\leqslant a\leqslant N^2-1)$ are the generators of $SU(N)$, and

$$\begin{aligned}&\mathscr{G}_{ab}=\frac{1}{2}\mathrm{Tr}(\tau_a\tau_b^\dagger),\\ &G^a_{\mu\nu}=\partial_\mu G^a_\nu-\partial_\nu G^a_\mu+\frac{g}{\hbar c}\lambda^a_{bc}G^b_\mu G^c_\nu.\end{aligned}$$

According to PID and PLD, for the action (6.5.17) the field equations are given by

$$\begin{aligned}&\frac{\delta L}{\delta G^a_\mu}=D_\mu\phi_a &&\text{by PID},\\ &\frac{\delta L}{\delta\Psi}=0 &&\text{by PLD},\end{aligned} \tag{6.5.19}$$

where D_μ is the PID gradient operator given by

$$D_\mu=\frac{1}{\hbar c}\left(\partial_\mu-\frac{1}{4}k^2x_\mu+\frac{g\alpha}{\hbar c}G_\mu+\frac{g\beta}{\hbar c}G^0_\mu\right),$$

G_μ is as in (6.5.10), α and k are parameters, k^{-1} stands for the range of attracting force of the interaction, and $\left(\frac{g\alpha}{\hbar c}\right)^{-1}$ is the range of the repelling force.

Thus, by (6.5.18) and (6.5.19) we derive the field equations of the N-particle system (6.5.15)-(6.5.16) as follows

$$\begin{aligned}
&\mathscr{G}_{ab}\left[\partial^\nu G^b_{\nu\mu} - \frac{g}{\hbar c}\lambda^b_{cd} g^{\alpha\beta} G^c_{\alpha\mu} G^d_\beta\right] - g\overline{\Psi}\gamma_\mu \tau_a \Psi \\
&= \left[\partial_\mu - \frac{1}{4}k^2 x_\mu + \frac{g\alpha}{\hbar c} G_\mu + \frac{g\beta}{\hbar c} G^0_\mu\right]\phi_a \qquad \text{for } 1 \leqslant a \leqslant N^2-1,
\end{aligned} \tag{6.5.20}$$

$$i\gamma^\mu\left[\partial_\mu + \frac{ig}{\hbar c}G^0_\mu + \frac{ig}{\hbar c}G^a_\mu \tau_a\right]\begin{pmatrix}\psi_1\\ \vdots \\ \psi_N\end{pmatrix} - \frac{c}{\hbar}M\begin{pmatrix}\psi_1\\ \vdots \\ \psi_N\end{pmatrix} = 0, \tag{6.5.21}$$

where $\gamma_\mu = g_{\mu\nu}\gamma^\nu$, and G^0_μ is the interaction field of external systems. It is by this field G^0_μ that we can couple external sub-systems to the model (6.5.20)-(6.5.21).

Remark 6.29 In the field equations of multi-particle systems there is a gauge fixing problem. In fact, we know that the action (6.5.17)-(6.5.18) is invariant under the gauge transformation

$$\left(\widetilde{\Psi}, \widetilde{G}^a_\mu \tau_a\right) = \left(e^{i\theta^a\tau_a}\Psi, G^a_\mu e^{i\theta^b\tau_b}\tau_a e^{-i\theta^b\tau_b} - \frac{1}{g}\partial_\mu\theta^b\tau_b\right). \tag{6.5.22}$$

Hence if (Ψ, G^a_μ) is a solution of

$$\delta L = 0, \tag{6.5.23}$$

then $(\widetilde{\Psi}, \widetilde{G}^a_\mu)$ is a solution of (6.5.23) as well. In (6.5.22) we see that $\widetilde{G}^a_\mu$ have N^2-1 free functions

$$\theta^a(x) \qquad \text{with } 1 \leqslant a \leqslant N^2-1. \tag{6.5.24}$$

In order to eliminate the N^2-1 freedom of (6.5.24), we have to supplement N^2-1 gauge fixing equations for the equation (6.5.23). Now, as we replace the PLD equation (6.5.23). By the PID equations (6.5.19), (6.5.22) breaks the gauge invariance. Therefore the N^2-1 freedom of (6.5.24) is eliminated. However, in the PID equations (6.5.19) there are additional N^2-1 new unknown functions ϕ_a $(1 \leqslant a \leqslant N^2-1)$. Hence, the gauge fixing problem still holds true. There are two possible ways to solve this problem:

1) there might exist some unknown fundamental principles, which can provide the all or some of the N^2-1 gauge fixing equations; and

2) there might be no general physical principles to determine the gauge fixing equations, and these equations will be determined by underlying physical system. □

Bosonic systems

Consider N bosons with charge g, the Klein-Gordon fields are

$$\Phi=(\varphi_1,\cdots,\varphi_N)^{\mathrm{T}},$$

and the mass matrix is given by (6.5.16). The action is

$$L=\int(\mathscr{L}_G+\mathscr{L}_{KG})dx \tag{6.5.25}$$

where $\mathscr{L}_G$ is as given by (6.5.18), and $\mathscr{L}_{KG}$ is the Klein-Gordon sector given by

$$\begin{aligned}
&\mathscr{L}_{KG}=\frac{1}{2}|D_\mu\Phi|^2+\frac{1}{2}\left(\frac{c}{\hbar}\right)^2|M\Phi|^2\\
&D_\mu=\partial_\mu+\frac{ig}{\hbar c}G_\mu^0+\frac{ig}{\hbar c}G_\mu^a\tau_a.
\end{aligned}$$

Then, the PID equations of (6.5.25) are as follows

$$\begin{aligned}
&\mathscr{G}_{ab}\left[\partial^\nu G^b_{\nu\mu}-\frac{g}{\hbar c}\lambda^b_{cd}g^{\alpha\beta}G^c_{\alpha\mu}G^d_\beta\right]+\frac{ig}{2}\left[(D_\mu\Phi)^\dagger(\tau_a\Phi)-(\tau_a\Phi)^\dagger(D_\mu\Phi)\right] \qquad (6.5.26)\\
&\quad=\left[\partial_\mu-\frac{1}{4}k^2x_\mu+\frac{g}{\hbar c}\alpha G_\mu+\frac{g}{\hbar c}\beta G^0_\mu\right]\phi_a \qquad \text{for } 1\leqslant a\leqslant N^2-1,\\
&D^\mu D_\mu\begin{pmatrix}\varphi_1\\ \vdots\\ \varphi_N\end{pmatrix}-\left(\frac{c}{\hbar}\right)^2M^2\begin{pmatrix}\varphi_1\\ \vdots\\ \varphi_N\end{pmatrix}=0. \qquad (6.5.27)
\end{aligned}$$

Mixed systems

Consider a maxed system consisting of N_1 fermions with n_1 charges g and N_2 bosons with n_2 charges g, and the fields are

$$\begin{aligned}
&\text{Dirac fields:} && \Psi=(\psi_1,\cdots,\psi_{N_1})^{\mathrm{T}},\\
&\text{Klein-Gordon fields:} && \Phi=(\varphi_1,\cdots,\varphi_{N_2})^{\mathrm{T}}.
\end{aligned}$$

The interaction fields of this system are $SU(N_1)\times SU(N_2)$ gauge fields, $SU(N_1)$ gauge fields are for fermions, and $SU(N_2)$ for bosons:

$$\begin{aligned}
&\{G^a_\mu\ |1\leqslant a\leqslant N_1^2-1\} && \text{for Dirac fields } \Psi,\\
&\{\widetilde{G}^k_\mu\ |\ 1\leqslant k\leqslant N_2^2-1\} && \text{for Klein-Gordon fields } \Phi.
\end{aligned}$$

The action is given by

$$L=\int\left[\mathscr{L}^1_G+\mathscr{L}^2_G+\mathscr{L}_D+\mathscr{L}_{KG}\right]dx, \tag{6.5.28}$$

where $\mathscr{L}_G^1$ and $\mathscr{L}_G^2$ are the sectors of $SU(N_1)$ and $SU(N_2)$ gauge fields as given in (6.5.18) with $N=N_1$ and $N=N_2$ respectively.

Define the two total gauge fields of $SU(N_1)$ and $SU(N_2)$, as defined by (6.5.9)-(6.5.10):

$$\begin{aligned} G_\mu &= \alpha_a^{N_1} G_\mu^a \quad \text{for } 1 \leqslant a \leqslant N_1^2-1, \\ \widetilde{G}_\mu &= \alpha_k^{N_2} \widetilde{G}_\mu^k \quad \text{for } 1 \leqslant k \leqslant N_2^2-1. \end{aligned} \tag{6.5.29}$$

Namely, $\mathscr{L}_D$ and $\mathscr{L}_{KG}$ are given by

$$\begin{aligned} \mathscr{L}_D &= \overline{\Psi}\left[i\gamma^\mu\left(\partial_\mu + \frac{in_1 g}{\hbar c}G_\mu^0 + \frac{in_1 g}{\hbar c}\widetilde{G}_\mu + \frac{in_1 g}{\hbar c}G_\mu^a \tau_a^1\right) - \frac{c}{\hbar}M_1\right]\Psi, \\ \mathscr{L}_{KG} &= \frac{1}{2}\left|\left(\partial_\mu + \frac{in_2 g}{\hbar c}G_\mu^0 + \frac{in_2 g}{\hbar c}G_\mu + \frac{in_2 g}{\hbar c}\widetilde{G}_\mu^k \tau_k^2\right)\Phi\right|^2 + \frac{1}{2}\left(\frac{c}{\hbar}\right)^2 |M_2\Phi|^2, \end{aligned}$$

where G_μ^0 is the external field, and $\widetilde{G}_\mu$ and G_μ are as in (6.5.29).

Thus, we derive the field equations for mixed multi-particle systems expressed in the following form

$$\begin{aligned} &\mathscr{G}_{ab}^1\left[\partial^\nu G_{\nu\mu}^b - \frac{n_1 g}{\hbar c}\lambda_{1cd}^b g^{\alpha\beta} G_{\alpha\mu}^c G_\beta^d\right] - \frac{n_1 g}{\hbar c}\overline{\Psi}\gamma_\mu \tau_a^1 \Psi \\ &\quad + \frac{in_2 g}{\hbar c}\left[(D_\mu\Phi)^*(\alpha_a^{N_1}\Phi) - (\alpha_a^{N_1}\Phi)^*(D_\mu\Phi)\right] \\ &= \left[\partial_\mu - \frac{1}{4}k_1^2 x_\mu + \frac{n_1 g}{\hbar c}\alpha_1 G_\mu + \frac{n_2 g}{\hbar c}\alpha_2 \widetilde{G}_\mu\right]\phi_a \qquad \text{for } 1 \leqslant a \leqslant N^2-1, \end{aligned} \tag{6.5.30}$$

$$\begin{aligned} &\mathscr{G}_{kl}^2\left[\partial^\nu \widetilde{G}_{\nu\mu}^l - \frac{n_2 g}{\hbar c}\lambda_{2ij}^l g^{\alpha\beta} \widetilde{G}_{\alpha\mu}^i \widetilde{G}_\beta^j\right] - \frac{n_1 g}{\hbar c}\alpha_k^{N_2}\overline{\Psi}\gamma_\mu \Psi \\ &\quad + \frac{in_1 g}{2\hbar c}\left[(D_\mu\Phi)^\dagger(\tau_k^2\Phi) - (\tau_k^2\Phi)^\dagger(D_\mu\Phi)\right] \\ &= \left[\partial_\mu - \frac{1}{4}k_2^2 x_\mu + \frac{n_1 g}{\hbar c}\beta_1 G_\mu + \frac{n_2 g}{\hbar c}\beta_2 \widetilde{G}_\mu\right]\widetilde{\phi}_k \qquad \text{for } 1 \leqslant k \leqslant N_2^2-1, \end{aligned} \tag{6.5.31}$$

$$i\gamma^\mu\left(\partial_\mu + \frac{in_1 g}{\hbar c}G_\mu^0 + \frac{in_1 g}{\hbar c}\widetilde{G}_\mu + \frac{in_1 g}{\hbar c}G_\mu^a \tau_a^1\right)\Psi - \frac{c}{\hbar}M_1\Psi = 0, \tag{6.5.32}$$

$$g^{\mu\nu}D_\mu D_\nu \Phi - \left(\frac{c}{\hbar}\right)^2 M_2^2 \Phi = 0, \tag{6.5.33}$$

where G_μ and $\widetilde{G}_\mu$ are as in (6.5.29), and D_μ is defined by

$$D_\mu = \partial_\mu + \frac{in_2 g}{\hbar c}G_\mu^0 + \frac{in_2 g}{\hbar c}G_\mu + \frac{in_1 g}{\hbar c}\widetilde{G}_\mu^k \tau_k^2.$$

We remark here that the coupling interaction between fermions and bosons is directly represented on the right hand side of gauge field equations (6.5.30) and (6.5.31), due to the presence of the dual interaction fields based on PID. Namely, the interactions between particles in an N-particle system are achieved through both the interaction gauge fields and the corresponding dual fields. This fact again validates the importance of PID.

Another remark is that the gauge actions $\mathscr{L}_G^1$ and $\mathscr{L}_G^2$ in (6.5.28) obey the gauge invariance, but sectors $\mathscr{L}_D$ and $\mathscr{L}_{KG}$ break the gauge symmetry, due to the coupling of different level physical systems. Namely, the Principle of Symmetry-Breaking 2.14 holds true here. In addition, the field equations (6.5.30) and (6.5.31) spontaneously break the gauge symmetry, due essentially to the fields G_μ and $\widetilde{G}_\mu$ on the right-sides of the field equations.

Layered systems

Let a system be layered consisting of two levels: 1) level A consists of K sub-systems $A_1, \cdots, A_K$, and 2) level B is level inside of each sub-system A_j, which consists of N particles $B_1^j, \cdots, B_N^j$:

$$\begin{aligned} &\text{at level } A: \quad A = \{A_1, \cdots, A_K\}, \\ &\text{at level } B: \quad A_j = \{B_1^j, \cdots, B_N^j\} \qquad \text{for } 1 \leqslant j \leqslant K. \end{aligned} \tag{6.5.34}$$

Each particle B_i^j carries n charges g.

Let the particle field functions be

$$\begin{aligned} &\text{at level } A: \qquad \Psi_A = (\psi_{A_1}, \cdots, \psi_{A_K}), \\ &\text{at level } B: \qquad \Psi_{B_j} = (\psi_{B_{j_1}}, \cdots, \psi_{B_{j_N}}) \qquad \text{for } 1 \leqslant j \leqslant K. \end{aligned}$$

The interaction is the $SU(K) \times SU(N)$ gauge fields:

$$\begin{aligned} &\text{at level } A: \qquad SU(K) \ \text{ gauge fields } A_\mu^a \qquad 1 \leqslant a \leqslant K^2 - 1, \\ &\text{at level } B: \qquad SU(N) \ \text{ gauge fields } (B_j)_\mu^k \qquad 1 \leqslant k \leqslant N^2 - 1. \end{aligned}$$

Without loss of generality, we assume A and B are the fermion systems. Thus the action of this layered system is

$$L = \int \left[\mathscr{L}_{AG} + \sum_{j=1}^K \mathscr{L}_{B_jG} + \mathscr{L}_{AD} + \sum_{j=1}^K \mathscr{L}_{B_jD} \right] dx, \tag{6.5.35}$$

where

$$\begin{aligned} &\mathscr{L}_{AG} = \text{ the sector of } SU(K) \ \text{ gauge fields}, \\ &\mathscr{L}_{AD} = \overline{\Psi}_A \left[i\gamma^\mu \left(\partial_\mu + \frac{inN}{\hbar c} g G_\mu^0 + \frac{inN}{\hbar c} B_\mu + \frac{inN}{\hbar c} g A_\mu^a \tau_a^K \right) - \frac{c}{\hbar} M_A \right] \Psi_A, \\ &\mathscr{L}_{B_jG} = \text{ the j-th the sector of } SU(N) \text{ gauge fields}, \\ &\mathscr{L}_{B_jD} = \overline{\Psi}_{B_j} \left[i\gamma^\mu \left(\partial_\mu + \frac{ing}{\hbar c} G_\mu^0 + \frac{ing}{\hbar c} A_\mu + \frac{ing}{\hbar c} (B_j)_\mu^k \tau_k^N \right) - \frac{c}{\hbar} M_{B_j} \right] \Psi_{B_j}, \end{aligned} \tag{6.5.36}$$

where G_μ^0 is the external field. The corresponding PID field equations of the layered multi-particle system (6.5.34) follow from (6.5.35) and (6.5.36), and here we omit the details.

Remark 6.30 Postulate 6.27 is essentially another expression of PRI, which is very crucial to couple all sub-systems together to form a complete set of field equations for a given multi-particle system. In particular, this approach is natural and unique to derive models for multi-particle systems, satisfying all fundamental principles of (6.5.4), the Principle of Symmetry-Breaking 2.14, and the gauge symmetry breaking principle (Principle 4.4). It is also a unique way to establish a unified field theory coupling the gravity and other interactions in various levels of multi-particle systems. In the next subsection we discuss this topic.

6.5.4 Unified field model coupling matter fields

In Chapter 4, we have discussed the unified field theory, in which we consider two aspects: 1) the interaction field particles, and 2) the interaction potentials. Hence, it restricted the unified field model to be the theory based on

$$\text{Einstein relativity} + U(1) \times SU(2) \times SU(3) \text{ symmetry.} \tag{6.5.37}$$

However, if we consider the interaction potentials between the particles of N-particle systems, then the unified field theory has to be based on the following symmetries instead of (6.5.37):

$$\text{Einstein relativity} + SU(N_1) \times \cdots \times SU(N_K) \text{ symmetry,} \tag{6.5.38}$$

where $N_1, \cdots, N_K$ are the particle numbers of various sub-systems and layered systems.

The two types of unified field models based on (6.5.37) and (6.5.38) are mutually complementary. They have different roles in revealing the essences of interactions and particle dynamic behaviors.

In this subsection, we shall establish the unified field model of multi-particle systems based on (6.5.38), which matches the vision of Einstein and Nambu. In his Nobel lecture (Nambu, 2008), Nambu stated that

> *Einstein used to express dissatisfaction with his famous equation of gravity*
>
> $$G_{\mu\nu} = 8\pi T_{\mu\nu}$$
>
> *His point was that, from an aesthetic point of view, the left hand side of the equation which describes the gravitational field is based on a beautiful geometrical principle, whereas the right hand side, which describes everything else, . . . looks arbitrary and ugly.*
>
> *... [today] Since gauge fields are based on a beautiful geometrical principle, one may shift them to the left hand side of Einstein s equation. What is left on the right are the matter fields which act as the source for the gauge fields ... Can one geometrize the matter fields and shift everything to the left?*

The gravity will be considered only in systems possessing huge amounts of particles, which we call gravitational systems. Many gravitational systems have very complicated structures. But they are composites of some simple systems. Here we only discuss two cases.

Systems with gravity and electromagnetism

Consider the system consisting of N_1 fermions with n_1 electric charges n_1e and N_2 bosons with n_2 charges n_2e:

$$\begin{aligned} &\Psi = (\psi_1, \cdots, \psi_{N_1}) && \text{for fermions,} \\ &\Phi = (\varphi_1, \cdots, \varphi_{N_2}) && \text{for bosons.} \end{aligned}$$

The action is given by

$$L = \int \left[\frac{c^4}{8\pi G} R + \mathscr{L}_A^{N_1} + \mathscr{L}_A^{N_2} + \hbar c \mathscr{L}_D + \hbar c \mathscr{L}_{KG} \right] \sqrt{-g} dx \tag{6.5.39}$$

where R is the scalar curvature, G is the gravitational constant, $g = \det(g_{\mu\nu})$, $\mathscr{L}_A^{N_1}$ and $\mathscr{L}_A^{N_2}$ are the sectors of $SU(N_1)$ and $SU(N_2)$ gauge fields for the electromagnetic interaction

$$\begin{aligned} &\mathscr{L}_A^{N_1} = -\frac{1}{4}\mathscr{G}_{ab} g^{\mu\alpha} g^{\nu\beta} A_{\mu\nu}^a A_{\alpha\beta}^b && 1 \leqslant a, b \leqslant N_1^2 - 1, \\ &\mathscr{L}_A^{N_2} = -\frac{1}{4}\widetilde{\mathscr{G}}_{kl} g^{\mu\alpha} g^{\nu\beta} \widetilde{A}_{\mu\nu}^k \widetilde{A}_{\alpha\beta}^l && 1 \leqslant k, l \leqslant N_2^2 - 1, \\ &A_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + \frac{n_1 e}{\hbar c} \lambda_{bc}^a A_\mu^b A_\nu^c && n_1 \in \mathbb{Z}, \\ &\widetilde{A}_{\mu\nu}^k = \partial_\mu \widetilde{A}_\nu^k - \partial_\nu \widetilde{A}_\mu^k + \frac{n_2 e}{\hbar c} \widetilde{\lambda}_{ij}^k \widetilde{A}_\mu^i \widetilde{A}_\nu^j && n_2 \in \mathbb{Z}, \end{aligned} \tag{6.5.40}$$

and $\mathscr{L}_D, \mathscr{L}_{KG}$ are the Dirac and Klein-Gordon sectors:

$$\begin{aligned} &\mathscr{L}_D = \overline{\Psi} \left[i\gamma^\mu \left(\partial_\mu + \frac{in_1 e}{\hbar c} A_\mu^0 + \frac{in_1 e}{\hbar c} A_\mu^a \tau_a \right) - \frac{c}{\hbar} M_1 \right] \Psi, \\ &\mathscr{L}_{KG} = \frac{1}{2} g^{\mu\nu} (D_\mu \Phi)^\dagger (D_\nu \Phi) + \frac{1}{2} \left(\frac{c}{\hbar} \right)^2 |M_2 \Phi|^2, \\ &D_\mu = \nabla_\mu + \frac{in_2 e}{\hbar c} A_\mu^0 + \frac{in_2 e}{\hbar c} \widetilde{A}_\mu^k \widetilde{\tau}_k, \end{aligned} \tag{6.5.41}$$

where M_1 and M_2 are the masses, ∇_μ is the covariant derivative, and A_μ^0 is the external electromagnetic field.

Based on PID and PLD, the field equations of (6.5.39) are given by

$$\begin{aligned}
&\frac{\delta}{\delta g_{\mu\nu}}L=\frac{c^4}{8\pi G}D^G_\mu\phi^g_\nu, && \text{(PID)}\\
&\frac{\delta}{\delta A^a_\mu}L=D^A_\mu\phi_a, && \text{(PID)}\\
&\frac{\delta}{\delta \widetilde{A}^k_\mu}L=D^{\widetilde{A}}_\mu\widetilde{\phi}_k, && \text{(PID)}\\
&\frac{\delta}{\delta\Psi}L=0, && \text{(PLD)}\\
&\frac{\delta}{\delta\Phi}L=0, && \text{(PLD)}
\end{aligned} \tag{6.5.42}$$

where

$$\begin{aligned}
D^G_\mu&=\nabla_\mu+\frac{n_1e}{\hbar c}A_\mu+\frac{n_2e}{\hbar c}\widetilde{A}_\mu,\\
D^A_\mu&=\partial_\mu-\frac{1}{4}k_1^2x_\mu+\frac{n_1e}{\hbar c}\alpha A_\mu+\frac{n_2e}{\hbar c}\widetilde{\alpha}\widetilde{A}_\mu,\\
D^{\widetilde{A}}_\mu&=\partial_\mu-\frac{1}{4}k_2^2x_\mu+\frac{n_1e}{\hbar c}\beta A_\mu+\frac{n_2e}{\hbar c}\widetilde{\beta}\widetilde{A}_\mu.
\end{aligned} \tag{6.5.43}$$

Here $A_\mu=\alpha^{N_1}_aA^a_\mu$ and $\widetilde{A}_\mu=\alpha^{N_2}_k\widetilde{A}^k_\mu$ are the total electromagnetic fields generated by the fermion system and the boson system.

By (6.5.39)-(6.5.41), the equations (6.5.42)-(6.5.43) are written as

$$R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R=-\frac{8\pi G}{c^4}T_{\mu\nu}+\left(\nabla_\mu+\frac{n_1e}{\hbar c}A_\mu+\frac{n_2e}{\hbar c}\widetilde{A}_\mu\right)\phi^g_\nu, \tag{6.5.44}$$

$$\begin{aligned}
&\mathscr{G}_{ab}\left[\partial^\nu A^b_{\nu\mu}-\frac{n_1e}{\hbar c}\lambda^b_{cd}g^{\alpha\beta}A^c_{\alpha\mu}A^d_\beta\right]-n_1e\overline{\Psi}\gamma_\mu\tau_a\Psi\\
=&\left[\partial_\mu-\frac{1}{4}k_1^2x_\mu+\frac{n_1e}{\hbar c}\alpha A_\mu+\frac{n_2e}{\hbar c}\widetilde{\alpha}\widetilde{A}_\mu\right]\phi_a,
\end{aligned} \tag{6.5.45}$$

$$\begin{aligned}
&\widetilde{\mathscr{G}}_{kl}\left[\partial^\nu\widetilde{A}^l_{\nu\mu}-\frac{n_2e}{\hbar c}\widetilde{\lambda}^l_{ij}g^{\alpha\beta}\widetilde{A}^i_{\alpha\mu}\widetilde{A}^j_\beta\right]+\frac{i}{2}n_2e\left[(D_\mu\Phi)^\dagger(\widetilde{\tau}_k\Phi)-(\widetilde{\tau}_k\Phi)^\dagger(D_\mu\Phi)\right],\\
=&\left[\partial_\mu-\frac{1}{4}k_2^2x_\mu+\frac{n_1e}{\hbar c}\beta A_\mu+\frac{n_2e}{\hbar c}\widetilde{\beta}\widetilde{A}_\mu\right]\widetilde{\phi}_k,
\end{aligned} \tag{6.5.46}$$

$$i\gamma^\mu\left[\partial_\mu+\frac{in_1e}{\hbar c}A^0_\mu+\frac{in_1e}{\hbar c}A^a_\mu\tau_a\right]\Psi-\frac{c}{\hbar}M_1\Psi=0, \tag{6.5.47}$$

$$g^{\mu\nu}D_\mu D_\nu\Phi-\left(\frac{c}{\hbar}\right)^2M_2^2\Phi=0, \tag{6.5.48}$$

where the energy-momentum tensor $T_{\mu\nu}$ in (6.5.44) is

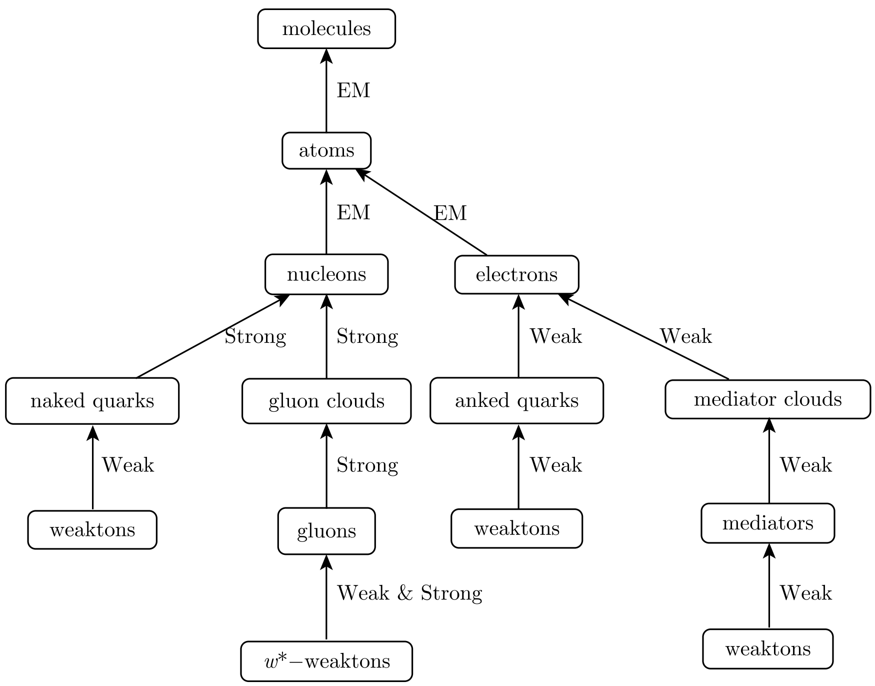
$$\begin{aligned}
T_{\mu\nu}=&-\frac{1}{2}g_{\mu\nu}(\mathscr{L}^{N_1}_A+\mathscr{L}^{N_2}_A+\hbar c\mathscr{L}_D+\hbar c\mathscr{L}_{KG})+\frac{1}{2}(D_\mu\Phi)^\dagger(D_\nu\Phi)\\
&-\frac{1}{4}\mathscr{G}_{ab}g^{\alpha\beta}A^a_{\mu\alpha}A^b_{\nu\beta}-\frac{1}{4}\widetilde{\mathscr{G}}_{kl}g^{\alpha\beta}\widetilde{A}^k_{\mu\alpha}\widetilde{A}^l_{\nu\beta}.
\end{aligned} \tag{6.5.49}$$

The energy-momentum tensor $T_{\mu\nu}$ contains the masses M_1, M_2, the kinetic energy and electromagnetic energy.

It is clear that both sides of the field equations (6.5.44)-(6.5.48) are all generated by the fundamental principles. It is the view presented by Einstein and Nambu and shared by many physicists that the Nature obeys simple beautiful laws based on a few first physical principles. In other words, the energy-momentum tensor $T_{\mu\nu}$ is now derived from first principles and is geometrized as Einstein and Nambu hoped.

Systems with four interactions

The above systems with gravity and electromagnetism in general describe the bodies in lower energy density. For the systems in higher energy density, we have to also consider the weak and strong interactions. The interactions are layered as shown below, which were derived in (Ma and Wang, 2015b, 2014g):



The layered systems and sub-systems above determine the action of the system with four interactions as follows:

$$L = \int \frac{c^4}{8\pi G} R\sqrt{-g}dx + \text{ actions of all levels,} \tag{6.5.50}$$

and the action of each layered level is as given by the manner as used in (6.5.35)-(6.5.36).

Hence, the unified field model of a multi-particle system is completely determined by the layered structure of this system, as given by (6.5.50). It is very natural that a rationale unified field theory must couple the matter fields and interaction fields together.

Remark 6.31 Once again we emphasize that, using PRI contractions as given by (6.5.10) and proper gauge fixing equations, from the unified field model (6.5.50) coupling matter fields for multi-particle system, we can easily deduce that the total electromagnetic field A_μ obtained from (6.5.50) satisfies the $U(1)$ electromagnetic gauge field equations, and derive the weak and strong interaction potentials as given in (Ma and Wang, 2015a, 2014h).

6.5.5 Atomic spectrum

Classical quantum mechanics is essentially a subject to deal with single particle systems. Hence, the hydrogen spectrum theory was perfect under the framework of the Dirac equations. But, for general atoms the spectrum theory was defective due to lack of precise field models of multi-particle systems.

In this subsection, we shall apply the field model of multi-particle systems to establish the spectrum equations for general atoms.

1. *Classical theory of atomic shell structure.* We recall that an atom with atomic number Z has energy spectrum

$$E_n = -\frac{Z^2}{n^2}\frac{me^4}{2\hbar^2}, \quad n = 1, 2, \cdots. \tag{6.5.51}$$

If we ignore the interactions between electrons, the orbital electrons of this atom have the idealized discrete energies (6.5.51). The integers n in (6.5.51) are known as principal quantum number, which characterizes the electron energy levels and orbital shell order:

$$\begin{array}{llllllll} n: & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \text{shell symbol:} & K & L & M & N & O & P & Q. \end{array} \tag{6.5.52}$$

Each orbital electron is in some shell of (6.5.52) and possesses the following four quantum numbers:

1) principle quantum number $n = 1, 2, \cdots$,

2) orbital quantum number $l = 0, 1, 2, \cdots, (n-1)$,

3) magnetic quantum number $m = 0, \pm 1, \cdots, \pm l$,

4) spin quantum number $J = \pm\frac{1}{2}$.

For each given shell n, there are sub-shells characterized by orbital quantum number l, whose symbols are:

$$\begin{array}{llllllll} l: & 0 & 1 & 2 & 3 & 4 & \cdots \\ \text{sub-shell:} & s & p & d & f & g & \cdots. \end{array} \tag{6.5.53}$$

By the Pauli exclusion principle, at a give sub-shell nl, there are at most the following electron numbers

$$N_{nl} = N_l = 2(2l+1) \qquad \text{for } 0 \leqslant l \leqslant n-1.$$

Namely, for the sub-shells $s(l=0), p(l=1), d(l=2), f(l=3), g(l=4)$, their maximal electron numbers are

$$N_s = 2,\ N_p = 6,\ N_d = 10,\ N_f = 14,\ N_g = 18.$$

Thus, on the n-th shell, the maximal electron number is

$$N_n = \sum_{l=0}^{n-1} N_l = 2n^2. \tag{6.5.54}$$

2. *Atomic field equations.* Based on the atomic shell structure, the electron system of an atom consists of shell systems as (6.5.52), which we denote by

$$S_n = \text{ the } n\text{-th shell system for } n = 1, 2, \cdots. \tag{6.5.55}$$

Each shell system S_n has n sub-shell systems as in (6.5.53), denoted by

$$S_{nl} = \text{ the } l\text{-th sub-shell system of } S_n \qquad \text{for } 0 \leqslant l \leqslant n-1. \tag{6.5.56}$$

Thus, we have two kinds of classifications (6.5.55) and (6.5.56) of sub-systems for atomic orbital electrons, which lead to two different sets of field equations.

A. FIELD EQUATION OF SYSTEM S_n. If we ignore the orbit-orbit interactions, then we take (6.5.55) as an N-particle system. Let S_n have K_n electrons:

$$S_n: \ \Psi_n = (\psi_n^1, \cdots, \psi_n^{K_n}), \quad K_n \leqslant N_n,\ 1 \leqslant n \leqslant N, \tag{6.5.57}$$

where N_n is as in (6.5.54). Hence, the model of (6.5.57) is reduced to the $SU(K_1) \times \cdots \times SU(K_N)$ gauge fields of fermions. Referring to the single fermion system (6.5.15)-(6.5.21), the action of (6.5.57) is

$$L = \int \sum_{n=1}^{N} (\mathscr{L}_{SU(K_n)} + \mathscr{L}_D^n) dx, \tag{6.5.58}$$

where

$$
\begin{aligned}
&\mathscr{L}_{SU(K_n)} = -\frac{1}{4\hbar c} g^{\mu\alpha} g^{\nu\beta} A^{a_n}_{\mu\nu} A^{a_n}_{\alpha\beta} && 1 \leqslant a_n \leqslant K_n,\\
&\mathscr{L}^n_D = \overline{\Psi}_n \left(i\gamma^\mu D_\mu - \frac{m_e c}{\hbar} \right) \Psi_n && 1 \leqslant n \leqslant N,\\
&A^{a_n}_{\mu\nu} = \partial_\mu A^{a_n}_\nu - \partial_\nu A^{a_n}_\mu - \frac{e}{\hbar c} \lambda^{a_n}_{b_n c_n} A^{b_n}_\mu A^{c_n}_\nu, \\
&D_\mu \Psi_n = \left(\partial_\mu - \frac{ie}{\hbar c} A^0_\mu - \frac{ie}{\hbar c} A^{a_n}_\mu \tau_{a_n} \right) \Psi_n,
\end{aligned}
\tag{6.5.59}
$$

where $A^{a_n}_\mu$ are the $SU(K_n)$ gauge fields representing the electromagnetic (EM) potential of the electrons in S_n, $\lambda^{a_n}_{b_n c_n}$ are the structure constants of $SU(K_n)$ such that $\mathscr{G}_{a_n b_n} = \frac{1}{2}\text{tr}(\tau_{a_n} \tau^\dagger_{b_n}) = \delta_{a_n b_n}$, A^0_μ is the EM potential generated by the nuclear, $g = -e$ $(e > 0)$ is the charge of an electron, and m_e is the electron mass.

The PID gradient operators for $SU(K_1) \times \cdots \times SU(K_N)$ in (6.5.19) are given by

$$
D^n_\mu = \frac{1}{\hbar c} \left[\partial_\mu + \frac{e}{\hbar c} \sum_{k \neq n} A^{(k)}_\mu \right] \qquad \text{for } 1 \leqslant n \leqslant N, \tag{6.5.60}
$$

where $A^{(k)}_\mu = \alpha^{K_k}_{a_K} A^{a_k}_\mu$ is the total EM potential of S_k shell as defined in (6.5.10).

Then by (6.5.58)-(6.5.60), the field equations of (6.5.57) can be written in the following form

$$
\begin{aligned}
&\partial^\nu A^{a_n}_{\nu\mu} + \frac{e}{\hbar c} \lambda^{a_n}_{b_n c_n} g^{\alpha\beta} A^{b_n}_{\alpha\mu} A^{c_n}_\beta + e\overline{\Psi}\gamma_\mu \tau^{a_n} \Psi_n \\
&= \left[\partial_\mu + \frac{e}{\hbar c} \sum_{k \neq n} A^{(k)}_\mu \right] \phi^{a_n} \qquad \text{for } 1 \leqslant a_n \leqslant K^2_n - 1,\ 1 \leqslant n \leqslant N,
\end{aligned}
\tag{6.5.61}
$$

$$
i\gamma^\mu \left[\partial_\mu - \frac{ie}{\hbar c} A^0_\mu - \frac{ie}{\hbar c} A^{a_n}_\mu \tau_{a_n} \right] \Psi_n - \frac{m_e c}{\hbar} \Psi_n = 0. \tag{6.5.62}
$$

B. Field equation of system S_{nl}. The precise model of atomic spectrum should take (6.5.56) as an N-particle system. Also, $S_n = \sum\limits_{l=0}^{n-1} S_{nl}$ is again divided into n sub-systems

$$
S_n :\ S_{n0}, \cdots, S_{nn-1}.
$$

Hence, the system S_{nl} has more sub-systems than S_n, i.e. if S_n has N sub-systems, then S_{nl} has $\dfrac{1}{2}N(N+1)$ sub-systems.

