Markov Processes and Markov Chain Monte Carlo

Zhang Ruiyang

Contents

Chapter 1

Introduction

Markov process is a class of stochastic processes, usually in continuous time, that satisfies the Markov property. A stochastic process is a sequence of random variables indexed by time, and here we will usually denote it as $\{X_t\}_{t\in\mathbb{R}^+}$. The subscript t denotes time, and if we are working in discrete time we will usually use the letter n instead to denote the step number. The Markov property states that the behaviour of the future, condition on the current behaviour, will be independent of the past. A more mathematical formulation of these things will appear later this chapter.

In the rest of this chapter, we will go over some background material on functional analysis, Markov process, and Kolmogorov equations as well as the Fokker-Planck equation. They will prepare ourselves with the discussions in the following chapters.

In Chapter 2, we will go over three basic types of Markov processes, i.e. the deterministic process, the jump process, and the diffusion process. A lot of the common Markov processes can be constructed using some of the three types. We will study some properties of these processes, such as their generator and their invariant. We will also mention two composite processes - the Langevin diffusion and the piecewise deterministic Markov process. These two processes are of special interest to us as they can be used to be the underlying Markov chain of some Markov chain Monte Carlo algorithms, which is the topic of the next chapter.

In Chapter 3, we will study the Markov chain Monte Carlo algorithms. Two classes of algorithms will be highlighted. The first class is algorithms using Langevin diffusions, while the second class is algorithms using piecewise deterministic Markov processes. We will discuss the design of these algorithms, and mention some basic theoretical properties of them. The first class of algorithms has been studied for decades, whereas the second class are only introduced fairly recently.

1.1 Functional Analysis Basics

In this section, we will present superficially some definitions and facts about basic functional analysis and operator theory to straighten out the notations. A more comprehensive discussion of functional analysis can be found in Conway (2019).

Let X be a vector space over $\mathbb{R}$.

Definition 1.1. *A **norm** is a function* $\|\cdot\| : X \to [0,\infty)$ *such that*

1. $\|x\| = 0 \iff x = 0$
2. $\|\lambda x\| = |\lambda|\|x\|$ *for all* $\lambda \in \mathbb{R}$ *and* $x \in X$
3. $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in X$.

Given this norm, we can then induce a ***metric*** d *if we let* $d(x, y) := \|x - y\|$ *for all* $x, y \in X$.

An important property that we would like to have is completeness. A normed space (i.e. a vector space equipped with a norm) is **complete** if every Cauchy sequence converges.

Definition 1.2. *A* ***Banach space*** $(X, \|\cdot\|)$ *is a normed space that becomes complete with respect to the induced metric* d.

Definition 1.3. *A vector space* X *is an* ***inner product space*** *if there exists a function, known as the* ***inner product****,* $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ *such that*

1. $\langle x, x \rangle \geq 0$ *for all* x *in*X
2. $\langle x, x \rangle = 0 \iff x = 0$
3. $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ *for all* $\lambda, \mu \in \mathbb{R}$ *and* $x, y, z \in X$.
4. $\langle x, y \rangle = \langle y, x \rangle$ *for all* $x, y \in X$.

Given an inner product $\langle \cdot, \cdot \rangle$ we can induce a norm $\|\cdot\|$ by setting $\|x\| = \sqrt{\langle x, x \rangle}$. To verify this function is indeed a norm, we would need to use the **Cauchy-Schwarz-Buniakovski inequality**, which states that $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$ for all $x, y \in X$.

A space is **complete** if every Cauchy sequence in this space converges. A normed space that is complete is known as the Banach space. Now that we have an inner product space, we would have the following definition.

Definition 1.4. *An inner product space* $(X, \langle \cdot, \cdot \rangle)$ *is a* ***Hilbert space*** *if it is complete in the induced norm* $\|\cdot\|$.

We will use H to denote an arbitrary Hilbert space. Among them, there are two Hilbert spaces that we will be using extensively in the following. The first is the space $L^2(\mathbb{R})$, defined by

$$L^2(\mathbb{R}) := \left\{ \text{measurable } f : \mathbb{R} \to \mathbb{R} \;\middle|\; \int f^2(x)dx < \infty \right\}$$

and it is equipped with inner product $\langle f, g \rangle = \int_{\mathbb{R}} f(x)g(x)dx$. The second is the space $L^2(\pi)$ for some (probability) measure π, defined by

$$L^2(\pi) := \left\{ \text{measurable } f : \mathbb{R} \to \mathbb{R} \;\middle|\; \int f^2 d\pi < \infty \right\}$$

and it is equipped with inner product $\langle f, g \rangle_\mu = \int_{\mathbb{R}} f g d\mu$.

Theorem 1.5. *Let* X, Y *be normed spaces with norms* $\|\cdot\|_X$ *and* $\|\cdot\|_Y$ *respectively, and* $A : X \to Y$. *The following are equivalent:*

1. A *is continuous*
2. A *is continuous at any point*
3. A *is continuous at* 0
4. *there exists constant* $c > 0$ *such that* $\|Ax\|_Y \leq c\|x\|_X$ *for all* $x \in X$.

We would also say A *is a* ***bounded*** *operator.*

For a space X, we will use $B(X)$ to denote the set of bounded operators $A : X \to X$. The last formulation in the above theorem helps us to define the norm of an operator.

Definition 1.6. *For a bounded operator A, we can define $\|A\|$ in the following equivalent ways:*

1. $\inf\{c > 0 \mid \|Ax\|_Y \leq c\|x\|_X \text{ for all } x \in X\}$
2. $\inf\{c > 0 \mid \|Ax\|_Y \leq c \text{ for all possible } \|x\| \leq 1\}$
3. $\sup_{x \in X, x \neq 0} \frac{\|Ax\|_Y}{\|x\|_X}$
4. $\sup_{x \in X, x \neq 0, \|x\| \leq 1} \|Ax\|_Y$.

In particular, we have $\|Ax\| \leq \|A\| \cdot \|x\|$ for all $x \in X$.

Definition 1.7. *Let $A \in B(H)$. Then there exists unique operator $A^* \in B(H)$ such that $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for all $x, y \in H$. We will call A^* the **adjoint** of A. An operator $A \in B(H)$ is called **self-adjoint** or **symmetric** if $A = A^*$. Note that self-adjoint and symmetric are possibly different when the operator is unbounded.*

1.2 Markov Processes and Semigroups

A Markov process is any stochastic process $\{X_t\}$ that satisfies the Markov property.

Definition 1.8. *A stochastic process $\{X_t\}$ is a **Markov process** if the increment $X_t - X_s$ (assuming $t > s$) is independent of X_u for all $u < s$.*

Each element of the sequence $\{X_t\}$ is a random variable. A random variable, using the language of measure-theoretic probability theory, is a real-valued measurable function defined on the sample space. Here, each X_t is a measurable function that maps from the sample space to the state space $\mathbb{R}$. Notice that we can also call a Markov process as a **Markov chain**, which is a broader term if we use it loosely, and a narrower term (a discrete-time Markov process with finite state space) if we use it specifically. In this note, we will use the terms 'Markov process' and 'Markov chain' interchangeably.

We are interested in the behaviour of the process over time.

Definition 1.9. *The **transition function** $p(B, t|x, s)$, with $t \geq s$, of the Markov process $\{X_t\}$ is defined as*

$$p(B, t|x, s) := \mathbb{P}(X_t \in B | X_s = x)$$

*where B is a Borel set of $\mathbb{R}$. This function with varying B is a probability measure, and it with varying x is a measurable function. Furthermore, the transition function is called a **transition kernel** if it admits a density, i.e.*

$$p(B, t|x, s) = \int_B p(dy, t|x, s) = \int_B p(y, t|x, s)\ dy.$$

An important property of the transition function is that it can be broken down into steps. Intuitively, assuming we have timestamps $a \leq b \leq c$, the movement from X_a to X_c should be the same as the combination of movements from X_a to all possible intermediate steps X_b then to X_c. Mathematically, we have

$$p(B, t|x, s) = \int_{\mathbb{R}} p(dy, u|x, s) \cdot p(B, t|dy, u), \qquad s \leq u \leq t$$

and this is known as the **Chapman-Kolmogorov equation**.

Definition 1.10. *A stochastic process $\{X_t\}$ is called **time-homogeneous**, or simply **homogeneous**, if the transition function depends only on the difference in time, i.e. $p(B, t|x, s) = p(B, t+k|x, s+k)$ for any constant k. A process without this property is called **time-inhomogeneous**.*

For a homogeneous Markov process, we will denote its transition kernel simply as $p(t, x, B)$, where

$$p(t, x, B) := p(B, t|x, 0) = p(B, t + k|x, k)$$

for any $t, k \geq 0$.

