Chapter 2

Optimal Markovian Couplings

This chapter introduces our first mathematical tool, the coupling methods, in the study of the topics in the book, and they will be used many times in the subsequent chapters. We introduce couplings, Markovian couplings (Section 2.1), and optimal Markovian couplings (Sections 2.2 and 2.3), mainly for time-continuous Markov processes. The study emphasizes analysis of the coupling operators rather than the processes. Some constructions of optimal Markovian couplings for Markov chains and diffusions are presented, which are often unexpected. Two general results of applications to the estimation of the first eigenvalue are proved in Section 2.4. Furthermore, some typical applications of the methods are illustrated through simple examples.

2.1 Couplings and Markovian couplings

Let us recall the simple definition of couplings.

Definition 2.1. Let μ_k be a probability on a measurable space $(E_k, \mathscr{E}_k)$, $k=1,2$. A probability measure $\widetilde{\mu}$ on the product measurable space $(E_1 \times E_2, \mathscr{E}_1 \times \mathscr{E}_2)$ is called a *coupling of* μ_1 *and* μ_2 if the following *marginality* condition holds:

$$\begin{aligned} \tilde{\mu}(A_1 \times E_2) &= \mu_1(A_1), \qquad A_1 \in \mathscr{E}_1, \\ \tilde{\mu}(E_1 \times A_2) &= \mu_2(A_2), \qquad A_2 \in \mathscr{E}_2. \end{aligned} \tag{M}$$

Example 2.2 (Independent coupling $\tilde{\mu}_0$). $\tilde{\mu}_0 = \mu_1 \times \mu_2$. That is, $\widetilde{\mu}_0$ is the independent product of μ_1 and μ_2.

This trivial coupling has already a nontrivial application. Let $\mu_k = \mu$ on $\mathbb{R}$, $k = 1, 2$. We say that μ satisfies the *FKG inequality* if

$$\int_{\mathbb{R}} f g \mathrm{d}\mu \geqslant \int_{\mathbb{R}} f \mathrm{d}\mu \int_{\mathbb{R}} g \mathrm{d}\mu, \qquad f, g \in \mathscr{M}, \tag{2.1}$$

where $\mathscr{M}$ is the set of bounded monotone functions on $\mathbb{R}$. Here is a one-line proof based on the independent coupling:

$$\iint \tilde{\mu}_0(\mathrm{d}x, \mathrm{d}y)[f(x) - f(y)][g(x) - g(y)] \geqslant 0, \qquad f, g \in \mathscr{M}.$$

We mention that a criterion of FKG inequality for higher-dimensional measures on $\mathbb{R}^d$ (more precisely, for diffusions) was obtained by Chen and F.Y. Wang (1993a). However, a criterion is still unknown for Markov chains.

Open Problem 2.3. What is the criterion of FKG inequality for Markov jump processes?

We will explain the meaning of the problem carefully at the end of this section and explain the term "Markov jump processes" soon. The next example is nontrivial.

Example 2.4 (Basic coupling $\tilde{\mu}_b$). Let $E_k = E$, $k = 1, 2$. Denote by Δ the diagonals in E: $\Delta = \{(x, x) : x \in E\}$. Take

$$\tilde{\mu}_b(\mathrm{d}x_1, \mathrm{d}x_2) = (\mu_1 \wedge \mu_2)(\mathrm{d}x_1) I_\Delta + \frac{(\mu_1 - \mu_2)^+(\mathrm{d}x_1)(\mu_1 - \mu_2)^-(\mathrm{d}x_2)}{(\mu_1 - \mu_2)^+(E)} I_{\Delta^c},$$

where $\nu^\pm$ is the Jordan–Hahn decomposition of a signed measure ν and $\nu_1 \wedge \nu_2 = \nu_1 - (\nu_1 - \nu_2)^+$.

Note that one may ignore I_{Δ^c} in the above formula, since $(\mu_1 - \mu_2)^+$ and $(\mu_1 - \mu_2)^-$ have different supports.

Actually, the basic coupling is optimal in the following sense. Let ρ be the discrete distance: $\rho(x, y)$=1 if $x \neq y$, and $= 0$ if $x = y$. Then a simple computation shows that

$$\tilde{\mu}_b(\rho) = \frac{1}{2}\|\mu_1 - \mu_2\|_{\mathrm{Var}}.$$

Thus, by Dobrushin's theorem (see Theorem 2.23 below), we have

$$\tilde{\mu}_b(\rho) = \inf_{\tilde{\mu}} \tilde{\mu}(\rho),$$

where $\tilde{\mu}$ varies over all couplings of μ_1 and μ_2. In other words, $\tilde{\mu}_b(\rho)$ is a *ρ-optimal coupling*. This indicates an optimality for couplings that we are going to study in this chapter.

Similarly, we can define a coupling process of two stochastic processes in terms of their distributions at each time t for fixed initial points. Of course, for given marginal Markov processes, the resulting coupled process may not be Markovian. Non-Markovian couplings are useful, especially in the time-discrete situation. However, in the time-continuous case, they are often not practical. Hence, we now restrict ourselves to the Markovian couplings.

Definition 2.5. Given two Markov processes with semigroups $P_k(t)$ or transition probabilities $P_k(t, x_k, \cdot)$ on $(E_k, \mathscr{E}_k)$, $k = 1,\ 2$, a *Markovian coupling* is a Markov process with semigroup $\widetilde{P}(t)$ or transition probability $\widetilde{P}(t; x_1, x_2; \cdot)$ on the product space $(E_1 \times E_2, \mathscr{E}_1 \times \mathscr{E}_2)$ having the *marginality*

$$\begin{aligned}&\widetilde{P}(t; x_1, x_2; A_1 \times E_2) = P_1(t, x_1, A_1),\\ &\widetilde{P}(t; x_1, x_2; E_1 \times A_2) = P_2(t, x_2, A_2), \quad t \geqslant 0, x_k \in E_k, A_k \in \mathscr{E}_k, k = 1, 2.\end{aligned} \tag{MP}$$

Equivalently,

$$\begin{aligned}&\widetilde{P}(t) f(x_1, x_2) = P_1(t) f(x_1),\\ &\widetilde{P}(t) f(x_1, x_2) = P_2(t) f(x_2), \qquad t \geqslant 0,\ x_k \in E_k,\ f \in {}_b\mathscr{E}_k,\ k = 1, 2,\end{aligned} \tag{MP}$$

where ${}_b\mathscr{E}$ is the set of all bounded $\mathscr{E}$-measurable functions. Here, on the left-hand side, f is regarded as a bivariate function.

We now consider Markov jump processes. For this, we need some notation. Let $(E, \mathscr{E})$ be a measurable space such that $\{(x, x) :\ x \in E\} \in \mathscr{E} \times \mathscr{E}$ and $\{x\} \in \mathscr{E}$ for all $x \in E$. It is well known that for a given sub-Markovian transition function $P(t, x, A)\, (t \geqslant 0, x \in E, A \in \mathscr{E})$, if it satisfies the *jump condition*

$$\lim_{t \to 0} P(t, x, \{x\}) = 1, \qquad x \in E, \tag{2.2}$$

then the limits

$$q(x) := \lim_{t \to 0} \frac{1 - P(t, x, \{x\})}{t} \quad \text{and} \quad q(x, A) := \lim_{t \to 0} \frac{P(t, x, A \setminus \{x\})}{t} \tag{2.3}$$

exist for all $x \in E$ and $A \in \mathscr{R}$, where

$$\mathscr{R} = \left\{ A \in \mathscr{E} :\ \lim_{t \to 0} \sup_{x \in A} \big[1 - P(t, x, \{x\})\big] = 0 \right\}.$$

Moreover, for each $A \in \mathscr{R}$, $q(\cdot),\ q(\cdot, A) \in \mathscr{E}$, for each $x \in E$, $q(x, \cdot)$ is a finite measure on $(E, \mathscr{R})$, and $0 \leqslant q(x, A) \leqslant q(x) \leqslant \infty$ for all $x \in E$ and $A \in \mathscr{R}$. The pair $(q(x),\ q(x, A))\, (x \in E,\ A \in \mathscr{R})$ is called a *q-pair* (also called the transition intensity or transition rate). The q-pair is said to be *totally stable* if $q(x) < \infty$ for all $x \in E$. Then $q(x, \cdot)$ can be uniquely extended to the whole space $\mathscr{E}$ as a finite measure. Next, the q-pair $\big(q(x), q(x, A)\big)$ is called *conservative* if $q(x, E) = q(x) < \infty$ for all $x \in E$ (Note that the conservativity here is different from the one often used in the context of diffusions). Because of the above facts, we often call the sub-Markovian transition $P(t, x, A)$ satisfying (2.3) a *jump process* or a *q-process*. Finally, a q-pair is called *regular* if it is not only totally stable and conservative but also determines uniquely a jump process (nonexplosive).