Let S_{nl} have K_{nl} electrons with wave functions:

$$
S_{nl} :\ \Psi_{nl} = (\psi^1_{nl}, \cdots, \psi^{K_{nl}}_{nl}), \quad 1 \leqslant n \leqslant N,\ 0 \leqslant l \leqslant n-1, \tag{6.5.63}
$$

and $K_{nl} \leqslant 2(2l+1)$. Then the action of (6.5.63) takes as

$$
L = \int \sum_{l=0}^{n-1} \sum_{n=1}^{N} (\mathscr{L}_{SU(K_{nl})} + \mathscr{L}^{nl}_D) dx, \tag{6.5.64}
$$

where $\mathscr{L}_{SU(K_{nl})}$ and $\mathscr{L}_D^{nl}$ are similar to that of (6.5.59). Thus, the field equation of the system (6.5.63) is determined by (6.5.64).

Remark 6.32 The reason why atomic spectrum can be divided into two systems (6.5.57) and (6.5.63) to be considered is that in the system (6.5.63) the electrons in each S_{nl} have the same energy, and in (6.5.57) the electons in each S_n have the same energy if we ignore the interaction energy between different l-orbital electrons of S_{nl}. Hence, the system of S_{nl} is precise and the system of S_n is approximative. □

3. *Atomic spectrum equations.* For simplicity, we only consider the system S_n, and for S_{nl} the case is similar. Since the electrons in each S_n have the same energy λ_n, the wave functions in (6.5.57) can take as

$$\psi_n^j = \varphi_n^j(x)e^{-i\lambda_n t/\hbar} \qquad \text{for } 1 \leqslant j \leqslant K_n. \tag{6.5.65}$$

It is known that the EM fields A_μ^a in atomic shells are independent of time t, i.e. $\partial_t A_\mu^a = 0$. Therefore, inserting (6.5.65) into (6.5.61) and (6.5.62) we derive the spectrum equation in the form

$$\lambda_n\Phi_n = ic\hbar(\vec{\alpha}\cdot D)\Phi_n - eV\Phi_n + m_ec^2\alpha_0\Phi_n + eA_0^{a_n}\tau_{a_n}\Phi_n \qquad \text{for } 1 \leqslant n \leqslant N, \tag{6.5.66}$$

$$\Delta A_0^{a_n} - \frac{e}{\hbar c}\lambda_{b_nc_n}^{a_n}\vec{A}^{b_n}\cdot(\nabla A_0^{c_n} + \frac{e}{\hbar c}\lambda_{d_nf_n}^{c_n}A^{d_n}\vec{A}^{f_n}) - e\Phi_n^\dagger\tau_{a_n}\Phi_n = \frac{e}{\hbar c}\sum_{k\neq n}A_0^{(k)}\phi^{a_n}, \tag{6.5.67}$$

$$\Delta\vec{A}^{a_n} - \nabla(\operatorname{div}\vec{A}^{a_n}) + \frac{e}{\hbar c}\lambda_{b_nc_n}^{a_n}g^{\alpha\alpha}\vec{A}_\alpha^{b_n}A_\alpha^{c_n} + e\overline{\Phi})_n\vec{\gamma}\tau_{a_n}\Phi_n = \left(\nabla + \frac{e}{\hbar c}\sum_{k\neq n}\vec{A}^{(k)}\right)\phi^{a_n}, \tag{6.5.68}$$

where $\Phi_n = (\varphi_n^1, \cdots, \varphi_n^{K_n})^{\mathrm{T}}, A_\mu^{a_n} = (A_0^{a_n}, \vec{A}^{a_n}), \vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and α_0 are as in (3.1.15), $\vec{\gamma} = (\gamma^1, \gamma^2, \gamma^3)$ is as in (6.2.8), $V = ze/r$ is the Coulomb potential of the nuclear, $\vec{A} = (A_1, A_2, A_3)$ is the magnetic potential of the nuclear, and

$$D\Phi_n = \left(\nabla - \frac{ie}{\hbar c}\vec{A} - \frac{ie}{\hbar c}\vec{A}^{a_n}\tau_{a_n}\right)\Phi_n,$$
$$\vec{A}_\alpha^{b_n} = \partial_\alpha\vec{A}^{b_n} - \nabla A_\alpha^{b_n} - \frac{e}{\hbar c}\lambda_{c_nd_n}^{b_n}A_\alpha^{c_n}\vec{A}^{d_n}.$$

The equations (6.5.66)-(6.5.68) need to be complemented with some gauge fixing equations; see Remark 6.29.

Chapter 7
Astrophysics and Cosmology

The aim of this chapter is to study fundamental issues of astrophysics and cosmology based on the first principles dictating the law of gravity, the cosmological principle, and the principle of symmetry-breaking. The study and the results presented in this chapter are based on recent papers (Ma and Wang, 2014e,a; Hernandez, Ma and Wang, 2015).

First, we have rigorously proved a basic Blackhole Theorem on the nature and structure of black holes, Theorem 7.15:

> *Assume the validity of the Einstein theory of general relativity, then black holes are closed, innate and incompressible.*

This theorem was originally discovered and proved in (Ma and Wang, 2014a, Theorem 4.1). One important part of the theorem is that all black holes are closed: matters can neither enter nor leave their interiors. Classical view was that nothing can get out of black holes, but matters can fall into blackholes. We show that nothing can get inside the blackhole either. This theorem offers a very different views on the geometric structure and the origin of our Universe, on the formation and stability of stars and galaxies, and on the mechanism of supernovae and active galactic nucleus (AGN) jets.

In particular, we rigorously show that our Universe is not originated from a Big-Bang, and is static, under the assumption of the Einstein general relativity and the cosmological principle. The redshift is then due mainly to the black hole effect, and the CMB is caused by the blackbody equilibrium radiation.

Also, we show that the dark energy is associated with the the negative pressure, the effect of the gravitational repelling force, and the dark matter is caused the space curved energy.

This chapter is organized as follows. Section 7.1 introduces the momentum representation of astrophysical fluid dynamical models. One important aspect of the models is the coupling between the gravitational field equations and the fluid dynamics equations, using the principle of symmetry-breaking. Section 7.4 proves the black hole theorem.

Stellar circulations are studied in Section 7.2, leading to the mechanism of supernovae explosion. Section 7.3 addresses galactic circulations, and provides the mechanism of AGN jets.

Based on the black hole theorem, theorems on the structure and origin of the Universe are proved in Section 7.5.

Finally, in Section 7.6, we have derived 1) the PID cosmological model of our Universe, 2) the nature of dark energy and dark matter, and 3) the gravitational force formulas.

7.1 Astrophysical Fluid Dynamics

The main objective of this section is to establish fluid dynamical models for astrophysical and cosmological objects such as stars, galaxies, and clusters of galaxies, based on the Principle of Symmetry-Breaking 2.14.

7.1.1 Fluid dynamic equations on Riemannian manifolds

To consider astrophysical fluid dynamics, we first need to discuss the Navier-Stokes equations on Riemannian manifolds.

Let $(\mathscr{M}, g_{ij})$ be an n-dimensional Riemannian manifold. The fluid motion on $\mathscr{M}$ are governed by the Navier-Stokes equations given by

$$\begin{aligned} &\frac{\partial u}{\partial t} + (u\cdot\nabla)u = \nu\Delta u - \frac{1}{\rho}\nabla p + f \qquad \text{for } x\in\mathscr{M}, \\ &\text{div}u = 0, \end{aligned} \tag{7.1.1}$$

where $u = (u^1, \cdots, u^n)$ is the velocity field, p is the pressure, f is the external force, ρ is the mass density, ν is the dynamic viscosity, and the differential operator Δ is the Laplace-Beltrami operator defined as $\Delta u = (\Delta u^1, \cdots, \Delta u^n)$ with

$$\Delta u^i = \text{div}(\nabla u^i) + g^{ij}R_{jk}u^k, \tag{7.1.2}$$

$$\begin{aligned} \text{div}(\nabla u^i) = g^{kl}\Bigg[&\frac{\partial}{\partial x^l}\left(\frac{\partial u^i}{\partial x^k} + \Gamma^i_{kj}u^j\right) + \Gamma^i_{lj}\left(\frac{\partial u^j}{\partial x^k} + \Gamma^j_{ks}u^s\right) \\ &-\Gamma^j_{kl}\left(\frac{\partial u^i}{\partial x^j} + \Gamma^i_{js}u^s\right)\Bigg]. \end{aligned} \tag{7.1.3}$$

Here R_{ij} is the Ricci curvature tensor and Γ^i_{kj} the Levi-Civita connection:

$$\begin{aligned} R_{ij} = &\frac{1}{2}g^{kl}\left(\frac{\partial^2 g_{kl}}{\partial x^i\partial x^j} + \frac{\partial^2 g_{ij}}{\partial x^k\partial x^l} - \frac{\partial^2 g_{kj}}{\partial x^i\partial x^l} - \frac{\partial^2 g_{li}}{\partial x^j\partial x^k}\right) \\ &+ g^{kl}g_{rs}\left(\Gamma^r_{kl}\Gamma^s_{ij} - \Gamma^r_{il}\Gamma^s_{kj}\right), \end{aligned} \tag{7.1.4}$$

$$\Gamma^i_{kj} = \frac{1}{2}g^{il}\left(\frac{\partial g_{kl}}{\partial x^j} + \frac{\partial g_{jl}}{\partial x^k} - \frac{\partial g_{kj}}{\partial x^l}\right). \tag{7.1.5}$$

The nonlinear convection term $(u\cdot\nabla)u$ in (7.1.1) is defined by

$$\begin{aligned} &(u\cdot\nabla)u = (u^iD_iu^1, \cdots, u^iD_iu^n), \\ &u^iD_iu^k = u^i\frac{\partial u^k}{\partial x^i} + \Gamma^k_{ij}u^iu^j \end{aligned} \tag{7.1.6}$$

the pressure term is

$$\nabla p = \left(g^{1k} \frac{\partial p}{\partial x^k}, \cdots, g^{nk} \frac{\partial p}{\partial x^k} \right), \tag{7.1.7}$$

the divergence of u is

$$\mathrm{div} u = \frac{\partial u^k}{\partial x^k} + \Gamma^k_{jk} u^j = \frac{1}{\sqrt{g}} \frac{\partial (\sqrt{g} u^k)}{\partial x^k}, \tag{7.1.8}$$

and $g = \det(g_{ij})$.

By (7.1.2) and (7.1.6)-(7.1.8), the Navier-Stokes equations (7.1.1) can be equivalently written as

$$\begin{aligned} & \frac{\partial u^i}{\partial t} + u^k \frac{\partial u^i}{\partial x^k} + \Gamma^i_{kj} u^k u^j = \nu \left[\mathrm{div}(\nabla u^i) + g^{ij} R_{jk} u^k \right] - \frac{1}{\rho} g^{ij} \frac{\partial p}{\partial x^j} + f^i, \\ & \frac{\partial u^k}{\partial x^k} + \Gamma^k_{jk} u^j = 0. \end{aligned} \tag{7.1.9}$$

Remark 7.1 In the Navier-Stokes equations (7.1.1), the Laplace operator Δ can be taken in two forms:

$$\Delta = d\delta + \delta d \qquad \text{the Laplace-Beltrami operator,} \tag{7.1.10}$$

$$\Delta = \mathrm{div} \cdot \nabla \qquad \text{the Laplace operator.} \tag{7.1.11}$$

Here we choose (7.1.10) instead of (7.1.11) to represent the viscous term in (7.1.1). The reason is that the Laplace-Beltrami operator

$$(d\delta + \delta d) u^i = \mathrm{div} \cdot \nabla u^i + g^{ij} R_{jk} u^k$$

gives rise to an additional term $g^{ij} R_{jk} u^k$. In fluid dynamics, the term $\mu \ \mathrm{div} \cdot \nabla u$ represents the viscous (frictional) force, and the term $g^{ij} R_{jk} u^k$ is the force generated by space curvature and gravitational interaction. Hence physically, it is more natural to take (7.1.10) instead of (7.1.11). □

Remark 7.2 In the fluid dynamic equations (7.1.9), the symmetry of general relativity breaks, and the space and time are treated independently. □

7.1.2 Schwarzschild and Tolman-Oppenheimer-Volkoff (TOV) metrics

We recall in this section the classical Schwarzschild and TOV metrics for centrally symmetric gravitational fields.

Schwarzschild metric

Many stars in the Universe are spherically-shaped, generating centrally symmetric gravitational fields. It is known that the Riemannian metric of a spherically symmetric gravitation field takes the following form:

$$ds^2 = -e^u c^2 dt^2 + e^v dr^2 + r^2 (d\theta^2 + \sin^2\theta d\varphi^2), \tag{7.1.12}$$

where (r,θ,φ) is the spherical coordinate system, and $u=u(r,t)$ and $v=v(r,t)$ are functions of r and t, which are determined by the gravitational field equations.

In the exterior of a ball, Schwarzschild first obtained an exact solution of the Einstein field equations in 1916, which describes the gravitational fields for the external vacuum state of a static spherically symmetric matter field.

Let m be the total mass of a centrally symmetric ball. Then the classical Newtonian gravitational potential of the ball reads

$$\varphi=-\frac{mG}{r}. \tag{7.1.13}$$

Based on the Einstein general theory of relativity, the time-component g_{00} of gravitational potential $g_{\mu\nu}$ and the Newton potential φ have the following relation

$$g_{00}=-\left(1+\frac{2}{c^2}\varphi\right) \tag{7.1.14}$$

Hence, by (7.1.13) and (7.1.14) for the ball we have

$$g_{00}=-1+\frac{2mG}{c^2r}. \tag{7.1.15}$$

Now we consider the gravitational field equations in the exterior of the ball. In the vacuum state,

$$T_{\mu\nu}=0. \tag{7.1.16}$$

On the other hand, by $R=g^{\mu\nu}R_{\mu\nu}$, the Einstein gravitational field equations

$$R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R=-\frac{8\pi G}{c^4}T_{\mu\nu}$$

can be equivalently written as

$$R_{\mu\nu}=-\frac{8\pi G}{c^4}(T_{\mu\nu}-\frac{1}{2}g_{\mu\nu}T), \quad T=g^{kl}T_{kl}.$$

Thus, by (7.1.16) the Einstein field equations become

$$R_{\mu\nu}=0 \tag{7.1.17}$$

The nonzero components of the metric (7.1.12) $g_{\mu\nu}$ are

$$g_{00}=-e^u, \quad g_{11}=e^v, \quad g_{22}=r^2, \quad g_{33}=r^2\sin^2\theta. \tag{7.1.18}$$

Since the gravitational resource is static, u and v only depend on r. By (7.1.18), all nonzero components of the Levi-Civita connection (7.1.5) are given by

$$\begin{aligned}
&\Gamma^1_{00}=\frac{1}{2}e^{u-v}u', && \Gamma^1_{11}=\frac{1}{2}v', && \Gamma^1_{22}=-re^{-v},\\
&\Gamma^1_{33}=-re^{-v}\sin^2\theta, && \Gamma^0_{10}=\frac{1}{2}u', && \Gamma^2_{12}=\frac{1}{r},\\
&\Gamma^2_{33}=-\sin\theta\cos\theta, && \Gamma^3_{13}=\frac{1}{r}, && \Gamma^3_{23}=\frac{\cos\theta}{\sin\theta}.
\end{aligned} \tag{7.1.19}$$

Thus, by (7.1.4) and (7.1.19) we get $R_{\mu\nu}$ as

$$\begin{aligned}
&R_{00} = e^{u-v}\left[-\frac{1}{2}u'' - \frac{1}{r}u' + \frac{1}{4}u'(v'-u')\right],\\
&R_{11} = \frac{1}{2}u'' - \frac{1}{r}v' - \frac{1}{4}u'(v'-u'),\\
&R_{22} = e^{-v}\left[1 - e^{v} + \frac{r}{2}(u'-v')\right],\\
&R_{33} = R_{22}\sin^2\theta,\\
&R_{\mu\nu} = 0, \qquad \forall \mu \neq \nu.
\end{aligned} \tag{7.1.20}$$

Therefore, the vacuum Einstein field equations (7.1.17) become the following system of ordinary differential equations

$$\frac{1}{2}u'' + \frac{1}{r}u' - \frac{1}{4}(v'-u')u' = 0, \tag{7.1.21}$$

$$\frac{1}{2}u'' - \frac{1}{r}v' - \frac{1}{4}(v'-u')u' = 0, \tag{7.1.22}$$

$$\frac{r}{2}(u'-v') - e^{v} + 1 = 0. \tag{7.1.23}$$

By the Bianchi identity, only two equations of (7.1.22)-(7.1.23) are independent. The difference of (7.1.21) and (7.1.22) leads to

$$u' + v' = 0,$$

which implies that

$$u + v = \beta \quad \text{(constant)}, \tag{7.1.24}$$

and (7.1.23) becomes

$$e^{v} + rv' - 1 = 0.$$

Namely

$$\frac{d}{dr}(re^{-v}) = 1.$$

It follows that

$$e^{-v} = 1 - \frac{b}{r},$$

where b is a to-be-determined constant.

Then it follows from (7.1.24) that

$$e^{u} = e^{\beta}e^{-v} = e^{\beta}\left(1 - \frac{b}{r}\right).$$

By scaling time t, we can take $e^{\beta} = 1$. Hence the solution of the Einstein field equations (7.1.17) is given by

$$\begin{aligned} g_{00} &= -e^{u} = -\left(1 - \frac{b}{r}\right), \\ g_{11} &= e^{v} = \left(1 - \frac{b}{r}\right)^{-1}. \end{aligned} \tag{7.1.25}$$

Comparing (7.1.25) with (7.1.15), we can obtain

$$b = \frac{2mG}{c^2}.$$

Thus, we get the solution of (7.1.17) as

$$g_{00} = -\left(1 - \frac{2mG}{c^2 r}\right),$$

$$g_{11} = \left(1 - \frac{2mG}{c^2 r}\right)^{-1},$$

and the metric reads

$$ds^2 = -\left(1 - \frac{2mG}{c^2 r}\right)c^2 dt^2 + \left(1 - \frac{2mG}{c^2 r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2, \tag{7.1.26}$$

We have in particular

$$e^{-v} = e^{u} = 1 - \frac{2mG}{c^2 r}.$$

which is called the Schwarzschild solution or metric.

TOV metric

The Schwarzschild metric (7.1.26) describes the exterior gravitational fields of a centrally symmetric ball. For the interior gravitational fields, the metric is given by the TOV solution.

Let m be the mass of a centrally symmetric ball, and R be the radius of this ball. In the interior of the ball, the variable r satisfies $0 \leqslant r < R$. Let the ball be a static liquid sphere consisting of idealized fluid, an approximation of stars. The energy-momentum tensor of an idealized fluid is in the form

$$T^{\mu\nu} = (c^2\rho + p)u^{\mu}u^{\nu} + pg^{\mu\nu},$$

where p is the pressure, ρ is the density, and u^{μ} is the 4-velocity. For a static fluid, u^{μ} is given by

$$u^{\mu} = \frac{1}{\sqrt{-g_{00}}}(1,0,0,0).$$

Hence, the (1,1)-type of the energy-momentum tensor is in the form

$$T_\mu^\nu = \begin{pmatrix} -\rho^2 & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}.$$

The Einstein gravitational field equations of an interior ball read

$$\begin{aligned} & R_\mu^\nu - \frac{1}{2}\delta_\mu^\nu R = -\frac{8\pi G}{c^4} T_\mu^\nu, \\ & D_\nu T_\mu^\nu = 0. \end{aligned} \tag{7.1.27}$$

By (7.1.20) we have

$$R = g^{\mu\nu} R_{\mu\nu} = -e^{-u} R_{00} + e^{-\nu} R_{11} + \frac{2}{r^2} R_{22}.$$

Then, by (7.1.20) and $R_\mu^\nu = g^{\nu\alpha} R_{\alpha\mu}$ we get

$$\begin{aligned} & R_0^0 - \frac{1}{2}R = -e^{-\nu}\left(\frac{1}{r^2} - \frac{\nu'}{r}\right) + \frac{1}{r^2}, \\ & R_1^1 - \frac{1}{2}R = -e^{-\nu}\left(\frac{u'}{r} + \frac{1}{r^2}\right) + \frac{1}{r^2}, \\ & R_2^2 - \frac{1}{2}R = -\frac{1}{2}e^{-\nu}\left(u'' - \frac{1}{2}u'\nu' + \frac{1}{2}u'^2 + \frac{1}{r}u' - \frac{1}{r}\nu'\right) \end{aligned}$$

In addition, we know that

$$\begin{aligned} & D_\alpha T_\mu^\alpha = D_0 T_\mu^0 + D_1 T_\mu^1 + D_2 T_\mu^2 + D_3 T_\mu^3, \\ & D_\alpha T_\mu^\beta = \frac{\partial T_\mu^\beta}{\partial x^\alpha} + \Gamma_{\alpha\nu}^\beta T_\mu^\nu - \Gamma_{\alpha\mu}^\nu T_\nu^\beta. \end{aligned}$$

Thus, by (7.1.19) we have

$$\begin{aligned} & D_\alpha T_1^\alpha = \frac{dp}{dr} + \frac{1}{2}(p + c^2\rho)u', \\ & D_\alpha T_\mu^\alpha = 0 \qquad \text{for } \mu \neq 1. \end{aligned}$$

Hence, the field equations (7.1.27) can be written as

$$e^{-\nu}\left(\frac{1}{r^2} - \frac{\nu'}{r}\right) - \frac{1}{r^2} = -\frac{8\pi G}{c^2}\rho, \tag{7.1.28}$$

$$e^{-\nu}\left(\frac{1}{r^2} + \frac{u'}{r}\right) - \frac{1}{r^2} = \frac{8\pi G}{c^4} p, \tag{7.1.29}$$

$$e^{-\nu}(u'' - \frac{1}{2}u'\nu' + \frac{1}{2}u'^2 + \frac{1}{r}u' - \frac{1}{r}\nu') = \frac{16\pi G}{c^4} p, \tag{7.1.30}$$

$$p' + \frac{1}{2}(p + c^2\rho)u' = 0. \tag{7.1.31}$$

By the Bianchi identity, only three equations of (7.1.28)-(7.1.31) are independent. Here we also regard p and ρ as unknown functions. Therefore, for the four unknown functions u, v, p, ρ, we have to add an equation of state to the system of (7.1.28)-(7.1.31):

$$\rho = f(p), \tag{7.1.32}$$

and the function f will be given according to physical conditions.

On the surface $r = R$ of the ball, $p = 0$ and u and v are given in terms of the Schwarzschild solution:

$$p(R) = 0, \quad u(R) = -v(R) = \ln\left(1 - \frac{2Gm}{Rc^2}\right). \tag{7.1.33}$$

We are now in position to discuss the solutions of problem (7.1.28)-(7.1.33). Let

$$M(r) = \frac{c^2 r}{2G}(1 - e^{-v}). \tag{7.1.34}$$

Then the equation (7.1.28) can be rewritten as

$$\frac{1}{r^2}\frac{dM}{dr} = 4\pi\rho,$$

whose solution is given by

$$M(r) = \int_0^r 4\pi r^2 \rho dr \qquad \text{for } 0 < r < R. \tag{7.1.35}$$

By (7.1.35), we see that $M(r)$ is the mass, contained in the ball B_r. It follows from (7.1.34) that

$$e^{-v} = 1 - \frac{2GM(r)}{c^2 r}. \tag{7.1.36}$$

Inserting (7.1.36) in (7.1.29) we obtain

$$u' = \frac{1}{r(c^2 r - 2MG)}\left[\frac{8\pi G}{c^2} p r^3 + 2GM(r)\right]. \tag{7.1.37}$$

Putting (7.1.37) into (7.1.31) we get

$$p' = -\frac{p + c^2\rho}{2r(c^2 r - 2MG)}\left[\frac{8\pi G}{c^2} p r^3 + 2GM(r)\right]. \tag{7.1.38}$$

Thus, it suffices for us to derive the solution p, M and ρ from (7.1.32)-(7.1.34) and (7.1.38), and then v and u will follow from (7.1.36)-(7.1.37) and (7.1.33).

The equation (7.1.38) is called the TOV equation, which was derived to describe the structure of neutron stars.

We note that (7.1.36) is the interior metric of a blackhole provided that $2GM(r)/(c^2 r) = 1$. Thus the TOV solution (7.1.36) gives a rigorous proof of the following theorem for the existence of black holes.

Theorem 7.3 *If the matter field in a ball B_R of radius R is spherically symmetric, and the mass M_R and the radius R satisfy*

$$\frac{2GM_R}{c^2R} = 1,$$

then the ball must be a blackhole.

An idealized model is that the density is homogeneous, i.e. (7.1.32) is given by

$$\rho = \rho_0 \quad \text{a constant.}$$

In this case, we have

$$M(r) = \frac{4\pi}{3}\rho_0 r^3 \qquad \text{for } 0 \leqslant r \leqslant R,$$
$$\rho_0 = \frac{3}{4\pi}\frac{m}{R^3}.$$

Thus we obtain the following solution of (7.1.36)-(7.1.38) with (7.1.33):

$$p(r) = \rho_0\left[\frac{\left(1-\frac{2Gmr^2}{c^2R^3}\right)^{1/2} - \left(1-\frac{2Gm}{c^2R}\right)^{1/2}}{3\left(1-\frac{2Gm}{c^2R}\right)^{1/2} - \left(1-\frac{2Gmr^2}{c^2R^3}\right)^{1/2}}\right], \tag{7.1.39}$$

$$e^u = \left[\frac{3}{2}\left(1-\frac{2Gm}{c^2R}\right)^{1/2} - \frac{1}{2}\left(1-\frac{2Gmr^2}{c^2R^3}\right)^{1/2}\right]^2, \tag{7.1.40}$$

$$e^v = \left[1-\frac{2Gmr^2}{c^2R^3}\right]^{-1}. \tag{7.1.41}$$

The functions (7.1.39)-(7.1.41) are the TOV solution. By (7.1.12), the solution (7.1.40) and (7.1.41) yields the metric

$$\begin{aligned} ds^2 = &-\left[\frac{3}{2}\left(1-\frac{2Gm}{c^2R}\right)^{1/2} - \frac{1}{2}\left(1-\frac{2Gmr^2}{c^2R^3}\right)^{1/2}\right]^2 c^2dt^2 \\ &+\left[1-\frac{2Gmr^2}{c^2R^3}\right]^{-1} dr^2 + r^2(d\theta^2+\sin^2\theta d\varphi^2), \end{aligned} \tag{7.1.42}$$

which is called the TOV metric.

7.1.3 Differential operators in spherical coordinates

In Subsection 7.1.1, we gave the Navier-Stokes equations on general Riemannian manifolds. For astrophysical fluid dynamics, we mainly concern the equations on $3D$ spheres. Hence in this subsection we discuss the basic differential operators (7.1.2)-(7.1.8) under the spherical coordinate systems (θ,φ,r).

For a $3D$ sphere M, the Riemannian metric is given by

$$ds^2=\alpha(r)dr^2+r^2d\theta^2+r^2\sin^2\theta d\varphi^2 \tag{7.1.43}$$

where $\alpha(r)>0$ represents the relativistic effects:

$$\alpha=\begin{cases}1 & \text{no relativistic effect,}\\ \left(1-\dfrac{2Gm}{c^2r}\right)^{-1} & \text{for the Schwarzschild metric (7.1.26),}\\ \left(1-\dfrac{2Gmr^2}{c^2R^3}\right)^{-1} & \text{for the TOV metric (7.1.42).}\end{cases} \tag{7.1.44}$$

In (7.1.43) we have

$$g_{11}=\alpha(r),\qquad g_{22}=r^2,\qquad g_{33}=r^2\sin^2\theta,\qquad g_{ij}=0 \text{ for } i\neq j.$$

By (7.1.5) we can get the Levi-Civita connection as

$$\begin{aligned}&\Gamma^2_{21}=\Gamma^2_{12}=\frac{1}{r}, && \Gamma^2_{33}=-\sin\theta\cos\theta, && \Gamma^3_{31}=\Gamma^3_{13}=\frac{1}{r},\\ &\Gamma^3_{32}=\Gamma^3_{32}=\frac{\cos\theta}{\sin\theta}, && \Gamma^1_{22}=-\frac{r}{\alpha}, && \Gamma^1_{33}=-\frac{r}{\alpha}\sin^2\theta,\\ &\Gamma^1_{11}=\frac{1}{2\alpha}\frac{d\alpha}{dr}, && \Gamma^k_{ij}=0 \ \text{ for others.}\end{aligned} \tag{7.1.45}$$

We deduce from (7.1.45) the explicit form of the Ricci curvature tensor (7.1.4):

$$\begin{aligned}&R_{11}=-\frac{1}{\alpha r}\frac{d\alpha}{dr}, \quad R_{22}=\frac{1}{\alpha}-\frac{r}{2\alpha^2}\frac{d\alpha}{dr}-1,\\ &R_{33}=R_{22}\sin^2\theta, \quad R_{ij}=0 \qquad \forall i\neq j.\end{aligned} \tag{7.1.46}$$

Based on (7.1.45) and (7.1.46) we can obtain the expressions of the differential operators (7.1.2)-(7.1.8) as follows:

1) The Laplace-Beltrami operator $\Delta u^k=(\Delta u_r,\Delta u_\theta,\Delta u_\varphi)$:

$$\begin{aligned}\Delta u_\theta=&\frac{1}{r^2}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u_\theta}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2u_\theta}{\partial\varphi^2}\right.\\ &\left.+\frac{1}{\alpha}\frac{\partial}{\partial r}\left(r^2\frac{\partial u_\theta}{\partial r}\right)+\frac{2}{r}\frac{\partial u_r}{\partial\theta}-\frac{2\cos\theta}{\sin\theta}\frac{\partial u_\varphi}{\partial\varphi}-\frac{1}{\sin^2\theta}u_\theta\right]\\ &+\frac{1}{\alpha r^2}\left[2\frac{\partial}{\partial r}(ru_\theta)-\frac{\alpha'}{2\alpha}\frac{\partial}{\partial r}(r^2u_\theta)\right],\end{aligned} \tag{7.1.47}$$

$$\Delta u_\varphi = \frac{1}{r^2}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u_\varphi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 u_\varphi}{\partial\varphi^2} + \frac{1}{\alpha}\frac{\partial}{\partial r}\left(r^2\frac{\partial u_\varphi}{\partial r}\right)\right. \tag{7.1.48}$$
$$\left. + \frac{2\cos\theta}{\sin\theta}\frac{\partial u_\varphi}{\partial\theta} - 2u_\varphi + \frac{2\cos\theta}{\sin^3\theta}\frac{\partial u_\theta}{\partial\varphi} + \frac{2}{r\sin^2\theta}\frac{\partial u_r}{\partial\varphi}\right]$$
$$+ \frac{1}{\alpha r^2}\left[2\frac{\partial}{\partial r}(ru_\varphi) - \frac{\alpha'}{2\alpha}\frac{\partial}{\partial r}(r^2 u_\varphi)\right],$$

$$\Delta u_r = \frac{1}{r^2}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u_r}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 u_r}{\partial\varphi^2} + \frac{1}{\alpha}\frac{\partial}{\partial r}\left(r^2\frac{\partial u_r}{\partial r}\right)\right] \tag{7.1.49}$$
$$- \frac{2}{\alpha r^2}\left[u_r + r\frac{\cos\theta}{\sin\theta}u_\theta + r\frac{\partial u_\theta}{\partial\theta} + r\frac{\partial u_\varphi}{\partial\varphi} - \frac{r^2}{2}\frac{\partial}{\partial r}\left(\frac{\alpha'}{2\alpha}u_r\right)\right].$$

2) By (7.1.6) and (7.1.45), $(u\cdot\nabla)u^k$ can be written as

$$u^k D_k u_\theta = u_r\frac{\partial u_\theta}{\partial r} + u_\theta\frac{\partial u_\theta}{\partial\theta} + u_\varphi\frac{\partial u_\theta}{\partial\varphi} + \frac{2}{r}u_\theta u_r - \sin\theta\cos\theta u_\varphi^2 \tag{7.1.50}$$

$$u^k D_k u_\varphi = u_r\frac{\partial u_\varphi}{\partial r} + u_\theta\frac{\partial u_\varphi}{\partial\theta} + u_\varphi\frac{\partial u_\varphi}{\partial\varphi} + \frac{2\cos\theta}{\sin\theta}u_\theta u_\varphi + \frac{2}{r}u_\varphi u_r, \tag{7.1.51}$$

$$u^k D_k u_r = u_r\frac{\partial u_r}{\partial r} + u_\theta\frac{\partial u_r}{\partial\theta} + u_\varphi\frac{\partial u_r}{\partial\varphi} - \frac{r}{\alpha}(u_\theta^2 + \sin^2\theta u_\varphi^2 - \frac{\alpha'}{2r}u_r^2). \tag{7.1.52}$$

3) The gradient operator:

$$\nabla p = \left(\frac{1}{\alpha}\frac{\partial p}{\partial r}, \frac{1}{r^2}\frac{\partial p}{\partial\theta}, \frac{1}{r^2\sin^2\theta}\frac{\partial p}{\partial\varphi}\right). \tag{7.1.53}$$

4) By (7.1.8) and $\sqrt{g} = r^2\sin\theta\sqrt{\alpha}$, the divergent operator div$u$ is

$$\text{div}u = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta u_\theta) + \frac{\partial u_\varphi}{\partial\varphi} + \frac{1}{r^2\sqrt{\alpha}}\frac{\partial}{\partial r}(r^2\sqrt{\alpha}u_r). \tag{7.1.54}$$

Remark 7.4 The expressions (7.1.47)-(7.1.54) are the differential operators appearing in the fluid dynamic equations describing the stellar fluids. However, we need to note that the two components u_θ and u_φ are the angular velocities of θ and φ, i.e.

$$u_\theta = \frac{d\theta}{dt}, \quad u_\varphi = \frac{d\varphi}{dt}.$$

In classical fluid dynamics, the velocity field $v = (v_\theta, v_\varphi, v_r)$ is the line velocity. The relation of u and v is given by

$$u_\theta = \frac{1}{r}v_\theta, \quad u_\varphi = \frac{1}{r\sin\theta}v_\varphi, \quad u_r = v_r. \tag{7.1.55}$$

Hence, inserting (7.1.55) into (7.1.1) with the expressions (7.1.47)-(7.1.54), we derive the Navier-Stokes equations in the usual spherical coordinate form as follows

$$\begin{aligned}
&\frac{\partial v_\theta}{\partial t}+(u\cdot\nabla)v_\theta=\nu\Delta v_\theta-\frac{1}{\rho r}\frac{\partial p}{\partial\theta}+f_\theta,\\
&\frac{\partial v_\varphi}{\partial t}+(u\cdot\nabla)v_\varphi=\nu\Delta v_\varphi-\frac{1}{\rho r\sin\theta}\frac{\partial p}{\partial\varphi}+f_\varphi,\\
&\frac{\partial v_r}{\partial t}+(u\cdot\nabla)v_r=\nu\Delta v_r-\frac{1}{\rho\alpha}\frac{\partial p}{\partial r}+f_r,\\
&\text{div } v=0,
\end{aligned}\tag{7.1.56}$$

where

$$\begin{aligned}
\Delta v_\theta=&\widetilde{\Delta}v_\theta+\frac{2}{r^2}\frac{\partial v_r}{\partial\theta}-\frac{2\cos\theta}{r^2\sin^2\theta}\frac{\partial v_\varphi}{\partial\varphi}-\frac{v_\theta}{r^2\sin^2\theta}-\frac{1}{2\alpha^2 r}\frac{d\alpha}{dr}\frac{\partial}{\partial r}(rv_\theta),\\
\Delta v_\varphi=&\widetilde{\Delta}v_\varphi+\frac{2}{r^2\sin\theta}\frac{\partial v_r}{\partial\varphi}+\frac{2\cos\theta}{r^2\sin^2\theta}\frac{\partial v_\theta}{\partial\varphi}-\frac{v_\varphi}{r^2\sin^2\theta}-\frac{1}{2\alpha^2 r}\frac{d\alpha}{dr}\frac{\partial}{\partial r}(rv_\varphi),\\
\Delta v_r=&\widetilde{\Delta}v_r-\frac{2}{\alpha r^2}\left(v_r+\frac{\partial v_\theta}{\partial\theta}+\frac{\cos\theta}{\sin\theta}v_\theta+\frac{1}{\sin\theta}\frac{\partial v_\varphi}{\partial\varphi}\right)+\frac{1}{2\alpha}\frac{\partial}{\partial r}\left(\frac{1}{\alpha}\frac{d\alpha}{dr}v_r\right),
\end{aligned}\tag{7.1.57}$$

$\widetilde{\Delta}$ is the Laplace operator for scalar fields given by

$$\widetilde{\Delta}T=\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial T}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 T}{\partial\varphi^2}+\frac{1}{\alpha r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial T}{\partial r}\right)-\frac{\alpha'}{2r\alpha^2}\frac{\partial T}{\partial r},\tag{7.1.58}$$

the nonlinear term $(u\cdot\nabla)v$ is

$$\begin{aligned}
(v\cdot\nabla)v_\theta=&\frac{v_\theta}{r}\frac{\partial v_\theta}{\partial\theta}+\frac{v_\varphi}{r\sin\theta}\frac{\partial v_\theta}{\partial\varphi}+v_r\frac{\partial v_\theta}{\partial r}+\frac{v_\theta v_r}{r}-\frac{\cos\theta v_\varphi^2}{r\sin\theta},\\
(v\cdot\nabla)v_\varphi=&\frac{v_\theta}{r}\frac{\partial v_\varphi}{\partial\theta}+\frac{v_\varphi}{r\sin\theta}\frac{\partial v_\varphi}{\partial\varphi}+v_r\frac{\partial v_\varphi}{\partial r}+\frac{v_\varphi v_r}{r}+\frac{\cos\theta v_\varphi v_\theta}{r\sin\theta},\\
(v\cdot\nabla)v_r=&\frac{v_\theta}{r}\frac{\partial v_r}{\partial\theta}+\frac{v_\varphi}{r\sin\theta}\frac{\partial v_r}{\partial\varphi}+v_r\frac{\partial v_r}{\partial r}-\frac{1}{\alpha r}(v_\theta^2+v_\varphi^2)+\frac{1}{2\alpha}\frac{d\alpha}{dr}v_r^2,
\end{aligned}\tag{7.1.59}$$

and the divergent term div v reads

$$\text{div }\ v=\frac{1}{r\sin\theta}\frac{\partial(\sin\theta v_\theta)}{\partial\theta}+\frac{1}{r\sin\theta}\frac{\partial v_\varphi}{\partial\varphi}+\frac{1}{r^2\sqrt{\alpha}}\frac{\partial(r^2\sqrt{\alpha}v_r)}{\partial r}.\tag{7.1.60}$$

7.1.4 Momentum representation

The Universe, galaxies and galactic clusters are composed of stars and interstellar nebulae. Their velocity fields are not continuous. Hence it is not appropriate that we model cosmic objects using continuous velocity field $v(x,t)$ as in the Navier-Stoks equations or by discrete position variables $x_k(t)$ as in the N-body problem.