The movement of a homogeneous Markov process over time can be viewed as an operator. For $t \geq 0$, we define the operator P_t on some nice class (to be specified later) of functions f as

$$P_t f(x) := \mathbb{E}_x[f(X_t)] = \mathbb{E}[f(X_t)|X_0 = x] = \int_{\mathbb{R}} f(y)p(t, x, dy)$$

Clearly, using this definition, we can notice that $P_0 = Id$. Some more, we have

$$\|P_t f\| = \left\| \int_{\mathbb{R}} f(y)p(t, x, dy) \right\| \leq \|f\| \left\| \int_{\mathbb{R}} f(y)p(t, x, dy) \right\| = \|f\|$$

so $\|P_t\| \leq 1$ and this operator is a contraction. Additionally, we notice that $P_s \circ P_t = P_{s+t}$, using the Chapman-Kolmogorov Equation. To see this, for any suitable f and x, we have

$$\begin{aligned}
P_s \circ P_t f(x) &= P_s \int_{y\in\mathbb{R}} f(y)p(t, x, dy) \\
&= \int_{z\in\mathbb{R}} p(s, x, dz) \int_{y\in\mathbb{R}} f(y)p(t, z, dy) \\
&= \int_{z\in\mathbb{R}} \int_{y\in\mathbb{R}} p(s, x, dz)f(y)p(t, z, dy) \\
&= \int_{y\in\mathbb{R}} \int_{z\in\mathbb{R}} f(y)p(s, x, dz)p(t, z, dy) \\
&= \int_{y\in\mathbb{R}} f(y) \int_{z\in\mathbb{R}} p(s, x, dz)p(t, z, dy) \\
&= \int_{y\in\mathbb{R}} f(y)p(t + s, x, dy) = P_{t+s} f(x).
\end{aligned}$$

The Markov process might begin behaving strangely, yet after some time it might start to have some nice pattern and become stabilised.

Definition 1.11. *Given a probability measure μ and a Markov process $\{P_t\}$, we say μ is **invariant** for $\{P_t\}$ if for any positive measurable function f and $t \geq 0$, we have*

$$\int P_t f d\mu = \int f d\mu.$$

*Then, we would also say $\{P_t\}$ is μ-**invariant**.*

If we have a Markov process $\{X_t\}$ starting at invariant distribution μ, we will always be following this distribution as time goes on. To see this, we have

$$\begin{aligned}
\mathbb{E}[f(X_t)] &= \mathbb{E}[\mathbb{E}_x[f(X_t)]] = \mathbb{E}[\mathbb{E}[f(X_t)|X_0 = x]] \\
&= \mathbb{E}[P_t f(x)] = \int P_t f(x) d\mu \\
&= \int f(x) d\mu = \mathbb{E}[f(X_0)].
\end{aligned}$$

These properties give $\{P_t\}$ a nice structure. In fact, these (along with a few others) imply that $\{P_t\}$ is a **Markov semigroup** (Bakry et al.; 2014).

Definition 1.12. *A family of operators $\{P_t\}_{t\in\mathbb{R}_+}$ defined on the bounded measurable functions on a state space $(E,\mathcal{F})$ with stationary measure μ is called a **Markov semigroup**. It satisfies the following properties: for all $p \geq 1$, $t,s \in \mathbb{R}_+$, $\alpha,\beta \in \mathbb{R}$, bounded measurable functions f,g, we have*

1. *For any t, P_t sends bounded measurable functions to bounded measurable functions.*
2. *(initial) $P_0 = Id$.*
3. *(conservation) $P_t 1 = 1$ almost surely.*
4. *(contraction) $\|P_t f\|_p \leq \|f\|_p$*
5. *(linearity) $P_t(\alpha f + \beta g) = \alpha P_t f + \beta P_t g$ almost surely*
6. *(semigroup) $P_{t+s} f = P_t P_s f$ almost surely.*
7. *(continuity) For every $f \in L^2(\mu)$, $P_t f$ converges to f in this space as $t \to 0$.*

Given a semigroup, we can define its generator, which is an important concept and its significance will be revealed later on.

Definition 1.13. *The **(infinitesimal) generator** $\mathcal{L}$ is defined as*

$$\mathcal{L}f := \lim_{t\to 0^+} \frac{P_t f - f}{t}$$

*for every $f \in L^2(\mu)$ for which the above limit exists in $L^2(\mu)$. The set of f for which this limit is well defined is called the **domain** of $\mathcal{L}$, or $Dom(\mathcal{L})$. So, $\mathcal{L}$ defines a linear operator from $Dom(\mathcal{L}) \subseteq L^2(\mu)$ to $L^2(\mu)$.*

Remark. *The domain is usually not the whole space. This is a technical detail that we will not worry about in this notes. Usually, any domain issue only requires some additional technicalities to be remedied.*

1.3 Kolmogorov Equations and Fokker-Planck Equation

Given an operator and the space it acts on, we can define its adjoint. Consider a Markov semigroup $\{P_t\}$ with generator $\mathcal{L}$, we can define the adjoint of the generator in $L^2(\mathbb{R})$ as, for any suitable $f,g \in L^2(\mathbb{R})$,

$$\langle \mathcal{L}f, g\rangle = \int \mathcal{L}f(x)g(x)\,dx = \int f(x)\mathcal{L}^* g(x)\,dx = \langle f, \mathcal{L}^* g\rangle.$$

Additionally, we have the transition kernel $p(t,x,B)$ for a time-homogeneous Markov process.

The **Kolmogorov Forward Equation** is, for any suitable function f,

$$\frac{d}{dt}P_t f = P_t \mathcal{L} f, \tag{1}$$

where as the **Kolmogorov Backward Equation** is

$$\frac{d}{dt}P_t f = \mathcal{L} P_t f. \tag{2}$$

The **Fokker-Planck Equation** is

$$\frac{\partial}{\partial t} p_t = \mathcal{L}^* p_t. \tag{3}$$

where p_t is the transition kernel $p(t, x, B)$.

These three (in fact two of them are equivalent) equations allow us to study the evolution over time of the Markov process. These partial differential equations have a profound impact on various fields, such as engineering and physics. There are also extensive studies of the Fokker-Planck equation, see the monograph Risken (1996).

It can be shown that the Fokker-Planck equation is equivalent to the Kolmogorov forward equation. Because of this, we would normally not distinguish the two. We will prove this for the time-homogeneous case.

Proposition 1.14. *The Fokker-Planck equation is equivalent to the Kolmogorov forward equation.*

Proof. Consider a time-homogeneous Markov process $\{X_t\}$ with associated Markov semigroup $\{P_t\}$ and generator $\mathcal{L}$. For suitable function f, we have operator $P_t f$ defined as

$$P_t f = \mathbb{E}[f(X_t)]$$

and

$$P_t f(x) = \mathbb{E}[f(X_t)|X_0 = x].$$

We will use p_t to denote the transition density of $\{X_t\}$. Also, $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$ in $L^2(\mathbb{R})$, so

$$\langle \mathcal{L}f, g\rangle = \int \mathcal{L}f(x)g(x)\ dx = \int f(x)\mathcal{L}^* g(x)\ dx = \langle f, \mathcal{L}^* g\rangle.$$

We have, using the Kolmogorov forward equation,

$$\frac{d}{dt}P_t f = P_t \mathcal{L} f = \mathbb{E}[\mathcal{L}f(X_t)] = \int \mathcal{L}f(x)p_t(x)\ dx = \int \mathcal{L}^* p_t(x) f(x)\ dx,$$
$$\frac{d}{dt}P_t f = \frac{d}{dt}\mathbb{E}[f(X_t)] = \frac{d}{dt}\int f(x)p_t(x)\ dx = \int \frac{\partial}{\partial t}p_t(x) f(x)\ dx.$$

Combining the two yields

$$\int \mathcal{L}^* p_t(x) f(x)\ dx = \int \frac{\partial}{\partial t}p_t(x) f(x)\ dx \iff \int [\mathcal{L}^* p_t(x) - \frac{\partial}{\partial t}p_t(x)] f(x)\ dx = 0$$

which holds for any suitable f. We claim that there would be a sufficiently large class of such functions for the derivation to hold. This then implies

$$\mathcal{L}^* p_t(x) - \frac{\partial}{\partial t}p_t(x) = 0 \iff \mathcal{L}^* p_t(x) = \frac{\partial}{\partial t}p_t(x),$$

which is the Fokker-Planck equation.

Thus, we have established the equivalence of the two equations, as each of the above steps is equivalence. □

We will revisit the Fokker-Planck equation when we discuss the diffusion process in the next chapter.

Chapter 2

Markov Processes

2.1 Deterministic Process

A deterministic process is a Markov process with its path fully characterised by an ordinary differential equation (ODE) and some initial condition. If we denote a deterministic process by $\{X_t\}$, we have

$$\frac{d}{dt}x(t) = h(x(t))$$

for all $t \geq 0$, and $X_0 = x_0$. The stationary measure of this process is denoted as μ.