When E is countable, conventionally we use the matrix $Q = (q_{ij} :\ i, j \in E)$ (called a *Q-matrix*) and $P(t) = (p_{ij}(t) : i, j \in E)$,

$$p'_{ij}(t)\big|_{t=0} = q_{ij},$$

instead of the q-pair and the jump process, respectively. Here $q_{ii} = -q_i$, $i \in E$. We also call $P(t) = (p_{ij}(t))$ a *Markov chain* (which is used throughout this book only for a discrete state space) or a *Q-process*.

In practice, what we know in advance is the q-pair $(q(x), q(x, \mathrm{d}y))$ but not $P(t, x, \mathrm{d}y)$. Hence, our real interest goes in the opposite direction. How does a q-pair determine the properties of $P(t, x, \mathrm{d}y)$? A large part of the book (Chen, 1992a) is devoted to the theory of jump processes. Here, we would like to mention that the theory now has a very nice application to quantum physics that was missed in the quoted book. Refer to the survey article by A.A. Konstantinov, U.P. Maslov, and A.M. Chebotarev (1990) and references within.

Clearly, there is a one-to-one correspondence between a q-pair and the operator Ω:

$$\Omega f(x) = \int_E q(x, \mathrm{d}y)[f(y) - f(x)] - [q(x) - q(x, E)]f(x), \qquad f \in {}_b\mathscr{E}.$$

Because of this correspondence, we will use both according to our convenience. Corresponding to a coupled Markov jump process, we have a q-pair $(\tilde{q}(x_1, x_2), \tilde{q}(x_1, x_2; \mathrm{d}y_1, \mathrm{d}y_2))$ as follows:

$$\begin{aligned}
&\tilde{q}(x_1, x_2) = \lim_{t \to 0} \frac{1 - \widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\})}{t}, \qquad (x_1, x_2) \in E_1 \times E_2,\\
&\tilde{q}(x_1, x_2;\, \tilde{A}) = \lim_{t \to 0} \frac{\widetilde{P}(t; x_1, x_2; \tilde{A})}{t}, \qquad (x_1, x_2) \notin \tilde{A} \in \widetilde{\mathscr{R}},\\
&\widetilde{\mathscr{R}} := \Big\{\tilde{A} \in \mathscr{E}_1 \times \mathscr{E}_2 : \lim_{t \to 0} \sup_{(x_1, x_2) \in \tilde{A}} \big[1 - \widetilde{P}(t; x_1, x_2; \{(x_1, x_2)\})\big] = 0\Big\}.
\end{aligned}$$

Concerning the total stability and conservativity of the q-pair of a coupling (or coupled) process, we have the following result.

Theorem 2.6. The following assertions hold:

(1) A (equivalently, any) Markovian coupling is a jump process iff so are their marginals.

(2) A (equivalently, any) coupling q-pair is totally stable iff so are the marginals.

(3) [Y. H. Zhang, 1994]. A (equivalently, any) coupling q-pair is conservative iff so are the marginals.

Proof of parts (1) and (2). To obtain a feeling for the proof, we prove here the easier part of the theorem. This proof is taken from Chen (1994b).

(a) First, we consider the *jump condition.* Let $P_k(t, x_k, \mathrm{d}y_k)$ and $\widetilde{P}(t; x_1, x_2; \mathrm{d}y_1, \mathrm{d}y_2)$ be the marginal and coupled Markov processes, respectively. By the marginality for processes, we have

$$\begin{aligned}
&\widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\}) \\
&\quad \geqslant \widetilde{P}(t; x_1, x_2; \{x_1\} \times E_2) - \widetilde{P}(t; x_1, x_2; E_1 \times (E_2 \setminus \{x_2\})) \\
&\quad \geqslant \widetilde{P}(t; x_1, x_2; \{x_1\} \times E_2) - 1 + \widetilde{P}(t; x_1, x_2; E_1 \times \{x_2\}) \\
&\quad = P_1(t, x_1, \{x_1\}) - 1 + P_2(t, x_2, \{x_2\}).
\end{aligned}$$

If both of the marginals are jump processes, then $\underline{\lim}_{t\to 0} \widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\}) \geqslant 1$. Thus, a Markovian coupling $\widetilde{P}(t)$ must be a jump process.

Conversely, since

$$\widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\}) \leqslant \widetilde{P}(t; x_1, x_2; \{x_1\} \times E_2) = P_1(t, x_1, \{x_1\}),$$

if $\widetilde{P}(t)$ is a jump process, then $\underline{\lim}_{t\to 0} P_1(t, x_1, \{x_1\}) \geqslant 1$, and so $P_1(t)$ is also a jump process. Symmetrically, so is $P_2(t)$.

(b) Next, we consider the *equivalence of total stability.* Assume that all the processes concerned are jump processes. Denote by $(q_k(x_k),\ q_k(x_k, \mathrm{d}y_k))$ the marginal q-pairs on $(E_k, \mathscr{R}_k)$, where

$$\mathscr{R}_k = \left\{ A \in \mathscr{E}_k : \ \lim_{t\to 0} \sup_{x \in A} \big[1 - P_k(t, x, \{x\})\big] = 0 \right\}, \qquad k = 1,\ 2.$$

Denote by $(\tilde{q}(x_1, x_2), \tilde{q}(x_1, x_2; \mathrm{d}y_1, \mathrm{d}y_2))$ a coupling q-pair on $(E_1 \times E_2, \widetilde{\mathscr{R}})$. We need to show that $\tilde{q}(\tilde{x}) < \infty$ for all $\tilde{x} \in E_1 \times E_2$ iff $q_1(x_1) \vee q_2(x_2) < \infty$ for all $x_1 \in E_1$ and $x_2 \in E_2$. Clearly, it suffices to show that

$$q_1(x_1) \vee q_2(x_2) \leqslant \tilde{q}(x_1, x_2) \leqslant q_1(x_1) + q_2(x_2).$$

Note that we cannot use either the conservativity or uniqueness of the processes at this step. But the last assertion follows from (a) and the first part of (2.3) immediately. $\square$

Due to Theorem 2.6, *from now on, assume that all coupling operators considered below are conservative.* Then we have

$$\begin{aligned}
\tilde{q}(x_1, x_2) &= \lim_{t\to 0} \frac{1 - \widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\})}{t}, \qquad (x_1, x_2) \in E_1 \times E_2, \\
\tilde{q}(x_1, x_2;\ \tilde{A}) &= \lim_{t\to 0} \frac{\widetilde{P}(t; x_1, x_2; \tilde{A})}{t}, \qquad (x_1, x_2) \notin \tilde{A} \in \mathscr{E}_1 \times \mathscr{E}_2.
\end{aligned}$$

Note that in the second line, the original set $\widetilde{\mathscr{R}}$ is replaced by $\mathscr{E}_1 \times \mathscr{E}_2$. Define

$$\Omega_1 f(x_1) = \int_{E_1} q_1(x_1, \mathrm{d}y_1)[f(y_1) - f(x_1)], \qquad f \in {}_b\mathscr{E}_1.$$

Similarly, we can define Ω_2. Corresponding to a coupling process $\widetilde{P}(t)$, we also have an operator $\widetilde{\Omega}$. Now, since the marginal q-pairs and the coupling

q-pairs are all conservative, it is not difficult to prove that (MP) implies the following:

$$\begin{aligned} \widetilde{\Omega} f(x_1, x_2) &= \Omega_1 f(x_1), \qquad f \in {}_b\mathscr{E}_1, \\ \widetilde{\Omega} f(x_1, x_2) &= \Omega_2 f(x_2), \qquad f \in {}_b\mathscr{E}_2,\ x_k \in E_k,\ k = 1, 2. \end{aligned} \tag{MO}$$

Again, on the left-hand side, f is regarded as a bivariate function. Refer to Chen (1986a) or Chen (1992a, Chapter 5). Here, "MO" means the marginality for operators.

Definition 2.7. Any operator $\widetilde{\Omega}$ satisfying (MO) is called a *coupling operator*.

Do there exist any coupling operators?

Examples of coupling operators for jump processes

The simplest example to answer the above question is the following.

Example 2.8 (Independent coupling $\widetilde{\Omega}_0$).

$$\widetilde{\Omega}_0 f(x_1, x_2) = [\Omega_1 f(\cdot, x_2)](x_1) + [\Omega_2 f(x_1, \cdot)](x_2), \qquad x_k \in E_k,\ k = 1, 2.$$

This coupling is trivial, but it does show that a coupling operator always exists.