The idea is that we use the momentum density field $P(x,t)$ to replace the velocity field $v(x,t)$ as the state function of cosmic objects. The main reason is that the momentum density field P is the energy flux containing the mass, the heat, and all interaction energy flux, and can be regarded as a continuous field. The aim of this section is to establish the momentum form of astrophysical fluid dynamics model.

The physical laws governing the dynamics of cosmic objects are as follows

$$\begin{aligned}&\text{Theory of General Relativity,}\\&\text{Newtonian Second Law,}\\&\text{Heat Conduction Law,}\\&\text{Energy-Momentum Conservation,}\\&\text{Equation of State.}\end{aligned} \tag{7.1.61}$$

The mathematical expressions of these laws are given respectively in the following:

1. *Gravitational field equations.*

$$\begin{aligned}&R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R=-\frac{8\pi G}{c^4}T_{\mu\nu}+\frac{1}{2}(\widetilde{D}_\mu\Phi_\nu+\widetilde{D}_\nu\Phi_\mu),\\&\widetilde{D}_\mu=D_\mu+\frac{e}{\hbar c}A_\mu,\end{aligned} \tag{7.1.62}$$

where A_μ is the electromagnetic potential, the time components of $g_{\mu\nu}$ are as

$$g_{00}=-\left(1+\frac{2}{c^2}\psi\right),\quad g_{0k}=g_{k0}=0 \qquad \text{for } 1\leqslant k\leqslant 3,$$

and ψ is the gravitational potential.

2. *Fluid dynamic equations.* The Newton's Second Law can be expressed as

$$\frac{dP}{d\tau}=\text{ Force,} \tag{7.1.63}$$

where τ is the proper time given by

$$d\tau=\sqrt{-g_{00}}dt, \tag{7.1.64}$$

P is the momentum density field, formally defined by

$$\frac{dx}{d\tau}=\frac{1}{\rho}P,$$

with ρ being the energy density,

$$\frac{dP}{d\tau}=\frac{\partial P}{\partial\tau}+\frac{\partial P}{\partial x^k}\frac{dx^k}{d\tau}=\frac{\partial P}{\partial\tau}+\frac{1}{\rho}(P\cdot\nabla)P,$$

and

$\nu\Delta P+\mu\nabla(\text{div}P)$	the frictional force,
$-\nabla p$	the pressure gradient,
$\frac{c^2}{2}\rho(1-\beta T)\nabla g_{00}=-\rho(1-\beta T)\nabla\psi$	the gravitational force.

Hence, the momentum form of the fluid dynamic equations (7.1.63) is written as

$$\frac{\partial P}{\partial\tau}+\frac{1}{\rho}(P\cdot\nabla)P=\nu\Delta P+\mu\nabla(\text{div}P)-\nabla p-\rho(1-\beta T)\nabla\psi, \tag{7.1.65}$$

where the differential operators Δ,∇ and $(P\cdot\nabla)$ are with respect to the space metric g_{ij} $(1\leqslant i,j\leqslant 3)$ determined by (7.1.62), as defined in (7.1.2)-(7.1.8).

3. *Heat conduction equation:*

$$\frac{\partial T}{\partial\tau}+\frac{1}{\rho}(P\cdot\nabla)T=\kappa\widetilde{\Delta}T+Q, \tag{7.1.66}$$

where $\widetilde{\Delta}$ is defined as

$$\widetilde{\Delta}T=-\frac{1}{\sqrt{g}}\frac{\partial}{\partial x^i}\left(\sqrt{g}g^{ij}\frac{\partial T}{\partial x^j}\right),$$

and $g=\ \det(g_{ij}),1\leqslant i,j\leqslant 3.$

4. *Energy-momentum conservation:*

$$\frac{\partial\rho}{\partial\tau}+\ \text{div}P=0, \tag{7.1.67}$$

where ρ is the energy density:

$$\rho=\text{ mass + electromagnetism + potential + heat.}$$

5. *Equation of state:*

$$p=f(\rho,T). \tag{7.1.68}$$

Remark 7.5 Both physical laws (7.1.63) and (7.1.67) are the more general form than the classical ones:

$$\begin{aligned} &m\frac{dv}{d\tau}=\text{Force} && \text{the Newton's Second Law,}\\ &\frac{\partial m}{\partial\tau}+\ \text{div}(mv)=0 && \text{the continuity equation,}\end{aligned} \tag{7.1.69}$$

where m is the mass density. Hence the momentum representation equations (7.1.65)-(7.1.67) can be applicable in general. The momentum P represents the energy density flux, consisting essentially of

$$P=mv+\ \text{radiation flux+heat flux.}$$

Hence in astrophysics, the momentum density P is a better candidate than the velocity field v, to serve as the continuous-media type of state function. □

7.1.5 Astrophysical Fluid Dynamics Equations

Dynamic equations of stellar atmosphere

Different from planets, stars are fluid spheres. Like the Sun, most of stars possess atmospheric layers. The atmospheric dynamics of stars is an important topic, and we are now ready to present the stellar atmospheric model.

The spatial domain is a spherical shell:

$$\mathscr{M} = \{x \in \mathbb{R}^3 \mid r_0 < r < r_1\}.$$

The stellar atmosphere consists of rarefied gas. For example, the solar corona has mass density about $\rho_m = 10^{-9}\rho_0$ where ρ_0 is the density of the earth atmosphere. Hence we use the Schwarzschild solution in (7.1.44) as the metric:

$$\alpha(r) = \left(1 - \frac{2mG}{c^2 r}\right)^{-1}, \quad r_0 > \frac{2mG}{c^2}. \tag{7.1.70}$$

where m is the total mass of the star, and the condition $r_0 > 2mG/c^2$ ensures that the star is not a black hole.

The stellar atmospheric model is the momentum form of the astrophysical fluid dynamical equations defined on the spherical shell $\mathscr{M}$:

$$\begin{aligned}
&\frac{\partial P}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)P = \nu\Delta P + \mu\nabla(\text{div}P) - \nabla p - \frac{mG\rho}{r^2}(1-\beta T)\vec{k},\\
&\frac{\partial T}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)T = \kappa\widetilde{\Delta}T,\\
&\frac{\partial \rho}{\partial \tau} + \text{div}P = 0,
\end{aligned} \tag{7.1.71}$$

where $P = (P_r, P_\theta, P_\varphi)$ is the momentum density field, T is the temperature, p is the pressure, ρ is the energy density, ν and μ is the viscosity coefficient, β is the coefficient of thermal expansion, κ is the thermal diffusivity, α is as in (7.1.70), $\Delta P, (P\cdot\nabla)P, \widetilde{\Delta}T, \text{div}\, P$ are as in (7.1.57)-(7.1.60), and

$$(P\cdot\nabla)T = \frac{P_\theta}{r}\frac{\partial T}{\partial \theta} + \frac{P_\varphi}{r\sin\theta}\frac{\partial T}{\partial \varphi} + P_r\frac{\partial T}{\partial r}. \tag{7.1.72}$$

The equations (7.1.71) are supplemented with the boundary conditions:

$$\begin{aligned}
&P_r = 0, \quad \frac{\partial P_\theta}{\partial r} = 0, \quad \frac{\partial P_\varphi}{\partial r} = 0 && \text{at } r = r_0, r_1,\\
&T = T_0 && \text{at } r = r_0,\\
&T = T_1 && \text{at } r = r_1,
\end{aligned} \tag{7.1.73}$$

where T_0 and T_1 are approximatively taken as constants and satisfy the physical condition

$$T_0 > T_1.$$

A few remarks are now in order:

Remark 7.6 First, there are three important parameters: the Rayleigh number Re, the Prandtl number Pr and the δ-factor δ, which play an important role in astrophysical fluid dynamics:

$$\mathrm{Re} = \frac{mGr_0r_1\beta}{\kappa\nu}\frac{T_0 - T_1}{h}, \qquad \mathrm{Pr} = \frac{\nu}{\kappa}, \qquad \delta = \frac{2mG}{c^2r_0}. \tag{7.1.74}$$

The δ-factor δ reflects the relativistic effect contained the Laplacian operator. □

Remark 7.7 Astronomic observations show that the Sun has three layers of atmospheres: the photosphere, the chromosphere, and the solar corona, where the solar atmospheric convections occur. It manifests that the thermal convection is a universal phenomenon for stellar atmospheres. In the classical fluid dynamics, the Rayleigh number dictates the Rayleign-Bénard convection. Here, however, both the Rayleigh number Re and the δ-factor defined by (7.1.74) play an important role in stellar atmospheric convections. □

Remark 7.8 For rotating stars with angular velocity $\vec{\Omega}$, we need add to the right hand side of (7.1.71) the Coriolis term:

$$-2\vec{\Omega}\times P = 2\Omega(\sin\theta P_r - \cos\theta P_\theta, \cos\theta P_\varphi, -\sin\theta P_\varphi),$$

where Ω is the magnitude of $\vec{\Omega}$. □

Fluid dynamical equations inside open balls

As the fluid density in a stellar atmosphere is small, the equations (7.1.71) can be regarded as a precise model governing the stellar atmospheric motion. However, for a fluid sphere with high density, the fluid dynamic equations have to couple the gravitational field equations.

The Universe and all stars are in the momentum-flow state, i.e. they are fluid spheres. To investigate the interiors of the Universe, galaxies and stars, we need to develop dynamic models for fluid spheres.

The precise equations of fluid sphere should be defined in the Riemannian metric space as follows:

$$ds^2 = g_{00}(x,t)c^2dt^2 + g_{ij}(x,t)dx^idx^j \qquad \text{for } x \in \mathscr{M}^3, \tag{7.1.75}$$

where $\mathscr{M}^3$ is the spherical space. The gravitational field equations are expressed as

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} - D_\mu\Phi_\mu, \quad (g_{j0} = g_{0j} = 0, 1 \leqslant j \leqslant 3), \tag{7.1.76}$$

where

$$T_{\mu\nu} = g_{\mu\alpha} g_{\nu\beta} \left[\varepsilon^\alpha \varepsilon^\beta + p g^{\alpha\beta} \right],$$

and ε^μ is the 4D energy-momentum vector.

For the fluid component of the system, it is necessary to simplify the model by making some physically sound approximations.

Approximation Hypothesis 7.9 *The metric (7.1.75) and the stationary solutions of the fluid dynamical equations are radially symmetric.*

Under Hypothesis 7.9, the metric (7.1.75) is as in (7.1.12), or is written in the following form

$$ds^2 = -\psi(r) c^2 dt^2 + \alpha(r) dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2, \tag{7.1.77}$$

and the fluid dynamic equations are rewritten as

$$\begin{aligned}
&\frac{\partial P}{\partial \tau} + \frac{1}{\rho}(P \cdot \nabla) P = \nu \Delta P + \mu \nabla (\text{div} P) - \nabla p - \frac{c^2 \rho}{2\alpha} \frac{d\psi}{dr} (1 - \beta(T - T_0)) \vec{k}, \\
&\frac{\partial T}{\partial \tau} + \frac{1}{\rho}(P \cdot \nabla) T = \kappa \widetilde{\Delta} T + Q(r), \\
&\frac{\partial \rho}{\partial \tau} + \text{div} P = 0,
\end{aligned} \tag{7.1.78}$$

where $P = (P_r, P_\theta, P_\varphi), \nabla P, \widetilde{\Delta} T, (P \cdot \nabla) P,$ div P are as in (7.1.57)-(7.1.60), ∇p is as in (7.1.56), $\vec{k} = (1, 0, 0)$, and

$$(P \cdot \nabla) T = P_r \frac{\partial T}{\partial r} + \frac{P_\theta}{r} \frac{\partial T}{\partial \theta} + \frac{P_\varphi}{r \sin\theta} \frac{\partial T}{\partial \varphi}.$$

The gravitational field equation (7.1.76) for the metric (7.1.77) is radially symmetric, therefore

$$\Phi_\nu = D_\nu \phi, \qquad \phi = \phi(r).$$

Thus we have

$$\begin{aligned}
&D_0 D^0 \phi = \frac{1}{2\alpha\psi} \psi' \phi', && D_1 D^1 \phi = \frac{1}{\alpha} \phi'' - \frac{1}{2\alpha^2} \alpha' \phi', \\
&D_2 D^2 \phi = D_3 D^3 \phi = \frac{1}{r\alpha} \phi', && D_\mu D^\nu \phi = 0 \qquad \text{for } \mu \neq \nu.
\end{aligned}$$

Then, in view of (7.1.28)-(7.1.31), the equation (7.1.76) can be expressed by

$$\begin{aligned}
&\frac{1}{\alpha}\left(\frac{1}{r^2} - \frac{\alpha'}{r\alpha}\right) - \frac{1}{r^2} = -\frac{8\pi G}{c^2} \rho_0 + \frac{1}{2\alpha\psi} \psi' \phi', \\
&\frac{1}{\alpha}\left(\frac{1}{r^2} + \frac{\psi'}{r\psi}\right) - \frac{1}{r^2} = \frac{8\pi G}{c^2} p + \frac{1}{\alpha} \phi'' - \frac{1}{2\alpha^2} \alpha' \phi', \\
&\frac{1}{\alpha}\left[\frac{\psi''}{\psi} - \frac{1}{2}\left(\frac{\psi'}{\psi}\right)^2 - \frac{\alpha'\psi'}{2\alpha\psi} + \frac{1}{r}\left(\frac{\psi'}{\psi} - \frac{\alpha'}{\alpha}\right)\right] = \frac{16\pi G}{c^2} p + \frac{2}{r\alpha} \phi',
\end{aligned} \tag{7.1.79}$$

where the pressure p satisfies the stationary equations of (7.1.78) with $P=0$ as follows

$$\begin{aligned} & p' = -\frac{c^2}{2}\psi'\rho\,[1-\beta(T-T_0)], \\ & \frac{\kappa}{\alpha r^2}\frac{d}{dr}\left(r^2\frac{dT}{dr}\right) = -Q(r). \end{aligned} \tag{7.1.80}$$

The functions ψ and α satisfy the boundary conditions (7.1.33), i.e.

$$\psi(r_0) = 1-\frac{2Gm}{c^2 r_0}, \quad \alpha(r_0) = \left(1-\frac{2Gm}{c^2 r_0}\right)^{-1}. \tag{7.1.81}$$

In addition, for the ordinary differential equations (7.1.79)-(7.1.81), we also need the boundary conditions for ψ', ϕ' and T. Since $-\frac{1}{2}c^2\psi'$ represents the gravitational force, the condition of ψ' at $r=r_0$ is given by

$$\psi'(r_0) = \frac{2mG}{c^2 r_0^2}. \tag{7.1.82}$$

Based on the Newton gravitational law, ϕ' is very small in the external sphere; also see (Ma and Wang, 2014e). Hence we can approximatively take that

$$\phi'(r_0) = 0. \tag{7.1.83}$$

Finally, it is rational to take the temperature gradient in the boundary condition as follows

$$\frac{\partial T}{\partial r}(r_0) = -A \qquad (A>0). \tag{7.1.84}$$

Let the stationary solution of the problem (7.1.79)-(7.1.81) be given by $\widetilde{p}, \widetilde{T}, \psi, \alpha, \phi'$. Make the translation transformation

$$P \to P, \quad p \to p+\widetilde{p}, \quad T \to T+\widetilde{T}.$$

Then equations (7.1.78) are rewritten in the form

$$\begin{aligned} & \frac{\partial P}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)P = \nu\Delta P - \nabla p + \frac{c^2\rho}{2\alpha}\frac{d\psi}{dr}\beta\vec{k}T, \\ & \frac{\partial T}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)T = \kappa\widetilde{\Delta}T - \frac{1}{\rho}\frac{d\widetilde{T}}{dr}P_r, \\ & \text{div}\, P = 0, \end{aligned} \tag{7.1.85}$$

supplemented with the boundary conditions:

$$\frac{\partial T}{\partial r} = 0, \quad P_r = 0, \quad \frac{\partial P_\theta}{\partial r} = \frac{\partial P_\varphi}{\partial r} = 0 \quad \text{at } r = r_0. \tag{7.1.86}$$

The model (7.1.85)-(7.1.86), we just derived describes interior dynamics of the Universe, galaxies and stars.

7.2 Stars

7.2.1 Basic knowledge

The large scale structure of the Universe consists of mainly the following levers:

stars, stellar clusters, galaxies, clusters of galaxies.

Star is the most elementary constituent of the Universe, and we now explore some of their basic properties.

1. *Mass m.* The mass of the Sun is $m_{\odot} = 2 \times 10^{30}$kg, and the range of the masses of the main-sequence stars is about

$$0.1m_{\odot} \sim 40m_{\odot}.$$

A few extreme stars have masses $m \simeq 60m_{\odot}$, and the least massive stars have masses around $m \simeq 0.07m_{\odot}$.

2. *Radius R.* The radius of the Sun is $R_{\odot} = 7 \times 10^5$km, and the radii of the main-sequence stars are

$$0.3R_{\odot} \sim 25R_{\odot}.$$

A neutron star has radius $R \simeq 10$km, and a red giant star has $R = 10^3 R_{\odot}$.

3. *Surface temperature T.* The surface temperature of the Sun is $T_{\odot} = 5800°K$, and the range of surface temperatures of stars in general is

$$2600°K \sim 35000°\text{K}.$$

4. *Luminosity L.* The Sun's luminosity is $L_{\odot} = 1(4 \times 10^{33}\text{erg}/s)$, and the luminosities of the main-sequence stars have ranges in

$$8 \times 10^{-3} L_{\odot} \sim 3.2 \times 10^5 L_{\odot}.$$

5. *Parameter relation.* Based on the radiation theory of black bodies, the above three parameters R, T, L enjoy

$$L = 4\pi\sigma R^2 T^4, \tag{7.2.1}$$

where σ is the Stefan-Boltzmann constant:

$$\sigma = 5.7 \times 10^{-5}\text{erg}/\text{cm}^2 \cdot \text{s} \cdot \text{k}^4.$$

For the main-sequence stars, we have the following the relation

$$L/L_{\odot} = (m/m_{\odot})^{\alpha}, \tag{7.2.2}$$

where α takes different values as

$$\alpha = \begin{cases} 1.8 & \text{for } m < 0.3m_\odot, \\ 4 & \text{for } 0.3m_\odot < m < 3m_\odot, \\ 2.8 & \text{for } 3m_\odot < m. \end{cases}$$

6. *Main-sequence stars.* In (7.2.1) we see that the luminosity L depends on T and R. However, the radius R is also related with the temperature T. In 1910, two astronomers E. Hertzsprung and H. R. Russell independently obtained the statistical law between L and T. They discovered that for most stars, called the main-sequence stars, their luminosity and temperature possess certain relation. This relation is illustrated by a diagram, called the HR diagram, which is shown in Figure 7.1.

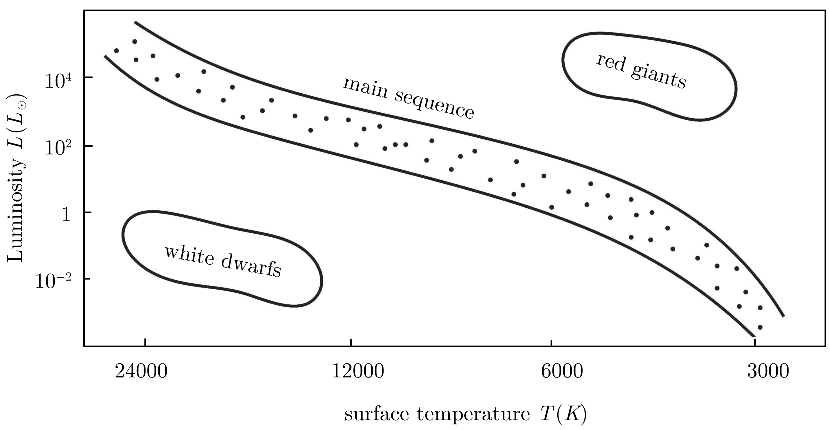


Figure 7.1 The HR diagram in which the luminosity L and the surface temperature T of stars are plotted.

Each star is plotted as a point in the HR diagram in Figure 7.1. The stars in the band are the main-sequence stars. The red giants are in the right-top region, and have lower temperatures and higher brightness, and the white dwarfs are in the left-bottom region, and have higher temperatures and low brightness.

The main-sequence stars are young and middle aged stars, the red giants are old aged, and the white dwarfs are dead stars.

7. *Variable stars.* Most stars shine with almost constant brightness, but a small number of stars, called variable stars, change periodically their brightness. Since the luminosity of a star depends on its radius (see (7.2.1)), variable stars are rhythmically expanding and contracting, pulsating in size and brightness.

8. *Red giants.* Stars are held together by gravity. The gravitational force pulling inward is opposed by a force pushing outward, consisting mainly the thermal pressure in the interior. Stars maintain their balance of pressure and gravity by the heat energy produced by burning hydrogen.

Stars less massive than the Sun evolve more slowly and stay on the main-sequence longer than 10^{10} years. But more massive stars evolve more quickly, and end their life becoming white dwarfs and neutron stars (or possibly black holes) as their final fate determined by their masses.

After a more massive star consumes its central supply of hydrogen, it leaves the main sequence and enters into the group of red giants in the HR diagram.

When the central nuclear reaction has ceased to generate heat, the interior pressure reduces and the star begins to contract, leading to the release of gravitational energy. Hence, this contraction causes the temperature to rise. Thus, the hydrogen outside the core begins to burn. The burning shell causes the star to expend. Thus, the star is luminous, large and cool, and becomes a red giant.

9. *White dwarfs.* For a star with mass in the range

$$1m_\odot \leqslant m < 5m_\odot, \tag{7.2.3}$$

when it is in the red-giant phase, its expanding shell is thrown off into space, and the naked core is all that remains. Contraction ceases, nuclear burning ends, and the core cools down as a white dwarf. A white dwarf has a size approximatively equal to that of the earth, with mass $m \leqslant 1.44M_\odot$.

10. *Neutron stars and pulsars.* A star with mass

$$5m_\odot < m, \tag{7.2.4}$$

will eventually evolve to a neutron star. A neutron star has a radius of about 10km with mass $m > 3m_\odot$, and has a high density of nearly $10^9\text{ton}/\text{cm}^3$. The neutron star has a magnetic field of 10^{12} gauss that is 10^{12} times stronger than the earth's magnetic field.

The pulsar is a special neutron star, which emits a pulse-like energy message.

11. *Supernovae.* When a large massive red giant ($m > 5m_\odot$) exhausts its nuclear fuels, it begins to collapse. This contraction will lead to explosion, and the explosive star is called a supernova. A neutron star is born after a huge explosion of a supernova.

7.2.2 Main driving force for stellar dynamics

Stars can be regarded as fluid balls. To investigate the stellar interior dynamic behavior, we need to use the fluid spherical models coupling the heat conductivity equation. There are two types of starts: stable and unstable. The sizes of stable stars do not change. The main-sequence stars, white dwarfs and neutron stars are stable stars. The radii of unstable stars may change; variable stars and expanding red giants are unstable stars. The dynamic equations governing the two types of stars are different, and will be addressed hereafter separately.

We note that the fluid dynamic equations (7.1.78) represent the Newton's second law, and their left-hand sides are the acceleration and their right-hand sides are the total force. The total force consists of four parts: the viscous friction, the pressure gradient, the relativistic effect, and the thermal expansion force, which are given as follows:

- The *viscous friction force* is caused by the electromagnetic, the weak and the strong interactions between the particles and the pressure, and is given by
$$F_\nu P = \nu \triangle P = \nu(\Delta P_\theta, \Delta P_\varphi, \Delta P_r), \tag{7.2.5}$$
as defined by (7.1.57).
- The *pressure gradient* is defined by:
$$-\nabla p = -\left(\frac{1}{r}\frac{\partial p}{\partial \theta}, \frac{1}{r\sin\theta}\frac{\partial p}{\partial \varphi}, \frac{1}{\alpha}\frac{\partial p}{\partial r}\right). \tag{7.2.6}$$
- The *relativistic effect* is reflected in the following terms:
$$F_G P = \left(-\frac{\nu}{2\alpha^2 r}\frac{d\alpha}{dr}\frac{\partial}{\partial r}(rP_\theta), -\frac{\nu}{2\alpha^2 r}\frac{d\alpha}{dr}\frac{\partial}{\partial r}(rP_\varphi), \frac{\nu}{2\alpha}\frac{\partial}{\partial r}\left(\frac{1}{\alpha}\frac{d\alpha}{dr}P_r\right)\right), \tag{7.2.7}$$
which can be regarded as the coupling interaction between the gravitational potential α and the electromagnetic, the weak and the strong potentials represented by the viscous coefficient ν.

  We shall see that the force (7.2.7) is responsible to the supernova's huge explosion.
- The *thermal expansion force* is due to the coupling between the gravity $\nabla\psi$ and the heat Q:
$$F_T = \left(0, 0, \frac{c^2}{2\alpha}\frac{d\psi}{dr}\beta T\right), \tag{7.2.8}$$
which is the main driving force for generating stellar interior circulations and nebular matter spurts of red giants.

The two forces (7.2.7) and (7.2.8) are the main driving forces for the stellar motion, and hereafter we derive their explicit formulas.

1. *Formula for thermal expansion force.* The thermal expansion force (7.2.8) is radially symmetric, which is simply written in the r-component form
$$f_T = \frac{c^2}{2\alpha}\frac{d\psi}{dr}\beta T.$$

In its nondimensional form, f_T is expressed as

$$f_T = \sigma(r)T, \tag{7.2.9}$$

and $\sigma(r)$ is called the thermal factor given by

$$\sigma(r) = -\frac{c^2 r_0^4 \beta}{2\kappa^2} \frac{1}{\alpha} \frac{dT}{dr} \frac{d\psi}{dr}. \tag{7.2.10}$$

Here α, ψ, T satisfy equations (7.1.79)-(7.1.80) with the boundary conditions (7.1.81)-(7.1.84). The detailed derivation of (7.2.9)-(7.2.10) will be given hereafter.

The σ-factor (7.2.10) can be expressed in the following form, to be deduced later:

$$\begin{aligned}\sigma(r) =& \frac{c^2 r_0^3 (1-\delta)\beta}{2\kappa^2 r^2} \frac{e^{\zeta(r)}}{e^{\zeta(1)}} \cdot (1-\delta r^2 - \eta) \cdot \left(\frac{1}{r^2}\frac{\delta r^2 + \eta}{1-\delta r^2 - \eta} + r\xi\right) \\ &\cdot \left(A - \frac{1}{\kappa}\int_r^1 \frac{r^2 Q}{1-\delta r^2 - \eta} dr\right) \qquad \text{for } 0 \leqslant r \leqslant 1,\end{aligned} \tag{7.2.11}$$

where

$$\begin{aligned}\eta &= \frac{1}{2r}\int_0^r \frac{r^2 \psi' \phi'}{\alpha \psi} dr, \\ \zeta &= \int_0^r \left(\frac{\alpha}{r} + r\xi\right) dr, \\ \xi &= \frac{8\pi G}{c^2}\alpha p + \phi'' - \frac{\alpha' \phi'}{2\alpha} \qquad \text{for } 0 \leqslant r \leqslant 1,\end{aligned} \tag{7.2.12}$$

δ is called the δ-factor given by

$$\delta = \frac{2mG}{c^2 r_0}, \tag{7.2.13}$$

and m, r_0 are the mass and radius of the star.

2. *Formula for the relativistic effect.* The term $F_G P$ in (7.2.7) can be expressed in the following form:

$$F_G P = \begin{pmatrix} -\nu\left(\delta + \dfrac{\eta'}{2r}\right)\dfrac{\partial}{\partial r}(rP_\theta) \\ -\nu\left(\delta + \dfrac{\eta'}{2r}\right)\dfrac{\partial}{\partial r}(rP_\varphi) \\ \dfrac{\nu}{2}\left(\dfrac{(2\delta r + \eta')^2}{1-\delta r^2 - \eta} + 2\delta + \eta''\right)P_r + \dfrac{\nu}{2}(2\delta r + \eta')\dfrac{\partial}{\partial r}P_r \end{pmatrix}, \tag{7.2.14}$$

where η, δ are as in (7.2.12) and (7.2.13).

The force $F_G P$ is of special importance in studying supernovae, black holes, and the galaxy cores. In fact, by the boundary conditions (7.1.81)-(7.1.83), we can reduce that the radial component of the force (7.2.14) on the stellar shell is as

$$f_r = \left(\frac{2\nu\delta^2}{1-\delta} + \delta + \phi''(1)\right)P_r + \delta\nu\frac{\partial P_r}{\partial r},$$

which has

$$f_r \sim \frac{2\nu\delta^2}{1-\delta}P_r \to \infty \quad \text{as} \quad \delta \to 1 \text{ (for } P_r > 0). \tag{7.2.15}$$

The property (7.2.15) will lead to a huge supernovae explosions as they collapse to the radii $r_0 \to 2mG/c^2$. It is the explosive force (7.2.15) that prevents the formation of black holes; see Sections 7.2.6 and 7.3.3.

3. *Derivation of formula (7.2.14).* To deduce (7.2.14) we have to derive the gravitational potential α. The first equation of (7.1.79) can be rewritten as

$$\frac{dM}{dr} = 4\pi r^2 \rho_0 - \frac{c^2}{4G}\frac{r^2\psi'\phi'}{\alpha\psi}, \qquad M = \frac{c^2 r}{2G}\left(1 - \frac{1}{\alpha}\right),$$

It gives the solution as

$$M = \frac{4}{3}\pi r^3 \rho_0 - \frac{c^2}{4G}\int_0^r \frac{r^2\psi'\phi'}{\alpha\psi}dr, \qquad \alpha = \left(1 - \frac{2MG}{c^2 r}\right)^{-1}.$$

By $\rho_0 = m/\frac{4}{3}\pi r_0^3$, and in the nondimensional form ($r \to r_0 r$), we get

$$\alpha = (1 - \delta r^2 - \eta)^{-1} \qquad \text{for } 0 \leqslant r \leqslant 1, \tag{7.2.16}$$

where η, δ are as in (7.2.12) and (7.2.13). By (7.1.81) we have

$$\eta(1) = 0, \qquad (\text{i.e. } \eta(r_0) = 0). \tag{7.2.17}$$

Then, the formula (7.2.14) follows from (7.2.16).

4. *Derivation of σ-factor (7.2.11).* By (7.2.10) we need to calculate T' and ψ'. By (7.1.80), T' can be expressed in the form

$$\frac{dT}{dr} = -\frac{1}{\kappa r^2}\int_0^r r^2 \alpha Q dr + \frac{a}{r^2},$$

where a is a determined constant. By (7.1.84) we obtain

$$a = -Ar_0^2 + \frac{1}{\kappa}\int_0^{r_0} r^2 \alpha Q dr.$$

In the nondimensional form, we have

$$\frac{dT}{dr} = -\frac{A}{r^2} + \frac{1}{\kappa r^2}\int_r^1 r^2 \alpha Q dr \qquad \text{for } 0 \leqslant r \leqslant 1, \quad A > 0. \tag{7.2.18}$$

To consider ψ', by the second equation of (7.1.79) we obtain

$$\psi = \frac{k}{r}e^{\zeta(r)}, \tag{7.2.19}$$

where $\zeta(r)$ is as in (7.2.12), k is a to-be-determined constant. In view of (7.1.81), i.e. $\psi(1) = 1-\delta$, we have

$$k = (1-\delta)e^{-\zeta(1)}.$$

Then, it follows from (7.2.19) that

$$\frac{d\psi}{dr} = \frac{(1-\delta)e^{\zeta(r)}}{e^{\zeta(1)r_0}}\left(\frac{\alpha-1}{r^2}+r\xi\right) \qquad \text{for } 0 \leqslant r \leqslant 1, \tag{7.2.20}$$

where ξ is as in (7.2.12). By (7.1.82) and (7.2.16)-(7.2.17), we can deduce that

$$\xi(1) = 0, \quad (\text{i.e. } \xi(r_0) = 0). \tag{7.2.21}$$

Thus, the σ-factor (7.2.11) follows from (7.2.16), (7.2.18) and (7.2.20).

5. *Thermal Force and (7.2.21).* The thermal expansion force acting on the stellar shell (i.e. at $r = r_0$) can be deduced from (7.2.9) and (7.2.11) in the following (nondimensional) form

$$f_T = \sigma_0 T, \quad \sigma_0 = \frac{c^2 r_0^3 \beta(1-\delta)\delta}{2\kappa^2}A \quad (A > 0). \tag{7.2.22}$$

By $0 < \delta < 1$, we have

$$\sigma_0 > 0 \quad (\sigma_0 = \sigma(1), \text{ i.e. } \sigma(r_0)).$$

Hence, it follows that there is an $\varepsilon \geqslant 0$ such that

$$\sigma(r) > 0, \quad \text{for } \varepsilon < r \leqslant 1. \tag{7.2.23}$$

The positiveness of $\sigma(r)$ in (7.2.23) shows that the thermal force f_T of (7.2.22) is an outward expanding force. It is this power that causes the swell and the nebular matter spurt of a red giant. We also remark that the temperature gradient A on the boundary is maintained by the heat source Q.

7.2.3 Stellar interior circulation

Recapitulation of dynamic transition theory

First we briefly recall the dynamic transition theory developed by the authors in (Ma and Wang, 2013b) and the references therein. Many dissipative systems, both finite and infinite dimensional, can be written in an abstract operator equation form as follows

$$\frac{du}{dt} = L_\lambda u + G(u,\lambda), \tag{7.2.24}$$

where L_λ is a linear operator, G is a nonlinear operator, and λ is the control parameter.

It is clear that $u = 0$ is a stationary solution of (7.2.24). We say that (7.2.24) undergoes a dynamic transition from $u = 0$ at $\lambda = \lambda_1$ if $u = 0$ is stable for $\lambda < \lambda_1$, and unstable for $\lambda > \lambda_1$. The dynamic transition of (7.2.24) depends on the linear eigenvalue problem:

$$L_\lambda \varphi = \beta(\lambda)\varphi.$$

Mathematically this eigenvalue problem has eigenvalues $\beta_k(\lambda) \in \mathbb{C}$ such that

$$\mathrm{Re}\beta_1(\lambda) > \mathrm{Re}\beta_2(\lambda) > \cdots .$$

The following are the main conclusions for the dynamic transition theory; see (Ma and Wang, 2013b) for details:

- Dynamic transitions of (7.2.24) take place at $(u,\lambda) = (0,\lambda_1)$ provided that λ_1 satisfies the following principle of exchange of stability (PES):

$$\begin{aligned} &\mathrm{Re}\beta_1 \begin{cases} < 0 & \text{for } \lambda < \lambda_1 \text{ (or } \lambda > \lambda_1), \\ = 0 & \text{for } \lambda = \lambda_1, \\ > 0 & \text{for } \lambda > \lambda_1 \text{ (or } \lambda < \lambda_1), \end{cases} \\ &\mathrm{Re}\beta_k(\lambda_1) < 0, \qquad \forall k \geqslant 2. \end{aligned} \tag{7.2.25}$$

- Dynamic transitions of all dissipative systems described by (7.2.24) can be classified into three categories: continuous, catastrophic, and random. Thanks to the universality, this classification is postulated in citeptd as a general principle called principle of dynamic transitions.

- Let u_λ be the first transition state. Then we can also use the same strategе outlined above to study the second transition by considering PES for the following linearized eigenvalue problem

$$L_\lambda \varphi + DG(u_\lambda,\lambda)\varphi = \beta^{(2)}(\lambda)\varphi.$$

  Also we know that successive transitions can lead to chaos.

Stellar interior circulation

The governing fluid component equations are (7.1.85). We first make the nondimensional. Let

$$\begin{aligned} &(r,\tau) = (r_0 r', r_0^2\tau'/\kappa), \\ &(P,T,p,\rho) = \left(\kappa\rho_0 P'/r_0, -\frac{d\widetilde{T}}{dr} r_0 T', \rho_0\kappa^2 p'/r_0^2, \rho_0\rho\right). \end{aligned}$$

Then the equations (7.1.85) are rewritten as (drop the primes):

$$\begin{aligned}
&\frac{\partial P}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)P=\mathrm{Pr}\Delta P+\frac{1}{\kappa}F_G P+\sigma(r)T\vec{k}-\nabla p,\\
&\frac{\partial T}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)T=\widetilde{\Delta}T+P_r,\\
&\mathrm{div}P=0,
\end{aligned}\tag{7.2.26}$$

where $P=(P_\theta,P_\varphi,P_r),\vec{k}=(0,0,1)$, $\sigma(r)$ and $F_G P$ are as in (7.2.11) and (7.2.14), $\mathrm{Pr}=\nu/\kappa$ is the Prandtl number, and the Δ is given by

$$\begin{aligned}
\Delta P_\theta &= \widetilde{\Delta}P_\theta+\frac{2}{r^2}\frac{\partial P_r}{\partial\theta}-\frac{2\cos\theta}{r^2\sin^2\theta}\frac{\partial P_\varphi}{\partial\varphi}-\frac{P_\theta}{r^2\sin^2\theta},\\
\Delta P_\varphi &= \widetilde{\Delta}P_\varphi+\frac{2}{r^2\sin\theta}\frac{\partial P_r}{\partial\varphi}+\frac{2\cos\theta}{r^2\sin\theta}\frac{\partial P_\theta}{\partial\varphi}-\frac{P_\varphi}{r^2\sin^2\theta},\\
\Delta P_r &= \widetilde{\Delta}P_r-\frac{2}{\alpha r^2}\left(P_r+\frac{\partial P_\theta}{\partial\theta}+\frac{\cos\theta}{\sin\theta}P_\theta+\frac{1}{\sin\theta}\frac{\partial P_\varphi}{\partial\varphi}\right),
\end{aligned}\tag{7.2.27}$$

Based on the dynamic transition theory introduced early in this section, the stellar circulation depends on the following three forces:

$$\mathrm{Pr}\Delta P,\quad \frac{1}{\kappa}F_G P,\quad \sigma T\vec{k},\tag{7.2.28}$$

where in general $\mathrm{Pr}\Delta P$ prevents/slows-down the circulation.