First, we should notice that this process satisfies the Markov property. Given $\{X_r\}_{r \leq s}$, we have

$$X_t = X_s + \int_s^t h(x(u))\ du$$

for all $t \geq s$, and this quantity only depends on the information at $t = s$ and is independent of the past, as desired. The path of this Markov process is simply smooth, as shown in Figure 1 below.

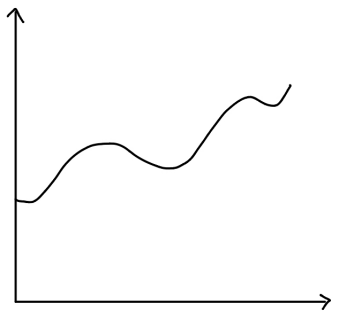

Figure 1: Path of a Deterministic Process

Based on the construction of $\{X_t\}$, we can define the operator P_t, which is, for some $f \in L^2(\mu)$,

$$P_t f(x) = f(x) + \int_0^t \frac{df}{ds}(x(s))\, ds, \qquad \frac{df(x(s))}{ds} = \frac{df}{dx} \cdot \frac{dx(s)}{ds}.$$

So, we can apply the definition of the generator to P_t, and get

$$\begin{aligned}
\mathcal{L}f(x) &= \lim_{t \to 0^+} \frac{P_t f(x) - f(x)}{t} \\
&= \lim_{t \to 0^+} \frac{1}{t} \int_0^t \frac{df}{ds}(x(s))\, ds \\
&= \frac{df}{dt}(x(t)) \\
&= \frac{df}{dx} \cdot \frac{dx(t)}{dt}.
\end{aligned}$$

Notice that the generator $\mathcal{L}$ is not well-defined for any $f \in L^2(\mu)$. Anything that is in $L^2(\mu)$ but not differentiable would be such a case, and an example would be $f(x) = \sin(1/x^2)$. This means, $\mathrm{Dom}(\mathcal{L}) \neq L^2(\mu)$.

Next, we will derive the adjoint $\mathcal{L}^*$ of $\mathcal{L}$ in $L^2(\mathbb{R})$. We have, for suitable[1] f, g with compact supports (so they decay to zero at infinities),

$$\begin{aligned}
\langle \mathcal{L}f, g \rangle &= \int \mathcal{L}f(x) g(x) dx \\
&= \frac{dx(t)}{dt} \cdot \frac{df}{dx} g(x) dx \\
&= \int g(x) h(x) \cdot \nabla f(x) dx \\
&= [ghf]_{-\infty}^{\infty} - \int f(x) \nabla [g(x)h(x)] dx \\
&= \int f(x) \nabla [-g(x)h(x)] dx = \langle f, \mathcal{L}^* g \rangle
\end{aligned}$$

so the adjoint is $\mathcal{L}^* g(x) = \nabla[-g(x)h(x)]$.

2.2 Jump Process

A jump process is a Markov process that stays constant when there is no event, and takes a "jump" (change in value) at event time T, which is an exponentially distributed random variable. To be more precise, here is how we would simulate one jump of a jump process with stationary measure μ, given a rate function $\lambda(\cdot)$, a transition kernel $J(\cdot, \cdot)$, and an initial distribution f.

1. Draw initial state $X_0 \sim f$,
2. Denote the current state by x.
3. Draw a time $T \sim \mathrm{Exp}[\lambda(x)]$.
4. Draw $X_{0+T} \sim J(x, \cdot)$.

[1] We will not worry about the existence of such suitable functions, as this could be verified but requires various technical results. We will hand-wave at such instances many times later too.

5. Set $X_t = x$ for all $t < T$.

The path to this process is, as one can infer from the construction, piecewise constant. The path will consist of pieces of right-continuous constants, as shown in Figure 2 below.

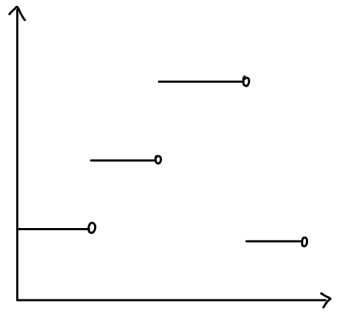

Figure 2: Path of a Jump Process

Based on the construction of $\{X_t\}$, we can derive its operator P_t for some $f \in L^2(\mu)$. Consider some small $\delta > 0$, we have

$$P_\delta f(x) = \begin{cases} \int f(y) J(x, dy) & \text{with probability } \delta\lambda(x) + o(\delta) \\ f(x) & \text{with probability } 1 - \delta\lambda(x) + o(\delta) \end{cases}$$

since at most one event can happen during a small enough interval (Norris; 1998). Note that here we say $f(x) = o(x)$ if $\lim_{x \to 0} f(x)/x = 0$. The first case above happens when $T \leq \delta$, so we have

$$\mathbb{P}(T \leq \delta) = \int_0^\delta \lambda(x) \exp[-\lambda(x)t] \, dt = 1 - \exp[-\lambda(x)\delta].$$

Using Taylor expansion, we have

$$1 - \exp[-\lambda(x)\delta] = 1 - \sum_{n=0}^{\infty} \frac{(-\lambda(x)\delta)^n}{n!} = \delta\lambda(x) + o(\delta).$$

The probability for the second case is

$$1 - [1 - \exp[-\lambda(x)\delta]] = \exp[-\lambda(x)\delta] = 1 - \delta\lambda(x) + o(\delta).$$

So, combining the two cases gives us

$$P_\delta f(x) = [1 - \delta\lambda(x) + o(\delta)] f(x) + [\delta\lambda(x) + o(\delta)] \int f(y) J(x, dy).$$

Using the operator P_δ, we can compute the generator of this semigroup.

$$\begin{aligned}
\mathcal{L}f(x) &= \lim_{\delta\to 0^+} \frac{P_\delta f(x) - f(x)}{\delta} \\
&= \lim_{\delta\to 0^+} \frac{1}{\delta}\left[[1-\delta\lambda(x)+o(\delta)]f(x) + [\delta\lambda(x)+o(\delta)]\int f(y)J(x,dy) - f(x)\right] \\
&= \lim_{\delta\to 0^+} \frac{1}{\delta}\left[[1-\delta\lambda(x)]f(x) + [\delta\lambda(x)]\int f(y)J(x,dy) - f(x)\right] \\
&= \lim_{\delta\to 0^+} \frac{1}{\delta}\left[-\delta\lambda(x)f(x) + \delta\lambda(x)\int f(y)J(x,dy)\right] \\
&= -\lambda(x)f(x) + \lambda(x)\int f(y)J(x,dy) \\
&= \lambda(x)\int [f(y)-f(x)]J(x,dy)
\end{aligned}$$

where the last equality is due to the fact that $\int J(x,dy) = 1$.

Next, we will derive the adjoint $\mathcal{L}^*$ of $\mathcal{L}$ in $L^2(\mathbb{R})$. We will assume that $J(x,dy) = j(x,y)dy$. We have, for suitably smooth f, g,

$$\begin{aligned}
\langle \mathcal{L}f, g\rangle &= \int_{x\in\mathbb{R}} \mathcal{L}f(x)g(x)dx \\
&= \int_{x\in\mathbb{R}} \lambda(x)\int_{y\in\mathbb{R}} [f(y)-f(x)]j(x,y)dyg(x)dx \\
&= \int_{x\in\mathbb{R}} \lambda(x)g(x)\int_{y\in\mathbb{R}} [f(y)-f(x)]j(x,y)dydx \\
&= \int_{x\in\mathbb{R}} \lambda(x)g(x)\int_{y\in\mathbb{R}} f(y)j(x,y)dydx - \int_{x\in\mathbb{R}} \lambda(x)g(x)\int_{y\in\mathbb{R}} f(x)j(x,y)dydx \\
&= \int_{y\in\mathbb{R}}\int_{x\in\mathbb{R}} \lambda(y)g(y)f(x)j(y,x)dxdy - \int_{x\in\mathbb{R}} \lambda(x)g(x)f(x)\int_{y\in\mathbb{R}} j(x,y)dydx \\
&= \int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}} \lambda(y)g(y)f(x)j(y,x)dxdy - \int_{x\in\mathbb{R}} \lambda(x)g(x)f(x)dx \\
&= \int_{x\in\mathbb{R}} f(x)\int_{y\in\mathbb{R}} \lambda(y)g(y)j(y,x)dydx - \int_{x\in\mathbb{R}} f(x)\lambda(x)g(x)dx \\
&= \int_{x\in\mathbb{R}} f(x)\left[\int_{y\in\mathbb{R}} \lambda(y)g(y)j(y,x)dy - \lambda(x)g(x)\right]dy = \langle f, \mathcal{L}^*g\rangle.
\end{aligned}$$

So, the adjoint $\mathcal{L}^*g(x) = \int_{y\in\mathbb{R}} \lambda(y)g(y)j(y,x)dy - \lambda(x)g(x)$.