To simplify our notation, in what follows, instead of writing down a coupling operator, we will use tables. For instance, a conservative q-pair can be expressed as follows:

$$x \to \mathrm{d}y \setminus \{x\} \qquad \text{at rate} \qquad q(x, \mathrm{d}y).$$

In particular, in the discrete case, a conservative Q-matrix can be expressed as

$$i \to j \neq i \qquad \text{at rate} \qquad q_{ij}.$$

Example 2.9 (Classical coupling $\widetilde{\Omega}_c$). Take $E_1 = E_2 = E$ and let $\Omega_1 = \Omega_2 = \Omega$. If $x_1 \neq x_2$, then take

$$\begin{aligned} (x_1,\, x_2) \quad &\to \quad (y_1,\, x_2) \quad \text{at rate } q(x_1, \mathrm{d}y_1) \\ &\to \quad (x_1,\, y_2) \quad \text{at rate } q(x_2, \mathrm{d}y_2). \end{aligned}$$

Otherwise,

$$(x,\, x) \to (y,\, y) \quad \text{at rate} \quad q(x, \mathrm{d}y).$$

Each coupling has its own character. The classical coupling means that the marginals evolve independently until they meet. Then they move together. A nice way to interpret this coupling is to use a Chinese idiom: fall in love at first sight. That is, a boy and a girl had independent paths of their lives before the first time they met each other. Once they meet, they are in love at once and will have the same path of their lives forever. When the marginal Q-matrices are the same, all couplings considered below will have the property listed in the last line, and hence we will omit the last line in what follows.

Example 2.10 (Basic coupling $\widetilde{\Omega}_b$). For $x_1, x_2 \in E$, take

$$\begin{aligned}(x_1, x_2) &\to (y, y) &&\text{at rate } \big[q_1(x_1, \cdot) \wedge q_2(x_2, \cdot)\big](\mathrm{d}y)\\ &\to (y_1, x_2) &&\text{at rate } \big[q_1(x_1, \cdot) - q_2(x_2, \cdot)\big]^+(\mathrm{d}y_1)\\ &\to (x_1, y_2) &&\text{at rate } \big[q_2(x_2, \cdot) - q_1(x_1, \cdot)\big]^+(\mathrm{d}y_2).\end{aligned}$$

The basic coupling means that the components jump to the same place at the greatest possible rate. This explains where the term $q_1(x_1, \mathrm{d}y_1) \wedge q_2(x_2, \mathrm{d}y_2)$ comes from, which is the biggest one to guarantee the marginality. This term is the key of the coupling. Note that whenever we have a term $A \wedge B$, we should have the other two terms $(A - B)^+$ and $(B - A)^+$ automatically, again, due to the marginality. Thus, in what follows, we will write down the term $A \wedge B$ only for simplicity.

Example 2.11 (Coupling of marching soldiers $\widetilde{\Omega}_m$). Assume that E is an addition group. Take

$$(x_1, x_2) \to (x_1 + y, x_2 + y) \qquad \text{at rate} \qquad q_1(x_1, x_1 + \mathrm{d}y) \wedge q_2(x_2, x_2 + \mathrm{d}y).$$

The word "marching" is a Chinese name, which is the command to soldiers to start marching. Thus, this coupling means that at each step, the components maintain the same length of jumps at the biggest possible rate.

In the time-discrete case, the classical coupling and the basic coupling are due to W. Doeblin (1938) (which was the first paper to study the convergence rate by coupling) and L.N. Wasserstein (1969), respectively. The coupling of marching soldiers is due to Chen (1986b). The original purpose of the last coupling is mainly to preserve the order.

Let us now consider a birth–death process with regular Q-matrix:

$$q_{i,i+1} = b_i, \quad i \geqslant 0; \qquad q_{i,i-1} = a_i, \quad i \geqslant 1.$$

Then for two copies of the process starting from i_1 and i_2, respectively, we have the following two examples taken from (Chen, 1990).

Example 2.12 (Modified coupling of marching soldiers $\widetilde{\Omega}_{cm}$). Take $\widetilde{\Omega}_{cm} = \widetilde{\Omega}_c$ if $|i_1 - i_2| \leqslant 1$ and $\widetilde{\Omega}_{cm} = \widetilde{\Omega}_m$ if $|i_1 - i_2| \geqslant 2$.

Example 2.13 (Coupling by inner reflection $\widetilde{\Omega}_{ir}$). Again, take $\widetilde{\Omega}_{ir} = \widetilde{\Omega}_c$ if $|i_1 - i_2| \leqslant 1$. For $i_2 \geqslant i_1 + 2$, take

$$\begin{aligned}(i_1, i_2) &\to (i_1 + 1, i_2 - 1) &&\text{at rate } b_{i_1} \wedge a_{i_2}\\ &\to (i_1 - 1, i_2) &&\text{at rate } a_{i_1}\\ &\to (i_1, i_2 + 1) &&\text{at rate } b_{i_2}.\end{aligned}$$

By exchanging i_1 and i_2, we can get the expression of $\widetilde{\Omega}_{ir}$ for the case that $i_1 \geqslant i_2$.

This coupling lets the components move to the closed place (not necessarily the same place as required by the basic coupling) at the biggest possible rate.

From these examples one sees that there are many choices of a coupling operator $\widetilde{\Omega}$. Indeed, there are infinitely many choices! Thus, in order to use the coupling technique, a basic problem we should study is the regularity (nonexplosive problem) of coupling operators, for which, fortunately, we have a complete answer [Chen (1986a) or Chen (1992a, Chapter 5)]. The following result can be regarded as a fundamental theorem for couplings of jump processes.

Theorem 2.14 (Chen, 1986a)**.**

(1) If a coupling operator is nonexplosive, then so are its marginals.
(2) If the marginals are both nonexplosive, then so is every coupling operator.
(3) In the nonexplosive case, (MP) and (MO) are equivalent.

Clearly, Theorem 2.14 simplifies greatly our study of couplings for general jump processes, since the marginality (MP) of a coupling process is reduced to the rather simpler marginality (MO) of the corresponding operators. The hard but most important part of the theorem is the second assertion, since there are infinitely many coupling operators having no unified expression.

Markovian couplings for diffusions

We now turn to study the couplings for diffusion processes in $\mathbb{R}^d$ with second-order differential operator

$$L = \frac{1}{2}\sum_{i,\,j=1}^{d} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i}.$$

For simplicity, we write $L \sim (a(x), b(x))$. Given two diffusions with operators

$$L_k \sim (a_k(x), b_k(x)), \qquad k = 1,\, 2,$$

respectively, an elliptic (may be degenerate) operator $\widetilde{L}$ on the product space $\mathbb{R}^d \times \mathbb{R}^d$ is called a *coupling of* L_1 and L_2 if it satisfies the following *marginality*:

$$\begin{aligned} &\widetilde{L}f(x,y) = L_1 f(x) \qquad \big(\text{respectively, } \widetilde{L}f(x,y) = L_2 f(y)\big), \\ &\qquad f \in C_b^2(\mathbb{R}^d), \quad x \neq y. \end{aligned} \tag{MO}$$

Again, on the left-hand side, f is regarded as a bivariate function. From this, it is clear that the coefficients of any coupling operator $\widetilde{L}$ should be of the form

$$a(x,y) = \begin{pmatrix} a_1(x) & c(x,y) \\ c(x,y)^* & a_2(y) \end{pmatrix}, \qquad b(x,y) = \begin{pmatrix} b_1(x) \\ b_2(y) \end{pmatrix},$$

where the matrix $c(x,y)^*$ is the conjugate of $c(x,y)$. This condition and the nonnegative definite property of $a(x,y)$ constitute the *marginality* in the context of diffusions. Obviously, the only freedom is the choice of $c(x,y)$.

As an analogue of jump processes, we have the following examples.

Example 2.15 (Classical coupling). $c(x,y) \equiv 0$ for all $x \neq y$.

Example 2.16 (Coupling of marching soldiers [Chen and S.F. Li 1989]**).** Let $a_k(x) = \sigma_k(x)\sigma_k(x)^*$, $k = 1, 2$. Take $c(x,y) = \sigma_1(x)\sigma_2(y)^*$.

The two choices given in the next example are due to T. Lindvall and L.C.G. Rogers (1986), Chen and S.F. Li (1989), respectively.

Example 2.17 (Coupling by reflection). Let $L_1 = L_2$ and $a(x) = \sigma(x)\sigma(x)^*$. We have two choices:

$$c(x,y) = \sigma(x)\left[\sigma(y)^* - 2\frac{\sigma(y)^{-1}\bar{u}\bar{u}^*}{|\sigma(y)^{-1}\bar{u}|^2}\right], \quad \det\sigma(y) \neq 0, \qquad x \neq y,$$
$$c(x,y) = \sigma(x)\big[I - 2\bar{u}\bar{u}^*\big]\sigma(y)^*, \qquad x \neq y,$$

where $\bar{u} = (x-y)/|x-y|$.

This coupling was generalized to Riemannian manifolds by W.S. Kendall (1986) and M. Cranston (1991).

In the case that $x = y$, the first and the third couplings here are defined to be the same as the second one.