By (7.2.14) we see that $F_G P$ depends on the δ-factor. The Sun's δ-factor is $\delta_\odot = 10^{-5}/2$, and in general the δ-factors for stars are of the order:

$$\begin{aligned}
&\delta\sim 10^{-8} && \text{for red giants,}\\
&\delta\sim 10^{-5} && \text{for main-sequence stars,}\\
&\delta\sim 10^{-3} && \text{for white dwarfs,}\\
&\delta\sim 10^{-1} && \text{for neutron stars,}\\
&\delta = 1 && \text{for black holes.}
\end{aligned}\tag{7.2.29}$$

Hence it is clear that all stars, except black holes, have small δ-factors.

On the other hand, for a small δ-factor, it follows from equations (7.1.79) and (7.1.80) that α,ψ,ϕ has the order

$$\begin{aligned}
&\alpha\sim 1+\delta+\eta, && \eta,\eta'\sim\delta^2, && \alpha'\sim\delta,\\
&\psi\sim r^\delta, && \psi'/\psi\sim\delta, && \phi',\phi''\sim\delta.
\end{aligned}$$

Hence, we deduce from (7.2.14) that

$$F_G P\sim\delta P\quad\text{for}\quad \delta\ll 1.$$

Thus, in view of (7.2.29) we conclude that the relativistic effect $F_G P$ is negligible on the stellar interior motion for all stars except supernovae.

Hence the main driving force for stellar circulations is the thermal expansion force characterized by the σ-factor σ_0 in (7.2.22). Due to $\delta \ll 1, \sigma_0$ can be approximately given by

$$\sigma_0 = \frac{r_0^2 mG\beta}{\kappa^2} A, \tag{7.2.30}$$

which plays the similar role as the Rayleigh number Re in the earth's atmospheric circulation. The value σ_0 of (7.2.30) is large enough to generate thermal convections for main-sequence stars and red giants.

We remark that the heat source Q is caused not only by nuclear reactions, but also by the pressure gradient, the density and the gravitational potential energy. Based on the σ-factor in (7.2.30), we obtain the following physical conclusions:

1. *Main-sequence stars.* Based on the dynamic transition theory, by (7.2.23), we deduce that there is a critical number $\sigma_c > 0$, independent of the parameter σ_0 in (7.2.30), such that equations (7.2.26) undergo no dynamic transition if $\sigma_0 < \sigma_c$, and a dynamic transition if $\sigma_0 > \sigma_c$:

$$\sigma_0 - \sigma_c \begin{cases} < 0 & \text{there is no stellar circulation,} \\ > 0 & \text{there exists stellar circulation.} \end{cases} \tag{7.2.31}$$

In particular, σ_c has the same order of magnitude as the first eigenvalue λ_1 of the the following equations in the unit ball B_1:

$$\begin{aligned} & -\text{Pr}\Delta P + \nabla p = \lambda_1 P && \text{for } x \in B_1 \subset \mathbb{R}^3, \\ & \left(P_r, \frac{\partial P_\theta}{\partial r}, \frac{\partial P_\varphi}{\partial r}\right) = 0 && \text{at } r = 1, \end{aligned}$$

where ΔP is as in (7.2.27).

For the main-sequence stars, the σ-factors are much larger than the first eigenvalue λ_1 of (7.2.32). For example, the Sun consists of hydrogen, and

$$r_0 = 7 \times 10^8 \text{m}, \quad m = 2 \times 10^{30} \text{kg}, \quad G = 6.7 \times 10^{-11} \text{m}^3/(\text{kg} \cdot \text{s}^2).$$

Using the average temperature $T = 10^6 K$, the parameter κ is given by

$$\kappa = 0.18 \left(\frac{T}{190k}\right)^{1.72} 10^{-4} \text{m}^2/\text{s} \simeq 50 \text{m}^2/\text{s}.$$

With thermal expansion coefficient β in the order $\beta \sim 10^{-4}/K$, the σ-factor of (7.2.30) for the Sun is about

$$\sigma_\odot \sim 10^{30} A \quad [\text{m/K}]. \tag{7.2.32}$$

Due to nuclear fusion, stars have a constant heat supply, which leads to a higher boundary temperature gradient A. Referring to (7.2.32), we conclude that there are always interior circulations and thermal motion in main-sequence stars and red giants, which has large σ-factors.

2. *Red giants.* The nuclear reaction of a red giant stops in its core, but does take place in the shell layer, which maintains a larger temperature gradient A on the boundary than the main-sequence phase. Therefore, the greater σ-factor makes the star to expand, and the increasing radius r_0 raises the σ-factor (7.2.30). The increasingly larger σ-factor provides a huge power to hurl large quantities of gases into space at very high speed.

3. *Neutron stars and pulsars.* Neutron stars are different from other stars, which have bigger δ-factors, higher rotation Ω and lower σ-factor (as the nuclear reaction stops). Instead of (7.2.26) the dynamic equations governing neutron stars are

$$
\begin{aligned}
&\frac{\partial P}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)P=\Pr\Delta P+\frac{1}{\kappa}F_G P-2\vec{\Omega}\times P-\nabla p+\sigma T\vec{k},\\
&\frac{\partial T}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)T=\widetilde{\Delta}T+P_r,\\
&\mathrm{div}P=0.
\end{aligned}
\tag{7.2.33}
$$

As the nuclear reaction ceases, the temperature gradient A tends to zero as time $t\to\infty$, and consequently

$$
\sigma\to 0 \quad \text{as} \quad t\to\infty. \tag{7.2.34}
$$

Based on the dynamic transition theory briefly recalled earlier in this section, we derive from (7.2.33) and (7.2.34) the following assertions:

- By (7.2.34), neutron stars will eventually stop convection.

- Due to the high rotation Ω, the convection of (7.2.33) for the early neutron star is time periodic, and its period $\mathscr{T}$ is inversely proportional to Ω, and its convection intensity B is proportional to $\sqrt{\sigma-\sigma_c}$, i.e.

$$
\begin{aligned}
&\text{period } \mathscr{T}\simeq\frac{C_1}{\Omega},\\
&\text{intensity } B\simeq C_2\sqrt{\sigma-\sigma_c},
\end{aligned}
\tag{7.2.35}
$$

  where σ_c is the critical value of the transition, and c_1, c_2 are constants. The properties of (7.2.35) explain that the early neutron stars are pulsars, and by (7.2.34) their pulsing radiation intensities decay at the rate of $\sqrt{\sigma-\sigma_c}$ or $\sqrt{kA-\sigma_c}$ (k is a constant).

7.2.4 Stellar atmospheric circulations

The model describing stellar atmosphere circulation without rotation is given by (7.1.71)-(7.1.73), and the eigenvalue equations read as

$$
\begin{aligned}
&\Pr\left[\widetilde{\Delta}P_\theta + \frac{2}{r^2}\frac{\partial P_r}{\partial\theta} - \frac{2\cos\theta}{r^2\sin^2\theta}\frac{\partial P_\varphi}{\partial\varphi} - \frac{P_\theta}{r^2\sin\theta} - \frac{c_0}{r^2\nu}P_\theta\right.\\
&\qquad\left. + \frac{\delta}{2r^2}\frac{\partial}{\partial r}(rP_\theta)\right] - \frac{1}{r}\frac{\partial\varphi}{\partial\theta} = \beta P_\theta,\\
&\Pr\left[\widetilde{\Delta}P_\varphi + \frac{2}{r^2\sin\theta}\frac{\partial P_r}{\partial\varphi} + \frac{2\cos\theta}{r^2\sin^2\theta}\frac{\partial P_\theta}{\partial\varphi} - \frac{P_\varphi}{r^2\sin^2\theta} - \frac{c_0}{r^2\nu}P_\varphi\right.\\
&\qquad\left. + \frac{\delta}{2r^2}\frac{\partial}{\partial r}(rP_\varphi)\right] - \frac{1}{r\sin\theta}\frac{\partial p}{\partial\varphi} = \beta P_\varphi,\\
&\Pr\left[\widetilde{\Delta}P_r - \frac{2}{r^2}(1-\delta)(P_r + \frac{\partial P_\theta}{\partial\theta} + \frac{\cos\theta}{\sin\theta}P_\theta + \frac{1}{\sin\theta}\frac{\partial P_\varphi}{\partial\varphi}) - \frac{c_1}{r^2\nu}P_r\right.\\
&\qquad\left. + \frac{\delta^2}{2(1-\delta)}\frac{1}{r^2}P_r - \frac{\delta}{2r}\frac{\partial P_r}{\partial r} + \frac{1}{r^2}\sqrt{\mathrm{Re}}T\right] - (1-\delta)\frac{\partial p}{\partial r} = \beta P_r,\\
&\widetilde{\Delta}T + \frac{1}{r^2}\sqrt{\mathrm{Re}}P_r = \beta T,\\
&\mathrm{div}P = 0,
\end{aligned}
\tag{7.2.36}
$$

with the boundary conditions

$$
T = 0,\ P_r = 0,\ \frac{\partial P_\theta}{\partial r} = \frac{\partial P_\varphi}{\partial r} = 0 \text{ at } r = r_0, r_0 + 1, \tag{7.2.37}
$$

where

$$
\widetilde{\Delta}f = \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial f}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial\varphi^2} + \frac{1-\delta}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right),
$$

and $\delta = 2mG/c^2rh$ is the δ-factor with $r_0 < r < r_0 + 1$ being nondimensional.

For stellar atmosphere circulations, the star radius r_0 is much greater than the convection height $h = r_1 - r_0$. Hence, we can approximatively take δ as the δ-factor as defined in (7.1.74).

Thus the eigenvalues β of (7.2.36)-(7.2.37) depend on δ and the Rayleigh number Re in (7.1.74):

$$
\beta = \beta(\delta, \mathrm{Re}), \quad (0 < \delta < 1,\ 0 < \mathrm{Re}).
$$

For (7.2.36)-(7.2.37), the following conclusions hold true:

1. The differential operator Δ in (7.2.36) is the Laplace-Beltrami operator, which is symmetric. Hence, the differential operator in (7.2.36)-(7.2.37) is symmetric. It implies that all eigenvalues β of (7.2.36)-(7.2.37) are real. Hence, the stellar atmospheric system (7.1.71)-(7.1.73) undergo the first dynamic transition to stationary solutions.

2. Let L be the differential operator on the left-hand side of (7.2.36); then the first eigenvalue β_1 satisfies

$$\beta_1 = \max_{||u||=1} \frac{1}{2}\langle Lu, u\rangle. \tag{7.2.38}$$

As the δ-factor of (7.1.74) is smaller than one, then by (7.2.38), β_1 satisfies the inequality

$$\beta_1 \leqslant -k_1 + k_2\sqrt{\mathrm{Re}}, \tag{7.2.39}$$

for some constants $k_1, k_2 > 0$. By (7.1.74) we see that

$$\mathrm{Re} \to \infty \text{ as } m \to \infty, \text{ or as } r_0 \to \infty, \text{ for } (T_0 - T_1) > 0.$$

By (7.2.39) we get

$$\beta_1 \to +\infty \text{ as } m \to \infty, \text{ or as } r_0 \to \infty.$$

In addition, it is known that

$$\beta_1^{(k)} \to +\infty \ (k \geqslant 2) \text{ as } \beta_1 \to +\infty.$$

Hence based on the dynamic transition theory, we derive the following physical conclusions.

Physical Conclusion 7.10 *For the stars with small $\delta \ll 1$, their atmosphere have thermal convections to occur. In particular for the stars with large mass and radius, the convections are in turbulent state.*

3. In the third equation of (7.2.36), the term

$$F = \frac{\delta^2}{2(1-\delta)}\frac{1}{r^2}P_r \tag{7.2.40}$$

represents the radially expanding force which is the relativistic gravitational effect. The force (7.2.40) satisfies

$$F \to +\infty \text{ as } \delta \to 1 \text{ for } P_r > 0.$$

It implies that the force F in (7.2.40) is explosive as $\delta \to 1$. In addition, by (7.2.38) the first eigenvalue β_1 satisfies the following inequality

$$\beta_1 \leqslant -k_1 + k_2\sqrt{\mathrm{Re}} + \frac{k_3}{1-\delta} \quad \text{for } \delta \to 1.$$

Thus we can deduce the following conclusions.

Physical Conclusion 7.11 *Due to the explosive force F of (7.2.40), the stars with $\delta \simeq 1(\delta = 1$ is the black hole) has no the stellar atmospheres, which are erupted into the outer space. However, as $1-\delta > 0$ is small, the stellar atmospheres are in the turbulent convection state,caused by the relatistic gravitational effect.*

4. Let l_τ and l_r be the convection scales in the horizontal direction and radial direction. The ratio

$$r = \frac{l_\tau}{l_r}$$

depends on the coefficient ratio of the horizontal components of the momentum (P_θ, P_φ) and the radial momentum component P_r in (7.2.36), i.e. the ratio

$$\eta = \frac{\delta - 1 - c_0/\nu}{-2(1-\delta) - c_1/\nu + \delta^2/2(1-\delta)}. \tag{7.2.41}$$

As $|\eta| \ll \infty, \gamma$ andη have the qualitative relation

$$\gamma \propto \begin{cases} \eta^{-1} & \text{for } \eta > 0, \\ |\eta| & \text{for } \eta < 0. \end{cases}$$

By (7.2.41) we see that

$$\begin{aligned} &\eta > 0 \text{ and } \eta^{-1} \sim o(1) && \text{for } \delta \ll 1 \text{ and large } \nu, \\ &\eta < 0 \text{ and } |\eta| \ll 1 && \text{for } \delta \to 1. \end{aligned}$$

Hence we deduce the following conclusions.

Physical Conclusion 7.12 *The convection scale ratio γ depends on δ for $\delta \ll 1$ and $\delta \to 1$, and on the ratio c_0/c_1 for $\delta < 1$ and $v \ll 1$. Moreover, γ possesses the following properties*

$$\gamma = \begin{cases} o(1) & \textit{for } \delta \ll 1 \textit{ and } v \textit{ large}, \\ \ll 1 & \textit{for } \delta \to 1. \end{cases} \tag{7.2.42}$$

The sun atmosphere convections satisfy the relation (7.2.42). For the Sun, $\delta \simeq \frac{1}{2} \times 10^{-5}$. The observations show that

$$\begin{aligned} &\text{the photosphere convection} && r \doteq 1 \sim 2, \\ &\text{the chromosphere convection} && r \doteq 2 \sim 3. \end{aligned}$$

For the solar corona, $v \ll 1$, we need to know the ratio c_0/c_1.

5. For the stars with rotation $\vec{\Omega}$, the corresponding eigenvalue equations with the Coriolis force are

$$\begin{aligned} & \Pr\left[\Delta P - \frac{f}{r^2} + \frac{\sqrt{\mathrm{Re}}}{r^2}\vec{k}T\right] - 2\vec{\Omega}\times P = \beta P, \\ & \widetilde{\Delta T} + \frac{\sqrt{\mathrm{Re}}}{r^2}P_r = \beta T, \\ & \operatorname{div} P = 0, \end{aligned} \tag{7.2.43}$$

with the boundary conditions (7.2.37).

In (Ma and Wang, 2013b) we showed that there is a lower bound Ω_0 such that as $\Omega > \Omega_0$ the first eigenvalue β_1 of (7.2.43) is complex in the critical state (7.2.25). Therefore, for the stars with the bigger rotation their atmosphere convections in the first phase transition are time periodic.

7.2.5 Dynamics of stars with variable radii

For stars with varying sizes and for supernovae, their radii expand and shrink periodically. Therefore, the metric in the interior of such stars is as follows:

$$ds^2 = -\psi c^2 dt^2 + R^2(t)\left[\alpha dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2)\right],$$

where $\psi = \psi(r,t), \alpha = \alpha(r,t)$, and $R(t)$ is the scalar factor representing the star radius. For convenience, we denote

$$\psi = e^{u(r,t)}, \quad \alpha = e^{v(r,t)}, \quad R^2(t) = e^{k(t)}, \quad 0 \leqslant r \leqslant 1.$$

Then the metric is rewritten as

$$ds^2 = -e^u c^2 dt^2 + e^k\left[e^v dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2)\right]. \tag{7.2.44}$$

The stars with variable radii are essentially in radial motion. Hence, the horizontal momentum (P_θ, P_φ) is assumed to be zero:

$$(P_\theta, P_\varphi) = 0. \tag{7.2.45}$$

In the following we develop dynamic models for astronomical objects with variable sizes.

1. *Gravitational field equations.* We recall the gravitational field equations (Ma and Wang, 2014e):

$$R_{\mu\nu} = -\frac{8\pi G}{c^4}(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T) - (D_{\mu\nu}\phi - \frac{1}{2}g_{\mu\nu}\Phi). \tag{7.2.46}$$

The nonzero components of the metric (7.2.44) are

$$g_{00} = -e^u, \quad g_{11} = e^{k+v}, \quad g_{22} = e^k r^2, \quad g_{33} = e^k r^2 \sin^2\theta,$$

the nonzero components of the Levi-civita connections are

$$\Gamma^0_{00} = \frac{1}{2c}u_t, \qquad \Gamma^0_{11} = \frac{1}{2c}e^{v-u}(k_t+v_t), \qquad \Gamma^0_{10} = \frac{1}{2}u_r,$$
$$\Gamma^0_{22} = \frac{r^2}{2c}e^{-u}k_t, \qquad \Gamma^0_{33} = \frac{r^2}{2c}e^{-u}k_t\sin^2\theta, \qquad \Gamma^1_{00} = \frac{1}{2}e^{u-v}u_r$$
$$\Gamma^1_{11} = \frac{1}{2}v_r, \qquad \Gamma^1_{10} = \frac{1}{2c}(k_t+v_t), \qquad \Gamma^1_{22} = -re^{-v},$$
$$\Gamma^1_{33} = -re^{-v}\sin^2\theta, \qquad \Gamma^2_{21} = \frac{1}{r}, \qquad \Gamma^2_{33} = \sin\theta\cos\theta,$$
$$\Gamma^3_{31} = \frac{1}{r}, \qquad \Gamma^3_{32} = \frac{\cos\theta}{\sin\theta},$$

and the nonzero components of the Ricci curvature tensor read

$$\begin{aligned}
R_{00} =& \frac{1}{2c^2}\left[3k_{tt} + \frac{3}{2}k_t^2 + v_{tt} + \frac{1}{2}v_t^2 + k_t v_t - u_t(k_t+v_t)\right] \\
&-\frac{1}{2}e^{u-k-v}\left[u_{rr} + \frac{1}{2}u_r^2 - \frac{1}{2}u_r v_r + \frac{2}{r}u_r\right], \\
R_{11} =& -\frac{e^{k+v-u}}{2c^2}\left[k_{tt} + \frac{3}{2}k_t^2 + v_{tt} + v_t^2 + 3k_t v_t - \frac{1}{2}u_t(k_t+v_t)\right] \\
&+\frac{1}{2}\left[u_{rr} + \frac{1}{2}u_r^2 - \frac{1}{2}u_r v_r - \frac{2}{r}v_r\right], \\
R_{22} =& -\frac{r^2e^{k-u}}{2c^2}\left[k_{tt} + \frac{3}{2}k_t^2 + \frac{1}{2}k_t(v_t-u_t)\right] - e^{-v}\left[e^v + \frac{r}{2}(k_r+v_r-u_r) - 1\right], \\
R_{33} =& R_{22}\sin^2\theta, \\
R_{10} =& -\frac{1}{cr}\left[(1+\frac{r}{2}u_r)k_t + v_r\right].
\end{aligned}$$

The energy-momentum tensor is in the form

$$T_{\mu\nu} = \begin{pmatrix} \rho & g_{00}g_{11}P_r c & 0 & 0 \\ g_{00}g_{11}P_r c & g_{11}p & 0 & 0 \\ 0 & 0 & g_{22}p & 0 \\ 0 & 0 & 0 & g_{33}p \end{pmatrix},$$

where ρ is the energy density, P_r is the radial component of the momentum density. Then direct computations imply that

$$T = g^{\mu\nu}T_{\mu\nu} = -\rho + 3p, \qquad T_{00} - \frac{1}{2}g_{00}T = \frac{1}{2}(\rho+3p),$$
$$T_{11} - \frac{1}{2}g_{11}T = \frac{1}{2}e^{k+v}(\rho - p), \qquad T_{22} - \frac{1}{2}g_{22}T = \frac{1}{2}e^k r^2(\rho-p),$$
$$T_{33} - \frac{1}{2}g_{33}T = (T_{22} - \frac{1}{2}g_{33}T)\sin^2\theta, \qquad T_{10} - \frac{1}{2}g_{10}T = g_{00}g_{11}P_r c.$$

To derive an explicit expression of (7.2.46), we need to compute the covariant derivatives of the dual gravitational field ϕ:

$$D_{\mu\nu}\phi = \frac{\partial^2\phi}{\partial x^\mu \partial x^\nu} - \Gamma^\lambda_{\mu\nu}\frac{\partial\phi}{\partial x^\lambda}.$$

Let $\phi = \phi(r,t)$. Then we have

$$\begin{aligned}
D_{00}\phi &= \frac{1}{c^2}\phi_{tt} - \frac{1}{2c^2}u_t\phi_t - \frac{1}{2}e^{u-v}u_r\phi_r,\\
D_{11}\phi &= \phi_{rr} - \frac{1}{2c^2}e^{v-u}(k_t+v_t)\phi_t - \frac{1}{2}v_r\phi_r,\\
D_{22}\phi &= -\frac{r^2}{2c^2}e^{-u}k_t\phi_t + re^{-v}\phi_r,\\
D_{33}\phi &= D_{22}\phi\sin^2\theta,\\
D_{10}\phi &= \phi_{rt} - \frac{1}{2c}(u_r\phi_t + \phi_r k_t + \phi_r v_t).
\end{aligned}$$

Thus, the field equations (7.2.46) are written as

$$\begin{aligned}
R_{10} &= -D_1D_0\phi,\\
R_{kk} &= -\frac{8\pi G}{c^4}(T_{kk} - \frac{1}{2}g_{kk}T) - (D_{kk}\phi - \frac{1}{2}g_{kk}\Phi) \qquad \text{for } k = 0,1,2,
\end{aligned}$$

which are expressed as

$$\left(1+\frac{ru_r}{2}\right)k_t + v_t = \frac{8\pi Gr}{c^2}e^{u+k+v}P_r + cr\phi_{rt} - \frac{r}{2}(u_r\phi_t + \phi_r k_z + \phi_r v_t), \tag{7.2.47}$$

$$\begin{aligned}
&3k_{tt} + \frac{3}{2}k_t^2 + v_{tt} + \frac{1}{2}v_t^2 + k_tv_t - u_t(k_t+v_t) \\
&\qquad - c^2e^{u-k-v}\left[u_{rr} + \frac{1}{2}u_r^2 - \frac{1}{2}u_rv_r + \frac{2}{r}u_r\right]\\
&\quad = -\frac{8\pi G}{c^2}(\rho+3p) - c^2\left(D_{00}\phi + e^{u-k-v}D_{11}\phi + \frac{2e^{u-k}}{r^2}D_{22}\phi\right),
\end{aligned} \tag{7.2.48}$$

$$\begin{aligned}
&k_{tt} + \frac{3}{2}k_t^2 + v_{tt} + v_t^2 + 3k_tv_t - \frac{1}{2}u_t(k_t+v_t)\\
&\qquad - c^2e^{u-k-v}\left[u_{rr} + \frac{1}{2}u_r^2 - \frac{1}{2}u_rv_r - \frac{2}{r}v_r\right]\\
&\quad = \frac{8\pi G}{c^2}e^u(\rho-p) + c^2(e^{u-k-v}D_{11}\phi - D_{00}\phi - \frac{2e^{u-k}}{r^2}D_{22}\phi),
\end{aligned} \tag{7.2.49}$$

$$\begin{aligned}
&k_{tt} + \frac{3}{2}k_t^2 + \frac{1}{2}k_t(v_t-u_t) + \frac{2c^2e^{u-k-v}}{r^2}\left[e^v + \frac{r}{2}(k_r+v_r-u_r) - 1\right]\\
&\quad = \frac{8\pi G}{c^2}e^u(\rho-p) + c^2(D_{00}\phi - e^{u-k-v}D_{11}\phi),
\end{aligned} \tag{7.2.50}$$

The equations (7.2.47)-(7.2.50) have seven unknown functions $u, v, k, \phi, P_r, \rho, p$, in which P_r, ρ, p satisfy the fluid dynamic equations and the equation of state introduced hereafter.

2. *Fluid dynamic model.* The fluid dynamic model takes the momentum representation equations coupling the heat equation. Under the condition (7.2.45) and the radial symmetry, they are given as follows:

$$\begin{aligned}
&\frac{\partial P_r}{\partial\tau}+\frac{1}{\rho}P_r\frac{\partial P_r}{\partial r}+\frac{1}{2}\frac{\partial v}{\partial r}P_r^2 \\
&=\nu e^{-v}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P_r}{\partial r}\right)-\frac{2}{r^2}P_r+\frac{1}{2}\frac{\partial}{\partial r}\left(\frac{\partial v}{\partial r}P_r\right)\right] \\
&\quad+\gamma e^{-v}\frac{\partial}{\partial r}\left[\frac{e^{-v/2}}{r^2}\frac{\partial}{\partial r}(r^2e^{v/2}P_r)\right]-e^{-v}\left[\frac{\partial p}{\partial r}-\frac{\rho}{2}(1-\beta T)\frac{\partial e^u}{\partial r}\right],
\end{aligned} \tag{7.2.51}$$

$$\frac{\partial T}{\partial\tau}+\frac{1}{\rho}P_r\frac{\partial T}{\partial r}=\frac{\kappa e^{-v}}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial T}{\partial r}\right)+Q(r), \tag{7.2.52}$$

$$\frac{\partial\rho}{\partial\tau}+\frac{e^{-v/2}}{r^2}\frac{\partial}{\partial r}(r^2e^{v/2}P_r)=0. \tag{7.2.53}$$

3. *Equation of state.* We know that the gravitational field equations represent the law of gravity, which essentially dictates the gravity related unknowns: $e^u, e^v, R=e^{k/2}, \phi$.

The laws for describing the matter field are the motion equation (7.2.51), the heat equation (7.2.52), and the energy conservation equation (7.2.53). To close the system, one needs to supplement an equation of state given by thermal dynamics, which provides a relation between temperature T, pressure p, and energy density ρ:

$$f(T,p,\rho)=0, \tag{7.2.54}$$

which depends on the underlying physical system.

In summary, we have derived a consistent model coupling the gravitational field equations, the fluid dynamic equations and the equation of state consists of eight equations solving for eight unknowns: $\psi=e^u$, $\alpha=e^v$, $R=e^{k/2}$, ϕ, P_r, T, p and ρ.

4. *Energy conservation formula.* From the energy conservation equation (7.2.53), we can deduce energy conservation in the following form

$$R^3r^2e^{v/2}P_r+\frac{1}{4\pi}\frac{d}{dt}E_r=0 \qquad \text{for } 0<r<1, \tag{7.2.55}$$

where $r=1$ stands for the boundary $R=e^{k/2}$ of the star, E_r is the total energy in the ball B_r with radius r.

To see this, we first note that the volume differential element of the Riemannian manifold is given by

$$dV=\sqrt{g_{11}g_{22}g_{33}}drd\theta d\varphi=e^{3k/2}r^2e^{v/2}\sin\theta drd\theta d\varphi.$$

Taking volume integral for (7.2.53) on B_r implies that

$$\frac{d}{dt}\int_{B_r}\rho dV + R^3\int_0^r \frac{\partial}{\partial r}\left(r^2 e^{\nu/2}P_r\right)\int_0^{\pi}\int_0^{2\pi}\sin\theta d\theta d\varphi = 0,$$

which leads to

$$\frac{dM_r}{dt} + 4\pi R^3 r^2 e^{\nu/2}P_r = 0,$$

and (7.2.55) follows.

5. *Shock wave.* As the total energy E_R of the star is invariant, we have

$$\frac{d}{dt}E_R = 0.$$

It follows from (7.2.55) that

$$P_r = 0 \quad \text{on } r = 1 \quad (\text{i.e. on the boundary } R). \tag{7.2.56}$$

On the other hand, the physically sound boundary condition for the star with variable radius is

$$\frac{\partial P_r}{\partial r} = 0 \quad \text{on } r = 1, \tag{7.2.57}$$

which means that there is no energy exchange between the star and its exterior. Thus, (7.2.56) and (7.2.57) imply that there is a shock wave outside the star near the boundary.

Remark 7.13 Formula (7.2.55) is very important. In fact, due to the boundary condition (7.1.81) and $e^{\nu/2} \simeq 1/\sqrt{1-\delta}$, in the star shell layer, (7.2.55) can be approximately written as

$$\rho P_r = -\frac{\sqrt{1-\delta}}{4\pi R^2}\frac{d}{dt}M_r \quad \text{for } R - r > 0 \text{ small}, \tag{7.2.58}$$

where $\delta = 2M_{r_0}G/c^2R$. This shows that a collapsing supernova is prohibited to shrink into a black hole $(\delta = 1)$. In fact, the strongest evidence for showing that black holes cannot be created comes from the relativistic effect of (7.2.14), which provides a huge explosive power in the star shell layer given by

$$\frac{\nu\delta^2}{1-\delta}P_r \to \infty \quad \text{as } \delta \to 1 \quad (P_r \neq 0). \tag{7.2.59}$$

Here P_r is the convective momentum different from the contracting momentum P_r in (7.2.58); see Section 7.3.3 for details. □

Remark 7.14 One difficulty encountered in the classical Einstein gravitational field equations is that the number of unknowns is less than the number of equations, and consequently the coupling between the field equations and fluid dynamic and heat equations become troublesome. □

7.2.6 Mechanism of supernova explosion

In its late stage of life, a massive red giant collapses, leading to a supernova's huge explosion. It was still a mystery where does the main source of driving force for the explosion come from, and the current viewpoint, that the blast is caused by the large amount of neutrinos erupted from the core, is not very convincing.

The stellar dynamic model (7.1.78)-(7.1.84) provides an alternative explanation for supernova explosions, and we proceed in a few steps as follows:

1. When a very massive red giant completely consumes its central supply of nuclear fuels, its core collapses. Its radius r_0 begins to decrease, and consequently the δ-factor increases:

$$r_0 \text{ decreases } \Rightarrow \ \delta = \frac{2mG}{c^2 r_0} \text{ increases.}$$

2. The huge mass m and the rapidly reduced radius r_0 make the δ-factor approaching one:

$$\delta \to 1 \quad \text{as } r_0 \to R_s$$

where $R_s = 2mG/c^2$ is the Schwarzschild radius.

3. By (7.2.58), the shrinking of the star slows down:

$$P_r \sim \sqrt{1-\delta},$$

and nearly stops as $\delta \to 1$.

4. Then the model (7.2.26) is valid, and the eigenvalue equations of (7.2.26) are given by

$$\begin{aligned} & \Pr\Delta P + \frac{1}{\kappa} F_G P + \sigma T\vec{k} - \nabla p = \beta P, \\ & \widetilde{\Delta} T + P_r = \beta T, \\ & \text{div} P = 0. \end{aligned} \tag{7.2.60}$$

The first eigenvalue β depends on the δ-factor, and by (7.2.14)

$$\beta_1 \sim \left(\frac{\Pr\delta^2}{1-\delta}\right)^{1/2} \quad \text{as } \delta \to 1. \tag{7.2.61}$$

Based on the transition criterion (7.2.25), the property (7.2.61) implies that the star has convection in the shell layer, i.e., the radial circulation momentum flux P_r satisfies

$$P_r > 0 \quad \text{in certain regions of the shell layer.}$$

5. The radial force (7.2.15) in the shell layer is

$$f_r \simeq \frac{2P_r\delta^2}{1-\delta} P_r \to \infty \quad \text{as } \delta \to 1 \text{ and } P_r > 0,$$

which provides a very riving force, resulting in the supernova explosion, as shown in Figure 7.2.

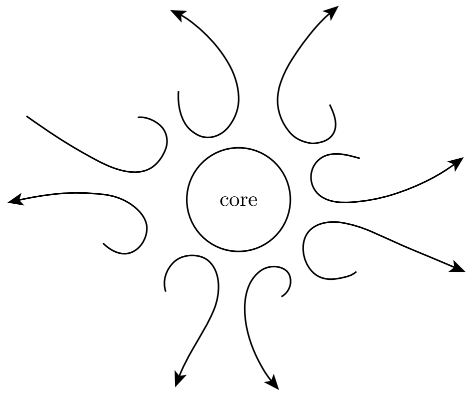


Figure 7.2 Circulation in a shell layer causing blast

6. Since $P_r = 0$ at $r = r_0$, the radial force of (7.2.14) is zero:

$$f_r = 0 \quad \text{at} \quad r = r_0.$$

Here r_0 is the radius of the blackhole core. Hence, the supernova's huge explosion preserves an interior core of smaller radius containing the blackhole core, which yields a neutron star. In particular, the huge explosion has no imploding force, and will not generate a new black hole.

The analysis in Steps 1-6 above provides the supernova exploding mechanism, and clearly provides the power resource of the explosion.

In addition, by (7.2.12) and (7.2.16) we have

$$\alpha = \frac{1}{1 - \delta r^2/r_0^2 - \eta(r)}, \qquad \eta(r) = \frac{1}{2r}\int_0^r \frac{r^2\psi'\phi'}{\alpha\psi} dr \qquad \text{for } 0 \leqslant r \leqslant r_0.$$

We can verify that

$$\eta(r) > 0 \qquad \text{for } 0 < r < r_0. \tag{7.2.62}$$

In fact, by (7.2.17) and (7.2.12) we have

$$\eta(0) = 0, \quad \eta(r_0) = 0. \tag{7.2.63}$$

Therefore, η has an extremum $\overline{r}$ $(0 < \overline{r} < r)$ satisfying

$$\eta'(\overline{r}) = 0.$$

Let $\eta = \frac{1}{r} f$. Then

$$\eta'(r) = 0 \Rightarrow f(r) = e^a r \quad (a = \text{constant}).$$

Hence, at the extremum $\overline{r}, \eta$ takes a positive value

$$\eta(\overline{r}) = \frac{1}{\overline{r}} f(\overline{r}) = e^a > 0 \qquad \text{for } 0 < \overline{r} < r_0. \tag{7.2.64}$$

Thus, (7.2.62) follows from (7.2.63) and (7.2.64).

The fact (7.2.62) implies that the critical δ-factor δ_c for the supernova explosion is less than one, i.e. $\delta_c < 1$.

7.3 Black Holes

7.3.1 Geometric realization of black holes

The concept of black holes was originated from the Einstein general theory of relativity. Based on the Einstein gravitational field equations, K. Schwarzschild derived in 1916 an exact exterior solution for a spherically symmetrical matter field, and Tolman-Oppenheimer-Volkoff derived in 1939 an interior solution; see Section 7.1.2. In both solutions if the radius R of the matter field with mass M is less than or equal to a critical radius R_s, called the Schwarzschild radius:

$$R \leqslant R_s = \frac{2MG}{c^2}, \tag{7.3.1}$$

then the matter field generates a singular spherical surface with radius R_s, where time stops and the spatial metric blows-up; see Figure 7.3. The spherical region with radius R_s is called the black hole.

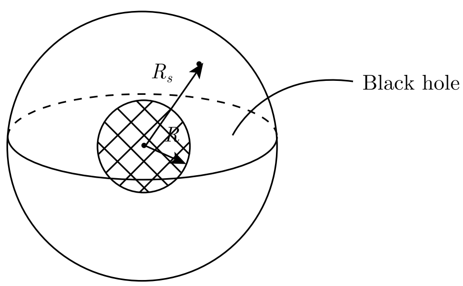


Figure 7.3 The spherical region enclosing a matter field with mass M and radius R satisfying (7.3.1) is called black hole.

We recall again the Schwarzschild metric in the exterior of a black hole written as

$$\begin{aligned} ds^2 &= g_{00}c^2dt^2 + g_{11}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \\ g_{00} &= -\left(1 - \frac{2MG}{c^2 r}\right), \quad g_{11} = \left(1 - \frac{2MG}{c^2 r}\right)^{-1}, \end{aligned} \tag{7.3.2}$$

where $r > R_s$ when the condition (7.3.1) is satisfied.

In (7.3.2) we see that at $r = R_s$, the time interval is zero, and the spatial metric blows up:

$$\sqrt{-g_{00}}dt = \left(1 - \frac{R_s}{r}\right)^{1/2} dt = 0 \qquad \text{at } r = R_s, \tag{7.3.3}$$

$$\sqrt{g_{11}}dr = \left(1 - \frac{R_s}{r}\right)^{-1/2} dr = \infty \qquad \text{at } r = R_s. \tag{7.3.4}$$

Physically, the proper time and distance for (7.3.2) are

$$\begin{aligned} &\text{proper time} = \sqrt{-g_{00}}\, t, \\ &\text{proper distance} = \sqrt{g_{11}dr^2 + r^2 d\theta^2 + r^2 \sin\theta d\varphi^2}. \end{aligned}$$

The coordinate system (t,x) with $x = (r,\theta,\varphi)$ represents the projection of the real world to the coordinate space. Therefore the radial motion speed dr/dt in the projected world differs from the proper speed v_r by a factor $\sqrt{-g_{00}/g_{11}}$, i.e.

$$\frac{dr}{dt} = \sqrt{-g_{00}/g_{11}}v_r.$$

Hence, the singularity (7.3.3) and (7.3.4) means that for an object moving toward to the boundary of a black hole, its projection speed vanishes:

$$\frac{dr}{dt} = 0 \qquad \text{at } r = R_s.$$

This implies that any object in the exterior of the black hole cannot pass through its boundary and enter into the interior. In the next subsection we shall rigorously prove that a black hole is a closed and innate system.