2.3 Diffusion Process

Studying the diffusion process is slightly more involved than the previous two. It requires some knowledge of Brownian motion, Itô integral and lemma, as well as stochastic differential equation (SDE). We will cover the essentials of these topics in order to derive the generator of a one-dimensional SDE, which is more general than a diffusion process. For a more comprehensive study of the material in this section, we direct the readers to Øksendal (2013) and E et al. (2021). Many definitions and results presented below follow these two texts quite closely.

2.3.1 Brownian Motion

First, let us define what a Brownian motion, or Wiener process, is.

Definition 2.1. *A stochastic process $\{B_t\}$ is called a **Brownian motion**, or a **Wiener process**, if it satisfies the following properties:*

1. $B_0 = 0$
2. $B_t - B_s \sim N(0, t-s)$ *for all* $0 \leq s < t$.
3. $\{B_t\}$ *has independent increment, i.e. for* $0 \leq s_1 < t_1 \leq s_2 < t_2$, $B_{t_2} - B_{s_2}$ *and* $B_{t_1} - B_{s_1}$ *are independent random variables.*
4. *The graph of the process is continuous almost everywhere, but differentiable almost nowhere.*

A Brownian motion is a scaling limit of random walk. Consider i.i.d. random variables $X_1, X_2, \ldots \sim \alpha N(0,1) = N(0, \alpha^2)$, and define

$$S_n = \sum_{i=1}^{n} X_i \sim N(0, n\alpha^2).$$

If we let $h := \alpha^2$ and define time $t_n := nh = n\alpha^2$, and then define $\tilde{B}(t_n) = S_n$, we will have the following properties about $\tilde{B}_t$, i.e.

- $\tilde{B}(t_n) \sim N(0, t_n)$
- $\tilde{B}(t_n) - \tilde{B}(s_n) = \sum_{i=s_n/h+1}^{t_n/h} X_i \sim N(0, t_n - s_n)$ where $0 \leq s_n < t_n$.

Now, if we take $h \to 0$ while keeping $t := t_n$ as a constant (so we are taking the step size to zero while the step number to infinity), we will get the desired Brownian motion by

$$B(t) = \lim_{h \to 0} \tilde{B}(t_n = nh = t).$$

This construction is justified by the Donsker's theorem (Karatzas et al.; 1991).

Notice that for some constant $c > 0$, if we defined a new stochastic process $\{X_t\}$ by

$$X_t = \frac{1}{\sqrt{c}} B_{ct},$$

then it can be shown that $\{X_t\}$ is a Brownian motion as well. This means, the Brownian motion is **self-similar**, so it is impossible to plot the trajectory of a Brownian motion exactly. Instead, we will always plot $\tilde{B}$ instead, for some suitably small h. This issue will pop up once again in the next chapter.

A Brownian motion has some nice properties.

1. $\mathbb{E}(B_t) = 0$
2. $\mathbb{E}(B_t^2) = t$
3. $\mathbb{E}(B_s B_t) = \min(s,t) = s \wedge t$

The first two properties can be easily obtained by realising that $B_t - B_0 = B_t \sim N(0, t-0) = N(0,t)$ for any $t > 0$. The third property is derived as follows: WLOG assume $s < t$, then

$$\mathbb{E}(B_s B_t) = \mathbb{E}[(B_t - B_s)B_s + B_s^2] = \mathbb{E}[(B_t - B_s)B_s] + \mathbb{E}[B_s^2] = \mathbb{E}[(B_t - B_s)]\mathbb{E}[B_s] + s = s.$$

2.3.2 Itô Calculus

A standard integral will have the form:

$$\int f(x)\ dx.$$

A stochastic integral, however, will have the following form:

$$\int f(B_s, s)\ dB_s.$$

A standard integral can be calculated by Riemann sum while f is Riemann integrable. For stochastic integrals, we will mimic the Riemann sum. One major difference between stochastic integrals and standard integrals is that the choice of endpoints of the Riemann sum does not affect the value of the integral, whereas the choice of endpoints will affect the eventual value for the case of stochastic integral.

The Itô integral is defined as follows:

$$\int_0^t f(B_s, s)\ dB_s = \lim_{h\to 0} \sum_{i=1}^{N} f(B_{t_i}, i)(B_{t_{i+1}} - B_{t_i})$$

where $t_i = (i-1)h$ and $N = t/h$.

According to the definition of Itô integral, we can derive that

$$\int_0^t f(B_s, s)\ (dB_s)^2 = \int_0^t f(B_s, s)\ ds.$$

The derivation of the above equality is slightly technical yet standard, so we will omit it here. It can be viewed as a corollary of the Itô isometry. This equality can be expressed as $(dB_t)^2 = dt$, and we also have $dB_t dt = dt^2 = 0$. The derivation of the second equality is also omitted here, though it can be derived similarly to the first one. To summarise, these two results, repeated once again below, are known as the **Itô increment rule**.

- $(dB_t)^2 = dt$
- $dB_t dt = dt^2 = 0$

With these, we can obtain the following important result.

Lemma 2.2 (Itô Lemma)**.** *For some suitable function $f(B_t, t)$, we have $df = (f_t + f_{xx}/2)dt + f_x dB_t$.*

Proof. Using the Taylor expansion, we have

$$\begin{aligned} f(B_{t+dt}, t+dt) = f(B_t, t) &+ \frac{\partial}{\partial x} f(B_t, t)\ dB_t + \frac{\partial}{\partial t} f(B_t, t)\ dt + \\ &\frac{1}{2}\left(\frac{\partial^2}{\partial x^2} f(B_t, t)\ (dB_t)^2 + \frac{\partial^2}{\partial x \partial t} f(B_t, t)\ dB_t dt + \frac{\partial}{\partial t} f(B_t, t)\ dt^2\right) + \cdots \\ &= f(B_t, t) + \frac{\partial}{\partial x} f(B_t, t)\ dB_t + \frac{\partial}{\partial t} f(B_t, t)\ dt + \frac{1}{2}\frac{\partial^2}{\partial x^2} f(B_t, t)\ dt, \end{aligned}$$

which after some rearranging will yield

$$\begin{aligned} df &= f(B_{t+dt}, t+dt) - f(B_t, t) \\ &= \frac{\partial}{\partial x} f(B_t, t)\ dB_t + \frac{\partial}{\partial t} f(B_t, t)\ dt + \frac{1}{2}\frac{\partial^2}{\partial x^2} f(B_t, t)\ dt \\ &= \left(f_t + \frac{1}{2} f_{xx} \right) dt + f_x dB_t. \end{aligned}$$

□

We will show without proof the result on the expectation of stochastic integrals. This result can be obtained by taking the expectation of the Itô integral directly.

Proposition 2.3. *For any suitably smooth function f, we have*

$$\mathbb{E}\left[\int_0^t f(B_s, s) dB_s \right] = 0.$$

2.3.3 Stochastic Differential Equation

A stochastic differential equation (SDE) is a stochastically perturbed differential equation. A one-dimensional SDE has the following general form:

$$dX_t = a(X_t, t)dt + b(X_t, t)dB_t$$

where $a(X_t, t)$ is the **drift** of the process, and $b^2(X_t, t)$ is the **diffusion** of the process. This characterises the stochastic process $\{X_t\}$. The above form is equivalent to the following:

$$X_t = X_0 + \int_0^t a(X_s, s)ds + \int_0^t b(X_s, s)dB_s.$$

If we consider $X_t = f(X_t, t)$, so $f_t = 0$, then we can obtain the Itô lemma for SDE.

$$\begin{aligned} f(X_{t+dt}, t+dt) = f(X_t, t) &+ \frac{\partial}{\partial x} f(X_t, t)\ dX_t + \frac{\partial}{\partial t} f(X_t, t)\ dt \\ &+ \frac{1}{2}\left(\frac{\partial^2}{\partial x^2} f(X_t, t)\ (dX_t)^2 + \frac{\partial^2}{\partial x \partial t} f(X_t, t)\ dX_t dt + \frac{\partial}{\partial t} f(X_t, t)\ dt^2 \right) \\ &+ \cdots \end{aligned}$$

According to the statement of the SDE, we have

$$dX_t dt = a(X_t, t)dt^2 + b(X_t, t)dB_t dt = 0$$

according to the Itô increment rule. Also, based on the definition, we have

$$\begin{aligned} (dX_t)^2 &= (a(X_t, t)dt + b(X_t, t)dB_t)^2 \\ &= a(X_t, t)^2 dt^2 + b(X_t, t)^2 (dB_t)^2 + 2a(X_t, t)b(X_t, t)dt dB_t \\ &= b(X_t, t)^2 dt \end{aligned}$$

by the Itô increment rule. Using these results, we can obtain the desired Itô lemma for SDE

$$\begin{aligned} df &= f_x dX_t + f_t dt + 1/2 f_{xx} (dX_t)^2, \\ df &= (f_t + a f_x + \frac{1}{2} b^2 f_{xx})dt + b f_x dB_t. \end{aligned}$$