In probabilistic language, suppose that the original process is given by the stochastic differential equation

$$\mathrm{d}X_t = \sqrt{2}\,\sigma(X_t)\mathrm{d}B_t + b(X_t)\mathrm{d}t,$$

where (B_t) is a Brownian motion. We want to construct a new process (X_t'),

$$\mathrm{d}X_t' = \sqrt{2}\,\sigma'(X_t)\mathrm{d}B_t' + b'(X_t)\mathrm{d}t,$$

on the same probability space, having the same distribution as that of (X_t). Then, what we need is only to choose a suitable Brownian motion (B_t'). Corresponding to the above three examples, we have

(1) *Classical coupling*: B_t' is a new Brownian motion, independent of B_t.
(2) *Coupling of marching soldiers*: $B_t' = B_t$.
(3) *Coupling by reflection*: $B_t' = [I - 2\bar{u}\bar{u}^*](X_t, X_t')B_t$, where $\bar{u}$ is given in Example 2.17.

It is important to remark that in the constructions, we need only consider the time $t < T$, where T is the *coupling time*,

$$T = \inf\{t \geqslant 0 : X_t = X_t'\},$$

since $X_t = X_t'$ for all $t \geqslant T$. This avoids the degeneration of the coupling operators.

Before moving further, let us mention a conjecture:

Conjecture 2.18. The fundamental theorem (Theorem 2.14) holds for diffusions.

The following facts strongly support the conjecture.

(a) A well known sufficient condition says that the operator L_k $(k = 1, 2)$ is well posed if there exists a function φ_k such that $\lim_{|x|\to\infty} \varphi_k(x) = \infty$ and $L_k\varphi_k \leqslant c\varphi_k$ for some constant c. Then the conclusion holds for all coupling operators, simply taking

$$\tilde{\varphi}(x_1, x_2) = \varphi_1(x_1) + \varphi_2(x_2).$$

(b) Let $\tau_{n,k}$ be the first time of leaving the cube with side length n of the kth process $(k = 1, 2)$ and let $\tilde{\tau}_n$ be the first time of leaving the product cube of coupled process. Then we have

$$\tau_{n,1} \vee \tau_{n,2} \leqslant \tilde{\tau}_n \leqslant \tau_{n,1} + \tau_{n,2}.$$

Moreover, a process, the kth one for instance, is well posed iff

$$\lim_{n\to\infty} \mathbb{P}_k[\tau_{n,k} < t] = 0.$$

Having studied the Markovian couplings for Markov jump processes and diffusions, it is natural to study the Lévy processes.

Open Problem 2.19. What should be the representation of Markovian coupling operators for Lévy processes?

2.2 Optimality with respect to distances

Since there are infinitely many Markovian couplings, we asked ourselves several times in the past years, does there exist an optimal one? Now another question arises: What is the optimality we are talking about? We now explain how we obtained a reasonable notion for optimal Markovian couplings. The first time we touched this problem was in Chen and S.F. Li (1989). It was proved there for Brownian motion that coupling by reflection is optimal with respect to the total variation, and moreover, for different probability metrics, the effective couplings can be different. The second time, in Chen (1990), it was proved that for birth–death processes, we have an order as follows:

$$\widetilde{\Omega}_{ir} \succ \widetilde{\Omega}_b \succ \widetilde{\Omega}_c \succ \widetilde{\Omega}_{cm} \succ \widetilde{\Omega}_m,$$

where $A \succ B$ means that A is better than B in some sense. However, only in 1992 it did become clear to the author how to optimize couplings.

To explain our optimal couplings, we need more preparation. As was mentioned several times in previous publications [Chen (1989a; 1989b; 1992a) and Chen and S. F. Li (1989)], it should be helpful to keep in mind the relation between couplings and the probability metrics. It will be clear soon that this is actually one of the key ideas of the study. As far as we know, there are more than 16 different probability distances, including the total variation and the Lévy–Prohorov distance for weak convergence. But we often are concerned with another distance. We now explain our understanding of how to introduce this distance.

As we know, in probability theory, we usually consider the types of convergence for real random variables on a probability space shown in Figure 2.1.

$$
\begin{tikzcd}
 & \text{convergence in } L^p \arrow[d, Rightarrow] & \\
\text{a.s. convergence} \arrow[r, Rightarrow] & \text{convergence in } \mathbb{P} \arrow[r, Rightarrow] & \text{vague convergence} \arrow[dl, Leftrightarrow] \\
 & \text{weak convergence} \arrow[ul, Rightarrow, dashed] &
\end{tikzcd}
$$

Figure 2.1 Typical types of convergence in probability theory

L^p-convergence, a.s. convergence, and convergence in $\mathbb{P}$ all depend on the reference frame, our probability space $(\Omega, \mathscr{F}, \mathbb{P})$. But vague (weak) convergence does not. By a result of Skorohod [cf. N. Ikeda and S. Watanabe (1988, p. 9 Theorem 2.7)], if P_n converges weakly to P, then we can choose a suitable reference frame $(\Omega, \mathscr{F}, \mathbb{P})$ such that $\xi_n \sim P_n$, $\xi \sim P$, and $\xi_n \to \xi$ a.s., where $\xi \sim P$ means that ξ has distribution P. Thus, all the types of convergence listed in Figure 2.1 are intrinsically the same, except L^p-convergence. In other words, if we want to find another intrinsic metric on the space of all probabilities, we should consider an analogue of L^p-convergence.

Let ξ_1, ξ_2: $(\Omega, \mathscr{F}, \mathbb{P}) \to (E, \rho, \mathscr{E})$. The usual L^p-metric is defined by

$$\|\xi_1 - \xi_2\|_p = \big\{\mathbb{E}\big[\rho(\xi_1, \xi_2)^p\big]\big\}^{1/p}.$$

Suppose that $\xi_i \sim P_i$, $i = 1, 2$, and $(\xi_1, \xi_2) \sim \widetilde{P}$. Then

$$\|\xi_1 - \xi_2\|_p = \bigg\{\int \rho(x_1, x_2)^p \widetilde{P}(\mathrm{d}x_1, \mathrm{d}x_2)\bigg\}^{1/p}.$$

Certainly, $\widetilde{P}$ is a coupling of P_1 and P_2. However, if we ignore our reference frame $(\Omega, \mathscr{F}, \mathbb{P})$, then there are many choices of $\widetilde{P}$ for given P_1 and P_2. Thus, the intrinsic metric should be defined as follows:

$$W_p(P_1, P_2) = \inf_{\widetilde{P}} \bigg\{\int \rho(x_1, x_2)^p \widetilde{P}(\mathrm{d}x_1, \mathrm{d}x_2)\bigg\}^{1/p}, \qquad p \geqslant 1,$$

where $\widetilde{P}$ varies over all couplings of P_1 and P_2.

Definition 2.20. The metric defined above is called the *W_p-distance* or pth *Wasserstein distance*. Briefly, we write $W = W_1$.

From the probabilistic point of view, the W_p-metrics have an intrinsic property that makes them more suitable for certain applications. For example, if (E, ρ) is the Euclidean space for P_2 obtained from P_1 by a translation, then $W_p(P_1, P_2)$ is just the length of the translation vector.

In general, it is quite hard to compute the W_p-distance exactly. Here are the main known results.

Theorem 2.21 (S.S. Vallender, 1973)**.** Let P_k be a probability on the real line with distribution function $F_k(x)$, $k = 1, 2$. Then

$$W(P_1, P_2) = \int_{-\infty}^{+\infty} |F_1(x) - F_2(x)| \mathrm{d}x.$$

Theorem 2.22 (D.C. Dowson and B.V. Landau (1982), C.R. Givens and R.M. Shortt (1984), I. Olkin and R. Pukelsheim (1982))**.** Let P_k be the normal distribution on $\big(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d)\big)$ $(d \geqslant 1)$ with mean value m_k and covariance matrix M_k, $k = 1, 2$. Then

$$\begin{aligned} W_2(P_1, P_2) = \big[&|m_1 - m_2|^2 + \mathrm{Trace}\, M_1 + \mathrm{Trace}\, M_2 \\ &- 2\ \mathrm{Trace}\big(\sqrt{M_1} M_2 \sqrt{M_1}\,\big)^{1/2}\big]^{1/2}, \end{aligned}$$

where $\mathrm{Trace}\, M$ denotes the trace of M.

Theorem 2.23 (R.L. Dobrushin, 1970)**.** (1) For bounded ρ, W is equivalent to the Lévy-Prohorov distance.
(2) For discrete distance ρ, $W = \|\cdot\|_{\mathrm{Var}}/2$.

Fortunately, in most cases, what we need is only certain estimates of an upper bound. Clearly, any coupling provides an upper bound of $W(P_1, P_2)$. Thus, it is very natural to introduce the following notion.