Mathematically, a Riemannian manifold $(\mathscr{M}, g_{ij})$ is called a geometric realization (i.e. isometric embedding) in $\mathbb{R}^N$, if there exists a one to one mapping

$$\vec{r} : \mathscr{M} \to \mathbb{R}^N,$$

such that

$$g_{ij} = \frac{d\vec{r}}{dx^i} \cdot \frac{d\vec{r}}{dx^j}.$$

The geometric realization provides a "visual" diagram of $\mathscr{M}$, the real world of our Universe.

In the following we present the geometric realization of a 3D metric space of a black hole near its boundary. By (7.3.2), the space metric of a black hole is given by

$$ds^2 = \left(1 - \frac{R_s}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2) \qquad \text{for } r > R_s = \frac{2MG}{c^2}. \tag{7.3.5}$$

It is easy to check that a geometric realization of (7.3.5) is given by $\vec{r} : \mathscr{M} \to \mathbb{R}^4$:

$$\vec{r}_{\text{ext}} = \left\{ r\sin\theta\cos\varphi, r\sin\theta\sin\varphi, r\cos\theta, 2\sqrt{R_s(r-R_s)} \right\} \qquad \text{for } r > R_s. \tag{7.3.6}$$

In the interior of a black hole, the Riemannian metric near the boundary is given by the TOV solution (7.1.42), and its space metric is in the form

$$ds^2 = \left(1 - \frac{r^2}{R_s^2}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2) \qquad \text{for } r < R_s, \tag{7.3.7}$$

A geometrical realization of (7.3.7) is

$$\vec{r}_{\text{int}}^{\pm} = \left\{ r\sin\theta\cos\varphi, r\sin\theta\sin\varphi, r\cos\theta, \pm\sqrt{R_s^2 - r^2} \right\}. \tag{7.3.8}$$

The diagrams of (7.3.6) and (7.3.8) are as shown in Figure 7.4, where case (a) is the embedding

$$\vec{r}^{+} = \begin{cases} \vec{r}_{\text{ext}} & \text{for } r > R_s, \\ \vec{r}_{\text{int}}^{+} & \text{for } r < R_s, \end{cases}$$

and case (b) is the embedding

$$\vec{r}^{-} = \begin{cases} \vec{r}_{\text{ext}} & \text{for } r > R_s, \\ \vec{r}_{\text{int}}^{-} & \text{for } r < R_s. \end{cases}$$

The base space marked as $\mathbb{R}^3$ in (a) and (b) are taken as the coordinate space (i.e. the projective space), and the surfaces marked by M represent the real world which are separated into two closed parts by the spherical surface of radius R_s: the black hole $(r < R_s)$ and the exterior world $(r > R_s)$.

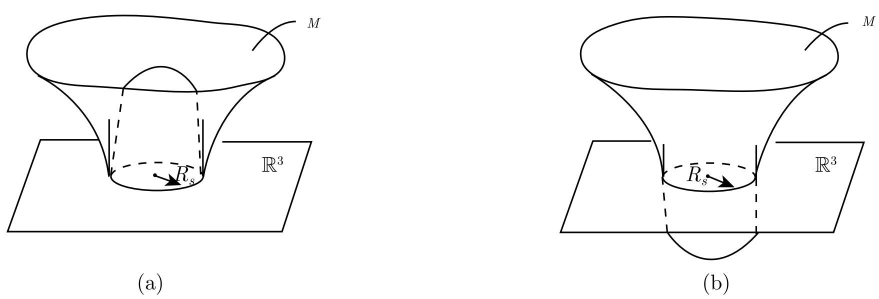


Figure 7.4 $\mathscr{M}$ is the real world with the metric (7.3.5) for $r > R_s$ and the metric (7.3.7) for $r < R_s$, and in the base space $\mathbb{R}^3$ the coordinate system is taken as spherical coordinates (r, θ, φ).

In particular, the geometric realization of (7.3.7) for a black hole clearly manifests that the real world in the black hole is a hemisphere with radius R_s embedded in $\mathbb{R}^4$; see Figure 7.4(a):

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = R_s^2 \qquad \text{for } 0 \leqslant |x_4| \leqslant R_s,$$

where

$$(x_1, x_2, x_3, x_4) = \left(r\sin\theta\cos\varphi,\ r\sin\theta\sin\varphi,\ r\cos\theta,\ \pm\sqrt{R_s^2 - r^2} \right).$$

We remark that the singularity of $\mathscr{M}$ at $r = R_s$, where the tangent space of $\mathscr{M}$ is perpendicular to the coordinate space $\mathbb{R}^3$, is essential, and cannot be removed by any coordinate transformations. The coordinate transformations such as those given by Eddington and Kruskal possess the singularity as well, and, consequently, cannot be used as the coordinate systems for the metrics (7.3.5) and (7.3.7).

7.3.2 Blackhole theorem

The main objective of this section is to prove the following blackhole theorem.

Theorem 7.15 (Blackhole Theorem) *Assume the validity of the Einstein theory of general relativity, then the following assertions hold true*:

1) *black holes are closed: matters can neither enter nor leave their interiors*;
2) *black holes are innate*: *they are neither born to explosion of cosmic objects, nor born to gravitational collapsing*; *and*
3) *black holes are filled and incompressible, and if the matter field is non-homogeneously distributed in a black hole, then there must be sub-blackholes in the interior of the black hole.*

We prove this theorem in three steps as follows.

Step 1. Closedness of black holes. First, it is classical that all matter, including photons, cannot escape from a black hole when they are within the Schwarzschild radius.

Step 2. Now we need to show that all external energy cannot enter into the interior of a black hole. By the energy-momentum conservation, we have

$$\frac{\partial E}{\partial \tau} + \operatorname{div} P = 0, \tag{7.3.9}$$

where E and P are the energy and momentum densities. Take the volume integral of (7.3.9) on $B = \{x \in \mathbb{R}^3 \mid R_s < |x| < R_1\}$:

$$\int_B \left[\frac{\partial E}{\partial \tau} + \operatorname{div} P\right] d\Omega = 0, \qquad d\Omega = \sqrt{g}\, dr d\theta d\varphi, \tag{7.3.10}$$

where $\operatorname{div}P$ is as in (7.1.60), and

$$g = \det(g_{ij}) = g_{11}g_{22}g_{33} = \alpha r^4 \sin^2\theta, \qquad \alpha = \left(1 - \frac{2MG}{c^2 r}\right)^{-1}.$$

Let $\mathscr{E}$ be the total energy in B, then the Gauss formula, we have

$$\int_B \frac{\partial E}{\partial \tau} d\Omega = \frac{d}{dt}\mathscr{E},$$
$$\int_B \operatorname{div} P d\Omega = \int_{S_{R_1}} \sqrt{\alpha(R_1)} P_r dS_{R_1} - \lim_{r \to R_s} \int_{S_r} \sqrt{\alpha} P_r dS_r.$$

Here $S_r = \{x \in \mathbb{R}^3 \mid |x| = r\}$. In view of (7.3.10) we deduce that

$$\frac{d\mathscr{E}}{dt} = \lim_{r\to R_s} \int_{S_r} \sqrt{\alpha} P_r dS_r - \sqrt{\alpha(R_1)} \int_{S_{R_1}} P_r dS_{R_1}. \tag{7.3.11}$$

The equality (7.3.11) can be rewritten as

$$\lim_{r\to R_s} \int_{S_r} P_r dS_r = \lim_{r\to R_s} \frac{1}{\sqrt{\alpha(r)}} \left[\frac{d\mathscr{E}}{dt} + \sqrt{\alpha(R_1)} \int_{S_{R_1}} P_r dS_{R_1} \right] = 0. \tag{7.3.12}$$

This together with no escaping of particles from the interior of the black hole shows that

$$\lim_{r\to R_s^+} P_r = 0.$$

In other words, there is no energy flux P_r on the Schwarzschild surface, and we have shown that no external energy can enter into a black hole.

In conclusion, we have shown that black holes are closed: no energy can penetrate the Schwarzschild surface.

Step 3. Innateness of black holes. The explosion mechanism introduced in Subsection 7.2.6 clearly manifests that any massive object cannot generate a new black hole. In other words, we conclude that black holes can neither be created nor be annihilated, and the total number of black holes in the Universe is conserved.

Step 4. Assertion 3) follows by applying conclusion (7.5.55) and the fact that sub black-holes are incompressible. The theorem is therefore proved. □

We remark again that the singularity on the boundary of black holes is essential and cannot be removed by any differentiable coordinate transformation with differentiable inverse. The Eddington and Kruskal coordinate transformations are non-differentiable, and are not valid.

Remark 7.16 The gravitational force F generated by a black hole in its exterior is given by

$$F = \frac{mc^2}{2} \nabla g_{00} = -mg^{11} \frac{\partial \psi}{\partial r},$$

where ψ is the gravitational potential. By (7.3.2) we have the following gravitational force:

$$F = -\left(1 - \frac{2MG}{c^2 r}\right) \frac{mMG}{r^2}. \tag{7.3.13}$$

Consequently, on the boundary of a black hole, the gravitational force is zero:

$$F = 0 \qquad \text{at } r = R_s.$$

7.3.3 Critical δ-factor

Black holes are a theoretical outcome. Although we cannot see them directly due to their invisibility, they are, however, strong evidences from many astronomical observations and theoretic studies.

In the following, we first briefly recall the Chandrasekhar limit of electron degeneracy pressure and the Oppenheimer limit of neutron degeneracy pressure; then we present new criterions to classify pure black holes, which do not contain other black holes in their interior, into two types: the quark and weakton black holes, by using the δ-factor.

1. *Electron and neutron degeneracy pressures.* Classically we know that there are two kinds of pressure to resist the gravitational pressure, called the electron degeneracy pressure and the neutron degeneracy pressure. These pressures prevent stars from gravitational collapsing with the following mass relation:

$$m < \begin{cases} 1.4M_{\odot} & \text{for electron pressure,} \\ \\ 3M_{\odot} & \text{for neutron pressure.} \end{cases} \tag{7.3.14}$$

Hence, by (7.3.14), we usually think that a dead star is a white dwarf if its mass $m < 1.4M_{\odot}$, and is a neutron star if its mass $m < 3M_{\odot}$. However, if the dead star has mass $m > 3M_{\odot}$, then it is regarded as a black hole. Hence the neutron pressure gradient is thought to be a final defense to prevent a star from collapsing into a black hole. Thus, $3M_{\odot}$ becomes a critical mass to determine the possible formation of a black hole.

2. *Interaction potential pressure.* However, thanks to the strong and weak interaction potentials established in (Ma and Wang, 2015a, 2014c), there still exist three kinds of potential pressures given by

$$\text{neutron potential, quark potential, weakton potential.} \tag{7.3.15}$$

These three potential pressures maintain three types of astronomical bodies:

$$\begin{aligned} &\text{neutron stars,} \\ &\text{quark black holes if they exist,} \\ &\text{weakton black holes if they exist.} \end{aligned} \tag{7.3.16}$$

We are now in position to discuss these potential pressures. By the theory of elementary particles, a neutron is made up of three quarks $n = uud$, and u, d quarks are made up of three weaktons as $u = w^* w_1 \overline{w}_1, d = w^* w_1 w_2$. The three levels of particles possess different potentials distinguished by their interaction charges:

$$\begin{aligned}
&\text{neutron charge} && g_n = 3\left(\frac{\rho_w}{\rho_n}\right)^3 g_s, \\
&\text{quark charge} && g_q = \left(\frac{\rho_w}{\rho_q}\right)^3 g_s, \\
&\text{weakton weak charge} && g_w,
\end{aligned} \tag{7.3.17}$$

where ρ_n, ρ_q, ρ_w are the radii of neutron, quark and weakton.

Let g be a specific charge in (7.3.17). Then by the interaction potentials obtained in (Ma and Wang, 2015a), the particle with charge g has a repulsive force:

$$f = \frac{g^2}{r^2}.$$

The force acts on particle's cross section with area $S = \pi r^2$, which yields the interaction potential pressure as

$$P = \frac{f}{S} = \frac{g^2}{\pi r^4}. \tag{7.3.18}$$

Let each ball B_r with radius r contain only one particle. Then the mass density ρ is given by

$$\rho = \frac{3m_0}{4\pi r^3}, \tag{7.3.19}$$

where m_0 is the particle mass. By the uncertainty relation, in B_r the particle energy ε_0 is

$$\varepsilon_0 = \frac{\hbar}{2t},$$

and $t = r/v$, where v is the particle velocity. Replacing v by the speed of light c, we have

$$\varepsilon_0 = \frac{\hbar c}{2r}.$$

By $m_0 = \varepsilon_0/c^2$, the density ρ of (7.3.19) is written as

$$\rho = \frac{3\hbar}{8\pi c r^4} \qquad \text{or equivalently} \qquad r^4 = \frac{3\hbar}{8\pi c \rho}. \tag{7.3.20}$$

Inserting r^4 of (7.3.20) into (7.3.18), we derive the interaction potential pressure P in the form

$$P = \frac{8c\rho g^2}{3\hbar}. \tag{7.3.21}$$

3. *Critical δ-factors.* It is known that the central pressure of a star with mass m and radius r_0 can be expressed as

$$P_M = \frac{Gm^2}{r_0^4} = \frac{2\pi c^2}{3}\rho\delta, \qquad \delta = \frac{2mG}{c^2 r_0}, \tag{7.3.22}$$

where δ is the δ-factor.

We infer from (7.3.21) and (7.3.22) the critical δ-factor as

$$\delta_c = \frac{4}{\pi}\frac{g^2}{\hbar c}, \tag{7.3.23}$$

where g is one of the interaction charges in (7.3.17).

The critical δ-factor in (7.3.23) provides criterions for the three types of astronomical bodies of (7.3.16).

4. *Physical significance of* δ_c. It is clear that for a star with $m > 1.4M_\odot$ if

$$\delta < \frac{4}{\pi}\frac{g_n^2}{\hbar c}, \tag{7.3.24}$$

then the neutron potential pressure P_n in (7.3.21) is greater than the star pressure P_M in (7.3.22):

$$P_n > P_M.$$

In this case, neutrons in the star cannot be crushed into quarks. Hence, (7.3.24) should be a criterion to determine if the body is a neutron star. It is known that

$$g_n^2 \sim \hbar c.$$

Thus, we take

$$g_n^2 = \frac{\pi}{4}\hbar c, \tag{7.3.25}$$

and (7.3.24) is just the black hole criterion.

If the δ-factor satisfies that

$$\frac{4}{\pi}\frac{g_n^2}{\hbar c} \leqslant \delta < \frac{4}{\pi}\frac{g_q^2}{\hbar c}, \tag{7.3.26}$$

then the neutrons will be crushed to become quarks and gluons. The equality (7.3.25) shows that the star satisfying (7.3.26) must be a black hole which is composed of quarks and gluons, and is called quark black hole.

If δ satisfies

$$\frac{4}{\pi}\frac{g_q^2}{\hbar c} \leqslant \delta < \frac{4}{\pi}\frac{g_w^2}{\hbar c}, \tag{7.3.27}$$

then the quarks are crushed into weaktons, and the body is called weakton black hole.

In summary, we infer from (7.3.24), (7.3.26) and (7.3.27) the following conclusions:

$$\text{a body} = \begin{cases} \text{a neutron star} & \text{if } \delta < \delta_n^c \text{ and } m > 1.4M_\odot, \\ \text{a quark black hole} & \text{if } \delta_n^c < \delta < \delta_q^c, \\ \text{a weakton black hole} & \text{if } \delta_q^c < \delta < \delta_w^c, \end{cases} \tag{7.3.28}$$

where

$$\delta_j^c = \frac{4}{\pi}\frac{g_j^2}{\hbar c} \qquad \text{for } j = n, q, w.$$

5. *Upper limit of the radius.* Weaktons are elementary particles, which cannot be crushed. Therefore, there is no star with δ-factor greater than δ_w^c. Thus there exists an upper limit for the radius r_c for astronomical bodies with mass m, determined by

$$\frac{2MG}{c^2 r_c} = \delta_w^c.$$

Namely, the upper limit of the radius r_c reads

$$r_c = \frac{\pi m G}{2c^2} \frac{\hbar c}{g_w^2}. \tag{7.3.29}$$

Finally we remark that since black holes cannot be compressed, the δ-factor of any astronomical object cannot be less than one: $\delta \geqslant 1$. Therefore, by (7.3.25) and (7.3.26), there exist no weakton black holes.

7.3.4 Origin of stars and galaxies

The closeness and innateness of black holes provide an excellent explanation for the origin of planets, stars and galaxies.

In fact, all black holes are inherent. Namely, black holes exist at the very beginning of the Universe. During the evolution of the Universe, each black hole forms a core and adsorbs a ball of gases around it. The globes of gases eventually evolve into planets, stars and galaxy nuclei, according to the radii or masses of the inner cores of black holes. Of course, it is possible that several black holes can bound together to form a core of a bulk of gases.

Due to the closedness of black holes, planets, stars and galaxy nuclei are stable, which cannot be absorbed into the inner cores of black holes and vanish.

1. *Jeans theory on the origin of stars and galaxies.* In the beginning of the twentieth century, J. Jeans presented a general theory for the formation of galaxies and stars. He thought that the Universe in the beginning was filled with chaotic gas, and various astronomical objects were formed in succession by a process of gas decomposition into bulks of clouds, consequently forming galaxies, stars, and planets.

According to the Jeans theory, a ball of clouds with homogeneous density ρ can be held together only if

$$V + K \leqslant 0, \tag{7.3.30}$$

where V is the total gravitational potential energy, and K is the total kinetic energy of all particles. The potential energy V is

$$V = -\int_0^R \frac{GM_r}{r} \times 4\pi r^2 \rho dr = -\frac{3GM^2}{5R}, \tag{7.3.31}$$

where M is the mass of the cloud, $M_r = 4\pi r^3\rho/3$, and R is the radius. The kinetic energy K is expressed as the sum of thermal kinetic energies of all particles:

$$K = \frac{3}{2}NkT,$$

where N is the particle number, T is the temperature, and k is the Boltzmann constant. Assume that all particles have the same mass m, then $N = M/m$, and we have

$$K = \frac{3M}{2m}kT. \tag{7.3.32}$$

Thus, by (7.3.30)-(7.3.32) we obtain that

$$\frac{GM}{R} \geqslant \frac{5}{2m}kT. \tag{7.3.33}$$

The inequality (7.3.33) is called the Jeans condition.

2. *Masses of astronomical objects.* The Jeans condition (7.3.33) guarantees only the gaseous clouds being held together, and does not imply that the gas clouds can contract to form an astronomical object. However, a black hole must attract the nebulae around it to form a compact body.

We consider the mass relation between an astronomical object and its black hole core. The mass M of the object is

$$M = M_b + M_1, \tag{7.3.34}$$

where M_b is the mass of the black hole, and M_1 is the mass of the material attracted by this black hole. The total binding potential energy V of this object is given by

$$V = -\int_{R_s}^{R} \frac{GM_r}{r} \times 4\pi r^2 \rho dr, \tag{7.3.35}$$

where R is the object radius, R_s is the radius of the black hole, ρ is the mass density outside the core, and

$$M_r = M_b + \int_{R_s}^{r} 4\pi r^2 \rho dr. \tag{7.3.36}$$

Since $R_s \ll R$, we take $R_s = 0$ in the integrals (7.3.35) and (7.3.36). We assume that the density ρ is a constant. Then, it follows from (7.3.35) and (7.3.36) that

$$V = -4\pi G\rho \int_0^R \left[M_b r + \frac{4\pi}{3}\rho r^4\right] dr = -4\pi G\rho \left(\frac{M_b R^2}{2} + \frac{4\pi}{3\times 5}\rho R^5\right).$$

By $\rho = M_1/\frac{4}{3}\pi R^3$ and $M_b = M - M_1$, we have

$$V=-\frac{G}{R}\left(\frac{3}{2}MM_1-\frac{9}{10}M_1^2\right). \tag{7.3.37}$$

The stability of the object requires that $-V$ takes its maximum at some M_1 such that $dV/dM_1=0$. Hence we derive from (7.3.37) that

$$M_1=\frac{5}{6}M, \qquad M_b=\frac{1}{6}M. \tag{7.3.38}$$

The relation (7.3.38) means that a black hole with mass M_b can form an astronomical object with mass $M=6M_b$.

3. *Relation between radius and temperature.* A black hole with mass M_b determines the mass M of the corresponding astronomical system: $M=6M_b$. Then, by the Jeans relation (7.3.33), the radius R and average temperature T satisfy

$$T=\frac{2\times 6GmM_b}{5kR} \tag{7.3.39}$$

where T is expressed as

$$T=\frac{3}{4\pi R^3}\int_{B_R}\tau(x)dx,$$

where B_R is the ball of this system, and $\tau(x)$ is the temperature distribution. Let $\tau=\tau(r)$ depend only on r, then we have

$$T=\frac{3}{R^3}\int_0^R r^2\tau(r)dr. \tag{7.3.40}$$

4. *Solar system.* For the Sun, $M=2\times 10^{30}$kg and $R=7\times 10^8$m. Hence the mass of the solar black hole core is about

$$M_{\odot b}=\frac{1}{3}\times 10^{30}\text{kg},$$

and the average temperature has an upper limit:

$$T=\frac{4}{5}\times\frac{6.7\times 10^{-11}\text{m}^3/\text{kg}\cdot\text{s}^2\times 10^{30}\text{kg}\times 1.7\times 10^{-27}\text{kg}}{1.4\times 10^{-24}\text{kg}\cdot\text{m}^2/\text{s}^2\cdot\text{K}\times 7\times 10^8\text{m}}\simeq 10^8\text{K}.$$

For the earth, $M=6\times 10^{24}\text{kg}, R=6.4\times 10^6\text{m}$. Thus we have

$$M_{eb}=10^{24}\text{kg}, \qquad T=3.3\times 10^4.$$

5. *The radii of the solar and earth's black hole cores.* The radius of solar black hole is given by

$$R_s^{\odot}=500\text{ m},$$

and the radius of black hole of the earth is as

$$R_s^e=\frac{3}{2}\text{ cm}.$$

7.4 Galaxies

7.4.1 Introduction

Galaxy is an elementary unit in the large scale structure of our Universe, which is composed of star clusters, stars, gases and dusts. In the following we introduce this topic.

1. *Levels of cosmical structure.* There are mainly five levels of cosmic structure: stars, star clusters, galaxies, clusters of galaxies, and the Universe, and their basic scales are given inTable 7.1.

Table 7.1 Levels of Cosmic Structure

levels	Sun	star clusters (globular)	galaxies (spiral)	clusters of galaxies	Universe
radius(pc)	10^{-8}	10	10^5	10^6	10^{10}
distance(pc)	1	10^3	10^6	10^8	
mass($M_\odot$)	1	10^6	10^{11}	10^{14}	10^{21}
density (kg/m^3)	10^3	10^{-18}	10^{-20}	10^{-23}	10^{-27}

2. *Galaxy types.* Galaxies consist of the following three types:

Elliptical (E), Spiral (S), irregular (Irr).

The spiral galaxies are divided into two different sequences:

$$\text{normal spirals, denoted by } S,$$
$$\text{barred spirals, denoted by } SB.$$

Elliptical galaxies have an oval appearance, which are classified into eight grades, denoted by $E_0, E_1, \cdots, E_7$, according to their apparent ellipticity $e = (a-b)/a$, where a is the major radius and b is the minor radius. The E_0-type of elliptical galaxies are spherical in shape with $e_0 = 0$, and the E_7-types are of the most ablate appearance, i.e. the apparent ellipticity e_k of E_k is arranged in order

$$e_0 < e_1 < \cdots < e_7.$$

The S-spirals form a sequence of three types: S_a, S_b, S_c. The S_a have large galaxy nuclei and tightly wound arms, the S_b have moderate galaxy nuclei and less tightly wound arms, and the S_c have the smallest nuclear bulges and the least wound arms. Our galaxy (the Milky Way) is type S_b.

About one third of all spiral galaxies are the SB-type. They also consist of three types: SB_a, SB_b, SB_c, according to the size of the galaxy nuclei and tightness of the spiral arms, exactly as in S-spiral sequence.

The galaxy classification given above was introduced by E. Hubble, who arranged the galaxies in an orderly diagram as shown in Figure 7.5.

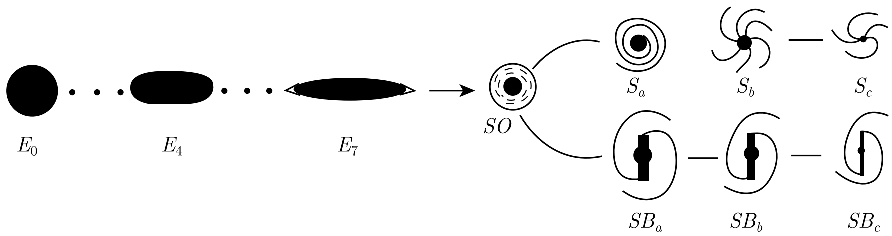


Figure 7.5 Hubble classification of galaxies, the ellipticals arranged on the left, and the spirals arranged on the right in two parallel sequences.

The *SO* (*S*-zero) galaxies have the characteristicsof ellipticals and spirals. They are disk shaped, like spirals, but lack spiral structure, and therefore like flat ellipticals.

The amount of interstellar gas, also called nebula, in galaxies increases from left to right in the diagram given by Figure 7.2. In ellipticals the amount of gas is very small, and it is also small in *SO* galaxies. However, the nebula increases progressively in the spirals from S_a and SB_a to S_c and SB_c. in addition, another important distinction between ellipticals and spirals is that clusters of galaxies contain 80% elliptics, and outside clusters, 80% of galaxies are spirals.

3. *Galaxy properties.* Various properties of galaxies are listed in Tables 7.2-7.4.

Table 7.2 Galaxy distribution

Types	E	$S0$	$S+SB$	Irr	Undetermined
%	13.0	21.5	61.1	3.1	0.9

Table 7.3 Average physical quantities

Types	E	S_a	S_b	S_c	Irr
mass ($M_\odot$)	2.0×10^{11}	1.6×10^{11}	1.3×10^{11}	1.6×10^{10}	1×10^{9}
mass/luminosity ($m_\odot/L_\odot$)	20-70	6.6	3.6	1.4	0.9
density ($M_\odot/pc^3$)	0.16	0.08	0.025	0.013	0.003
gas mass (%)	$\leqslant 0.2$	1.3	3	20	40

Table 7.4 Properties of ellipticals and spirals

Properties	ellipticals	spirals
interstellarmatter	non	plentiful
young stars	non	yes
appearance	elliptical shape	disk shape
stellar motion	random	rotation
color	red	blue

4. *Active galaxies.* According to their active extent, galaxies are divided into two types: normal galaxies and active galaxies. Active galaxies mainly include

Starbursts, Seyferts, Quasars, Radio-galaxies

Active galaxies are usually characterized by their extraordinary energy emision. The starburst galaxies emit a large amount of infrared radiation, which is caused by a great number of newly formed stars. But the three types of galaxies: Seyferts, Quasars and Radio-galaxies are known as active galaxies because they possess a compact region at their centers that has a much higher radiation than normal luminosity, which is called the active galactic nucleus (AGN).

AGN generates and emits immense quantities of energy over a wide range of electromagnetic wavelengths. Today, it is widely believed that AGN is an accretion disk generated by a supermassive black hole. Though we have not completely confirmed evidence to show the existence of black holes, many astronomical observations strongly suggest their existence. The main features of active galaxies are listed in Table 7.5.

Table 7.5 Features of active galaxies

Galaxy Type	Active nuclei	strong radiation	Jets
Normal	no	no	no
Starburst	no	yes	no
Seyfert	yes	yes	yes
Quasar	yes	yes	yes
Radio galaxy	yes	yes	yes

5. *The Milky Way.* Our own galaxy is known as the Milky Way, which is of S_b type and has 10^{11} stars and an enormous quantities of clouds of gases and dusts. Its radius is about 10^5 ly.

The Milky Way consists of two parts: the disk and the halo. The disk has the radius of about 5×10^5 ly and the thick of about 5×10^4 ly. The disk is composed of stars and nebulae, rotating around the center of the galaxy. The Sun is 3×10^4 ly from the nucleus, it moves at 300km/s and has a period of 2×10^8 years.

The halo of the Milky Way is spherically-shaped, centered on the nucleus of the galaxy, and has the radius of 10^5ly. The halo consists of less gas and roughly 120 globular clusters, each of which has hundreds of thousands of stars and moves in an elliptical orbit around the nucleus of the galaxy. The halo does not rotate with the disk.

6. *Galaxy clusters.* Galaxies are not uniformly distributed in the Universe, but aggregate in clusters of different size. Clusters of galaxies are the largest known astronomical systems bound by gravitational attraction. They form dense regions in the large scale structure of our Universe. The clusters are associated with much larger, non-gravitationally bounded groups, called superclusters.

The great regular clusters of galaxies are spherically-shaped, have roughly thousands of galaxies, almost all of which are of E and $S0$ types. The regular clusters of galaxies are typically 5×10^6ly in radius, and have no clear outer boundaries. Clusters are locked as their galaxies held together by mutual gravitational attraction. However, their velocities are too large, exceeding 10^3km/s, for them to remain gravitationally bound by their mutual attractions. It implies the presence of either an additional invisible mass component, or an additional attractive force besides the Newtonian gravity. This is the so called dark matter phenomenon. Astronomical observations manifest that there are large amounts of intergalactic gas which is very hot between $10^7 \sim 10^8$ K. The total mass of the gas is greater than that of all galaxies in the cluster. The wind of intergalactic gas streaming through these fast moving galaxies is strong enough to strip away their interstellar gas. This explains why the E and $S0$ types of galaxies have less gas because their gas has been swept out by intergalactic winds.

The other galaxy clusters are irregular. They have various sizes ranged from thousands of members to a few tens of members. The smaller clusters are also called galaxy groups. For example, our galaxy is a member of a galaxy group known as the local group which possesses about 40 galaxies. The irregular clusters lack spherical symmetry in shape and contain a mixture of all types of galaxies.

7.4.2 Galaxy dynamics

Galaxies are mainly either spiral or elliptical. Each galaxy possesses a compact core, known as galactic nucleus, which is supermassive and spherical-shaped. Thus, the galactic dynamic model is defined in an annular domain:

$$r_0 < r < r_1,$$

where r_0 is the radius of galaxy nucleus and r_1 the galaxy radius. In the following we develop models for spiral and elliptical galaxies, and provide their basic consequences on galactic dynamics.

1. *Spiral galaxies.* Spiral galaxies are disc-shaped, as shown in Figure 7.6. We model the galaxy in a disc domain as

$$D = \{x \in \mathbb{R}^2 \mid r_0 < |x| < r_1\}, \tag{7.4.1}$$

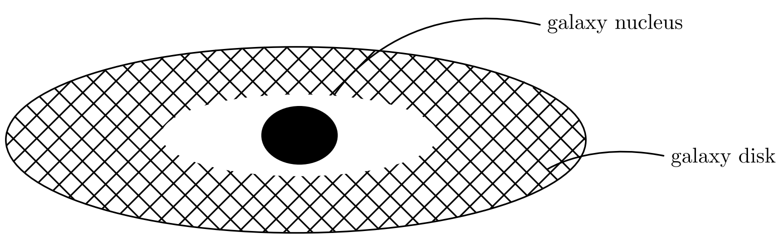


Figure 7.6 A schematic diagram of spiral galaxy.

for which the spherical coordinates reduce to the polar coordinate system (φ, r):

$$(\theta,\varphi,r)=\left(\frac{\pi}{2},\varphi,r\right) \qquad \text{for } 0\leqslant\varphi\leqslant 2\pi, \quad r_0<r<r_1. \tag{7.4.2}$$

The metric satisfying the gravitational field equations (7.1.62) of the galaxy nucleus is the Schwarzschild solution:

$$\begin{aligned} g_{00} &= -\left(1+\frac{2}{c^2}\psi\right), & \psi &= -\frac{M_0G}{r}, \\ g_{11} = \alpha(r) &= \left(1-\frac{\delta r_0}{r}\right)^{-1}, & \delta &= \frac{2M_0G}{c^2 r_0}, \end{aligned} \tag{7.4.3}$$

where $r_0<r<r_1$ and M_0 is the mass of galactic nucleus.

With (7.4.2) and (7.4.3), the 2D fluid equations (7.1.65)-(7.1.68) are written as

$$\begin{aligned} &\frac{\partial P_\varphi}{\partial\tau}+\frac{1}{\rho}(P\cdot\nabla)P_\varphi=\nu\Delta P_\varphi-\frac{1}{r}\frac{\partial p}{\partial\varphi}, \\ &\frac{\partial P_r}{\partial\tau}+\frac{1}{\rho}(P\cdot\nabla)P_r=\nu\Delta P_r-\frac{1}{\alpha}\frac{\partial p}{\partial r}-\rho(1-\beta T)\frac{M_rG}{\alpha r^2}, \\ &\frac{\partial T}{\partial\tau}+\frac{1}{\rho}(P\cdot\nabla)T=\kappa\widetilde{\Delta}T+Q, \\ &\frac{\partial\rho}{\partial\tau}+\text{div}P=0, \end{aligned} \tag{7.4.4}$$

supplemented with boundary conditions:

$$\begin{aligned} &P_\varphi(r_0)=\zeta_0, \quad P_r(r_0)=0, \quad T(r_0)=T_0, \\ &P_\varphi(r_1)=\zeta_1, \quad P_r(r_1)=0, \quad T(r_1)=T_1. \end{aligned} \tag{7.4.5}$$

Here α is as in (7.4.3), and M_r is the total mass in the ball B_r.

2. *Elliptical galaxies.* Elliptical galaxies are spherically-shaped, defined in a spherical-annular domain, as shown in Figure 7.7:

$$\Omega=\{x\in\mathbb{R}^3 \mid r_0<|x|<r_1\} \tag{7.4.6}$$

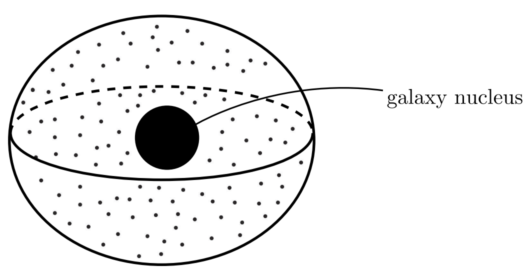


Figure 7.7 A schematic diagram of elliptical galaxy.

The metric is as in (7.4.3), and the corresponding fluid equations (7.1.65)-(7.1.68) are in the form:

$$\begin{aligned} &\frac{\partial P}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)P = \nu\Delta P - \nabla p - \rho(1-\beta T)\frac{M_0 G}{\alpha r^2}\vec{k}, \\ &\frac{\partial T}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)T = \kappa\widetilde{\Delta}T + Q, \\ &\frac{\partial \rho}{\partial \tau} + \text{div}\, P = 0, \end{aligned} \tag{7.4.7}$$

supplemented with the physically sound conditions:

$$\begin{aligned} &P_r = 0, \quad \frac{\partial P_\theta}{\partial r} = 0, \quad \frac{\partial P_\varphi}{\partial r} = 0 \quad \text{at } r = r_0, r_1, \\ &T(r_0) = T_0, \quad T(r_1) = T_1. \end{aligned} \tag{7.4.8}$$

3. *Galaxy dynamics.* Based on both models (7.4.4)-(7.4.5) and (7.4.7)-(7.4.8), we outline below the large scale dynamics of both spiral and elliptical galaxies.

Let the models be abstractly written in the following form

$$\frac{du}{dt} = F(u,\rho), \tag{7.4.9}$$

where $u = (P, T, p)$ is the unknown function, and ρ is the initial density distribution, which is used as a control parameter representing different physical conditions.

First, we consider the stationary equation of (7.4.9) given by

$$F(u,\rho) = 0. \tag{7.4.10}$$

Let u_0 be a solution of (7.4.10), and consider the deviation from u_0 as

$$u = v + u_0.$$

Thus, (7.4.9) becomes the following form

$$\frac{dv}{dt} = L_\lambda v + G(v,\lambda,\rho), \tag{7.4.11}$$

where $\lambda = (\delta, \text{Re})$, and the δ-factor and the Rayleigh number are defined by

$$\delta = \frac{2M_0 G}{c^2 r_0}, \qquad \text{Re} = \frac{M_0 G r_0 r_1 \beta}{\kappa\nu}\frac{T_0 - T_1}{r_1 - r_0}. \tag{7.4.12}$$

The L_λ is the derivative operator (i.e. the linearized operator) of $F(u, p)$ at u_0:

$$L_\lambda = DF(u_0,\rho),$$

and G is the higher order operator.

Then, we consider the dynamic transition of (7.4.11). Let $\widetilde{v}_\lambda$ be a stable transition solution of (7.4.11). Then the function

$$\widetilde{u} = u_0 + \widetilde{v}_\lambda \tag{7.4.13}$$

provides the physical information of the galaxy.

7.4.3 Spiral galaxies

Spiral galaxies are divided into two types: normal spirals (S-type) and barred spirals (SB-type). We are now ready to discuss these two sequences of galaxies by using the spiral galaxy model (7.4.4)-(7.4.5).