We have established sufficient background material to obtain the generator of the stochastic process $\{X_t\}$. For some suitable function $g(x)$, we know that $g_t = 0$ and we have

$$\begin{aligned}
\mathcal{L}g(x) &= \lim_{s\to 0} \frac{1}{s} \left[\mathbb{E}[g(X_{t+s}) \mid X_t = x] - g(x)\right] \\
&= \lim_{s\to 0} \frac{1}{s} \mathbb{E}[g(X_{t+s}) - g(X_t) \mid X_t = x] \\
&= \frac{1}{dt} \mathbb{E}[g(X_{t+dt}) - g(X_t) \mid X_t = x] \\
&= \frac{1}{dt} \mathbb{E}[dg(X_t) \mid X_t = x] \\
&= \frac{1}{dt} \mathbb{E}[(a(X_t,t)g_x + \frac{1}{2}b^2(X_t,t)g_{xx})dt + b(X_t,t)g_x dB_t \mid X_t = x] \\
&= a(x,t)g_x + \frac{1}{2}b^2(x,t)g_{xx}
\end{aligned}$$

where the last equality is due to the fact that $\mathbb{E}[b(X_t,t)g_x dB_t] = 0$.

A **diffusion process** is a stochastic process $\{X_t\}$ characterised by the SDE

$$dX_t = \sigma(X_t,t)dB_t,$$

and thus it will have the generator

$$\mathcal{L}f(x) = \frac{1}{2}\sigma^2(x,t)f''(x).$$

2.3.4 Fokker-Planck Equation

We will derive the Fokker-Planck equation of a one-dimensional SDE. Consider the SDE

$$dX_t = a(X_t,t)dt + b(X_t,t)dB_t$$

with initial condition $X_0 = y$, we would like to derive an evolution equation for the transition density $p(x,t) = p(x,t|y,0)$.

For some smooth function f that depends only on X_t, using Itô lemma for SDE, $df = f_x dX_t + f_t dt + 1/2 f_{xx}(dX_t)^2$, we have

$$\begin{aligned}
df &= f_x dX_t + \frac{1}{2}f_{xx}(dX_t)^2 \\
&= f_x(adt + bdB_t) + \frac{1}{2}f_{xx}(b^2 dt) \\
&= \left(f_x a + \frac{1}{2}f_{xx}b^2\right)dt + f_x b dB_t,
\end{aligned}$$

and integrating the above equation yields

$$f(X_t) = f(X_0) + \int_0^t \left[f_x a + \frac{1}{2}f_{xx}b^2\right]dt + \int_0^t f_x b dB_t.$$

Next, consider $\mathbb{E}[f(X_t)]$, we have

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[f(X_t)] &= \frac{d}{dt}\mathbb{E}\left[f(X_0) + \int_0^t \left[f_x a + \frac{1}{2}f_{xx}b^2\right]dt + \int_0^t f_x b dB_t\right] \\
&= \frac{d}{dt}\mathbb{E}[f(X_0)] + \frac{d}{dt}\mathbb{E}\left[\int_0^t \left[f_x a + \frac{1}{2}f_{xx}b^2\right]dt\right] + \mathbb{E}\left[\int_0^t f_x b dB_t\right] \\
&= \frac{d}{dt}\mathbb{E}\left[\int_0^t \left[f_x a + \frac{1}{2}f_{xx}b^2\right]dt\right] \\
&= \mathbb{E}\left[f_x a + \frac{1}{2}f_{xx}b^2\right] \\
&= \int_{\mathbb{R}} \left(a(x,t)f_x(x) + \frac{1}{2}b^2(x,t)f_{xx}(x)\right)p(x,t)dx \\
&= \int a(x,t)f_x(x)p(x,t)dx + \int \frac{1}{2}b^2(x,t)f_{xx}(x)p(x,t)dx \\
&= \int \frac{\partial}{\partial x}[-a(x,t)p(x,t)]f(x)dx + \int \frac{1}{2}\frac{\partial^2}{\partial xx}[b^2(x,t)p(x,t)]f(x)dx \\
&= \int f(x)\left[\frac{\partial}{\partial x}[-a(x,t)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial xx}[b^2(x,t)p(x,t)]\right]dx
\end{aligned}$$

where the second last step above requires various integration by parts as well as the fact that $p, p_x \to 0$ as $x \to \pm\infty$ since p is a density.

Then, by definition, we have

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = \frac{d}{dt}\int_{\mathbb{R}} f(x)p(x,t)dx = \int f(x)\frac{\partial}{\partial t}p(x,t)dx$$

and this, along with the result above gives us

$$\frac{\partial}{\partial t}p(x,t) = \frac{\partial}{\partial x}[-a(x,t)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial xx}[b^2(x,t)p(x,t)]$$

by noticing that the equality of the two integrals holds for any f. This is the Fokker-Planck equation, and it is the same as the form of Equation 3 in Section 1.3. To see this, as Equation 3 is

$$\frac{\partial}{\partial t}p(x,t) = \mathcal{L}^* p(x,t),$$

we just need to check the adjoint $\mathcal{L}^*$ in $L^2(\mathbb{R})$. Here we will derive the adjoint using $f, g \in L^2(\mathbb{R})$ with compact support, so $f(x), g(x) \to 0$ as $x \to \pm\infty$. We have

$$\begin{aligned}
\langle \mathcal{L}f, g\rangle &= \int_{\mathbb{R}} \mathcal{L}f(x)g(x)dx \\
&= \int \left(af_x + \frac{1}{2}b^2 f_{xx}\right)gdx \\
&= \int agf_x dx + \frac{1}{2}\int b^2 g f_{xx}dx \\
&= -\int f(ag)_x dx + \frac{1}{2}f(b^2 g)_{xx}dx \\
&= \int f\left[-(ag)_x + \frac{1}{2}(b^2 g)_{xx}\right]dx = \langle f, \mathcal{L}^* g\rangle
\end{aligned}$$

where the second last line is obtained by doing multiple integrations by parts as well as using the fact that f, g have compact supports so they vanish at infinities. The above derivation yields the adjoint $\mathcal{L}^*$ is

$$\mathcal{L}^* g = -(ag)_x + \frac{1}{2}(b^2 g)_{xx}$$

so

$$\mathcal{L}^* p(x,t) = -(ap)_x + \frac{1}{2}(b^2 p)_{xx} = \frac{\partial}{\partial t} p(x,t),$$

as desired.

One of the many things the Fokker-Planck equation tells us is its steady solution. If $p_s(x)$ is a solution to the equation

$$-(ap_s)_x + \frac{1}{2}(b^2 p_s)_{xx} = 0$$

with some boundary conditions (for differential equations we always need to note the boundary conditions) and $p(x,t) \to p_s(x)$ as $t \to \infty$ for all x, then $p_s(x)$ is the **invariant density** of the SDE

$$dX_t = a(X_t, t)dt + b(X_t, t)dB_t.$$

Thus, we can find the invariant density by solving the equation

$$(ag)_x = \frac{1}{2}(b^2 g)_{xx}, \tag{4}$$

and we will use this fact multiple times in the following section, as well as study any Markov chain Monte Carlo algorithm when we know their underlying stochastic processes.

2.4 Mix-N-Match

There are a lot of widely used Markov processes consisting of a combination of the three types of processes discussed above. Among them are geometric Brownian motion (GBM) which has a profound impact on mathematical finance as it is used to model stock prices in the Black–Scholes model, and Ornstein-Uhlenbeck (O-U) process which has been used to model things in finance, physics, and evolutionary biology. Here, we will focus on two particular models - the Langevin diffusion and the piecewise deterministic Markov process (PDMP). This choice is motivated by the strong relevance of these two processes with Markov chain Monte Carlo algorithms. This link will be explained in detail in the next chapter.

2.4.1 Langevin Diffusion

The Langevin diffusion is an SDE with the following form

$$dL_t = \frac{1}{2}\nabla \log \pi(L_t)dt + dB_t \tag{5}$$

where π is differentiable and $\pi > 0$.

This equation characterises the stochastic (and Markov) process $\{L_t\}$. It has a deterministic drift term and a diffusion term. Using previous derivations, we can easily obtain its generator and the adjoint of the generator. We have

$$\mathcal{L}g = a(x,t)g_x + \frac{1}{2}b^2(x,t)g_{xx} = \frac{1}{2}\nabla \log \pi(x) g_x + \frac{1}{2}g_{xx}$$

$$\mathcal{L}^* g = -(ag)_x + \frac{1}{2}(b^2 g)_{xx} = -\left(\frac{1}{2}\nabla \log \pi(x) g\right)_x + \frac{1}{2}g_{xx}$$

for some suitable function g.