Definition 2.24. A coupling $\overline{P}$ of P_1 and P_2 is called *ρ-optimal* if

$$\int \rho(x_1, x_2) \overline{P}(\mathrm{d}x_1, \mathrm{d}x_2) = W(P_1, P_2).$$

Now, it is natural to define the optimal coupling for time-discrete Markov processes without restriction to the Markovian class. In the special case of ρ being the discrete metric (or equivalently, restricted to the total variation), it is just the *maximal coupling*, introduced by D. Griffeath (1978). However, the maximal couplings constructed in the quoted paper are usually non-Markovian. Even though the maximal couplings as well as other non-Markovian couplings now constitute an important part of the theory and have been widely studied in the literature (cf. T. Lindvall (1992), J.G. Propp and D.B. Wilson (1996), H. Thorrison (2000), and references therein), they are difficult to handle, especially when we come to the time-continuous situation. Moreover, it will be clear soon that in the context of diffusions, in dealing with the optimal Markovian coupling in terms of their operators, the discrete metric will lose its meaning. Thus, our optimal Markovian couplings are essentially different from the maximal ones. It should also be pointed out that the sharp estimates introduced in Chapter 1 were obtained from the exponential rate in the W-metric with respect to some much more refined metric ρ rather than the discrete one. Replacing P_k and $\widetilde{P}$ with $P_k(t)$ and $\widetilde{P}(t)$, respectively, and then going to the operators, it is not difficult to arrive at the following notion [cf. Chen (1994b; 1994a) for details].

Definition 2.25. A coupling operator $\overline{\Omega}$ is called *ρ-optimal* if

$$\overline{\Omega}\,\rho(x_1,x_2)=\inf_{\widetilde{\Omega}}\widetilde{\Omega}\,\rho(x_1,x_2)\qquad\text{for all } x_1\neq x_2,$$

where $\widetilde{\Omega}$ varies over all coupling operators.

To see that the notion is useful, let us introduce one more coupling.

Example 2.26 (Coupling by reflection $\widetilde{\Omega}_r$). Given a birth–death process with birth rates b_i and death rates a_i, this coupling evolves in the following way. If $i_2=i_1+1$, then

$$\begin{array}{llll}(i_1,i_2) & \to & (i_1-1,i_2+1) & \text{at rate} \quad a_{i_1}\wedge b_{i_2}\\ & \to & (i_1+1,i_2) & \text{at rate} \quad b_{i_1}\\ & \to & (i_1,i_2-1) & \text{at rate} \quad a_{i_2}.\end{array}$$

If $i_2\geqslant i_1+2$, then

$$\begin{array}{llll}(i_1,i_2) & \to & (i_1-1,i_2+1) & \text{at rate} \quad a_{i_1}\wedge b_{i_2}\\ & \to & (i_1+1,i_2-1) & \text{at rate} \quad b_{i_1}\wedge a_{i_2}.\end{array}$$

By symmetry, we can write down the rates for the other case that $i_1>i_2$.

Intuitively, the reflection in the outside direction is quite strange, since it separates the components by distance 2 but not by 1. For this reason, even though the coupling came to our attention years ago, we never believed that it could be better than the coupling by inner reflection. But the next result changed our mind.

Theorem 2.27 (Chen, 1994a). For birth–death processes, the coupling by reflection is ρ-optimal for any translation-invariant metric ρ on $\mathbb{Z}_+$ having the property that

$$u_k:=\rho(0,k+1)-\rho(0,k),\quad k\geqslant 0$$

is nonincreasing in k.

To see that the optimal coupling depends heavily on the metric ρ, note that the above metric ρ can be rewritten as

$$\rho(i,j) = \sum_{k<|i-j|} u_k$$

for some positive nonincreasing sequence (u_k). In this way, for any positive sequence (u_k), we can introduce another metric as follows:

$$\tilde{\rho}(i,j) = \bigg|\sum_{k<i} u_k - \sum_{k<j} u_k\bigg|.$$

Because $(u_k > 0)$ is arbitrary, this class of metrics is still quite large. Now, among the couplings listed above, which are $\tilde{\rho}$-optimal?

Theorem 2.28 (Chen, 1994a)**.** For birth–death processes, every coupling mentioned above except the trivial (independent) one is $\tilde{\rho}$-optimal.

This result is again quite surprising, far from our probabilistic intuition. Thus, our optimality does produce some unexpected results.

We are now ready to study the optimal couplings for diffusion processes.

Definition 2.29. Given $\rho \in C^2(\mathbb{R}^d \times \mathbb{R}^d \setminus \{(x,x) : x \in \mathbb{R}^d\})$, a coupling operator $\overline{L}$ is called ρ-*optimal* if

$$\overline{L}\rho(x,y) = \inf_{\widetilde{L}} \widetilde{L}\rho(x,y), \qquad x \neq y,$$

where $\widetilde{L}$ varies over all coupling operators.

For the underlying Euclidean distance $|\cdot|$ in $\mathbb{R}^d$, we introduce a family of distances as follows:

$$\rho(x,y) = f(|x-y|), \text{ where } f(0) = 0,\ f' > 0, \text{ and } f'' \leqslant 0. \tag{2.4}$$

In order to make ρ a distance, the first two conditions of f are necessary and the third condition guarantees the triangle inequality. For this class of distance, as mentioned in the paper quoted below, the existence of ρ-optimal coupling for diffusion is not a serious problem. Here we introduce only some explicit constructions.

Theorem 2.30 (Chen, 1994a)**.** Let $\rho(x,y) = f(|x-y|)$ for some $f \in C^2(\mathbb{R}_+;\mathbb{R}_+)$ satisfying (2.4). Then the ρ-optimal solution $c(x,y)$ is given as follows:

(1) If $d = 1$, then $c(x,y) = -\sqrt{a_1(x)a_2(y)}$, and moreover,

$$\begin{aligned}\overline{L}f(|x-y|) = &\ \frac{1}{2}\big(\sqrt{a_1(x)} + \sqrt{a_2(y)}\,\big)^2 f''(|x-y|)\\ &+ \frac{(x-y)(b_1(x)-b_2(y))}{|x-y|} f'(|x-y|).\end{aligned}$$

Next, suppose that $a_k = \sigma_k^2$ $(k = 1, 2)$ is nondegenerate and write

$$c(x, y) = \sigma_1(x)H^*(x, y)\sigma_2(y).$$

(2) If $f''(r) < 0$ for all $r > 0$, then $H(x, y) = U(\gamma)^{-1}\big[U(\gamma)U(\gamma)^*\big]^{1/2}$, where

$$\gamma = 1 - \frac{|x-y|f''(|x-y|)}{f'(|x-y|)}, \qquad U(\gamma) = \sigma_1(x)(I - \gamma\bar{u}\bar{u}^*)\sigma_2(y).$$

(3) If $f(r) = r$, then $H(x, y)$ is a solution to the equation

$$U(1)H = \big(U(1)U(1)^*\big)^{1/2}.$$

In particular, if $a_k(x) = \varphi_k(x)\sigma^2$ for some positive function φ_k $(k = 1, 2)$, where σ is independent of x and $\det\sigma > 0$, then

(4) $H(x, y) = I - 2\sigma^{-1}\bar{u}\bar{u}^*\sigma^{-1}/|\sigma^{-1}\bar{u}|^2$ if $\rho(x, y) = |x - y|$. Moreover,

$$\begin{aligned}\overline{L}f(|x - y|) = \frac{1}{2|x - y|}\Big\{&\big(\sqrt{\varphi_1(x)} - \sqrt{\varphi_2(y)}\,\big)^2\big[\operatorname{Trace}\sigma^2 - |\sigma\bar{u}|^2\big]\\ &+ 2\big\langle x - y,\, b_1(x) - b_2(y)\big\rangle\Big\}.\end{aligned}$$

(5) H is the same as in the last assertion if $\rho(x, y) = |x - y|$ is replaced by $\rho(x, y) = f(|\sigma^{-1}(x - y)|)$. Furthermore,

$$\begin{aligned}\overline{L}\rho(x, y) = {}&\frac{1}{2}\big(\sqrt{\varphi_1(x)} + \sqrt{\varphi_2(y)}\,\big)^2 f''(|\sigma^{-1}(x - y)|)\\ &+ \Big\{(d - 1)\big(\sqrt{\varphi_1(x)} - \sqrt{\varphi_2(y)}\,\big)^2\\ &\qquad + 2\big\langle\sigma^{-1}(x - y),\, \sigma^{-1}(b_1(x) - b_2(y))\big\rangle\Big\}\\ &\times \frac{f'(|\sigma^{-1}(x - y)|)}{2|\sigma^{-1}(x - y)|}.\end{aligned}$$

2.3 Optimality with respect to closed functions

As an extension of the optimal couplings with respect to distances, we can consider the optimal couplings with respect to a more general, nonnegative, closed (= lower semicontinuous) function φ.

Definition 2.31. Given a metric space $(E, \rho, \mathscr{E})$, let φ be a nonnegative, closed function on $(E, \rho, \mathscr{E})$. A coupling is called a φ-*optimal* (Markovian) coupling if in the definitions given in the last section, the distance function ρ is replaced by φ.