Let the stationary solutions of (7.4.4)-(7.4.5) be independent of φ, given by

$$P_r = 0, \quad P_\varphi = \widetilde{P}_\varphi(r), \quad T = \widetilde{T}(r), \quad p = \widetilde{p}(r).$$

The heat source is approximatively taken as $Q = 0$. Then the stationary equations of (7.4.4)-(7.4.5) are

$$\begin{aligned}
&r\widetilde{P}''_\varphi + 2\widetilde{P}'_\varphi - \frac{1}{r}\widetilde{P}_\varphi - \frac{\delta r_0}{r}\left(r\widetilde{P}''_\varphi + \frac{3}{2}\widetilde{P}'_\varphi - \frac{\widetilde{P}_\varphi}{2r}\right) = 0,\\
&\frac{\partial \widetilde{p}}{\partial r} = \frac{1}{r\rho}\widetilde{P}^2_\varphi - \frac{1}{r^2}\rho(1-\beta\widetilde{T})M_r G,\\
&\frac{d}{dr}\left(r^2\frac{d\widetilde{T}}{dr}\right) = 0,\\
&\widetilde{P}_\varphi(r_0) = \zeta_0, \quad \widetilde{P}_\varphi(r_1) = \zeta_1, \quad \widetilde{T}(r_0) = T_0, \quad \widetilde{T}(r_1) = T_1.
\end{aligned} \tag{7.4.14}$$

The first equation of (7.4.14) is an elliptic boundary value problem, which has a unique solution P_φ. Since $\delta r_0/r$ is small in the domain (7.4.1), the first equation of (7.4.14) can be approximated by

$$\widetilde{P}''_\varphi + \frac{2}{r}P'_\varphi - \frac{1}{r^2}\widetilde{P}_\varphi = 0,$$

which has an analytic solution as

$$\widetilde{P}_\varphi = \beta_1 r^{k_1} + \beta_2 r^{k_2}, \quad k_1 = \frac{\sqrt{5}-1}{2}, \quad k_2 = -\frac{\sqrt{5}+1}{2}. \tag{7.4.15}$$

By the boundary conditions in (7.4.14), we obtain that

$$\beta_1 = \frac{r_0^{k_2}\zeta_1 - r_1^{k_2}\zeta_0}{r_1^{k_1}r_0^{k_2} - r_0^{k_1}r_1^{k_2}}, \qquad \beta_2 = \frac{r_1^{k_1}\zeta_0 - r_0^{k_1}\zeta_1}{r_1^{k_1}r_0^{k_2} - r_0^{k_1}r_1^{k_2}}. \tag{7.4.16}$$

Thus we derive the solution of (7.4.14) as

$$\widetilde{P}_\varphi, \quad \widetilde{T} = T_0 + \frac{T_0 - T_1}{r_1 - r_0}r_1\left(\frac{r_0}{r} - 1\right), \quad \widetilde{p} = \int\left[\frac{\widetilde{P}^2_\varphi}{r\rho} - \frac{\rho G}{r^2}(1-\beta\widetilde{T})M_r\right]dr.$$

Make the translation

$$P_r \to P_r, \quad P_\varphi \to P_\varphi + \widetilde{P}_\varphi, \quad T \to T + \widetilde{T}, \quad p \to p + \widetilde{p};$$

then the equations (7.4.4) and boundary conditions (7.4.5) become

$$\begin{aligned}
&\frac{\partial P_\varphi}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)P_\varphi=\nu\Delta P_\varphi-(\frac{\widetilde{P}_\varphi}{r}+\frac{d\widetilde{P}_\varphi}{dr})P_r-\frac{1}{r}\frac{\partial p}{\partial \varphi},\\
&\frac{\partial P_r}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)P_r=\nu\Delta P_r+\frac{2\widetilde{P}_\varphi}{\alpha r}P_\varphi+\frac{\rho\beta M_r G}{\alpha r^2}T-\frac{1}{\alpha}\frac{\partial p}{\partial r},\\
&\frac{\partial T}{\partial \tau}+\frac{1}{\rho}(P\cdot\nabla)T=\kappa\widetilde{\Delta}T+\frac{r_0 r_1}{\rho r^2}\gamma P_r-\frac{1}{\rho r}\widetilde{P}_\varphi\frac{\partial T}{\partial \varphi},\\
&\text{div}\,P=0,\\
&P=0,\quad T=0\qquad \text{at } r=r_0,r_1,
\end{aligned}\tag{7.4.17}$$

where $r=(T_0-T_1)/(r_1-r_0)$.

The eigenvalue equations of (7.4.17) are given by

$$\begin{aligned}
&-\Delta P_\varphi+\frac{1}{\nu}(\frac{\widetilde{P}_\varphi}{r}+\frac{d\widetilde{P}_\varphi}{dr})P_r+\frac{1}{r\nu}\frac{\partial p}{\partial \varphi}=\lambda P_\varphi,\\
&-\Delta P_r-\frac{2\widetilde{P}_\varphi}{\alpha\nu r}P_\varphi-\frac{\rho\beta M_r G}{\alpha\nu r^2}T+\frac{1}{\alpha\nu}\frac{\partial p}{\partial r}=\lambda P_r,\\
&-\Delta T+\frac{1}{\rho r}\widetilde{P}_\varphi\frac{\partial T}{\partial \varphi}-\frac{r_0 r_1\gamma}{\kappa\rho r^2}P_r=\lambda T,\\
&\text{div}\,P=0,\\
&P=0,\quad T=0\qquad \text{at } r=r_0,r_1.
\end{aligned}\tag{7.4.18}$$

The eigenvalues λ of (7.4.18) are discrete (not counting multiplicity):

$$\lambda_1>\lambda_2>\cdots>\lambda_k>\cdots,\qquad \lambda_k\to-\infty \text{ as } k\to\infty.$$

The first eigenvalue λ_1 and first eigenfunctions

$$\Phi=(P_\varphi^0,P_r^0,T^0)\tag{7.4.19}$$

dictate the dynamic behaviors of spiral galaxies, which are determined by the physical parameters:

$$\zeta_0,\zeta_1,r_0,r_1,\kappa,\nu,\beta,\gamma=\frac{T_0-T_1}{r_1-r_0},\ \delta=\frac{2M_0G}{c^2r_0},\ M_r=M_0+4\pi\int_{r_0}^{r_1}r^2\rho dr.\tag{7.4.20}$$

Based on the dynamic transition theory in (Ma and Wang, 2013b), we have the following physical conclusions:

- If the parameters in (7.4.20) make the first eigenvalue $\lambda_1<0$, then the spiral galaxy is of S0-type.

- If $\lambda_1 > 0$, then the galaxy is one of the types $S_a, S_b, S_c, SB_a, SB_b, SB_c$, depending on the structure of (P^0_φ, P^0_r) in (7.4.19).
- Let $\lambda_1 > 0$ and the first eigenvector (P^0_φ, P^0_r) of (7.4.19) have the vortex structure as shown in Figure 7.8. Then the number of spiral arms of the galaxy is $k/2$, where k is the vortex number of (P^0_φ, P^0_r). Hence, if $k = 2$, the galaxy is of the SB_c-type.

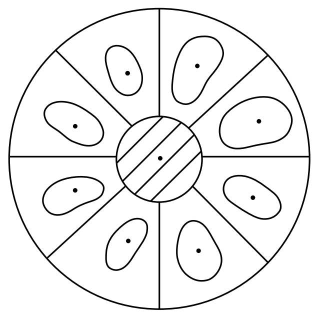


Figure 7.8 The vortex structure of the first eigenvector (P^0_φ, P^0_r).

The reason behind the number of spiral arms being $k/2$ is as follows. First the number of vortices in Figure 7.8 is even, and each pair of vortices have reversed orientations. Second, if the orientation of a vortex matches that of the stationary solution $P_\varphi(r)$ of (7.4.14), then the superposition of $P_\varphi(r)$ and P^0_φ of (7.4.19) gives rise to an arm; otherwise, the counteraction of $P_\varphi(r)$ and P^0_φ with reversed orientations reduces the energy momentum density, and the region becomes nearly void.

Remark 7.17 There are three terms in (7.4.18), which may generate the transition of (7.4.17):

$$F_1 = \left(0, -\frac{k_1 T}{r^2}, -\frac{k_2 P_r}{r^2}\right), \qquad k_2 = \frac{\rho\beta M_r G}{\alpha\nu}, \quad k_2 = \frac{r_0 r_1 \gamma}{\kappa\rho},$$

$$F_2 = \left(\frac{1}{\nu}\left(\frac{\widetilde{P}_\varphi}{r} + \frac{d\widetilde{P}_\varphi}{dr}\right)P_r, -\frac{2\widetilde{P}_\varphi}{\alpha\nu r}P_\varphi, 0\right),$$

$$F_3 = \left(-\frac{1}{2\alpha^2 r}\frac{d\alpha}{dr}\frac{\partial}{\partial r}(rP_\varphi), \frac{1}{2\alpha}\frac{\partial}{\partial r}(\frac{1}{\alpha}\frac{d\alpha}{dr}P_r), 0\right).$$

The term F_1 corresponds to the Rayleigh-Bénard convection with the Rayleigh number

$$R = k_1 k_2 = \frac{\beta M_r G r_0 r_1 \gamma}{\alpha\nu\kappa},$$

the term F_2 corresponds to the Taylor rotation which causes the instability of the basic flow $(P_\varphi, P_r) = (\widetilde{P}_\varphi, 0)$, and F_3 is the relativistic effect which only plays a role in the case where $\delta \simeq 1$. □

7.4.4 Active galactic nuclei (AGN) and jets

The black hole core of a galaxy attracts a large amounts of gases around it, forming a galactic nucleus. The mass of a galactic nucleus is usually in the range

$$10^5 M_\odot \sim 10^9 M_\odot. \tag{7.4.21}$$

Galactic nuclei are divided into two types: normal and active. In particular, an active galactic nucleus emits huge quantities of energy, called jets. We focus in this section the mechanism of AGN jets.

1. *Model for AGN.* The domain of an galactic nucleus is a spherical annulus:

$$B = \left\{ x \in \mathbb{R}^3 \mid R_s < |x| < R_1 \right\}, \tag{7.4.22}$$

where R_s is the Schwarzschild radius of the black hole core, and R_1 is the radius of the galaxy nucleus.

The model governing the galaxy nucleus is given by (7.4.7)-(7.4.8), defined in the domain (7.4.22) with boundary conditions:

$$\begin{aligned} &P_r = 0,\ \frac{\partial P_\theta}{\partial r} = 0,\ P_\varphi = P_0,\ T = T_0 \quad \text{for } r = R_s, \\ &P_r = 0,\ \frac{\partial P_\theta}{\partial r} = 0,\ P_\varphi = P_1,\ T = T_1 \quad \text{for } r = R_1. \end{aligned} \tag{7.4.23}$$

Let the stationary solution of the model be as

$$P_\theta = 0, \quad P_r = 0, \quad P_\varphi = P_\varphi(r,\theta),$$

and p, ρ, T be independent of φ. Then the stationary equations for the four unknown functions P_φ, T, p, ρ are in the form

$$\begin{aligned} &\frac{\partial p}{\partial \theta} = \frac{1}{\rho}\frac{\cos\theta}{\sin\theta}P_\varphi^2, \\ &\frac{\partial p}{\partial r} = \frac{1}{\rho r}P_\varphi^2 - \rho(1-\beta T)\frac{M_b G}{r^2}, \\ &-\nu\widetilde{\Delta}P_\varphi + \frac{P_\varphi}{r^2\sin\theta} + \frac{1}{2\alpha^2 r}\frac{d\alpha}{dr}\frac{\partial}{\partial r}(rP_\varphi) = 0, \\ &-\frac{\kappa}{\alpha r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)T = Q(r), \end{aligned} \tag{7.4.24}$$

where M_b is the mass of the black hole core, Q is the heat source generated by the nuclear burning, and

$$\begin{aligned} &\widetilde{\Delta} = \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\alpha r^2}\frac{\partial}{\partial r}(r^2\frac{\partial}{\partial r}), \\ &\alpha = \left(1 - \frac{2M_bG}{c^2 r}\right)^{-1}. \end{aligned}$$

The boundary conditions of (7.4.23) become

$$P_\varphi(R_s) = R_s\Omega_0, \quad P_\varphi(R_1) = R_1\Omega_1, \quad T(R_s) = T_0, \quad T(R_1) = T_1, \tag{7.4.25}$$

where Ω_0, Ω_1 only depend on θ, T_0, T_1 are constants.

Make the translation

$$P_r \to P_r, \quad P_\theta \to P_\theta, \quad P_\varphi \to P_\varphi + \widetilde{P}_\varphi, \quad T \to T + \widetilde{T}, \quad p \to p + \widetilde{p},$$

where $(\widetilde{P}_\varphi, \widetilde{T}, \widetilde{p}, \rho)$ is the solution of (7.4.24)-(7.4.25). Then the equations (7.4.7) are rewritten as

$$\begin{aligned}
&\frac{\partial P_\theta}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)P_\theta = \nu\Delta P_\theta - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial P_\theta}{\partial\varphi} + \frac{2\cos\theta\widetilde{P}_\varphi}{\rho r\sin\theta}P_\varphi - \frac{1}{r}\frac{\partial p}{\partial\theta},\\
&\frac{\partial P_\varphi}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)P_\varphi = \nu\Delta P_\varphi - \frac{1}{\rho r}\frac{\partial\widetilde{P}_\varphi}{\partial\theta}P_\theta - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial P_\varphi}{\partial\varphi} - \frac{1}{\rho}\frac{\partial\widetilde{P}_\varphi}{\partial r}P_r\\
&\qquad\qquad - \frac{\widetilde{P}_\varphi}{\rho r}P_r - \frac{\cos\theta\widetilde{P}_\varphi}{\rho r\sin\theta}P_\theta - \frac{1}{r\sin\theta}\frac{\partial p}{\partial\varphi},\\
&\frac{\partial P_r}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)P_r = \nu\Delta P_r - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial P_r}{\partial\varphi} + \frac{2\widetilde{P}_\varphi}{\rho\alpha r}P_\varphi - \frac{1}{\alpha}\frac{\partial p}{\partial r} + \beta\rho\frac{M_bG}{\alpha r^2}T,\\
&\frac{\partial T}{\partial \tau} + \frac{1}{\rho}(P\cdot\nabla)T = \kappa\widetilde{\Delta}T - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial T}{\partial\varphi} - \frac{1}{\rho}\frac{d\widetilde{T}}{dr}P_r,\\
&\mathrm{div}P = 0,
\end{aligned} \tag{7.4.26}$$

with the boundary conditions

$$P_r = 0, \quad P_\varphi = 0, \quad \frac{\partial P_\theta}{\partial r} = 0, \quad T = 0 \quad \text{at } r = R_s, R_1. \tag{7.4.27}$$

2. *Taylor instability.* By the conservation of angular momentum and $R_1 \gg R_s$, the angular momentums Ω_0 and Ω_1 in (7.4.25) satisfy that

$$\Omega_0 \gg \Omega_1, \tag{7.4.28}$$

This property leads to the instability of the rotating flow represented by the stationary solution:

$$(P_r, P_\theta, P_\varphi) = (0, 0, \widetilde{P}_\varphi), \tag{7.4.29}$$

which is similar to the Taylor-Couette flow in a rotating cylinder. The rotating instability can generate a circulation in the galactic nucleus, as the Taylor vortices in a rotating cylinder, as shown in Figure 7.9. The instability is caused by the force $F=(F_r,F_\theta,F_\varphi,T)$ in the equations of (7.4.26) given by

$$\begin{aligned}
F_r &= \frac{2\widetilde{P}_\varphi}{\rho\alpha r}P_\varphi - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial P_r}{\partial\varphi},\\
F_\theta &= \frac{2\cos\theta\widetilde{P}_\varphi}{\rho r\sin\theta}P_\varphi - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial P_\theta}{\partial\varphi},\\
F_\varphi &= -\frac{1}{\rho}\left(\frac{\widetilde{P}_\varphi}{r}+\frac{\partial\widetilde{P}_\varphi}{\partial r}\right)P_r - \frac{1}{\rho r}\left(\frac{\cos\theta}{\sin\theta}\widetilde{P}_\varphi+\frac{\partial\widetilde{P}_\varphi}{\partial\theta}\right)P_\theta - \frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial P_\varphi}{\partial\varphi},\\
T &= -\frac{\widetilde{P}_\varphi}{\rho r\sin\theta}\frac{\partial T}{\partial\theta}.
\end{aligned} \tag{7.4.30}$$

3. *Rayleigh-Bénard instability.* Due to the nuclear reaction (fusion and fission) and the large pressure gradient, the galactic nucleus possesses a very large temperature gradient in (7.4.25) as

$$DT = T_0 - T_1, \tag{7.4.31}$$

which yields the following thermal expansion force in (7.4.26), and gives rise to the Rayleigh-Bénard convection:

$$F_r = \beta\rho\frac{M_bG}{\alpha r^2}T, \qquad T = \frac{1}{\rho}\frac{d\widetilde{T}}{dr}P_r. \tag{7.4.32}$$

4. *Instability due to the gravitational effects.* Similar to (7.2.7), there is a radial force in the term $\nu\Delta u_r$ of the third equation of (7.4.26):

$$F_r = \frac{\nu}{2\alpha}\frac{\partial}{\partial r}\left(\frac{1}{\alpha}\frac{d\alpha}{dr}P_r\right), \tag{7.4.33}$$

where

$$\alpha = (1-R_s/r)^{-1}, \qquad R_s < r < R_1. \tag{7.4.34}$$

In (7.4.33) and (7.4.34), we see the term

$$f_r = \frac{\nu}{1-R_s/r}\frac{R_s^2}{r^4}P_r, \tag{7.4.35}$$

which has the property that

$$f_r = \begin{cases} +\infty & \text{for } P_r > 0 \quad \text{at } r = R_s,\\ -\infty & \text{for } P_r < 0 \quad \text{at } r = R_s. \end{cases} \tag{7.4.36}$$

It is the force (7.4.36) that not only causes the instability of the basic flow (7.4.29), but also generates jets of the galaxy nucleus.

5. *Latitudinal circulation.* The above three types of forces: the rotating force (7.4.30), the thermal expansion force (7.4.32), and the gravitational effect (7.4.35), cause the instability of the basic flow (7.4.29) and lead to the latitudinal circulation of the galactic nucleus, as shown in Figure 7.9.

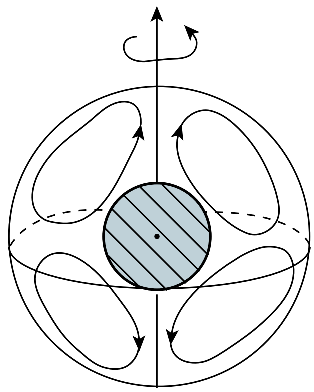


Figure 7.9 The latitudinal circulation with $k = 2$ cells.

6. *Jets and accretions.* Each circulation cell has an exit as shown in Figure 7.10, where the circulating gas is pushed up by the radial force (7.4.35)-(7.4.36), and erupts leading to a jet. The cell has an entrance as shown in Figure 7.10, where the exterior gas is pulled into the nucleus, is cyclo-accelerated by the force (7.4.35), goes down to the inner boundary $r = R_s$, and then is pushed by F_θ of (7.4.30) toward to the exit. Thus the circulation cells form jets in their exits and accretions in their entrances. In Figure 7.11(a), we see that there is a jet in the latitudinal circulation with $k = 1$ cell, and in Figure 7.11(b) there are two jets in the circulation with $k = 2$ cells in its south and north poles, and an accretion disk near its equator.

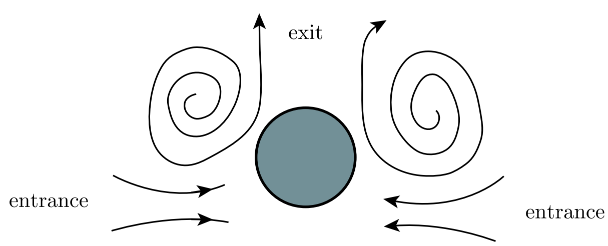


Figure 7.10

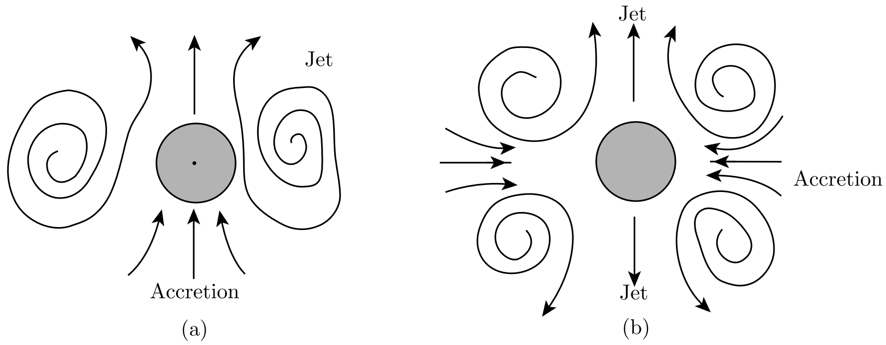


Figure 7.11 (a) A jet in the latitudinal circulation with $k=1$ cell, two jets in the latitudinal circulation with $k=2$ cells.

7. *Condition for jet generation.* The main power to generate jets comes from the gravitational effect of (7.4.35)-(7.4.36) by the black hole. The radial momentum P_r in (7.4.35) is the bifurcated solution of (7.4.26), which can be expressed as

$$P_r = R_s^2 Q_r,$$

where Q_r is independent of R_s. Thus, the radial force (7.4.35) near $r=R_s$ is approximatively written as

$$f_r = \frac{\nu}{1-R_s/r} Q_{R_s}, \quad r = R_s + \widetilde{r} \qquad \text{for } 0 < \widetilde{r} \ll R_s. \tag{7.4.37}$$

Let f_E be the lower limit of the effective force, which is defined as that the total radial force F_r in the third equation of (7.4.26) is positive provided $f_r > f_E$:

$$F_r > 0 \quad \text{if} \quad f_r > f_E.$$

Let R_E be the effective distance:

$$f_r > f_E \quad \text{if} \quad R_s < r < R_s + R_E.$$

Then, it follows from (7.4.37) that

$$R_E = kR_s \qquad (k = \nu Q_{R_s}/f_E). \tag{7.4.38}$$

It is clear that there is a critical distance R_c such that

$$\begin{aligned} &\text{a jet forms} \quad \text{if } R_E > R_c \quad \text{or } R_s > k^{-1}R_c, \\ &\text{no jet forms} \quad \text{if } R_E < R_c \quad \text{or } R_s < k^{-1}R_c. \end{aligned} \tag{7.4.39}$$

The criterion (7.4.39) is the condition for jet generation.

The condition (7.4.39) can be equivalently rewritten as that there is a critical mass M_c such that the galactic nucleus is active if its mass M is bigger than M_c, i.e. $M > M_c$. By (7.4.21), we have

$$10^5 M_\odot < M_c \qquad \text{or} \qquad 10^6 M_\odot < M_c.$$

Remark 7.18 The jets shown in Figures 7.9 and 7.10 are column-shaped. If the cell number $k \geqslant 3$ for the latitudinal circulation of galaxy nucleus, then there are jets which are disc-shaped. We don't know if there exist such galaxy nuclei which have the disc-shaped jets in the Universe. Theoretically, it appears to be possible. □

Remark 7.19 Galactic nucleus are made up of plasm. The precise description of AGN jets requires to take into consideration of the magnetic effect in the modeling. However the essential mechanism does not change and an explosive magnetic energy as in (7.4.37) will contribute to the supernovae explosion. □

7.5 The Universe

7.5.1 Classical theory of the Universe

In this section, we recall some basic aspects of modern cosmology, including the Hubble Law, the expanding universe, and the origin of our Universe, together with their experimental justifications.

1. *The Hubble Law.* In 1929, American astronomer Edwin Hubble discovered an approximatively linear relation between the recession velocity v and the distance R of remote galaxies, which is now called the Hubble Law:

$$v = HR, \tag{7.5.1}$$

where H is called the Hubble constant, depends on time, and its present-time value is

$$H = 70\ \text{km/s} \cdot \text{Mpc}, \qquad \text{Mpc} = 10^6 \text{pc}\ (1\ \text{pc} = 3.26\ \text{ly}). \tag{7.5.2}$$

Formula (7.5.1) implies that the farther away the galaxy is from our galaxy, the greater its velocity is.

2. *Expansion of the Universe.* An important physical conclusion from the Hubble Law (7.5.1)-(7.5.2) is that our Universe is expanding.

If we regard our Universe as a 3-dimensional sphere:

$$M = S^3_r = \{x \in \mathbb{R}^4 \mid x_1^2 + x_2^2 + x_3^2 + x_4^2 = r^2\}. \tag{7.5.3}$$

Each point on the sphere S^3_r can be regarded as a center, as the radius r increases, all points on the sphere are moving away from the point. Moreover, the farther away a remote object is from the point, the faster the object appears to be moving. For example, for any two points p_1 and p_2 on a sphere S_r^3, θ is the angle between p_1 and p_2; see Figure 7.12. As the radius r varies from r_1 to r_2, the distance R between P_1 and P_2 varies from R_1 to R_2:

$$R_1 = \theta r_1 \rightarrow R_2 = \theta r_2.$$

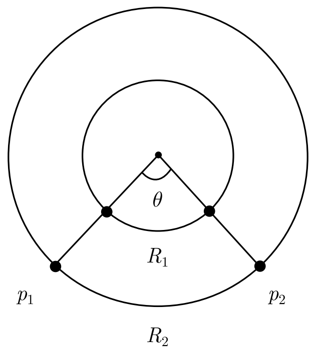


Figure 7.12

The velocity of separation of these two points is

$$v = \frac{dR}{dt} = \theta \dot{r}, \qquad (R = \theta r). \tag{7.5.4}$$

Assume that $\dot{r}$ is proportional to r, i.e.

$$\dot{r} = Hr, \qquad (H \text{ is a constant}). \tag{7.5.5}$$

Then the Hubble law (7.5.1) follows from (7.5.4) and (7.5.5).

The Hubble Law (7.5.1) also manifests that the recession velocity is independent of the direction. Hence, it implies that the Universe is isotropic. This very property gives rise to a strong restriction on the global structure of the Universe: there are only two types of geometry for our Universe: either the 3D sphere (7.5.3) or the 3D Euclidian space $\mathbb{R}^3$. Namely, we have reached the following physical conclusion.

Physical Conclusion 7.20 *The Hubble law allows only two types of topological structure, either S^3 or R^3, as our Universe. Further, the Hubble Law requires that the Universe is expanding.*

3. *Big-Bang theory.* In 1927, the Belgium cosmologist G. Lemaître first proposed that the Universe begins with a big explosion, known as the Big-Bang theory. After the Hubble law was discovered, in 1931, G. Lemaître regarded the recessions of remote galaxies as the

expanding of the Universe, and he thought that as the time inverses, the early universe must be in a high temperature and dense state, and the stage is considered as the beginning of the Big-Bang.

The Big-Bang theory appears to be consistent with the three most important astronomical facts: the Hubble law, the cosmic microwave background, discovered in 1965 by two physicists A. Penzias and R. Wilson, and the abundance of helium.

4. *Age of the Universe.* Based on the Big-Bang theory, the age of the Universe is finite. Let T be the age, then the present-day distance between any two points is

$$R = vT.$$

It follows from (7.5.1) that

$$T = \frac{1}{H}, \tag{7.5.6}$$

where H is the Hubble constant. We infer then from (7.5.2) and (7.5.6) the age T of the Universe as

$$T = 1.4 \times 10^{10} \text{ year.} \tag{7.5.7}$$

As we approximatively regard the recession velocity v of remote galaxies as the speed of light, i.e. $v \simeq c$, the age (7.5.7) is also considered as the radius of the Universe.

5. *Structure of universes.* A central topic in modern cosmology is to investigate both the global topological and geometrical structure of the Universe. A fundamental principle of cosmology, called cosmological principle, states that ignoring local irregularities, the Universe is homogeneous and isotropic.

The cosmological principle is compatible with the Hubble law: Physical Conclusion 7.20, and can be equivalently stated in the following form.

Theorem 7.21 (Topological Structure of the Universe) *The global topological structure of the Universe is either S^3 or $\mathbb{R}^3$.*

6. *Global geometrical structure.* Based on the Cosmological Principle 7.21, three physicists A. Friedmann (1922), G. Lemaître (1927), H. P. Robertson (1935), and a mathematician A. G. Walker (1936), independently derived the globally geometrical structure of the Universe.

Theorem 7.22 (Geometrical Structure of the Universe) *The Riemannian metric of 4D space-time manifold satisfying the Cosmological Principle 7.21 takes the following form*

$$ds^2 = -c^2dt^2 + R(t)\left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\varphi^2)\right], \tag{7.5.8}$$

where $R(t)$ is the scalar factor representing the radius of the Universe, which depends only on time t, and $k = 1, 0, -1$ stand for the sign of space scalar curvature in the following sense:

$$
\begin{aligned}
&k = 1: && M = S^3 \text{ the 3D sphere with scalar curvature} = \frac{1}{R^2},\\
&k = 0: && M = \mathbb{R}^3 \text{ the 3D flat Euclidean space},\\
&k = -1: && M = L^3 \text{ the Lobachevsky space with scalar curvature} = -\frac{1}{R^2},
\end{aligned}
\tag{7.5.9}
$$

where the Lobachevsky space L^3 has the same topological structure as $\mathbb{R}^3$.

Historically, Theorem 7.22 was rigorously proved by Robertson and Walker, and was assumed by Friedmann and Lemaître to derive the field equations satisfied by $R(t)$.

7. *The Newton cosmology.* The Newton cosmology is based on the Newton Gravitational Law. By Cosmological Principle (Roos, 2003), the universe is spherically symmetric. For any reference point $p \in \mathscr{M}$, the motion equation of an object with distance r from p is

$$\frac{d^2 r}{dt^2} = -\frac{GM(r)}{r^2}, \tag{7.5.10}$$

where $M(r) = 4\pi r^3 \rho/3$, and ρ is the mass density. Thus, (7.5.10) can be rewritten as follows

$$r'' = -\frac{4}{3}\pi G r \rho. \tag{7.5.11}$$

Make the nondimensional

$$r = R(t) r_0,$$

where $R(t)$ is the scalar factor, which is the same as in the FLRW metric (Ma and Wang, 2014e). Let ρ_0 be the density at $R = 1$. Then we have

$$\rho = \rho_0 / R^3. \tag{7.5.12}$$

Thus, equation (7.5.11) is expressed as

$$R'' = -\frac{4\pi G}{3}\frac{\rho_0}{R^2}, \tag{7.5.13}$$

which is the dynamic equation of Newtonian cosmology.

Multiplying both sides of (7.5.13) by R' we have

$$\frac{d}{dt}\left(\dot{R}^2 - \frac{8\pi G}{3}\frac{\rho_0}{R}\right) = 0.$$

Hence, (7.5.13) is equivalent to the equation

$$\dot{R}^2 = \frac{8\pi G}{3}\frac{\rho_0}{R} - \kappa, \tag{7.5.14}$$

where κ is a constant, and we shall see that $\kappa = kc^2$, and $k = -1, 0$, or 1.

8. *The Friedmann cosmology.* The nonzero components of the Friedmann metric are

$$g_{00} = -1, \quad g_{11} = \frac{R^2}{1-kr^2}, \quad g_{22} = R^2 r^2, \quad g_{33} = R^2 r^2 \sin^2\theta.$$

Again by the Cosmological Principle (Roos, 2003), the energy-momentum tensor of the Universe is in the form

$$T_{\mu\nu} = \begin{pmatrix} \rho c^2 & 0 & 0 & 0 \\ 0 & g_{11}p & 0 & 0 \\ 0 & 0 & g_{22}p & 0 \\ 0 & 0 & 0 & g_{33}p \end{pmatrix}.$$

By the Einstein gravitational field equations

$$\begin{aligned} &R_{\mu\nu} = -\frac{8\pi G}{c^4}(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T), \\ &D^\mu T_{\mu\nu} = 0, \end{aligned}$$

we derive three independent equations

$$\ddot{R} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right)R, \tag{7.5.15}$$

$$R\ddot{R} + 2\dot{R}^2 + 2kc^2 = 4\pi G\left(\rho - \frac{p}{c^2}\right)R^2, \tag{7.5.16}$$

$$\dot{\rho} = -3\left(\frac{\dot{R}}{R}\right)\left(\rho + \frac{p}{c^2}\right), \tag{7.5.17}$$

where R, ρ, p are the unknown functions.

Equations (7.5.15)-(7.5.17) are called the Friedmann cosmological model, from which we can derive the Newtonian cosmology equations (7.5.14). To see this, by (7.5.15) and (7.5.16), we have

$$\left(\frac{\dot{R}}{R}\right)^2 = -\frac{kc^2}{R^2} + \frac{8\pi G}{3}\rho. \tag{7.5.18}$$

By the approximate $p/c^2 \simeq 0$, (7.5.12) follows from (7.5.17). Then we deduce (7.5.14) from (7.5.18) and (7.5.12).

From the equation (7.5.18), the density ρ_c corresponding to the case $k = 0$ is

$$\rho_c = \frac{3}{8\pi G}\left(\frac{\dot{R}}{R}\right)^2 = \frac{3}{8\pi G}H^2, \tag{7.5.19}$$

where $H = \dot{R}/R$ is the Hubble constant, and by (7.5.2) we have

$$\rho_c = 10^{-26}\text{kg/m}^3. \tag{7.5.20}$$

Thus, by the Friedmann model we can deduce the following conclusions.

Conclusions of Friedmann Cosmology 7.23

1) *By (7.5.18) we can see that*

$$\begin{aligned} &\rho > \rho_c \quad \Leftrightarrow k=1 \quad \text{the Universe is closed}: \ M=S^3, \\ &\rho = \rho_c \quad \Leftrightarrow k=0 \quad \text{the Universe is open}: \ M=\mathbb{R}^3, \\ &\rho < \rho_c \quad \Leftrightarrow k=-1 \quad \text{the Universe is open}: \ M=L^3. \end{aligned} \tag{7.5.21}$$

2) *Let E_0 be the total kinetic energy of the Universe, M is the mass, then we have*

$$E_0 = \begin{cases} \frac{3}{5}\frac{GM^2}{R} & \text{for } k=0, \\ \frac{2}{3\pi}\frac{GM^2}{R} - \frac{1}{2}Mc^2 & \text{for } k=1, \end{cases} \tag{7.5.22}$$

where the first term represents the total gravitational bound potential energy, and the second term is the energy resisting curvature tensor.

3) *By (7.5.18), $\dot{R} \not\equiv 0$, and consequently the universe is dynamic.*

4) *By (7.5.15), $\ddot{R} < 0$, the dynamic universe is decelerating.*

9. *The Lemaître cosmology.* Consider the Einstein gravitational field equations with the cosmological constant Λ term:

$$R_{\mu\nu} = -\frac{8\pi G}{c^4}(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T) + \Lambda g_{\mu\nu}, \quad \Lambda > 0, \tag{7.5.23}$$

then the metric (7.5.8) satisfies the following equations

$$\ddot{R} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right)R + \frac{\Lambda c^2}{3}R, \tag{7.5.24}$$

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda c^2}{3} - \frac{kc^2}{R^2}, \tag{7.5.25}$$

$$\dot{\rho} + 3\left(\frac{\dot{R}}{R}\right)\left(\rho + \frac{p}{c^2}\right) = 0. \tag{7.5.26}$$

The equations (7.5.24)-(7.5.26) are known as the Lemaître cosmological model, or the Λ-cosmological model, which leads to the following conclusions of Λ-cosmology:

Conclusions of Λ-Cosmology 7.24

1) *The Universe is temporally open*: $R \to \infty$ *as* $t \to \infty$;

2) *There is a critical radius R_c, such that*

$$\begin{aligned} &\text{the universe is decelerating} && \text{for } R < R_c, \\ &\text{the universe is accelerating} && \text{for } R > R_c, \end{aligned}$$

where $R_c \simeq \left(\frac{4\pi G\rho_0}{\Lambda c^2}\right)^{1/3}$;

3) *As $t \to \infty$ and $\rho \to 0$, then we deduce from (7.5.25) that the cosmological radius R has the asymptotic relation*

$$R \sim e^{\sqrt{\Lambda c^2/3}t} \qquad \text{as } t \to \infty. \tag{7.5.27}$$

Namely

$$R(t)/e^{\sqrt{\Lambda c^2/3}t} = \text{const.} \qquad \text{as } t \to \infty;$$

4) *The total kinetic energy E is given by*

$$E = \begin{cases} \dfrac{3}{5}\dfrac{GM^2}{R} + \dfrac{\Lambda}{10}Mc^2R^2 & \text{for } k = 0, \\ \dfrac{2}{3\pi}\dfrac{GM^2}{R} + \dfrac{\Lambda}{6}Mc^2R^2 - \dfrac{1}{2}Mc^2 & \text{for } k = 1. \end{cases} \tag{7.5.28}$$

Remark 7.25 The field equations (7.5.23) with a cosmological constant Λ lead to a special conclusion that in the expansion process, there are a large quantities of energy to be created, and the added energy in (7.5.28) is generated by Λ is as

$$\frac{1}{6}Mc^2\Lambda R^2 (k = 1) \quad \text{and} \quad \frac{\Lambda}{10}Mc^2R^2(k = 0).$$

It implies that the total energy is not conserved in the Λ-model.

7.5.2 Globular universe with boundary

If the spatial geometry of a universe is open, then by our theory of black holes developed in Section 7.3, we have shown that the universe must be in a ball of a black hole with a fixed radius. In fact, according to the basic cosmological principle that the universe is homogeneous and isotropic (Roos, 2003), given the energy density $\rho_0 > 0$ of the universe, by Theorem 7.3, the universe will always be bounded in a black hole of open ball with the Schwarzschild radius:

$$R_s = \sqrt{\frac{3c^2}{8\pi G\rho_0}},$$

as the mass in the ball B_{R_s} is given by $M_{R_s} = 4\pi R_s^3 \rho_0/3$. This argument also clearly shows that

$$\boxed{\text{there is no unbounded universe.}}$$

In addition, since a black hole is unable to expand and shrink, by property (7.5.55) of black holes, all globular universes must be static.

Globular universe

We have shown that the universe is bounded, and suppose that the universe is open, i.e. its topological structure is homeomorphic to $\mathbb{R}^3$, and it begins with a ball. Let E be its total energy:

$$E = \text{mass} + \text{kinetic} + \text{thermal} + \Psi, \tag{7.5.29}$$

where Ψ is the energy of all interaction fields. Let

$$M = E/c^2. \tag{7.5.30}$$

At the initial stage, all energy is concentrated in a ball with radius R_0. By the theory of black holes, the energy contained in the ball generates a black hole in $\mathbb{R}^3$ with radius

$$R_s = \frac{2MG}{c^2}, \tag{7.5.31}$$

provided $R_s \geqslant R_0$; see (7.3.1) and Figure 7.3.