The SDE has π as its invariant. To verify this, we know that by the Fokker-Planck equation, if π satisfies

$$(a\pi)_x = \frac{1}{2}(b^2\pi)_{xx}$$

it would indeed be the invariant if we also have convergence to this density. We have

$$\frac{\partial}{\partial x}a\pi = \frac{\partial}{\partial x}\frac{1}{2}\nabla \log \pi(L_t)\pi = \frac{\partial}{\partial x}\frac{1}{2}\nabla\pi(L_t)\frac{1}{\pi}\pi = \frac{\partial}{\partial x}\frac{1}{2}\nabla\pi(L_t) = \frac{1}{2}\nabla^2\pi(L_t)$$

and

$$\frac{\partial^2}{\partial xx}\frac{1}{2}b^2\pi = \frac{\partial^2}{\partial xx}\frac{1}{2}b^2\pi = \frac{1}{2}\nabla^2\pi(L_t) = \frac{\partial}{\partial x}a\pi,$$

as required. Here, π is a distribution so $\int \pi = 1$. In fact, $c\pi$ will satisfy this equation for any constant c. This property would be quite handy when we consider MCMC algorithms that revolve around Langevin diffusion.

To show that π is indeed the invariant density of the equation, we would also need to make sure that L_t converges to π. This is indeed true under some mild conditions on π, as proved in Roberts and Tweedie (1996).

Theorem 2.4 (Roberts and Tweedie (1996) Theorem 2.1)**.** *Suppose that $\nabla \log \pi(x)$ is continuously differentiable and that, for some $N, a, b < \infty$,*

$$\nabla \log \pi(x) \cdot x \leq a|x|^2 + b, \qquad |x| > N.$$

Then the Langevin diffusion L_t satisfies the following:

1. *The diffusion is non-explosive, μ^{Leb}-irreducible, aperiodic, strong Feller and all compact sets are small.*
2. *The measure π is invariant for $\{L_t\}$ and, moreover, for all x,*

$$\|P_L^t(x, \cdot) - \pi\| \to 0.$$

Remark. *Here, $P_L^t(x, A) = \mathbb{P}(L_t \in A | L_0 = x)$, and the metric $\|\cdot\|$ is the total variation norm.*

2.4.2 Piecewise Deterministic Markov Process

The following discussion is based on Fearnhead et al. (2018). For a monograph-length discussion of this process, the reader can check Davis (2018).

The d-dimensional piecewise deterministic Markov process (PDMP) $\{Z_t\}$ is a jump process with deterministic motion between jumps. Between jumps, we have

$$\frac{d}{dt}z_t^{(i)} = \phi_i(z_t)$$

where $\phi = (\phi_1, \phi_2, \ldots, \phi_d)$ is a known function. The solution of this ordinary differential equation is denoted as

$$z_{s+t} = \psi(z_t, s)$$

for some function ψ. The jump will occur with rate $\lambda(z_t)$ when we are at z_t. The probability of having an event in interval $[t, t+h]$ is therefore $\lambda(z_t)h + o(h)$, where $f(h) = o(h)$ if $lim_{h\to 0} f(h)/h = 0$. If there is a jump at time τ, there would be a change in value. If we use $z_{\tau-}$

to denote the position right before the jump, the position after the jump will be drawn from the distribution $q(\cdot|z_{\tau-})$.

We will first derive the generator of $\{Z_t\}$, which we denote by $\mathcal{A}$. For suitable f, we have

$$\begin{aligned} Af(z) &= \lim_{h\to 0} \frac{\mathbb{E}[f(Z_{t+h})|Z_t = z] - f(z)}{h} \\ &= \phi(z) \cdot \nabla f(z) + \lambda(z) \int [f(z') - f(z)] q(z'|z) dz'. \end{aligned}$$

The adjoint $\mathcal{A}^*$ in $L^2(\mathbb{R})$ can be obtained by considering the adjoint of each of the two terms and then adding them up.

Proposition 2.5. $(A_1 + A_2)^* = A_1^* + A_2^*$.

Proof. We will prove this proposition in $\mathbb{R}$. The case in $\mathbb{C}$ is equally simple. We have, for any suitable x, y,

$$\begin{aligned} \langle (A_1 + A_2)^* x, y\rangle &= \langle x, (A_1 + A_2) y\rangle \\ &= \langle x, A_1 y\rangle + \langle x, A_2 y\rangle \\ &= \langle A_1^* x, y\rangle + \langle A_2^* x, y\rangle \\ &= \langle (A_1^* + A_2^*) x, y\rangle. \end{aligned}$$

So $\langle (A_1 + A_2)^* x - (A_1^* + A_2^*) x, y\rangle = 0$ for any x, y, thus $(A_1 + A_2)^* = A_1^* + A_2^*$, as desired. □

So, we will let

$$A_1 f(z) := \phi(z) \cdot \nabla f(z), \qquad A_2 f(z) := \lambda(z) \int [f(z') - f(z)] q(z'|z) dz'.$$

Then, as derived in previous sections, we have

$$\mathcal{A}_1^* f(z) = \nabla[-f(z)\phi(z)] = -\sum_{i=1}^{d} \frac{\partial \phi_i(z) f(z)}{\partial z^{(i)}}$$

and

$$\mathcal{A}_2^* f(z) = \int \lambda(z') f(z') q(z|z') dz' - \lambda(z) f(z).$$

Thus,

$$A^* f(z) = \mathcal{A}_1^* f(z) + \mathcal{A}_2^* f(z) = -\sum_{i=1}^{d} \frac{\partial \phi_i(z) f(z)}{\partial z^{(i)}} + \int_{z' \in \mathbb{R}} \lambda(z') f(z') q(z|z') dz' - \lambda(z) f(z).$$

Additionally, if $p(z)$ is an invariant distribution of this PDMP, it then must satisfy the condition that $A^* p(z) = 0$, i.e.

$$-\sum_{i=1}^{d} \frac{\partial \phi_i(z) p(z)}{\partial z^{(i)}} + \int_{z' \in \mathbb{R}} \lambda(z') p(z') q(z|z') dz' - \lambda(z) p(z) = 0.$$

Chapter 3

Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) algorithms have proved to be extremely effective in various computation-intensive settings, such as Bayesian statistics (Diaconis; 2009), statistical mechanics (Faulkner and Livingstone; 2022), and machine learning (Andrieu et al.; 2003).

Usually, MCMC algorithms can be implemented to do two things - estimating and sampling. When we would like to compute an integral that is almost impossible to do by hand, say it is of high dimension and has a complicated form, we have to turn to an approximate solution and use Monte Carlo methods instead. Other times we might have a probability distribution in mind, and we would like to generate identical and independently distributed (i.i.d.) samples from this distribution, and MCMC algorithms are good at it, especially when the target distribution is too complex to be sampled from using standard methods (e.g. inverse CDF). The second goal is harder to achieve than the first goal, and we can compute good approximations once we have obtained good samples using ergodic averages. For example, if we would like to estimate the integral of the form

$$\int f(x)\pi(x)dx =: \mathbb{E}_\pi[f(X)],$$

where π is a probability distribution and f is an arbitrary function, we could generate i.i.d. samples $X_1, X_2, \ldots, X_n$ following π and estimate the integral by

$$\mathbb{E}_\pi[f(X)] \approx \sum_{i=1}^{n} f(X_i).$$

The convergence of this estimate is guaranteed by the central limit theorem, with a convergence rate of $1/\sqrt{n}$ (Robert and Casella; 1999). This estimation scheme is known as the **Monte Carlo method**. Note that this scheme works for any integral, as we can have

$$\int f(x)dx = \int \frac{f(x)}{\pi(x)}\pi(x)dx = \mathbb{E}_\pi\left[\frac{f}{\pi}(X)\right]$$

for some function f and probability distribution π.

The rough underlying idea of MCMC is as follows. An ergodic[1] Markov chain, after running for several steps, will converge to a particular invariant distribution regardless of the initial position.

[1]The definition of this property is slightly involved in the more generalised Markov process setting. A key thing about ergodicity is that it guarantees the convergence of the process to its unique invariant. See Meyn and Tweedie (2012) for a detailed discussion on this concept.

This means, once we have reached equilibrium, every new step made by the chain will follow that particular invariant distribution, and thus we can easily obtain samples and compute estimations using the Monte Carlo method afterwards. This also explains the name, i.e. we use the Markov chain to generate samples and then use Monte Carlo to approximate.