Here are some typical examples of φ.

Example 2.32. (1) φ is a distance of the form $f \circ \rho$ for some f having the properties $f(0) = 0$, $f' > 0$, and $f'' \leqslant 0$.
(2) φ is the discrete distance: $\varphi(x, y) = 1$ iff $x \neq y$; otherwise, $\varphi(x, y) = 0$.
(3) Let E be endowed with a measurable semiorder "$\prec$" and set $F = \{(x, y) : x \prec y\}$. Then F is a closed set. Take $\varphi = I_{F^c}$.

Before moving further, let us recall the definition of stochastic comparability.

Definition 2.33. Let $\mathscr{M}$ be the set of bounded monotone functions f: $x \prec y \Longrightarrow f(x) \leqslant f(y)$.

(1) We write $\mu_1 \prec \mu_2$ if $\mu_1(f) \leqslant \mu_2(f)$ for all $f \in \mathscr{M}$.
(2) Let P_1 and P_2 be transition probabilities. We write $P_1 \prec P_2$ if $P_1(f)(x_1) \leqslant P_2(f)(x_2)$ for all $x_1 \prec x_2$ and $f \in \mathscr{M}$.
(3) Let $P_1(t)$ and $P_2(t)$ be transition semigroups. We write $P_1(t) \prec P_2(t)$ if $P_1(t)(f)(x_1) \leqslant P_2(t)(f)(x_2)$ for all $t \geqslant 0$, $x_1 \prec x_2$, and $f \in \mathscr{M}$.

Here is a famous result about stochastic comparability.

Theorem 2.34 (V. Strassen, 1965)**.** For a Polish space, $\mu_1 \prec \mu_2$ iff there exists a coupling measure $\bar{\mu}$ such that $\bar{\mu}(F^c) = 0$.

Usually, in practice, it is not easy to compare two measures directly. For this reason, one introduces stochastic comparability for processes. First, one constructs two processes with stationary distributions μ_1 and μ_2. Then the stochastic comparability of the two measures can be reduced to that of the processes. The advantage for the latter comparison comes from the intuition of the stochastic dynamics. One can even see the answer from the coefficients of the operators. See Examples 2.44–2.46 below.

A general result for φ-optimal coupling is the following.

Theorem 2.35 (S.Y. Zhang, 2000a)**.** Let $(E, \rho, \mathscr{E})$ be Polish and $\varphi \geqslant 0$ be a closed function.

(1) Given $P_k(x_k, \mathrm{d}y_k)$, $k = 1, 2$, there exists a transition probability $\overline{P}(x_1, x_2; \mathrm{d}y_1, \mathrm{d}y_2)$ such that $\overline{P}\varphi(x_1, x_2) = \inf_{\widetilde{P}^{(x_1,x_2)}} \widetilde{P}^{(x_1,x_2)}\varphi(x_1, x_2)$, where for fixed (x_1, x_2), $\widetilde{P}^{(x_1,x_2)}$ varies over all couplings of $P_1(x_1, \mathrm{d}y_1)$ and $P_2(x_2, \mathrm{d}y_2)$.
(2) Given operators Ω_k of regular jump processes, $k = 1, 2$, there exists a coupling operator $\overline{\Omega}$ of jump process such that $\overline{\Omega}\,\varphi = \inf_{\widetilde{\Omega}} \widetilde{\Omega}\,\varphi$, where $\widetilde{\Omega}$ varies over all coupling operators of Ω_1 and Ω_2.

According to Theorem 2.35 (1), Strassen's theorem can be restated as follows: the I_{F^c}-optimal Markovian coupling satisfies $\bar{\mu}(F^c) = 0$. This shows that Theorem 2.35 (1) is an extension of Strassen's theorem. Even though the proof of Theorem 2.35 is quite technical, the main root is still clear. Consider first finite state spaces. Then the conclusion follows from an existence theorem of linear programming regarding the marginality as a constraint. Next, pass to the general Polish space by using a tightness argument (a generalized Prohorov theorem) plus an approximation of φ by bounded Lipschitz functions.

Concerning stochastic comparability, we have the following result.

Theorem 2.36 (Chen (1992a, Chapter 5), Zhang (2000b))**.** For jump processes on a Polish space, under a mild assumption, $P_1(t) \prec P_2(t)$ iff

$$\Omega_1 I_B(x_1) \leqslant \Omega_2 I_B(x_2),$$

for all $x_1 \prec x_2$ and B with $I_B \in \mathscr{M}$.

Here we mention an additional result, which provides us the optimal solutions within the class of order-preserving couplings.

Theorem 2.37 (T. Lindvall, 1999)**.** Again, let Δ denote the diagonals.

(1) Let $\mu_1 \prec \mu_2$. Then

$$\inf_{\tilde{\mu}(F^c)=0} \tilde{\mu}(\Delta^c) = \frac{1}{2}\|\mu_1 - \mu_2\|_{\text{Var}}.$$

(2) Let P_1 and P_2 be transition probabilities that satisfy $P_1 \prec P_2$. Then

$$\inf_{\widetilde{P}(x_1,x_2;\, F^c)=0} \widetilde{P}(x_1, x_2; \Delta^c) = \frac{1}{2}\|P_1(x_1, \cdot) - P_2(x_2, \cdot)\|_{\text{Var}}$$

for all $x_1 \prec x_2$.

In fact, the left-hand sides of the formulas in Theorem 2.37 can be replaced, respectively, by the I_{F^c}-optimal coupling given in Theorem 2.35 (1).

For order-preserving Markovian coupling for diffusions, refer to F.Y. Wang and M.P. Xu (1997).

Open Problem 2.38. Let $\varphi \in C^2(\mathbb{R}^{2d} \setminus \Delta)$. Prove the existence of φ-optimal Markovian couplings for diffusions under some reasonable hypotheses.

Open Problem 2.39. Construct φ-optimal Markovian couplings.

2.4 Applications of coupling methods

It should be helpful for readers, especially newcomers, to see some applications of couplings. Of course, the applications discussed below cannot be complete, and additional applications will be presented in Chapters 3, 5, and 9. One may refer to T.M. Liggett (1985), T. Lindvall (1992), and H. Thorrison (2000) for much more information. The coupling method is now a powerful tool in statistics, called "copulas" (cf. R.B. Nelssen (1999)). It is also an active research topic in PDE and related fields, named "optimal transportation" (cf. S.T. Rachev and L. Ruschendorf (1998), L. Ambrosio et al. (2003), C. Villani (2003)).

Spectral gap; exponential L^2-convergence

We introduce two general results, due to Chen and F.Y. Wang (1993b) [see also Chen (1994a)], on the estimation of the first nontrivial eigenvalue (spectral gap) by couplings.

Definition 2.40. Let L be an operator of a Markov process $(X_t)_{t\geqslant 0}$. We say that a function f is in the weak domain of L, denoted by $\mathscr{D}_w(L)$, if f satisfies the forward Kolmogorov equation

$$\mathbb{E}^x f(X_t) = f(x) + \int_0^t \mathbb{E}^x Lf(X_s)\mathrm{d}s,$$

or equivalently, if

$$f(X_t) - \int_0^t Lf(X_s)\mathrm{d}s$$

is a $\mathbb{P}^x$-martingale with respect to the natural flow of σ-algebras $\{\mathscr{F}_t\}_{t\geqslant 0}$, where $\mathscr{F}_t = \sigma\{X_s : s \leqslant t\}$.

Definition 2.41. We say that g is an eigenfuction of L corresponding to λ in the weak sense if g satisfies the eigenequation $Lg = -\lambda g$ pointwise.

Note that the eigenfunction defined above may not belong to $L^2(\pi)$, where π is the stationary distribution of $(X_t)_{t\geqslant 0}$. In the reversible case, all of the eigenvalues are nonnegative and all of the eigenfunctions are real.

The next two results remain true in the irreversible case (where λ and g are often complex), provided λ is replaced by $|\lambda|$.

Theorem 2.42. Let (E, ρ) be a metric space and let $\{X_t\}_{t\geqslant 0}$ be a reversible Markov process with operator L. Denote by g the eigenfunction corresponding to $\lambda \neq 0$ in the weak sense. Next, let (X_t, Y_t) be a coupled process, starting from (x, y), with coupling operator $\widetilde{L}$, and let $\gamma : E \times E \to [0, \infty)$ satisfy $\gamma(x, y) = 0$ iff $x = y$. Suppose that

(1) $g \in \mathscr{D}_w(L)$,
(2) $\gamma \in \mathscr{D}_w(\widetilde{L})$,
(3) $\widetilde{L}\gamma(x, y) \leqslant -\alpha\gamma(x, y)$ for all $x \neq y$ and some constant $\alpha \geqslant 0$,
(4) g is Lipschitz with respect to γ in the sense that

$$c_{g,\gamma} := \sup_{y\neq x} \gamma(y, x)^{-1}|g(y) - g(x)| < \infty.$$

Then we have $\lambda \geqslant \alpha$.