Thus, if the universe is born to a ball, then it is immediately trapped in its own black hole with the Schwarzschild radius R_s of (7.5.31). The 4D metric inside the black hole of the static universe is given by

$$ds^2 = -\psi(r)c^2dt^2 + \alpha(r)dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \tag{7.5.32}$$

where ψ and α satisfy the equations (7.1.79) and (7.1.80) with boundary conditions:

$$\psi \to \widetilde{\psi}, \quad \alpha \to \widetilde{\alpha} \quad \text{as} \quad r \to R_s,$$

where R_s is given by (7.5.31). Also, $\widetilde{\psi}, \widetilde{\alpha}$ are given via the TOV metric (7.1.40)-(7.1.41):

$$\widetilde{\psi} = \frac{1}{4}\left(1 - \frac{r^2}{R_s^2}\right), \quad \widetilde{\alpha} = \left(1 - \frac{r^2}{R_s^2}\right)^{-1} \qquad \text{for } 0 \leqslant r < R_s. \tag{7.5.33}$$

Basic problems

A static universe is confined in a ball with fixed radius R_s in (7.5.31), and the ball behaves like a black hole. We need to examine a few basic problems for a static universe, including the cosmic edge, the flatness, the horizon, the redshift, and the cosmic microwave background (CMB) radiation problems.

1. *The cosmic edge problem.* In the ancient Greece, the cosmic-edge riddle was proposed by the philosopher Archytas, a friend of Plato, who used "what happens when a spear is thrown across the outer boundary of the Universe?" The problem appears to be very difficult to answer. Hence, for a long time physicists always believe that the Universe is boundless.

Our theory of black holes presented in Section 7.3 shows that all objects in a globular universe cannot reach its boundary $r = R_s$. In particular, an observer in any position of the globular universe looking toward to the boundary will see no boundary due to the openness of the ball and the relativistic effect near the Schwarzschild surface. Hence the cosmic-edge riddle is no longer a problem.

2. *The flatness problem.* In modern cosmology, the flatness problem means that $k = 0$ in the FLRW metric (7.5.8)-(7.5.9). It is common to think that the flatness of the universe is equivalent to the fact that the present energy density ρ must be equal to the critical value given by (7.5.20). In fact, mathematically the flatness means that any geodesic triangle has the inner angular sum $\pi = 180^\circ$.

Measurements by the WMAP (Wilson Microwave Anisotropy Probe) spacecraft in the last ten years indicated that the Universe is nearly flat. The present radius of the Universe is about

$$R = 10^{26}\text{m}. \tag{7.5.34}$$

If the Universe is static, then (7.5.34) gives the Schwarzschild radius (7.5.31), from which it follows that the density ρ of our Universe is just the critical density of (7.5.20):

$$\rho = \rho_c = 10^{-26}\text{kg}/\text{m}^3. \tag{7.5.35}$$

Thus, we deduce that if the universe is globular, then it is static. In addition, we have shown that any universe is bounded and confined in a 3D hemisphere of a black hole or in a 3D sphere as shown in Figure 7.14. Hence as the radius is sufficiently large, the universe is nearly flat.

3. *The horizon problem.* The cosmic horizon problem can be simply stated as that all places in a universe look as the same. It seems as if the static Universe with boundary violates the horizon problem. However, due to the gravitational lensing effect, the light bents around a massive object. Hence, the boundary of a globular universe is like a concave spherical mirror, and all lights reaching close to it will be reflected back, as shown in Figure 7.13. It is this lensing effect that makes the globular universe looks as if everywhere is the same, and is horizontal. In Figure 7.13, if we are in position x, then we can also see a star as if it is in position $\widetilde{y}$, which is actually a virtual image of the star at y.

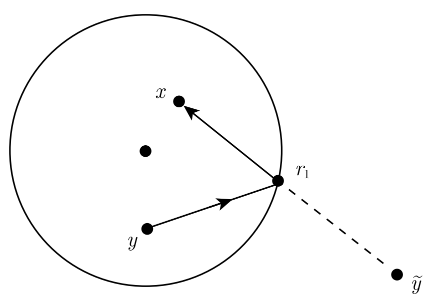


Figure 7.13 Due to the lensing effect, one at x can also see the star at y as if it is another star at $\widetilde{y}$.

4. *The redshift problem.* Observations show that light coming from a remote galaxy is redshifted, and the farther away the galaxy is, the larger the redshift is. In astronomy, it is

customary to characterize redshift by a dimensionless quantity z in the formula

$$1+z=\frac{\lambda_{\text{observ}}}{\lambda_{\text{emit}}},$$

where λ_{observ} and λ_{emit} represent the observed and emitted wavelenths. There are three redshift types:

Doppler effect, cosmological redshift, gravitational redshift.

The gravitational redshift in a black hole are caused by both the gravitational fields of the emitting object and the black hole.

The first type of redshift, due to the gravitational field, is formulated as

$$1+z=\frac{\sqrt{1-\frac{2mG}{c^2r}}}{\sqrt{1-\frac{2mG}{c^2r_0}}}, \tag{7.5.36}$$

where m is the mass of the **emitting object**, r_0 is its radius, and r is the distance between the object and the observer.

The second type of redshifts, due to the cosmological effect or black hole effect, is

$$1+z=\frac{\sqrt{-g_{00}(r_0)}}{\sqrt{-g_{00}(r_1)}}, \tag{7.5.37}$$

where g_{00} is the time-component of the black hole gravitational metric, r_0 and r_1 are the positions of the observer and the emitting object (including virtual images).

If a universe is not considered as a black hole, then the gravitational redshift is simply given by (7.5.36) and is very small for remote objects. Likewise, the cosmological redshift is also too small to be significant. Hence, astronomers have to think the main portion of the redshift is due to the Doppler effect:

$$1+z=\frac{\sqrt{1+v/c}}{\sqrt{1-v/c}}. \tag{7.5.38}$$

When v/c is small, (7.5.38) can be approximatively expressed as

$$z\simeq v/c. \tag{7.5.39}$$

In addition, Hubble discovered that the redshift has an approximatively linear relation with the distance:

$$z\simeq kR, \qquad k \text{ is a constant.} \tag{7.5.40}$$

Thus, the Hubble Law (7.5.1) follows from (7.5.39) and (7.5.40). It is the Hubble Law (7.5.1) that leads to the conclusion that our Universe is expanding.

However, if we adopt the view that the globular universe is in a black hole with the Schwarzschild radius R_s as in (7.5.31), the black hole redshift (7.5.37) cannot be ignored. By (7.5.32) and (7.5.33), the time-component g_{00} for the black hole can approximatively take the TOV solution as r near R_s:

$$g_{00} = -\frac{1}{4}\left(1 - \frac{r^2}{R_s^2}\right), \qquad \text{for } r \text{ near } R_s.$$

Hence, the redshift (7.5.37) is as

$$1 + z = \frac{\sqrt{1 - r_0^2/R_s^2}}{\sqrt{1 - r_1^2/R_s^2}}, \qquad \text{for } r_0, r_1 < R_s. \tag{7.5.41}$$

It is known that for a remote galaxy, r_1 is close to the boundary $r = R_s$. Therefore by (7.5.41) we have

$$z \to +\infty \quad \text{as} \quad r_1 \to R_s.$$

It reflects the redshifts observed from most remote objects. If the object is a virtual image as shown in Figure 7.13, then its position is the reflection point r_1. Thus, we see that even if the remote object is not moving, its redshift can still be very large.

5. *CMB problem.* In 1965, two physicists A. Penzias and R. Wilson discovered the low-temperature cosmic microwave background (CMB) radiation, which fills our Universe, and it is ever regarded as the Big-Bang product. However, for a static closed Universe, it is the most natural thing that there exists a CMB, because the Universe is a black-body and CMB is a result of black-body radiation.

6. *None expanding Universe.* As the energy of the Universe is given, the maximal radius, i.e. the Schwarzschild radius R_s, is determined, and the boundary is invariant. In fact, a globular universe must fill the ball with the Schwarzschild radius, although the distribution of the matter in this ball may be slightly non-homogenous. The main reason is that if the universe has a radius R smaller than R_s, then it must contain at least a sub-black hole with radius R_0 as follows

$$R_0 = \sqrt{\frac{R}{R_s}}R.$$

In Section 7.5.4 we shall discuss this topic.

7.5.3 Spherical Universe without boundary

Bounded universe has finite energy and space, and our Universe is bounded as we have demonstrated in the last section. Besides the globular universe, another type of bounded universe is the spherically-shaped corresponding to the $k = 1$ case in the Friedmann model (7.5.15)-(7.5.17).

A globular universe must be static. With the same argument, a spherical closed universe have to be static as well. In this subsection, we are devoted to investigate the spherical cosmology.

1. *Cosmic radius.* For a static spherical universe, its radius R_c satisfies that

$$\dot{R}_c = 0, \qquad \ddot{R}_c = 0.$$

By the Friedmann equation (7.5.18), it leads to that

$$R_c^2 = \frac{3c^2}{8\pi G\rho}. \tag{7.5.42}$$

For a 3D sphere, its volume V is given by

$$V = 2\pi^2 R^3.$$

Thus, $\rho = M/2\pi^2 R_c^3$, and by (7.5.43) we get the radius R_c as

$$R_c = \frac{4MG}{3\pi c^2}. \tag{7.5.43}$$

This value (7.5.43) is also the maximal radius for a (possibly) oscillatory spherical universe.

2. *Negative pressure.* By (7.5.15) and $\ddot{R}_c = 0$, the pressure is negative:

$$p = -\frac{\rho c^2}{3}. \tag{7.5.44}$$

In order to resist the gravitational pulling, it is natural that there is a negative pressure in a static universe, which originates from three sources:

thermal expanding, radiation pressure, and dark energy.

These three types of forces are repulsive, and therefore yield the negative pressure as given by (7.5.44).

In fact, in our Universe both thermal and radiation (microwave radiation) pressures are very small. The main negative pressure is generated by the so called "dark energy". In (Ma and Wang, 2014e), we have shown that the dark energy is the repulsive gravitational effect for a remote object of great distance. From the field theoretical point view, dark energy is an effect of the dual gravitational field Ψ_μ in the PID-induced gravitational field equations (4.4.10) discovered by the authors.

3. *Equivalence.* It seems that both spherical and globular geometries are very different. However, in the following we show that they are equivalent in cosmology. In fact, as the space-time curvature is caused by gravitation, a globular universe must be a 3D hemisphere as shown in Figure 7.14(a), and a spherical universe is as shown in Figure 7.14(b), which is a 3D sphere piecing the upper and lower hemispheres together.

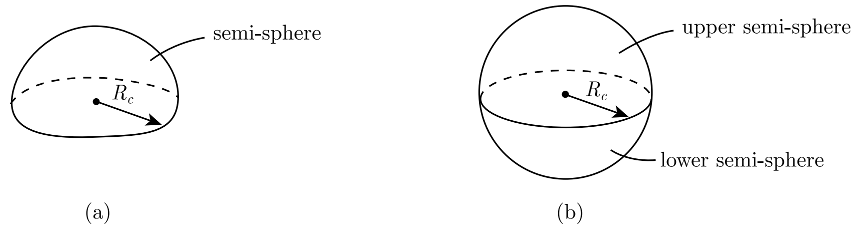


Figure 7.14 (a) A 3D hemisphere, and (b) a 3D sphere piecing the upper and lower hemispheres together.

In cosmology, the globular universe is a black hole, which likes as a 3D hemisphere, and the spherical universe can be regarded as if there were two hemispheres of black holes attached together.

We show this version from the cosmological dynamics.

First, by the Newtonian cosmological equation (7.5.14), i.e.

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{R^2}. \tag{7.5.45}$$

For a static universe in a black hole with maximal radius R_c, the equation (7.5.45) becomes

$$\dot{R}_c = 0 \quad \Leftrightarrow \quad \frac{8\pi G}{3}\rho = \frac{kc^2}{R_c^2}. \tag{7.5.46}$$

The volume of the hemisphere is

$$V_0 = \int_0^{2\pi} d\theta \int_0^{\pi} d\varphi \int_0^{\frac{\pi}{2}} R_c^3 \sin\theta \sin^2\psi d\psi = \pi^2 R_c^3,$$

where $x \in \mathbb{R}^4$ takes the spherical coordinate:

$$(x_1, x_2, x_3, x_4) = (R_c \sin\psi \sin\theta \cos\varphi, R_c \sin\psi \sin\theta \sin\varphi, R_c \sin\psi \cos\theta, R_c \cos\psi). \tag{7.5.47}$$

Thus, the mass density is

$$\rho = M_{\text{total}}/\pi^2 R_c^3, \tag{7.5.48}$$

Then it follows from (7.5.46) that

$$R_c = \frac{8GM_{\text{total}}}{3\pi c^2} \quad \text{for } k = 1. \tag{7.5.49}$$

Remark 7.26 The mass M_{total} in (7.5.48) contains the energy contributed by the space curvature, i.e.

$$M_{\text{total}} = M + \text{space curved energy},$$

where M is the mass of the flat space. By the invariance of density,

$$M/\frac{4\pi}{3}R_c^3 = M_{\text{total}}/\pi^2 R_c^3,$$

we get the relation

$$M_{\text{total}} = \frac{3\pi}{4}M. \tag{7.5.50}$$

With the flat space mass (7.5.50), from (7.5.49) we get the Schwarzschild radius $R_s = R_c$ for the cosmic black hole as follows

$$R_s = 2GM/c^2.$$

It means that the globular universe is essentially hemispherically-shaped. In particular the relation (7.5.50) can be generated to an arbitrary region $\Omega \subset \mathbb{R}^3$, i.e.

$$M_{\Omega;\text{total}} = \frac{V_\Omega}{|\Omega|}M_\Omega, \tag{7.5.51}$$

where M_Ω is the flat space mass in $\Omega, M_{\Omega;\text{total}}$ is the curved space mass, $|\Omega|$ is the volume of flat Ω,

$$V_\Omega = \int_\Omega \sqrt{g}dx, \quad g = \det(g_{ij}),$$

and g_{ij} $(1 \leqslant i, j \leqslant 3)$ is the spatial gravitational metric. □

Now, we return to the Friedmann model (7.5.18) with $k = 1$, which has the same form as that of the globular dynamic equation (7.5.45), and is of the same maximal radius R_c as that in (7.5.49). Hence, it is natural that a static spherical universe is considered as if there were two hemispherical black holes attached together. In fact, the static spherical universe forms an entire black hole as a closed space.

4. *Basic problems.* Since a static spherical universe is equivalent to two globular universes to be pieced together along with their boundary, an observer in its hemisphere is as if one is in a globular universe. Hence, the basic problems – the cosmic edge problem, flatness problem, horizon problem, and CMB problem– can be explained in the same fashion.

The redshift problem is slightly different, and the gravitational redshift is given by

$$1 + z = \frac{1}{\sqrt{-g_{00}(r)}}, \tag{7.5.52}$$

where r is the distance between the light source and the observer, and g_{00} is the time-component of the gravitational metric.

Due to the horizon of sphere, for an arbitrary point on a spherical universe, its opposite hemisphere relative to the point plays a similar role as a black hole. Hence, in the redshift formula (7.5.52), g_{00} can be approximatively taken as the Schwarzschild solution for distant objects as follows

$$-g_{00} = 1 - \frac{R_s}{\widetilde{r}}, \qquad R_s = \frac{2MG}{c^2}, \qquad \widetilde{r} = 2R_s - r \qquad \text{for } 0 \leqslant r < R_s,$$

where M is the cosmic mass of hemisphere, and $\widetilde{r}$ is the distance from the light source to the opposite radial point, and r is from the light source to the point. Hence, formula (7.5.52) can be approximatively written as

$$1 + z = \frac{1}{\sqrt{\alpha(r)\left(1 - \frac{R_s}{\widetilde{r}}\right)}} = \frac{\sqrt{2R_s - r}}{\sqrt{\alpha(r)(R_s - r)}} \qquad \text{for } 0 < r < R_s. \tag{7.5.53}$$

where

$$\alpha(0) = 2, \qquad \alpha(R_s) = 1, \qquad \alpha'(r) < 0.$$

We see from (7.5.53) that the redshift $z \to \infty$ as $r \to R_s$, and, consequently, we cannot see objects in the opposite hemisphere. Intuitively, $\alpha(r)$ represents the gravitational effect of the matter in the hemisphere of the observer.

5. *Physical conclusions.* In either case, globular or spherical, the universe is equivalent to globular universe(s). It is not originated from a Big-bang, is static, and confined in a black hole in the sense as addressed above. The observed mass M and the implicit mass M_{total} have the relation

$$M_{\text{total}} = 2 \times \frac{3\pi}{4} M = 3\pi M/2, \tag{7.5.54}$$

which is derived by (7.5.50) adding the mass of another hemisphere.

The implicit mass M_{total} of (7.5.54) contains the dark matter. In (Ma and Wang, 2014e), both the dark matter $M_{\text{total}} - M$ and the dark energy (i.e. the negative pressure (7.5.44)) are just a property of gravity.

7.5.4 New cosmology

We start with two difficulties encountered in modern cosmology.

First, if the Universe were born to a Big-Bang and expanded continuously, then in the expansion process it would generate successively a large number of black holes, whose radii vary as follows:

$$\sqrt{\frac{R_0}{R_s}} R_0 \leqslant r \leqslant \sqrt{\frac{R}{R_s}} R, \quad R_0 < R \leqslant R_s = \frac{2MG}{c^2}, \tag{7.5.55}$$

where M is the total mass in the universe, R_0 is the initial radius, R is the expanding radius, and r is the radius of sub-black holes.

To see this, we consider a homogeneous universe with radius $R < R_s$. Then the mass density ρ is given by

$$\rho = \frac{3M}{4\pi R^3}. \tag{7.5.56}$$

On the other hand, by Theorem 7.3, the condition for a ball B_r with radius r to form a black hole is that the mass M_r in B_r satisfies that

$$\frac{M_r}{r} = \frac{c^2}{2G}. \tag{7.5.57}$$

By (7.5.56), we have

$$M_r = \frac{4\pi}{3} r^3 \rho = \frac{r^3}{R^3} M.$$

Then it follows from (7.5.57) that

$$r = \sqrt{\frac{R}{R_s}} R. \tag{7.5.58}$$

Actually, in general for a ball B_r in a universe with radius $R < R_s$, if its mass M_r satisfies (7.5.57) then it will form a black hole, and its radius r satisfies that

$$r \leqslant \sqrt{\frac{R}{R_s}} R.$$

In particular, there must exist a black hole whose radius r is as in (7.5.58). Thus, we derive the conclusion (7.5.55).

Based on (7.5.55) we can deduce that if the Universe were born to a Big-Bang and continuously expands, then it would contain many black holes with smaller ones being embedded in the larger ones. In particular, the Universe would contain a huge black hole whose radius r is almost equal to the cosmic radius R_s. This is not what we observed in our Universe.

The second difficulty of modern cosmology concerns with the Hubble Law (7.5.1), which is restated as $v = HR$, where $c/H = R_s$. Consider a remote object with mass M_0, then the observed mass M_{obser} is given by $M_{\text{obser}} = \dfrac{M_0}{\sqrt{1 - \frac{v^2}{c^2}}}$. Consequently, the corresponding gravitational force F to the observer with mass m is

$$F = -\frac{mM_{\text{obser}}G}{r^2} = -\frac{mM_0 G}{r^2\sqrt{1 - \frac{v^2}{c^2}}} = -\frac{mM_0 G}{r^2\sqrt{1 - \frac{H^2}{c^2} r^2}} = -\frac{mM_0 G}{r^2\sqrt{1 - \frac{r^2}{R_s^2}}}.$$

It is clear then that as $r \to R_s$, $F \to -\infty$. This is clearly not what is observed.

In conclusion, we have rigorously derived the following new theory of cosmology:

Theorem 7.27 *Assume a) the Einstein theory of general relativity, and b) the principle of cosmological principle that the universe is homogeneous and isotropic. Then the following assertions hold true for our Universe:*

1) *All universes are bounded, are not originated from a Big-Bang, and are static; and*
2) *The topological structure of our Universe can only be the 3D sphere such that to each observer, the corresponding equator with the observer at the center of the hemisphere can be viewed as the black hole horizon.*

Theorem 7.28 *If we assume only a) the Einstein theory of general relativity, and b') the universe is homogeneous. Then all universes can only be either a 3D sphere as given in Theorem 7.27, or a globular universe, which is a 3D open ball B_{R_s} of radius R_s, forming the interior of a black hole with R_s as its Schwarzschild radius. In the later case, the Universe is also static, is not originated from a Big-Bang, and the matter fills the entire Universe. Also, the following assertions hold true:*

1) *The cosmic observable mass M and the total mass M_{total}, which includes both M and the non-observable mass due to the space curvature energy, satisfy the following relation*
$$M_{total} = \begin{cases} 3\pi M/2 & \text{for the spherical structure,} \\ 3\pi M/4 & \text{for the globular structure.} \end{cases} \tag{7.5.59}$$
The difference $M_{total} - M$ can be regarded as the dark matter;
2) *The static Universe has to possess a negative pressure to balance the gravitational attracting force. The negative pressure is actually the effect of the gravitational repelling force, also called dark energy; and*
3) *Both dark matter and dark energy are a property of gravity, which is reflected in both space-time curvature, and the attracting and repulsive gravitational forces in different scales of the Universe. This law of gravity is precisely described by the new gravitational field equations (4.4.10); see also (Ma and Wang, 2014e).*

We end this section with three remarks and observations.

First, astronomical observations have shown that the measurable mass M is about one fifth of total mass M_{total}. By (7.5.59), for the spherical universe,
$$M_{\text{total}} = 4.7M.$$
This relation also suggest that the spherical universe case fits better the current understanding for our Universe.

Second, due to the horizon of sphere, for an arbitrary point in a spherical universe, its opposite hemisphere relative to the point is as if it is a black hole. Hence the main contribution to the redshifts is from the effect of the black hole, as explicitly given by (7.5.53).

Third, in modern cosmology, the view of expanding universe was based essentially on the Friedmann model and the Hubble Law. The observations can accurately measure the distances and redshifts for some galaxies, which allowed astronomers to get both measured and theoretical data, and their deviation led to the conclusion that the expanding universe is accelerating. The misunderstanding comes from the perception that the Doppler redshift is the main source of redshifts.

7.6 Theory of Dark Matter and Dark Energy

7.6.1 Dark energy and dark matter phenomena

1. *Dark matter and Rubin rotational curve.* In astrophysics, dark matter is an unknown form of matter, which appears only participating in gravitational interaction, but does not emit nor absorb electromagnetic radiations.

Dark matter was first postulated in 1932 by Holland astronomer Jan Oort, who noted that the orbital velocities of stars in the Milky Way don't match their measured masses. Namely, the orbital velocity v and the gravity should satisfy the equilibrium relation

$$\frac{v^2}{r} = \frac{M_r G}{r^2}, \tag{7.6.1}$$

where M_r is the total mass in the ball B_r with radius r. But the observed mass M_0 was less than the theoretic mass M_r in (7.6.1), and the difference $M_r - M_0$ was explained as the presence of dark matter. The phenomenon was also discovered by Fritz Zwicky in 1933 for the missing mass in the orbital velocities of galaxies in clusters. Subsequently, other observations have manifested the existence of dark matter in the Universe, including the rotational velocities of galaxies, gravitational lensing, and the temperature distribution of hot gaseous.

A strong support to the existence of dark matter is the Rubin rotational curves for galactic rotational velocity. The rotational curve of a galaxy is the rotational velocity of visible stars or gases in the galaxy on their radial distance from the center of the galaxy. The Rubin rotational curve amounts to saying that most stars in spiral galaxies orbit at roughly the same speed. If a galaxy had a mass distribution as the observed distribution of visible astronomical objects, the rotational velocity would decrease at large distances. Hence, the Rubin curve demonstrates the existence of additional gravitational effect to the gravity by the visible matter in the galaxy.

More precisely, the orbital velocity $v(r)$ of the stars located at radius r from the center of galaxies is almost a constant:

$$v(r) \cong \text{a constant for a given galaxy}, \tag{7.6.2}$$

as illustrated typically by Figure 7.15 (a), where the vertical axis represents the velocity (Km/s), and the horizontal axis is the distance from the galaxy center (extending to the galaxy radius). However, the calculation from (7.6.1) gives a theoretic curve as shown in Figure f6.15(b), showing discrepancies between the mass determined from the gravitational effect and the mass calculated from the visible matter. The missing mass suggests the presence of dark matter in the Universe.

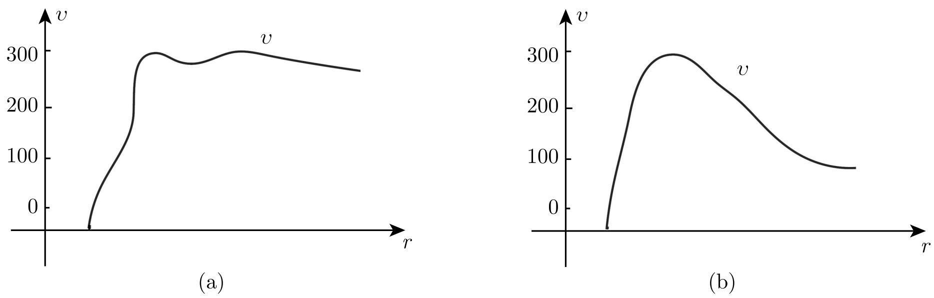


Figure 7.15 (a) Typical galactic rotational curve by Rubin, and (b) theoretic curve based on the Newtonian gravitational law.

In fact, we have seen in Section 7.5.4 that the dark matter is a space curved energy, or equivalently a gravitational effect, which is also reflected in the revised gravitational force formula in which there is an additional attracting force to the classical Newtonian gravity.

2. *Dark energy.* Dark energy was first proposed in 1990's, which was based on the hypotheses that the Universe is expanding.

The High-z Supernova Search Team in 1998 and the Supernova Cosmology Project in 1999 published their precisely measured data of the distances of supernovas and the redshifts. The observations indicated that the measured and theoretical data have a deviation, which was explained, based on the Hubble Law and the Friedmann model, as the acceleration of the expanding universe. The accelerating expansion is widely accepted as an evidence of the existence of dark energy.

However, based on the new cosmology postulated in the last section, the dark energy is a field energy form of gravitation which balances the gravitational attraction to maintain the homogeneity and stability of the Universe.

7.6.2 PID cosmological model and dark energy

We have shown in (Ma and Wang, 2014e) that both dark matter and dark energy are a property of gravity. Dark matter and dark energy are reflected in two aspects: a) the large scale space curved structure of the Universe caused by gravity, and b) the gravitational attracting and repelling aspects of gravity. In this section, we mainly explore the nature of dark energy in aspect a) using the PID-induced cosmological model.

PID cosmological model

According to Theorem 7.22, the metric of a homogeneous spherical universe is of the form

$$ds^2 = -c^2dt^2 + R^2\left[\frac{dr^2}{1-r^2} + r^2(d\theta^2 + \sin^2\theta d\varphi^2)\right], \tag{7.6.3}$$

where $R = R(t)$ is the cosmic radius. The PID induced gravitational field equations are given by

$$R_{\mu\nu} = -\frac{8\pi G}{c^4}\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right) - \left(\nabla_{\mu\nu}\phi - \frac{1}{2}g_{\mu\nu}\Phi\right), \tag{7.6.4}$$

where $\Phi = g^{\alpha\beta}D_{\alpha\beta}\phi$, and ϕ depends only on t.

The nonzero components of $R_{\mu\nu}$ read as

$$\begin{aligned} R_{00} &= \frac{3}{c^2}\frac{1}{R}R_{tt}, \\ R_{kk} &= -\frac{1}{c^2R^2}g_{kk}(RR''_{tt} + 2R_t^2 + 2c^2) \qquad \text{for } 1 \leqslant k \leqslant 3, \end{aligned}$$

and by $T_{\mu\nu} = \text{diag}(c^2\rho, g_{11}p, g_{22}p, g_{33}p)$, we have

$$\begin{aligned} T_{00} - \frac{1}{2}g_{00}T &= \frac{c^2}{2}\left(\rho + \frac{3p}{c^2}\right), \\ T_{kk} - \frac{1}{2}g_{kk}T &= \frac{c^2}{2}g_{kk}\left(\rho - \frac{p}{c^2}\right) \qquad \text{for} 1 \leqslant k \leqslant 3, \\ \phi_{00} - \frac{1}{2}g_{00}\Phi &= \frac{1}{2c^2}\left(\phi_{tt} - \frac{3R_t}{R}\phi_t\right), \\ \phi_{kk} - \frac{1}{2}g_{kk}\Phi &= \frac{1}{2c^2}g_{kk}\left(\phi_{tt} + \frac{R_t}{R}\phi_t\right) \qquad \text{for} 1 \leqslant k \leqslant 3. \end{aligned}$$

Thus, we derive from (7.6.4) two independent field equations as

$$R'' = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right)R - \frac{1}{6}\phi''R + \frac{1}{2}R'\phi', \tag{7.6.5}$$

$$\frac{R''}{R} + 2\left(\frac{R'}{R}\right)^2 + \frac{2c^2}{R^2} = 4\pi G\left(\rho - \frac{p}{c^2}\right) + \frac{1}{2}\phi'' + \frac{1}{2}\frac{R'}{R}\phi'. \tag{7.6.6}$$

We infer from (7.6.5) and (7.6.6) that

$$(R')^2 = \frac{8\pi G}{3}R^2\rho + \frac{1}{3}R^2\phi'' - c^2. \tag{7.6.7}$$

By the Bianchi identity:

$$\nabla^\mu(\nabla_{\mu\nu}\phi + \frac{8\pi G}{c^4}T_{\mu\nu}) = 0,$$

we deduce that

$$\phi''' + \frac{3R'}{R}\phi'' = -8\pi G\left(\rho' + \frac{3R'}{R}\rho + \frac{3R'}{R}\frac{p}{c^2}\right). \tag{7.6.8}$$

It is known that the energy density ρ and the cosmic radius R (also called the scale factor) satisfy the relation:

$$\rho = \frac{\rho_0}{R^3}, \qquad \rho_0 \text{ the density at } R = 1. \tag{7.6.9}$$

Hence, it follows from (7.6.9) that

$$\rho' = -3\rho R'/R.$$

Thus, (7.6.8) is rewritten as

$$\phi''' + \frac{3R'}{R}\phi'' = -\frac{24\pi G}{c^2}\frac{R'}{R}p. \tag{7.6.10}$$

In addition, making the transformation

$$\phi'' = \frac{\psi}{R^3}, \tag{7.6.11}$$

then, from (7.6.5), (7.6.7) and (7.6.9)-(7.6.11) we can deduce that

$$(R')^2\phi' = 0. \tag{7.6.12}$$

Denote $\varphi = \phi''$, by (7.6.12), the equations (7.6.5), (7.6.7) and (7.6.10) can be rewritten in the form

$$\begin{aligned} R'' &= -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2} + \frac{\varphi}{8\pi G}\right)R, \\ (R')^2 &= \frac{1}{3}(8\pi G\rho + \varphi)R^2 - c^2, \\ \varphi' + \frac{3R'}{R}\varphi &= -\frac{24\pi G}{c^2}\frac{R'}{R}p. \end{aligned} \tag{7.6.13}$$

Only two equations in (7.6.13) are independent. However, there are three unknown functions R, φ, p in (7.6.13). Hence, we need to add an additional equation, the equation of state, as follows:

$$p = f(\rho, \varphi). \tag{7.6.14}$$

Based on Theorem 7.27, the model describing the static Universe is the equation (7.6.14) together with the stationary equations of (7.6.13), which are equivalent to the form

$$\begin{aligned} \varphi &= -8\pi G\left(\rho + \frac{3p}{c^2}\right), \\ p &= -\frac{c^4}{8\pi G R^2}. \end{aligned} \tag{7.6.15}$$

The equations (7.6.14) and (7.6.15) provide a theoretic basis for the static Universe, including the dark energy.

Now, we need to determine the explicit expression for the equation (7.6.14) of state. It is natural to postulate that the equation of state is linear. Hence, (7.6.14) can be written as

$$p = \frac{c^2}{G}(\alpha_1 \varphi - \alpha_2 G\rho), \tag{7.6.16}$$

where α_1 and α_2 are nondimensional parameter, which will be determined by the observed data.

The equations (7.6.15) and (7.6.16) are the PID cosmological model, where the cosmological significants of R, p, φ, ρ are as follows:

$$\begin{aligned}
&R \quad \text{the cosmic radius (of the 3D spherical universe)},\\
&p \quad \text{the negative pressure, generated by the repulsive aspect of gravity},\\
&\varphi \quad \text{represents the dual gravitational potential},\\
&\rho \quad \text{the cosmic density, given by } \frac{3M}{4\pi R^3} = \frac{M_{\text{total}}}{\pi^2 R^3},
\end{aligned} \tag{7.6.17}$$

where M and M_{total} are as in Remark 7.26.

Here, we remark that in the classical Einstein field equations where $\phi = 0$, the relation (7.6.9) still holds true, by which we can deduce that $R'p = 0$.

Theory of dark energy

In the static cosmology, dark energy is defined in the following manner. Let E_{ob} be the observed energy, and R be the cosmic radius. We define the observable mass and the total mass as follows:

$$M_{\text{ob}} = \frac{E_{\text{ob}}}{c^2}, \tag{7.6.18}$$

$$M_{\text{T}} = \frac{Rc^2}{2G}. \tag{7.6.19}$$

If $M_{\text{T}} > M_{\text{ob}}$, then the difference

$$\Delta E = E_{\text{T}} - E_{\text{ob}} \tag{7.6.20}$$

is called the dark energy.

The CMB measurement and the WMAP analysis indicate that the difference ΔE in (7.6.20) is positive,

$$\Delta E > 0,$$

which is considered as another evidence for the presence of dark energy.

From the PID cosmological model (7.6.15)-(7.6.17), we see that the dark energy ΔE in (7.6.20) is essentially due to the dual gravitational potential φ. In fact, we infer from (7.6.15) that

$$\begin{aligned} &\varphi = 0 \quad \Leftrightarrow \quad R = 2M_{\text{ob}}G/c^2 \quad (\text{i.e. } \Delta E = 0), \\ &\varphi > 0 \quad \Leftrightarrow \quad \Delta E > 0. \end{aligned} \tag{7.6.21}$$

Hence, dark energy is generated by the dual gravitational field. This fact is reflected in the PID gravitational force formula derived in subsections hereafter.

If we can measure precisely, with astronomical observations, the energy (7.6.18) and the cosmic radius R (i.e. M_{T} of (7.6.19)), then we can obtain a relation between the parameters α_1 and α_2 in (7.6.16). In fact, we deduce from (7.6.15) and (7.6.16) that

$$\rho + \frac{\beta p}{c^2} = 0, \qquad \beta = \frac{1 + 24\pi\alpha_1}{\alpha_2 + 8\pi\alpha_1}. \tag{7.6.22}$$

As we get

$$\frac{\Delta M}{M_{\text{ob}}} = \frac{M_{\text{T}} - M_{\text{ob}}}{M_{\text{ob}}} = k \quad (k > 0). \tag{7.6.23}$$

Then by (7.6.22) and

$$\rho = \frac{3M_{\text{ob}}}{4\pi R^3}, \quad p = -\frac{c^4}{8\pi G R^2},$$

we obtain from (7.6.22) that

$$3\alpha_2 = 24k\pi\alpha_1 + k + 1. \tag{7.6.24}$$

By the relation (7.6.24) from (7.6.22), we can also derive, in the same fashion as above, the dark energy formula (7.6.23).

7.6.3 PID gravitational interaction formula

Consider a central gravitational field generated by a ball B_{r_0} with radius r_0 and mass M. It is known that the metric of the central field at $r > r_0$ can be written in the form

$$ds^2 = -e^u c^2 dt^2 + e^v dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \tag{7.6.25}$$

and $u = u(r), v = v(r)$.

In the exterior of B_{r_0}, the energy-momentum is zero, i.e.

$$T_{\mu\nu} = 0, \quad \text{for } r > r_0.$$

Hence, the PID gravitational field equation for the metric (7.6.25) is given by

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\nabla_{\mu\nu}\phi, \quad r > r_0. \tag{7.6.26}$$

where $\phi = \phi(r)$ is a scalar function of r.

By (7.1.25) and (7.1.26), we have

$$
\begin{aligned}
&R_{00}-\frac{1}{2}g_{00}R=-\frac{1}{r}e^{u-v}\left[v'+\frac{1}{r}(e^v-1)\right],\\
&R_{11}-\frac{1}{2}g_{11}R=-\frac{1}{r}\left[u'-\frac{1}{r}(e^v-1)\right],\\
&R_{22}-\frac{1}{2}g_{22}R=-\frac{r^2}{2}e^{-v}\left[u''+\left(\frac{1}{2}u'+\frac{1}{r}\right)(u'-v')\right],\\
&\nabla_{00}\phi=-\frac{1}{2}e^{u-v}u'\phi',\\
&\nabla_{11}\phi=\phi''-\frac{1}{2}v'\phi',\\
&\nabla_{22}\phi=-re^{-v}\phi'.
\end{aligned}
$$

Thus, the fields equations (7.6.26) are as follows

$$
\begin{aligned}
&v'+\frac{1}{r}(e^v-1)=-\frac{r}{2}u'\phi',\\
&u'-\frac{1}{r}(e^v-1)=r(\phi''-\frac{1}{2}v'\phi'),\\
&u''+\left(\frac{1}{2}u'+\frac{1}{r}\right)(u'-v')=-\frac{2}{r}\phi'.
\end{aligned}
\tag{7.6.27}
$$

Now we are ready to deduce from (7.6.27) the PID gravitational interaction formula as follows.