The main obstacle of the above approach is to figure out an efficient scheme to construct an ergodic Markov chain with invariant distribution being our target distribution. We will focus on two types of MCMC algorithms in this chapter, which use Langevin diffusion and PDMP as their underlying Markov chain respectively.

3.1 MCMC using Langevin Diffusion

Recall from the previous chapter that the Langevin diffusion is an SDE with the following form

$$dL_t = \frac{1}{2}\nabla \log \pi(L_t)dt + dB_t$$

where π is differentiable and $\pi > 0$.

The fact that π is the invariant distribution of the SDE is essential here. If the target distribution of the MCMC algorithm is π (which we know prior to running the algorithm), then as long as we can simulate the diffusion $\{L_t\}$ long enough, we would get samples following the desired distribution, thus achieving the goals of MCMC.

However, as mentioned earlier when we introduced the Brownian motion, the path of a diffusion cannot be obtained exactly, and we would always require some discretisation and approximations while doing so. One way to approximate the solution numerically is by using the Euler-Maruyama method, which we will introduce next. There are, of course, ways to remedy this issue. One way is to do data assimilation, which is a common strategy in the weather forecast community (Law et al.; 2015). Another approach is to use Metropolis adjustment, which we will introduce in a bit.

3.1.1 Euler-Maruyama and Unadjusted Langevin Algorithm

Numerical solutions of differential equations have been extensively studied in the mathematical community. One of the early methods to approximate the solution of an ordinary differential equation (ODE) is that of the **Euler method** (Sauer; 2011). Consider the following ODE with initial condition

$$\frac{d}{dt}y = f(t, y), \qquad y(0) = y_0,$$

then the Euler method will yield the approximate solutions w_i at time t_i with these values defined by

$$\begin{aligned} w_0 &= y_0, \quad t_0 = 0, \\ w_{i+1} &= w_i + hy(t_i, w_i), \quad t_{i+1} = t_i + h \end{aligned}$$

for $i = 0, 1, 2, \ldots$. Here, h is a tuning parameter and it denotes the step size of the update. Naturally, the approximation will be better for smaller values of h.

When we have an SDE instead of an ODE, we will have both the deterministic drift term and a stochastic diffusion term. The scheme to approximate the solution of an SDE, therefore, needs

to be adjusted. This is known as the **Euler-Maruyama Method** (Sauer; 2011). Consider the following SDE with initial condition

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t, \qquad X_0 = x_0,$$

then the Euler-Maruyama method will yield the following approximate solutions w_i at time t_i with these values defined by

$$\begin{aligned} w_0 &= x_0, \quad t_0 = 0, \\ w_{i+1} &= w_i + ha(t_i, w_i) + z_i b(t_i, w_i), \quad t_{i+1} = t_i + h \end{aligned}$$

where $z_i \sim N(0, h)$ are i.i.d. noises and h is the tuning parameter denoting the step size.

So, using the Euler-Maruyama method, we can have the following approximation of the k-dimensional Langevin diffusion with initial value x_0 using step size h :

$$\begin{aligned} U_0 &= x_0 \\ U_{n+1} &= U_n + \frac{h}{2}\nabla \log \pi(U_n) + \epsilon_n, \quad \epsilon_n \sim N(0, hI_k) \end{aligned}$$

for $n = 0, 1, 2, \ldots$. In fact, we could simply write $U_{n+1} \sim N(U_n + \frac{h}{2}\nabla \log \pi(U_n), hI_k)$. This yields a sequence $\{U_n\}$ which can be used as the output of the MCMC algorithm.

In the Statistics literature, this method is also known as the **Unadjusted Langevin Algorithm** (ULA). The reason for this name will become obvious once we introduce its adjusted counterpart in the next subsection.

Extensive research has been conducted on ULA to study its various theoretical properties of it. Roberts and Tweedie (1996) studied the rate of convergence (if at all) of the approximation to the target distribution. Dalalyan (2017) obtained a non-asymptotic bound on the convergence of the approximation of ULA samples, assuming that the target distributions are smooth and log-concave. Durmus and Moulines (2019) provided further theoretical results on the convergence, as well as proposed a decaying over dimension scheme for the selection of tuning parameter h of the algorithm in order to have good convergence properties.

3.1.2 Metropolis-Hastings and Metropolis Adjusted Langevin Algorithm

One of the first MCMC algorithms is the **Metropolis-Hastings Algorithm** (MH) (Hastings; 1970). Given a target distribution π, a proposal kernel $Q(\cdot, \cdot)$ with $Q(x, A) = \int_A q(x, y)dy$ where $q(x, y)$ is the rate of moving from x to y, and a starting position x, we have

1. $X_0 = x$
   For $i = 0, 1, 2, \cdots,$
2. $X_{curr} = X_i$
3. $X_{prop} \sim q(X_{curr}, \cdot)$
4. Accept X_{prop} with probability $\alpha(X_{curr}, X_{prop})$ and $X_{i+1} = X_{prop}$. Else, $X_i = X_{curr}$.

Here, $\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\}$ is the acceptance probability. The above algorithm will output a sequence $\{X_n\}$, and under some conditions on Q the distribution of X_n will converge to π.

The overall transition kernel $P(x, dy)$ of the above distribution is

$$P(x, dy) = Q(x, dy)\alpha(x, y) + \delta_x(dy) \int (1 - \alpha(x, u))Q(x, du)$$

where the delta function $\delta_a(b) = 1$ when $a = b$ and 0 otherwise. The above kernel consists of two parts. The first part is when we accept the proposal which moves us from x to y, and the second part is when our proposal is rejected but we have $x = y$ to begin with.

We would want the transition kernel to have π as its invariant. It turns out that if the kernel satisfies the detailed balance equation, the kernel will be π-reversible, or simply **reversible**, and the π-invariance is guaranteed. Recall that for a Markov chain with transition kernel P to be π-invariant, it means that we have

$$\int_x \pi(dx)P(x, dy) = \pi(dy).$$

Definition 3.1. *A Markov chain with transition kernel P is π-**reversible** for some distribution π when it satisfies the **detailed balanced equation***

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$$

for all possible x, y.

Proposition 3.2. *If a Markov chain with transition kernel P is π-reversible, then it is π-invariant.*

Proof. Using the detailed balanced equation, we have

$$\int_x \pi(dx)P(x, dy) = \int_x \pi(dy)P(y, dx) = \pi(dy) \int_x P(y, dx) = \pi(dy),$$

as desired. □

Reversibility is a nice property to have, not only because it makes checking for the right invariance easier, but also because there are other various neat consequences of it. See Sherlock (2018) for a more detailed discussion on this. However, this does not mean that non-reversibility is bad. In fact, many non-reversible MCMC algorithms have exhibited better efficiencies than their reversible counterparts (Diaconis et al.; 2000). PDMP algorithms that we will introduce in the next section are examples of such non-reversible algorithms, whereas Metropolis-Hastings algorithms are reversible.

Theorem 3.3. *The Metropolis-Hastings algorithm, as constructed earlier, produces a Markov chain $\{X_n\}$ that is π-reversible.*

Proof. We just need to show that the transition kernel $P(x, dy)$ of the algorithm satisfies the detailed balance equation

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx).$$

The equation is trivial when $x = y$, so we will only consider $x \neq y$. We have

$$\begin{aligned}
\pi(dx)P(x,dy) &= \pi(dx)\left[Q(x,dy)\alpha(x,y) + \delta_x(dy)\int(1-\alpha(x,u))Q(x,du)\right] \\
&= \pi(x)dxq(x,y)dy\alpha(x,y) \\
&= \pi(x)dxq(x,y)dy\min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \\
&= \min\left\{\pi(x)q(x,y), \pi(y)q(y,x)\right\}dxdy \\
&= \pi(y)q(y,x)\min\left\{\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}, 1\right\}dxdy \\
&= \pi(dy)Q(y,dx)\alpha(y,x) = \pi(dy)P(y,dx),
\end{aligned}$$

as desired. □

Even though the Metropolis-Hastings algorithm will generate a Markov chain that respects π as its invariant, we do not know whether it will actually converge to π as the chain goes. This question of convergence (and the rate of convergence) is a big and active area of research in MCMC theory, and we direct the readers to Roberts and Rosenthal (2004) for a survey on various existing results.

The Metropolis-Hastings algorithm does not specify the choice of proposal kernel Q, and a different choice of Q will, naturally, yield a very different result. One choice of Q is simply the normal distribution $N(0, \sigma^2 I)$, and that algorithm is called **Random Walk Metropolis** (RWM). Notice that in this case $q(x,y) = q(y,x)$ for all x, y due to the symmetry of centred normal distribution, so the acceptance function is just $\alpha(x,y) = \min\{1, \pi(y)/\pi(x)\}$. Another choice of Q is using the discretised Langevin diffusion, which is called **Metropolis Adjusted Langevin Algorithm** (MALA).