Proof. Without loss of generality, assume that $\alpha > 0$. Otherwise, the conclusion is trivial. By conditions (2), (3) and Lemma A.6, we have

$$\widetilde{\mathbb{E}}^{x,y}\gamma(X_t, Y_t) \leqslant \gamma(x, y)e^{-\alpha t}, \qquad t \geqslant 0.$$

Next, by condition (1) and the definition of g,

$$g(X_t) - \int_0^t Lg(X_s)\mathrm{d}s = g(X_t) + \lambda \int_0^t g(X_s)\mathrm{d}s$$

is a $\mathbb{P}^x$-martingale with respect to the natural flow of σ-algebras $\{\mathscr{F}_t\}_{t\geqslant 0}$. In particular,

$$g(x) = \mathbb{E}^x\bigg[g(X_t) + \lambda \int_0^t g(X_s)\mathrm{d}s\bigg].$$

Because of the coupling property,

$$\mathbb{E}^x\bigg[g(X_t) + \lambda \int_0^t g(X_s)\mathrm{d}s\bigg] = \widetilde{\mathbb{E}}^{x,y}\bigg[g(X_t) + \lambda \int_0^t g(X_s)\mathrm{d}s\bigg].$$

Thus, we obtain

$$g(x) - g(y) = \widetilde{\mathbb{E}}^{x,y}\bigg[g(X_t) - g(Y_t) + \lambda \int_0^t [g(X_s) - g(Y_s)]\mathrm{d}s\bigg].$$

Therefore

$$\begin{aligned}
|g(x) - g(y)| &\leqslant \widetilde{\mathbb{E}}^{x,y}\big|g(X_t) - g(Y_t)\big| + \lambda\widetilde{\mathbb{E}}^{x,y}\int_0^t |g(X_s) - g(Y_s)|\mathrm{d}s \\
&\leqslant c_{g,\gamma}\widetilde{\mathbb{E}}^{x,y}\gamma(X_t, Y_t) + \lambda c_{g,\gamma}\widetilde{\mathbb{E}}^{x,y}\int_0^{t\wedge T} \gamma(X_s, Y_s)\mathrm{d}s \\
&\leqslant c_{g,\gamma}\gamma(x,y)e^{-\alpha t} + \lambda c_{g,\gamma}\gamma(x,y)\int_0^t e^{-\alpha s}\mathrm{d}s.
\end{aligned}$$

Noting that g is not a constant, since $\lambda \neq 0$, we have $c_{g,\gamma} \neq 0$. Dividing both sides by $\gamma(x,y)$ and choosing a sequence (x_n, y_n) such that

$$|g(y_n) - g(x_n)|/\gamma(y_n, x_n) \to c_{g,\gamma},$$

we obtain

$$1 \leqslant e^{-\alpha t} + \lambda \int_0^t e^{-\alpha s}\mathrm{d}s = e^{-\alpha t} + \lambda\big(1 - e^{-\alpha t}\big)/\alpha$$

for all t. This implies that $\lambda \geqslant \alpha$ as required. $\square$

One may compare this probabilistic proof with the analytic one sketched in Section 1.2.

When γ is a distance, $\widetilde{\mathbb{E}}^{x,y}\gamma(X_t, Y_t)$ is nothing but the Wasserstein metric $W = W_1$ with respect to γ of the distributions at time t. The above proof shows that W_1 can be used to study the Poincaré inequality (i.e., λ_1). Noting that W_2 is stronger than W_1 for a fixed underframe distance ρ, it is natural to study the stronger logarithmic Sobolev inequality in terms of W_2 with respect to the Euclidean distance, for instance. To deal with the inequalities themselves, it is helpful but not necessary to go to the dynamics, since they are mainly concerned with measures. Next, it was discovered in the 1990s that in many cases, for two given probability densities, the optimal coupling for W_2 exists uniquely, and the mass of the coupling measure is concentrated on the set $\{(x, T(x)) : x \in \mathbb{R}^d\}$. Moreover, the optimal transport T can be expressed by $T = \nabla\Psi$ for some convex function Ψ that solves a nonlinear Monge–Ampère equation. It turns out that this transportation solution provides a new way to prove a class of logarithmic Sobolev (or even more general) inequalities in $\mathbb{R}^d$. This explains roughly the interaction between probability distances (couplings) and PDE. Considerable progress has been made recently in this field, as shown in the last two books mentioned in the first paragraph of this section.

Condition (3) in Theorem 2.42 is essential. The other conditions can often be relaxed or avoided by using a localizing procedure. Define the *coupling time* $T = \inf\{t \geqslant 0 : X_t = Y_t\}$. The next, weaker, result is useful. It has a different meaning, as will be explained in Section 5.6. Indeed, the condition "$\sup_{x\neq y} \widetilde{\mathbb{E}}^{x,y} T < \infty$" used in the next theorem is closely related to the strong ergodicity of the process rather than $\lambda_1 > 0$.

Theorem 2.43. Let $\{X_t\}_{t\geqslant 0}$, L, λ, and g be the same as in the last theorem. Suppose that

(1) $g \in \mathscr{D}_w(L)$,
(2) $\sup_{x\neq y} |g(x) - g(y)| < \infty$.

Then for every coupling $\widetilde{\mathbb{P}}^{x,y}$, we have $\lambda \geqslant \big(\sup_{x\neq y} \widetilde{\mathbb{E}}^{x,y} T\big)^{-1}$.

Proof. Set $f(x,y) = g(x) - g(y)$. By the martingale formulation as in the last proof, we have

$$\begin{aligned} f(x,y) &= \widetilde{\mathbb{E}}^{x,y} f\big(X_{t\wedge T}, Y_{t\wedge T}\big) - \widetilde{\mathbb{E}}^{x,y} \int_0^{t\wedge T} \widetilde{L} f\big(X_s, Y_s\big) \mathrm{d}s \\ &= \widetilde{\mathbb{E}}^{x,y} f\big(X_{t\wedge T}, Y_{t\wedge T}\big) + \lambda \widetilde{\mathbb{E}}^{x,y} \int_0^{t\wedge T} f\big(X_s, X_s\big) \mathrm{d}s. \end{aligned}$$

Hence

$$|g(x) - g(y)| \leqslant \widetilde{\mathbb{E}}^{x,y} \big|g\big(X_{t\wedge T}\big) - g\big(Y_{t\wedge T}\big)\big| + \lambda \widetilde{\mathbb{E}}^{x,y} \int_0^{t\wedge T} \big|g\big(X_s\big) - g\big(Y_s\big)\big| \mathrm{d}s.$$

Assume $\sup_{x\neq y} \widetilde{\mathbb{E}}^{x,y} T < \infty$, and so $\widetilde{\mathbb{P}}^{x,y}[T < \infty] = 1$. Letting $t \uparrow \infty$, we obtain

$$|g(x) - g(y)| \leqslant \lambda \widetilde{\mathbb{E}}^{x,y} \int_0^{T} \big|g\big(X_s\big) - g\big(Y_s\big)\big| \mathrm{d}s.$$

Choose x_n and y_n such that

$$\lim_{n\to\infty} |g(x_n) - g(y_n)| = \sup_{x,\,y} |g(x) - g(y)|.$$

Without loss of generality, assume that $\sup_{x,\,y} |g(x) - g(y)| = 1$. Then

$$1 \leqslant \lambda \varliminf_{n\to\infty} \widetilde{\mathbb{E}}^{x_n, y_n}(T).$$

Therefore $1 \leqslant \lambda \sup_{x\neq y} \widetilde{\mathbb{E}}^{x,y} T$. $\square$

For the remainder of this section, we emphasize the main ideas by using some simple examples. In particular, from now on, the metric is taken to be $\rho(x, y) = |x - y|$. That is, $f(r) = r$. In view of Theorem 2.30, this metric may not be optimal, since $f'' = 0$. Thus, in practice, additional work is often needed to figure out an effective metric ρ. The details will be discussed in the next chapter. Additional discrete examples are included in Appendix B.

To conclude this subsection, let us consider the Ornstein–Uhlenbeck process in $\mathbb{R}^d$. By Theorem 2.30 (4), we have $\overline{L}\rho(x, y) \leqslant -\rho(x, y)$, and so

$$\overline{\mathbb{E}}^{x,y} \rho(X_t, Y_t) \leqslant \rho(x, y) e^{-t}. \tag{2.5}$$

By using Theorem 2.42 with the help of a localizing procedure, this gives us $\lambda_1 \geqslant 1$, which is indeed exact!