First, we infer from (7.6.27) that

$$
\begin{aligned}
&u'+v'=\frac{r\phi''}{1+\frac{r}{2}\phi'},\\
&u'-v'=\frac{1}{1-\frac{r}{2}\phi'}\left[\frac{2}{r}(e^v-1)+r\phi''\right].
\end{aligned}
$$

Consequently,

$$
u'=\frac{1}{1-\frac{r}{2}\phi'}\frac{1}{r}(e^v-1)+\frac{r\phi''}{1-(\frac{r}{2}\phi')^2}. \tag{7.6.28}
$$

It is known that the interaction force F is given by

$$
F=-m\nabla\psi,\quad \psi=\frac{c^2}{2}(e^u-1).
$$

Then, it follows from (7.6.28) that

$$
F=\frac{mc^2}{2}e^u\left[-\frac{1}{1-\frac{r}{2}\phi'}\frac{1}{r}(e^v-1)-\frac{r\phi''}{1-\left(\frac{r}{2}\phi'\right)^2}\right]. \tag{7.6.29}
$$

The formula (7.6.29) provides the precise gravitational interaction force exerted on an object with mass m in a spherically symmetric gravitation field.

In classical physics, the field functions u and v in (7.6.29) are taken by the Schwarzschild solution:

$$e^u = 1 - \frac{2GM}{c^2 r}, \quad e^v = \left(1 - \frac{2GM}{c^2 r}\right)^{-1}, \tag{7.6.30}$$

and $\phi' = \phi'' = 0$, which leads to the Newton gravitation.

However, due to the presence of dark matter and dark energy, the field functions u, v, ϕ in (7.6.29) should be an approximation of the Schwarzschild solution (7.6.30). Hence we have

$$|r\phi'| \ll 1 \quad \text{for } r > r_0. \tag{7.6.31}$$

Under the condition (7.6.31), formula (7.6.29) can be approximatively expressed as

$$F = \frac{mc^2}{2} e^u \left[-\frac{1}{r}(e^v - 1) - r\phi'' \right]. \tag{7.6.32}$$

7.6.4 Asymptotic repulsion of gravity

In this section, we shall consider the asymptotic properties of gravity, and rigorously prove that the interaction force given by (7.6.32) is repulsive at very large distance.

To this end, we need to make the following transformation to convert the field equations (7.6.27) into a first order autonomous system:

$$\begin{aligned} r &= e^s, \\ w &= e^v - 1, \\ x_1(s) &= e^s u'(e^s), \\ x_2(s) &= w(e^s), \\ x_3(s) &= e^s \phi'(e^s). \end{aligned} \tag{7.6.33}$$

Then the equations (7.6.27) can be rewritten as

$$\begin{aligned} x_1' &= -x_2 + 2x_3 - \frac{1}{2}x_1^2 - \frac{1}{2}x_1x_3 - \frac{1}{2}x_1x_2 - \frac{1}{4}x_1^2x_3, \\ x_2' &= -x_2 - \frac{1}{2}x_1x_3 - x_2^2 - \frac{1}{2}x_1x_2x_3, \\ x_3' &= x_1 - x_2 + x_3 - \frac{1}{2}x_2x_3 - \frac{1}{4}x_1x_3^2. \end{aligned} \tag{7.6.34}$$

The system (7.6.34) is supplemented with an initial condition

$$(x_1, x_2, x_3)(s_0) = (\alpha_1, \alpha_2, \alpha_3) \qquad \text{with } r_0 = e^{s_0}. \tag{7.6.35}$$

We now study the problem (7.6.34)-(7.6.35) in a few steps as follows.

Step 1. Asymptotic flatness. For a globular matter distribution, its gravitational field should be asymptotically flat, i.e.

$$g_{00} \to -1, \quad g_{11} \to 1 \quad \text{if} \quad r \to \infty.$$

It implies that $x = 0$ is the uniquely physical equilibrium point of (7.6.34) and the following holds true:

$$x(s) \to 0 \quad \text{if} \quad s \to \infty \quad (\text{i.e. } r \to \infty). \tag{7.6.36}$$

Step 2. Physical initial values. The physically meaningful initial values $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ in (7.6.35) have to satisfy the following two conditions:

(a) The solutions $x(s,\alpha)$ of (7.6.34)-(7.6.35) are asymptotically flat in the sense of (7.6.36). Namely, the initial values α are in the stable manifold E^s of $x = 0$, defined by

$$E^s = \{\alpha \in \mathbb{R}^3 |\, x(s,\alpha) \to 0 \text{ for } s \to \infty\}; \tag{7.6.37}$$

(b) The solutions $x(s,\alpha)$ are near the Schwarzschild solution:

$$|x_1 - e^s u'|, \quad |x_2 + 1 - e^v|, \quad |x_3| \ll 1, \tag{7.6.38}$$

where u, v are as in (7.6.30).

In fact, by (7.6.31) and (7.6.33) we can see that all Schwarzschild solutions lie on the line

$$L = \{(x_1, x_2, 0) |\, x_1 = x_2,\ x_1, x_2 > 0\}. \tag{7.6.39}$$

In particular, the line L is on the stable manifold E^s of (7.6.37).

Step 3. Stable manifold E^s. The equations (7.6.34) can be written as

$$\dot{x} = Ax + O(|x|^2),$$

where

$$A = \begin{pmatrix} 0 & -1 & 2 \\ 0 & -1 & 0 \\ 1 & -1 & 1 \end{pmatrix}. \tag{7.6.40}$$

The dimension of the stable manifold E^s is the number of negative eigenvalues of the matrix A. It is easy to see that the eigenvalues of A are given by

$$\lambda_1 = -1, \quad \lambda_2 = -1, \quad \lambda_3 = 2.$$

Hence, the dimension of E^s is two:

$$\dim E^s = 2.$$

Consequently, the initial value α of an asymptotically flat solution has only two independent components due to $\alpha \in E^s$, which is of two dimensional. Namely, we arrive at the following conclusion.

Physical Conclusion 7.29 *In the gravitation formula (7.6.29) there are two free parameters to be determined by experiments (or by astronomical measurements).*

In fact, the two free parameters will be determined by the Rubin rotational curve and the repulsive property of gravity at large distance.

Step 4. Local expression of E^s. In order to derive the asymptotic property of the gravitational force F of (7.6.29), we need to derive the local expression of the stable manifold E^s near $x = 0$. Since the tangent space of E^s at $x = 0$ is spanned by the two eigenvectors $(1,1,0)^t$ and $(1,-1,-1)^t$ corresponding to the two negative eigenvalues $\lambda_1 = \lambda_2 = -1$, the coordinate vector $(0,0,1)$ of x_3 is not contained in E^s. This implies that the stable manifold can be expressed near $x = 0$ in the form

$$x_3 = h(x_1, x_2). \tag{7.6.41}$$

Inserting the Taylor expansion for (7.6.41) into (7.6.34), and comparing the coefficients, we derive the following local expression of (7.6.41) of the stable manifold function:

$$h(x_1,x_2) = -\frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{16}x_1^2 - \frac{1}{16}x_2^2 + O(|x|^3). \tag{7.6.42}$$

Inserting (7.6.41)-(7.6.42) into the first two equations of (7.6.34), we deduce that

$$\begin{aligned} x_1' &= -x_1 - \frac{1}{8}x_1^2 - \frac{1}{8}x_2^2 - \frac{3}{4}x_1x_2 + O(|x|^3), \\ x_2' &= -x_2 + \frac{1}{4}x_1^2 - x_2^2 - \frac{1}{4}x_1x_2 + O(|x|^3). \end{aligned} \tag{7.6.43}$$

The system (7.6.43) is the system (7.6.34) restricted on the stable manifold E^s. Hence, its asymptotic behavior represents that of the interaction force F in (7.6.29).

Step 5. Phase diagram of system (7.6.43). In order to obtain the phase diagram of (7.6.43) near $x = 0$, we consider the ratio: $x_2'/x_1' = dx_2/dx_1$. Omitting the terms $O(|x|^3)$, we infer from (7.6.43) that

$$\frac{dx_2}{dx_1} = \frac{x_2 + x_2^2 + \frac{1}{4}x_2x_1 - \frac{1}{4}x_1^2}{x_1 + \frac{1}{8}x_2^2 + \frac{3}{4}x_2x_1 + \frac{1}{8}x_1^2}. \tag{7.6.44}$$

Let k be the slope of an orbit reaching to $x = 0$:

$$k = \frac{x_2}{x_1} \quad \text{as} \quad (x_2, x_1) \to 0.$$

Then (7.6.44) can be expressed as

$$k = \frac{k + k^2 x_1 + \frac{1}{4} k x_1 - \frac{1}{4} x_1}{1 + \frac{1}{8} k^2 x_1 + \frac{3}{4} k x_1 + \frac{1}{8} x_1},$$

which leads to the equation

$$k^3 - 2k^2 - k + 2 = 0.$$

This equation has three solutions:

$$k = \pm 1, \quad k = 2,$$

giving rise to three line orbits:

$$x_2 = x_1, \quad x_2 = 2x_1, \quad x_2 = -x_1,$$

which divide the neighborhood of $x = 0$ into six invariant regions. It is clear that all physically meaningful orbits can only be in the following three regions:

$$\begin{aligned} \Omega_1 &= \left\{ -x_2 < x_1 < \frac{1}{2} x_2,\ x_2 > 0 \right\}, \\ \Omega_2 &= \left\{ \frac{1}{2} x_2 \leqslant x_1 \leqslant x_2,\ x_2 > 0 \right\}, \\ \Omega_3 &= \{ x_2 < x_1,\ x_2 > 0 \}. \end{aligned} \tag{7.6.45}$$

On the positive x_2-axis (i.e. $x_1 = 0, x_2 > 0$), which lies in Ω_1, the equations (7.6.43) take the form

$$\begin{aligned} x_1' &= -\frac{1}{8} x_2^2 + O(|x|^3), \\ x_2' &= -x_2 - x_2^2 + O(|x|^3). \end{aligned}$$

It is easy to show that the orbits in Ω_1 with $x_1 > 0$ will eventually cross the x_2-axis. Thus, using the three invariant sets in (7.6.45), we obtain the phase diagram of (7.6.43) on $x_2 > 0$ as shown in Figure 7.16. In this diagram, we see that, the orbits in Ω_2 and Ω_3 will not cross the x_2-axis, but these in Ω_1 with $x_1 > 0$ will do.

Step 6. Asymptotic repulsion theorem of gravity. We now derive an asymptotic repulsion theorem of gravity, based on the phase diagram in Figure 7.16. In fact, by (7.6.28) and (7.6.29), the gravitational force F reads as

$$F = -\frac{mc^2}{2} e^u u'. \tag{7.6.46}$$

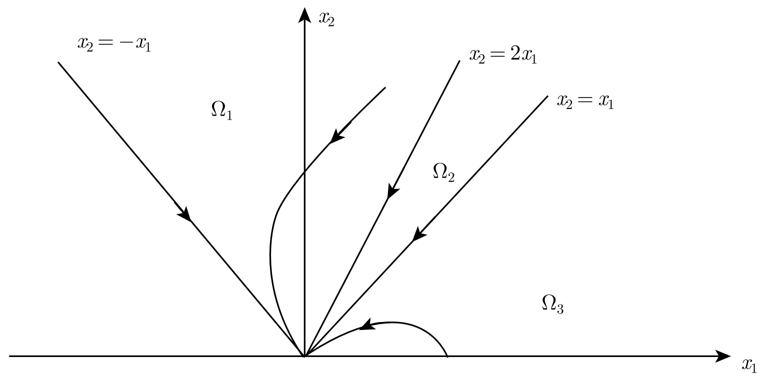


Figure 7.16 Only the orbits on Ω_1 with $x_1>0$ will eventually cross the x_2-axis, leading to the sign change of x_1, and to a repelling gravitational force corresponding to dark energy.

It is known that

$$F<0 \qquad \text{represents attraction,}$$
$$F>0 \qquad \text{represents repelling.}$$

Hence, by $x_1=ru'(r)$ and (7.6.46), the phase diagram shows that an orbit in Ω_1, starting with $x_1>0$, will cross the x_2-axis, and the sign of x_1 changes from positive to negative, leading consequently to a repulsive gravitational force F. Namely, we have obtained the following theorem.

Theorem 7.30 (Asymptotic Repulsion of Gravitation) *For a central gravitational field, the following assertions hold true:*

1) *The gravitational force F is given by (7.6.29), and is asymptotic zero*:
$$F\to 0 \quad \text{if} \quad r\to\infty. \tag{7.6.47}$$

2) *If the initial value α in (7.6.35) is near the Schwarzschild solution (7.6.31) with $0<\alpha_1<\alpha_2/2$, then there exists a sufficiently large r_1 such that the gravitational force F is repulsive for $r>r_1$*:
$$F>0 \quad \text{for} \quad r>r_1. \tag{7.6.48}$$

We remark that Theorem 7.30 is valid provided the initial value α is small because the diagram given by Figure 7.16 is in a neighborhood of $x=0$. However, all physically meaningful central fields satisfy the condition (note that any a black hole is enclosed by a huge quantity of matter with radius $r>0\gg 2MG/c^2$). In fact, the Schwarzschild initial values are as
$$x_1(r_0)=x_2(r_0)=\frac{\delta}{1-\delta}, \quad \delta=\frac{2MG}{c^2r_0}. \tag{7.6.49}$$

For example see (7.6.29), where the δ-factors are of the order $\delta \leqslant 10^{-1}$, sufficient for the requirements of Theorem 7.30.

The most important cases are for galaxies and clusters of galaxies. For these two types of astronomical objects, we have

$$\begin{aligned} &\text{galaxy}: && M = 10^{11} M_{\odot}, && r_0 = 3\times 10^5 \text{ly}, \\ &\text{cluster of galaxies}: && M = 10^{14} M_{\odot}, && r_0 = 3\times 10^6 \text{ly}. \end{aligned}$$

Thus the δ-factors are

$$\text{galaxies } \delta = 10^{-7}, \quad \text{cluster of galaxies } \delta = 10^{-5}. \tag{7.6.50}$$

In fact, the dark energy phenomenon is mainly evident between galaxies and between clusters of galaxies. Hence, (7.6.50) shows that Theorem 7.30 is valid for both central gravitational fields of galaxies and clusters of galaxies. The asymptotic repulsion of gravity plays the role to stabilize the large scale homogeneous structure of the Universe.

7.6.5 Simplified gravitational formula

We have shown that all four fundamental interactions are layered. Namely, each interaction has distinct attracting and repelling behaviors in different scales and levels. The dark matter and dark energy represent the layered property of gravity.

In this section, we simplify the gravitational formula (7.6.32) to clearly exhibit the layered phenomena of gravity.

In (7.6.32) the field functions u and v can be approximatively replaced by the Schwarzschild solution (7.6.30). Since $2MG/c^2r$ is very small for $r > r_0$ as indicted in (7.6.29) and (7.6.50), the formula (7.6.32) can be expressed as

$$F = mMG\left[-\frac{1}{r^2} - \frac{r}{\delta r_0}\phi''\right], \quad r > r_0. \tag{7.6.51}$$

By the field equation (7.6.26), we have

$$R = \Phi \quad \text{for } r > r_0, \tag{7.6.52}$$

where R is the scalar curvature, and

$$\Phi = g^{\mu\nu}D_{\mu\nu}\phi = e^{-v}\left[-\phi'' + \frac{1}{2}(u'-v')\phi' + \frac{2}{r}\varphi'\right].$$

In view of (7.6.52), we obtain that

$$\phi'' = -e^v R + \frac{2}{r}\phi' + \frac{1}{2}(u'-v')\phi'$$

Again by the Schwarzschild approximation, we have

$$\phi'' = \left(\frac{2}{r} + \frac{\delta r_0}{r^2}\right)\phi' - R. \tag{7.6.53}$$

Integrating (7.6.53) and omitting $e^{\pm\delta r_0/r}$, we derive that

$$\phi' = -r^2\left[\varepsilon + \int r^{-2}Rdr\right],$$

where ε is a constant. Thus (7.6.51) can be rewritten as

$$F = mMG\left[-\frac{1}{r^2} + \frac{r}{\delta r_0}R + \left(1 + \frac{2r}{\delta r_0}\right)\left(\varepsilon r + r\int \frac{R}{r^2}dr\right)\right]. \tag{7.6.54}$$

The solutions of (7.6.34) can be Taylor expanded. Hence by (7.6.33) we see that

$$u'(r) = \frac{1}{r^2}\sum_{k=0}^{\infty} a_k(r - r_0)^k.$$

By (7.6.46), the gravitational force F takes the following form

$$F = \frac{1}{r^2}\sum_{k=0}^{\infty} b_k r^k, \quad b_0 = -mMG.$$

In view of (7.6.54), it implies that R can be expanded as

$$R = \frac{\varepsilon_0}{r} - \varepsilon_1 + O(r),$$

and by Physical Conclusion 7.29, ε_0 and ε_1 are two to-be-determined free parameters. Inserting R into (7.6.54) we obtain that

$$F = mMG\left[-\frac{1}{r^2} - \frac{k_0}{r} + \varphi(r)\right] \qquad \text{for } r > r_0. \tag{7.6.55}$$

where $k_0 = \frac{1}{2}\varepsilon_0$, and

$$\varphi(r) = \varepsilon_1 + k_1 r + O(r), \quad k_1 = \varepsilon + \frac{\varepsilon_1}{\delta r_0}.$$

The nature of dark matter and dark energy suggests that

$$k_0 > 0, \qquad k_1 > 0.$$

Based on Theorem 7.30, $\varphi(r) \to 0$ as $r \to \infty$, and (7.6.55) can be further simplified as in the form for $r_0 < r < r_1$,

$$F = mMG\left[-\frac{1}{r^2} - \frac{k_0}{r} + k_1 r\right], \tag{7.6.56}$$

where k_0 and k_1 will be determined by the Rubin rotational curve and the astronomical data for clusters of galaxies in the next section, where we obtain that

$$k_0 = 4 \times 10^{-18}\text{km}^{-2}, \quad k_1 = 10^{-57}\text{km}^{-3}. \tag{7.6.57}$$

The formula (7.6.56) is valid only in the interval

$$r_0 < r < r_1,$$

and r_1 is the distance at which F changes its sign:

$$F(r_1) = 0.$$

Both observational evidence on dark energy and Theorem 7.30 show that the distance r_1 exists. The formula (7.6.56) with (7.6.57) clearly displays the layered property of gravity: attracting at short distance and repelling at large distance.

7.6.6 Nature of dark matter and dark energy

As mentioned in Section 7.6.2, both dark matter and dark energy are a property of gravitational effect, reflected in two aspects, which will be addressed in detail in this section:

a) spatially geometrical structure, and

b) gravitational attracting and repelling as in (7.6.56).

Space curved energy and negative pressure

Gravitational potential causes space curvature and the spherical structure of the Universe, and displays two types of energies: a) dark matter contributed by the curvature of space, and b) dark energy generated by the dual gravitational potential in (7.6.17). We address each type of energy as follows.

1. *Dark matter: the space curved energy.* In Section 7.5.2, we have introduced the space curved energy M_{total} for the 3D spherical Universe as follows:

$$M_{\text{total}} = \frac{3\pi}{2}M, \quad M \text{ is the observed mass in the hemisphere.}$$

Now, we consider a galaxy with an observed mass M_Ω. By (7.5.51), we have shown that the space curved energy $M_{\text{total};\Omega}$ is

$$M_{\text{total};\Omega} = \frac{V_\Omega}{|\Omega|}M_\Omega, \tag{7.6.58}$$

where Ω is the domain occupied by the galaxy, V_Ω and $|\Omega|$ are the volumes of curved and flat Ω. V_Ω contain two parts:

$$V_\Omega = \text{cosmic spherical } V_\Omega^1 + \text{local bump } V_\Omega^2. \tag{7.6.59}$$

It is known that

$$V_\Omega^1 = \frac{3\pi}{4}|\Omega|.$$

For V_Ω^2, we propose that

$$V_\Omega^2 = \pi^2 r_0^3, \qquad r_0 \text{ the galaxy radius.}$$

In fact, the formula is precise for the galaxy nucleus.

By $|\Omega| = \frac{4}{3}\pi r_0^3$, we infer from (7.6.58) and (7.6.59) that

$$M_{\text{total};\Omega} = \frac{3\pi}{2} M_\Omega, \tag{7.6.60}$$

which gives rise to the relation between the masses of dark matter and observable matter.

2. *Dark energy: the dual gravitational potential.* The static universe is described by the stationary solution of (7.6.13)-(7.6.14), which is given by (7.6.15)-(7.6.16). In the solution a negative pressure presents, which prevents galaxies and clusters of galaxies from gravitational contraction to form a void universe, and maintains the homogeneous distribution of the Universe. The negative pressure contains two parts:

$$p = -\frac{1}{3}\rho c^2 - \frac{c^2}{24\pi G}\varphi \quad (\text{see } (7.6.15)), \tag{7.6.61}$$

where the first term is contributed by the observable energy, and the second term is the dark energy generated by the dual gravitational potential φ; see also (7.6.21).

By the Blackhole Theorem, Theorem 7.15, black holes are incompressible in their interiors. Hence, in (7.6.61) the negative pressure

$$p = -\frac{1}{3}\rho c^2, \tag{7.6.62}$$

is essentially the incompressible pressure of the black hole generated by the normal energy.

By the cosmology theorem, Theorem 7.27, the Universe is a $3D$ sphere with a blackhole radius. However, the *CMB* and the *WMAP* measurements manifest that the cosmic radius R is greater than the blackhole radius of normal energy. By (7.6.21), the deficient energy is compensated by the dual gravitational potential, i.e. by the second term of (7.6.61).

Attraction and repulsion of gravity

Based on Theorem 7.30, gravity possesses additional attraction and repelling to the Newtonian gravity, as shown in the revised gravitational formula:

$$F = mMG\left(-\frac{1}{r^2} - \frac{k_0}{r} + k_1 r\right). \tag{7.6.63}$$

By using this formula we can explain the dark matter and dark energy phenomena. In particular, based on the Rubin rotational curve and astronomical data, we can determine an approximation of the parameters k_0 and k_1 in (7.6.63).

1. *Dark matter: the additional attraction.* Let M_r be the total mass in the ball with radius r of galaxy, and V be the constant galactic rotational velocity. By the force equilibrium, we infer from (7.6.63) that

$$\frac{V^2}{r} = M_r G\left(\frac{1}{r^2} + \frac{k_0}{r} - k_1 r\right), \tag{7.6.64}$$

which implies that

$$M_r = \frac{V^2}{G}\frac{r}{1 + k_0 r - k_1 r^3}. \tag{7.6.65}$$

The mass distribution (7.6.65) is derived based on both the Rubin rotational curve and the revised formula (7.6.63). In the following we show that the mass distribution M_r fits the observed data.

We know that the theoretic rotational curve given by Figure 7.15(b) is derived by using the observed mass M_{ob} and the Newton formula

$$F = -\frac{mM_{ob}G}{r^2}.$$

Hence, to show that $M_r = M_{ob}$, we only need to calculate the rotational curve v_r by the Newton formula from the mass M_r, and to verify that v_r is consistent with the theoretic curve. To this end, we have

$$\frac{v_r^2}{r} = \frac{M_r G}{r^2},$$

which, by (7.6.65), leads to

$$v_r = \frac{V}{\sqrt{1 - k_0 r - k_1 r^2}}.$$

As $k_1 \ll k_0 \ll 1, v_r$ can be approximatively written as

$$v_r = V\left(1 - \frac{1}{2}k_0 r + \frac{1}{4}k_0^2 r^2\right). \tag{7.6.66}$$

It is clear that the rotational curve described by (7.6.66) is consistent with the theoretic rotational curve as illustrated by Figure 7.15(b). Therefore, it implies that

$$M_r = M_{ob}. \tag{7.6.67}$$

The facts of (7.6.64) and (7.6.67) are strong evidence to show that the revised formula (7.6.63) is in agreement with the astronomical observations.

We now determine the constant k_0 in (7.6.63). According to astronomical data, the average mass M_{r_1} and radius r_1 of galaxies are about

$$\begin{aligned} &M_{r_1} = 10^{11} M_\odot \cong 2\times 10^{41}\text{Kg},\\ &r_1 = 10^4 \sim 10^5\text{pc} \cong 10^{18}\text{Km}. \end{aligned} \tag{7.6.68}$$

The observations show that the constant velocity V in the Rubin rotational curve is about $V = 300\text{km/s}$. Then we have

$$\frac{V^2}{G} = 10^{24}\text{kg/km}$$

Based on physical considerations,

$$k_0 \gg k_1 r_1 \quad (r_1 \text{ as in } (7.6.68)). \tag{7.6.69}$$

Then, we deduce from (7.6.65) that

$$k_0 = \frac{V^2}{G}\frac{1}{M_{r_1}} - \frac{1}{r_1} = 4\times 10^{-18} K_m^{-1}. \tag{7.6.70}$$

We can explain the dark matter by the revised formula (7.6.63). As the matter distribution M_r is in the form

$$M_r = \frac{V^2}{G}\frac{r}{1+k_0 r},$$

then the Rubin law holds true. In addition, the revised gravitation produces an excessive mass $\widetilde{M}$ as

$$\widetilde{M} = M_{\text{T}} - M_{r_1} = \frac{V^2}{G} r_1 - \frac{V^2}{G}\frac{r_1}{1+k_0 r_1},$$

where $M_1 = V^2 r_1/G$ is the Newton theoretic value of the total mass. Hence we have

$$\frac{\widetilde{M}}{M_{\text{T}}} = \frac{k_0 r_1}{1+k_0 r_1} = \frac{4}{5} \quad \text{or} \quad \frac{\widetilde{M}}{M_{r_1}} = 5.$$

Namely, the additional mass $\widetilde{M}$ is four time the visible matter $M_{r_1} = M_T - \widetilde{M}$. Thus, it gives an explanation for the dark matter.

We remark that the ratio $\widetilde{M}/M_{\text{ob}} = 5$ is essentially the same as in (7.6.60). It shows that the dark matter is a gravitational effect, reflected in both the space curvature and the additional gravitational attraction.

2. *Dark energy: asymptotic repulsion of gravity.* If gravity is always attracting as given by the Newton formula, then the cosmic homogeneity is unstable. In fact, It is known that the average mass M and distance for the clusters of galaxies are as

$$\begin{aligned} &M = 10^{14} M_\odot \cong 10^{44}\text{Kg}\\ &r \cong 10^8\text{pc} \cong 10^{20} \sim 10^{21}\text{Km}. \end{aligned} \tag{7.6.71}$$

Then the Newton gravitation between two clusters of galaxies is

$$F = -\frac{M^2 G}{r^2} \cong 10^{29} N = 10^{28} \text{Kg}. \tag{7.6.72}$$

However, astronomical observations indicate that no gravitational interaction between clusters of galaxies. The Universe is isotropic, therefore no rotation to balance the huge force of (7.6.72) in the clusters.

Thus, the new cosmology theorem, Theorem 7.27, suggests that gravity should be asymptotically repulsive. Theorem 7.30 offers a solid theoretic foundation for the property, based on which we derive the simplified gravitational force formula (7.6.63).

Now we consider the constant k_1 in (7.6.63). Due to the astronomical fact that no gravitational force between clusters of galaxies, we have

$$F(\overline{r}) = 0, \quad \overline{r} = \text{the average distance between galactic clusters.}$$

By (7.6.71), we take

$$\overline{r} = \sqrt{\frac{2}{5}} \times 10^{20} \text{km}. \tag{7.6.73}$$

Then we deduce from (7.6.63) that

$$k_1 \overline{r} - \frac{k_0}{\overline{r}} - \frac{1}{\overline{r}^2} = 0,$$

which, by (7.6.70) and (7.6.73), leads to

$$k_1 = 10^{-57} \text{km}^{-3}. \tag{7.6.74}$$

In summary, we conclude that the dark matter and dark energy are essentially gravitational effect generated by the gravitational potential field $g_{\mu\nu}$, its dual field Φ_μ and their nonlinear interactions. They can be regarded as the gravitational field energy caused by $g_{\mu\nu}$ and Φ_μ.

Bibliography

Davis, R., Harmer, D. S., and Hoffman, K. C. (1968). Search for neutrinos from the sun. *Phys. Rev. Lett.* **20**, 1205–1209, 10.1103/Phys. Rev. Lett., 20.1205.

de Broglie, L. (1924). Recherches sur la théorie des quanta, Thesis Paris.

Einstein, A. (1905). Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, **322**, 132–148, doi:10.1002/andp.19053220607.

Einstein, A. (1915). Die feldgleichungen der gravitation. *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin*, 844–847.

Einstein, A. (1916). The foundation of the general theory of relativity. *Annalen der Physik*, **354 (7)**, 769–822.

Englert, F. and Brout, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, **13 (9)**, 321–323.

Greiner, W. (2000). Relativistic Quantum Mechanics: Ware Equation. Berlin, Heidelberg, New York: Springer.

Griffiths, D. (2008). Introduction to Elementary Particles. New York: Wiley-Vch.

Gross, D. J. and Wilczek, F. (1973). Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.*, **30**, 1343–1346, 10.1103/Phys. Rev. Lett., 30.1343.

Guralnik, G., Hagen, C. R., and Kibble, T. W. B. (1964). Global conservation laws and massless particles. *Phys. Rev. Lett.*, **13 (20)**, 585–587.

Halzen, F. and Martin, A. D. (1984). Quarks and Leptons: An Introductory Course in Modern Particle Physics. New York: John Wiley and Sons.

Harrison, E. (2000). Cosmology, the Science of the Universe. 2nd ed. Cambridge University Press.

Hernandez, M., Ma, T., and Wang, S. (2015). Theory of dark energy and dark matter. To appear.

Higgs, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, **13**, 508–509.

Kaku, M. (1993). Quantum Field Theory, A Modern Introduction. London: Oxford University Press.

Kane, G. (1987). Modern Elementary Particle Physics, Vol. 2. Reading: Addison-Wesley.

Klein, O. (1938). *New Theories in Physics, 77* (Intern. Inst. of Intellectural Co-operation, League of nations).

Kutner, M. L. (2003). Astronomy: A Physical Perspective. Cambridge University Press.

Landau, L. (1957). On the conservation laws for weak interactions. *Nuclear Physics*, **3**, 1, 127–131, doi:http://dx.doi.org/10.1016/0029-5582(57)90061-5.

Landau, L. D. and Lifshitz, E. M. (1975). The Classical Theory of Fields. Oxford: Pergamon Press. Translated from the Russian by Morton Hamermesh.

Lee, T. D. and Yang, C. N. (1956). Question of parity conservation in weak interactions. *Phys. Rev.*, **104**, 254–258, doi:10.1103/PhysRev.104.254.

Lee, T. D. and Yang, C. N. (1957). Parity nonconservation and a two-component theory of the neutrino. *Phys. Rev.*, **105**, 1671.

Ma, T. (2010). Manifold Topology. Beijing: Science Press in Chinese.

Ma, T. (2011). Theory and Methods of Partial Differential Equations. Beijing: Science Press in Chinese.

Ma, T. (2012). Physical World From the Mathematical Point of View: Geometrical Analysis, Gravitational Field, and Relativity. Beijing: Science Press in Chinese.

Ma, T. (2014). Physical World From the Mathematical Point of View: Theory of Elementary Particles and Unified Fields. Beijing: Science Press in Chinese.

Ma, T. and Wang, S. (2005). *Bifurcation theory and applications*, *World Scientific Series on Nonlinear Science. Series A: Monographs and Treatises*, Vol. 53 (World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ),.

Ma, T. and Wang, S. (2013a). Duality theory of weak interaction, IU ISCAM Preprint Series, #1302: `http://www.indiana.edu/~iscam/preprint/1302.pdf`.

Ma, T. and Wang, S. (2013b). Phase Transition Dynamics. Springer-Verlag.

Ma, T. and Wang, S. (2013c). Structure and stability of matter, IU ISCAM Preprint Series, #1303: `http://www.indiana.edu/~iscam/preprint/1303.pdf`.

Ma, T. and Wang, S. (2014a). Astrophysical dynamics and cosmology. *Journal of Mathematical Study*, **47:4**, 305–378.

Ma, T. and Wang, S. (2014b). Color algebra in quantum chromodynamics, Isaac Newton Institute Preprint # NI14001: `https://www.newton.ac.uk/files/preprints/ni14001.pdf`.

Ma, T. and Wang, S. (2014c). Duality theory of strong interaction. *Electronic Journal of Theoretical Physics* **11:31**, 101–124, iU ISCAM Preprint Series, #1301: `http://www.indiana.edu/~iscam/preprint/1301.pdf`.

Ma, T. and Wang, S. (2014d). Field theory for multi-particle system, Isaac Newton Institute Preprint # NI14057: `https://www.newton.ac.uk/files/preprints/ni14057.pdf`.

Ma, T. and Wang, S. (2014e). Gravitational field equations and theory of dark matter and dark energy, *Discrete and Continuous Dynamical Systems, Ser. A* **34:2**, 335–366, preprint, `http://arxiv.org/abs/1206.5078`.

Ma, T. and Wang, S. (2014f). On solar neutrino problem, Isaac Newton Institute Preprint # NI14058: `https://www.newton.ac.uk/files/preprints/ni14058.pdf`.

Ma, T. and Wang, S. (2014g). Spectral theory of differential operators and energy levels of subatomic particles, Isaac Newton Institute Preprint # NI14002: `https://www.newton.ac.uk/files/preprints/ni14002.pdf`.

Ma, T. and Wang, S. (2014h). Unified field theory and principle of representation invariance. *Applied Mathematics and Optimization* **69:3**, 359–392, doi:0.1007/s00245-013-9226-0, preprint, `http://arxiv.org/abs/1212.4893`.

Ma, T. and Wang, S. (2015a). Unified field equations coupling four forces and principle of interaction dynamics. *Discrete and Continuous Dynamical Systems, Ser. A*, **35:3**, 1103–1138, preprint, `http://arxiv.org/abs/1210.0448`.

Ma, T. and Wang, S. (2015b). Weakton model of elementary particles and decay mechanisms. *Electronic Journal of Theoretical Physics* **12:32**, 139–178, iU ISCAM Preprint Series, #1304: `http://www.indiana.edu/~iscam/preprint/1304.pdf`.

Maki, Z., Nakagawa, M., and Sakata, S. (1962). Remarks on the unified model of elementary particles. *Progress of Theoretical Physics* **28**, 870–880.

Mikheev, S. P. and Smirnov, A. Y. (1986). Resonant amplification of ν oscillation in matter and solar-neutrino spectroscopy, *IL Nuovo Cimento C*, **9:1**, 17–26.

Milgrom, M. (1983). A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *Astrophysical Journal* **270**, 365–370, doi:10.1086/161130.

Milgrom, M. (2014). MOND laws of galactic dynamics. *MNRAS*, **437**, 2531–2541, doi:10.1093/mnras/stt2066, eprint 1212.2568.

Nambu, Y. (1960). Quasi-particles and gauge invariance in the theory of superconductivity. *Phys. Rev.*, **117**, 648–663, doi:10.1103/Phys.Rev.117.648.

Nambu, Y. (2008). Spontaneous symmetry breaking in particle physics: a case of cross fertilization, Slides for his Nobel Lecture.

Nambu, Y. and Jona-Lasinio, G. (1961a). Dynamical model of elementary particles based on an analogy with superconductivity. *Phys. Rev.* I, **122**, 345–358, doi:10.1103/PhysRev.122.345.

Nambu, Y. and Jona-Lasinio, G. (1961b). Dynamical model of elementary particles based on an analogy with superconductivity. *Phys. Rev.* II, **124**, 246–254, doi:10.1103/PhysRev.124.246.

Perlmutter, S. and et al. (1999). Measurements of ω and λ from 42 high-redshift supernovae. *Astrophys. J.*, **517**, 565–586.

Planck, M. (1901). Ueber das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik*, **309**, 553–563, doi:10.1002/andp.19013090310.

Politzer, H. D. (1973). Reliable perturbative results for strong interactions? *Phys. Rev. Lett.*, **30**, 1346–1349, doi:10.1103/PhysRevLett.30.1346.

Pontecorvo, B. (1957). Mesonium and anti-mesonium. *Sov.Phys.JETP*, **6**, 429.

Pontecorvo, B. (1968). Neutrino experiments and the problem of conservation of leptonic charge. *Sov.Phys.JETP*, **26**, 984.

Popławski, N. (2012). Nonsingular, big-bounce cosmology from spinor-torsion coupling. *Phys. Rev. D*, **85**, 107502, doi:10.1103/PhysRevD.85.107502.

Quigg, C. (2013). Gauge Theories of the Strong, Weak, and Electromagnetic Interactions, 2nd ed. Princeton Unversity Press.

Riess, A. G. and et al. (1989). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.*, **116**, 1009–1038.

Roos, M. (2003). *Introduction to Cosmology, 3rd Ed.*. Wiley.

Rubin, V. and Ford, J. W. K. (1970). Rotation of the andromeda nebula from a spectroscopic survey of emission regions. *Astrophysical Journal* **159**, 379–404.

Sakurai, J. J. (1994). *Modern Quantum Mechanics*, revised ed. Addison-Wesley Publishing Company, Inc.

Salam, A. (1957). On parity conservation and neutrino mass, *Nuovo Cimento*, **5**, 299–301.

Sokolov, A. A., Loskutov, Y. M., and Ternov, I. M. (1966). Quantum Mechanics. New York: Holt, Rinehart & Winston, Inc.

Weisskopf, V. (1972). Physics in the Twentieth Century: Selected Essays. Cambridge, Massachusetts: MIT Press.

Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.* **71**, 4, 441–479, doi:10.1007/BF01456804.

Weyl, H. (1919). Raum, Zeit, Materie. Verlag von Julius Springer.

Weyl, H. (1929). Electron and gravitation. *Z. Phys.*, **56**, 330–352.

Wolfenstein, L. (1978). Neutrino oscillations in matter. *Phys. Rev. D*, **17**, 2369–2374, doi:10.1103/PhysRevD.17.2369.

Wu, C. S., Ambler, E., Hayward, R. W., et al. (1957). Experimental test of parity conservation in beta decay. *Phys. Rev.*, **105**, 1413–1415, doi:10.1103/PhysRev.105.1413.

Yang, C. N. and Mills, R. (1954). Conservation of isotopic spin and isotopic gauge invariance. *Phys. Rev*. **96**, 191–195.

Zwicky, F. (1937). On the masses of nebulae and of clusters of nebulae. *Astrophysical Journal*, **86**, 217–246.

Index

C

D

G

H

Q

R

S

T

Y