Ergodicity

Coupling methods are often used to study the ergodicity of Markov processes. For instance, for an Ornstein–Uhlenbeck process, from (2.5), it follows that

$$W(P(t, x, \cdot), \pi) \leqslant C(x) e^{-t}, \qquad t \geqslant 0, \tag{2.6}$$

where π is the stationary distribution of the process. The estimate (2.6) simply means that the process is exponentially ergodic with respect to W.

Recall that $T = \inf\{t \geqslant 0 : X_t = Y_t\}$. Starting from time T, we can adopt the coupling of marching soldiers so that the two components will move together. Then we have

$$\|P(t, x, \cdot) - P(t, y, \cdot)\|_{\text{Var}} \leqslant 2\, \widetilde{\mathbb{E}}^{x,y} I_{[X_t \neq Y_t]} = 2\, \widetilde{\mathbb{P}}^{x,y}[T > t]. \tag{2.7}$$

Thus, if $\widetilde{\mathbb{P}}^{x,y}[T > t] \to 0$ as $t \to \infty$, then the existence of a stationary distribution plus (2.7) gives us the ergodicity with respect to the total variation. See T. Lindvall (1992) for details and references on this topic. Actually, for Brownian motion, as pointed out in Chen and S.F. Li (1989), coupling by reflection provides a sharp estimate for the total variation. We will come back to this topic in Chapter 5.

Gradient estimate

Recall that for every suitable function f, we have

$$f(x) - f(y) = \widetilde{\mathbb{E}}^{x,y}\big[f\big(X_{t\wedge T}\big) - f\big(Y_{t\wedge T}\big)\big] - \widetilde{\mathbb{E}}^{x,y}\int_0^{t\wedge T}\big[Lf\big(X_s\big) - Lf\big(Y_s\big)\big]\mathrm{d}s.$$

Thus, if f is L-harmonic, i.e., $Lf = 0$, then we have

$$f(x) - f(y) = \widetilde{\mathbb{E}}^{x,y}\big[f\big(X_{t\wedge T}\big) - f\big(Y_{t\wedge T}\big)\big].$$

Hence

$$|f(x) - f(y)| \leqslant 2\,\|f\|_\infty\,\widetilde{\mathbb{P}}^{x,y}[T > t].$$

Letting $t \to \infty$, we obtain

$$|f(x) - f(y)| \leqslant 2\,\|f\|_\infty\,\widetilde{\mathbb{P}}^{x,y}[T = \infty].$$

Now, if f is bounded and $\widetilde{\mathbb{P}}^{x,y}[T = \infty] = 0$, then $f =$constant. Otherwise, if $\widetilde{\mathbb{P}}^{x,y}[T = \infty] \leqslant$constant$\cdot\,\rho(x,y)$, then we get

$$\|\nabla f\|_\infty \leqslant \text{constant}\cdot\|f\|_\infty,$$

which is the gradient estimate we are looking for [cf. M. Cranston (1991; 1992) and F.Y. Wang (1994a; 1994b)]. For Brownian motion in $\mathbb{R}^d$, the optimal coupling gives us $\widetilde{\mathbb{P}}^{x,y}[T < \infty] = 1$, and so $f =$constant. We have thus proved a well-known result: every bounded harmonic function should be constant.

Comparison results

The stochastic order occupies a crucial position in the study of probability theory, since the usual order relation is a fundamental structure in mathematics.

The coupling method provides a natural way to study the order-preserving property (i.e., stochastic comparability). Refer to Chen (1992a, Chapter 5) for a study on jump processes. Here is an example for diffusions.

Example 2.44. Consider two diffusions in $\mathbb{R}$ with

$$a_1(x) = a_2(x) = a(x), \qquad b_1(x) \leqslant b_2(x). \tag{2.8}$$

Then we have $P_1(t) \prec P_2(t)$.

The conclusion was proved in N. Ikeda and S. Watanabe (1988, Section 6.1), using stochastic differential equations. The same proof with a slight modification works if we adopt the coupling of marching soldiers.

A criterion for order preservation for multidimensional diffusion processes was presented in Chen and F.Y. Wang (1993a), from which we see that condition (2.8) is not only sufficient but also necessary. A related topic, the preservation of positive correlations for diffusions, was also solved in the same paper, as mentioned at the beginning of this chapter.

To illustrate an application of the study, let us introduce a simple example.

Example 2.45. Let μ^λ be the Poisson measure on $\mathbb{Z}_+$ with parameter λ:

$$\mu^\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \qquad k \geqslant 0.$$

Then we have $\mu^\lambda \prec \mu^{\lambda'}$ whenever $\lambda \leqslant \lambda'$.

In some publications, one proves such a result by constructing a coupling measure $\tilde{\mu}$ such that $\tilde{\mu}\{(x, y) : x \prec y\} = 1$. Of course, such a proof is lengthy. So we now introduce a very short proof based on the coupling argument.

Consider a birth–death process with rate

$$a(k) \equiv 1, \qquad b^\lambda(k) = \frac{\mu^\lambda(k+1)}{\mu^\lambda(k)} = \frac{\lambda}{k+1} \uparrow \text{ as } \lambda \uparrow .$$

Denote by $P^\lambda(t)$ the corresponding process. It should be clear that

$$P^\lambda(t) \prec P^{\lambda'}(t) \quad \text{whenever} \quad \lambda \leqslant \lambda'$$

[cf. Chen (1992a, Theorem 5.26; Theorem 5.41 in the 2nd edition)]. Then, by the ergodic theorem,

$$\mu^\lambda(f) = \lim_{t\to\infty} P^\lambda(t)f \leqslant \lim_{t\to\infty} P^{\lambda'}(t)f = \mu^{\lambda'}(f)$$

for all $f \in \mathscr{M}$. Clearly, the technique using stochastic processes [goes back to R. Holley (1974)] provides an intrinsic insight into order preservation for probability measures.

We now return to the FKG inequality mentioned at the beginning of this chapter. Clearly, the inequality is meaningful in the higher-dimensional space $\mathbb{R}^d$ with respect to the ordinary partial ordering. The inequality for a Markov semigroup $P(t)$ becomes

$$P(t)(fg) \geqslant P(t)fP(t)g, \qquad t \geqslant 0,\ f, g \in \mathscr{M}.$$

The study of the FKG inequality in terms of semigroups is exactly the same as above. Choose a Markov process having the given measure as a stationary distribution. Then, study the inequality for the dynamics. Finally, passing to the limit as $t \to \infty$, we return to (2.1).

An aspect of the applications of coupling methods is to compare a rather complicated process with a simpler one. To provide an impression, we introduce an example that was used by Chen and Y.G. Lu (1990) in the study of large deviations for Markov chains.

Example 2.46. Consider a single birth Q-matrix $Q = (q_{ij})$, which means that

$$q_{i,i+1} > 0 \qquad \text{and} \quad q_{ij} = 0 \quad \text{for all } j > i+1,$$

and a birth–death Q-matrix $\overline{Q} = (\bar{q}_{ij})$ with $\bar{q}_{i,i-1} = \sum_{j<i} q_{ij}$. If $\bar{q}_{i,i+1} \geqslant q_{i,i+1}$ for all $i \geqslant 0$. Then $P(t) \prec \overline{P}(t)$.

The conclusion can be easily deduced by the following coupling:

$$\begin{array}{llll}(i_1, i_2) & \to & (i_1 - k, i_2 - 1) & \text{at rate} \quad q_{i_1, i_1 - k} \wedge q_{i_2, i_2 - k} \\ & \to & (i_1 - k, i_2) & \text{at rate} \quad (q_{i_1, i_1 - k} - q_{i_2, i_2 - k})^+ \\ & \to & (i_1, i_2 - 1) & \text{at rate} \quad (q_{i_2, i_2 - k} - q_{i_1, i_1 - k})^+ \\ & \to & (i_1 + 1, i_2 + 1) & \text{at rate} \quad q_{i_1, i_1 + 1} \wedge \bar{q}_{i_2, i_2 + 1} \\ & \to & (i_1 + 1, i_2) & \text{at rate} \quad (q_{i_1, i_1 + 1} - \bar{q}_{i_2, i_2 + 1})^+ \\ & \to & (i_1, i_2 + 1) & \text{at rate} \quad (\bar{q}_{i_2, i_2 + 1} - q_{i_1, i_1 + 1})^+,\end{array}$$

where we have used the convention $q_{ij} = 0$ if $j < 0$. Refer to Chen (1992a, Theorem 8.24) for details. This example illustrates the flexibility in the application of couplings.

The details of this chapter, except for diffusions, are included in Chapter 5 of the second edition of Chen (1992a).

Finally, we mention that the coupling methods are also powerful for time-inhomogeneous Markov processes, not touched on in this book. In fact, the fundamental theorem 2.14 is valid for Markov jump processes valued in Polish spaces [cf. J.L. Zheng (1993)]. For estimation of convergence rate, refer to A.I. Zeifman (1995), B.L. Granovsky and A.I. Zeifman (1997).