The algorithm of k-dimensional MALA is as follows: given a target distribution π and a starting position x, we have

1. $X_0 = x$
   For $i = 0, 1, 2, \cdots$,
2. $X_{curr} = X_i$
3. $X_{prop} \sim X_{curr} + \frac{h}{2}\nabla \log \pi(X_{curr}) + N(0, hI_k)$
4. Accept X_{prop} with probability $\alpha(X_{curr}, X_{prop})$ and $X_{i+1} = X_{prop}$. Else, $X_i = X_{curr}$.

Here, the transition kernel is $q(x,\cdot) = x + \frac{h}{2}\nabla \log \pi(x) + (2\pi h)^{-k/2}\exp[-|x|^2/(2h)]$.

Notice that because of the additional Metropolis adjustment step in the algorithm, the Markov chain produced by MALA has π as invariant, whereas the Markov chain produced by ULA does not, due to the error that occurred in the discretisation.

3.2 MCMC using PDMP

A large part of this section is based on the survey article Fearnhead et al. (2018).

The dynamics of PDMP, as described in the previous chapter, consists of a stochastic jump and deterministic motion between jumps. Essentially, for a d-dimensional PDMP $\{Z_t\}$, we have the following:

1. (Deterministic dynamics)
$$\frac{dz_t^{(i)}}{dt} = \phi_i(z_t)$$
for $i = 1, 2, \ldots, d$ and $\phi = (\phi_1, \ldots, \phi_d)$. So, if we are starting at z_t, we will be at z_{s+t} after time with length s and
$$z_{s+t} = \psi(z_t, s)$$
for some function ψ.
2. (Jump rate) Jumps will occur at a rate $\lambda(z_t)$ that depends on the current position. The probability of an event in interval $[t, t+h]$ given that we are at z_t at time t is $\lambda(z_t)h + o(h)$.
3. (Jump) At each jump, the state of the process will change. If the jump happens at τ, and we denote the state immediately before it as $z_{\tau-}$, then the new state after the jump will be drawn from $q(\cdot|z_{\tau-})$ for some kernel q.

We would also require an initial condition, which we would commonly assume to be drawn from some know distribution, i.e. $Z_0 \sim p_0(\cdot)$.

In order to design MCMC algorithms with an underlying Markov chain following a PDMP, we would need to figure out a way to simulate from it, which is not easy. One general approach is as follows:

1. Given current time t and state z_t, simulate the next jump time, denoted by τ.
2. Calculate $z_{\tau-}$ based on the deterministic dynamics, i.e.
$$z_{\tau-} = \psi(z_t, \tau - t).$$
3. Draw the new position after jump from $q(\cdot|z_{\tau-})$.

The above procedure hand-waves at some key steps. For example, we assume that ψ has a nice form (say, analytic) and q is easy to be simulated from. Another key thing is the simulation of the jump time τ. The jump occurs at a varying rate that depends on the current position. This obstacle will not be discussed in this notes, and we direct the readers to Lewis and Shedler (1979) for a more detailed discussion on this issue.

We would like our PDMP to admit the target distribution π as its invariant. So, we would need to design the PDMP such that π is a solution to the Fokker-Planck equation $\mathcal{A}^* f = 0$ where $\mathcal{A}^*$ is the adjoint in $L^2(\mathbb{R})$ of the generator $\mathcal{A}$ of the PDMP. This means we need to have
$$-\sum_{i=1}^{d} \frac{\partial \phi_i(z)\pi(z)}{\partial z^{(i)}} + \int_{z' \in \mathbb{R}} \lambda(z')\pi(z')q(z|z')dz' - \lambda(z)\pi(z) = 0.$$

3.2.1 Continuous Time Scaling Limit

The MCMC algorithms introduced in the previous section (RWM and MALA) produce a discrete-time Markov chain as the output. In the case of Langevin diffusion driven algorithms, the discretisation occurs due to the self-similar nature of the path of a Brownian motion. For PDMP, however, there is no diffusion component and only deterministic and jump components, which means the trajectory of a simulated PDMP could be continuous in time, and this is the case here for PDMP driven MCMC algorithms.

Continuous time MCMC algorithms can be obtained by taking a discrete time algorithm and doing a scaling limit, similar to the construction of Brownian motion from random walk via the Donsker's theorem described in Section 2.3.1.

Here, we will describe how one can obtain a continuous-time non-reversible MCMC algorithm from its discrete-time counterpart in general.

The desired continuous-time algorithm will target a joint distribution of (x, v), where we can interpret x as the position and v as the velocity which we keep its magnitude fixed here so that v only represents the direction of the motion. So, the target distribution will be of the form $\pi(x)p_u(v)$ where $p_u(v)$ is the uniform distribution over all possible directions v.

The discrete-time algorithm consists of two types of moves. The first type involves a two-part deterministic proposal

(1-a) Propose a reversible move from (x, v) to $(x + hv, -v)$, i.e. move towards the direction of v for h units of time then reverse direction. Accept it with standard Metropolis adjustment, i.e. with probability $\min\{1, \pi(x + hv)/\pi(x)\}$.

(1-b) Flip the current velocity v' to $-v'$ while keeping the position unchanged.

The above proposal is reversible, and we will keep the velocity unchanged if we accept the move in position while we will reverse the velocity if the move is rejected. These dynamics are the same as that of Hamiltonian Monte Carlo (HMC) (Neal et al.; 2011).

An algorithm using only this type of movement will yield a reducible[2] Markov chain when the dimension of x is beyond 1 as we will only move, towards to or away from, the direction of v, which is a one-dimensional motion. Reducibility will refrain the chain from converging to the invariant. So, in order to avoid reducibility, we have the second type of motion of updating v from some transition kernel with $p_u(v)$ as invariant.

Now, we are ready to take the scaling limit of this discrete-time algorithm. We will assume we have applied the first type of motion for N times between two consecutive motions of the second type. We will let $h \to 0$ while keeping $t = hN$ as a constant. The i-th movement of the first type will occur at time ih, and we let (x_t, v_t) to be the state of the algorithm at time $ih \le t < (i+1)h$.

For small h, the proposal of step (1-a) will be rejected with the following probability:

$$\begin{aligned}
1 - \min\{1, \pi(x + hv)/\pi(x)\} &= \max\{0, 1 - \pi(x + hv)/\pi(x)\} \\
&= \max\{0, 1 - \exp[\log \pi(x + hv) - \log \pi(x)]\} \\
&= \max\{0, 1 - \exp[hv \cdot \nabla \log \pi(x) + o(h)]\} \\
&= \max\{0, -hv \cdot \nabla \log \pi(x) + o(h)\} \\
&= \max\{0, -hv \cdot \nabla \log \pi(x)\} + o(h),
\end{aligned}$$

where we say $f(h) = o(h)$ if $\lim_{h\to 0} f(h)/h = 0$ and used Taylor expansion at various steps, assuming π is at least twice differentiable. This means as h tends to 0, step (1-a) will happen as events of a Poisson process with rate $\lambda(x_t, v_t) = \max\{0, -v_t \cdot \nabla \log \pi(x_t)\}$, which is a continuous time process.

To summarise, we have thus obtained a continuous time algorithm of the following form after taking the scaling limit. If we denote the state of our d-dimensional PDMP by $Z_t = (X_t, V_t)$ with some jump rate function λ and jump transition kernel q, we have

1. (Deterministic dynamics) For $i = 1, 2, \ldots, d$,

$$\frac{dx_t^{(i)}}{dt} = v_t^{(i)}, \quad \frac{dv_t^{(i)}}{dt} = 0.$$

[2] Roughly speaking, a reducible Markov chain will sometimes be stuck in a portion of the state space.

The solution of the dynamics is simply $(x_{t+s}, v_{t+s}) = (x_t + sv_t, v_t)$ for any $s > 0$.
2. (Jump rate) Jump occurs with rate $\lambda(x_t, v_t)$.
3. (Jump update) If the jump occurs at time τ, let $x_\tau = x_{\tau-}$ and $v_\tau \sim q(\cdot|x_{\tau-}, v_{\tau-})$.

The choice of λ and q thus determines the exact dynamics of this continuous time MCMC algorithm.

Remark. *As shown in the derivation above, the Metropolis adjustment in the above algorithm is not removed but transformed into a Poisson process rate. Therefore it would be inaccurate to say that PDMP driven MCMC algorithms are 'rejection-free', as was falsely remarked by some people.*

Now that we have a more precious form of the algorithm, we can obtain a more obvious conditions on λ and q to enforce the π-invariance. As derived earlier in this section, we have

$$-\sum_{i=1}^{d} \frac{\partial \phi_i(z)\pi(z)}{\partial z^{(i)}} + \int_{z' \in \mathbb{R}} \lambda(z')\pi(z')q(z|z')dz' - \lambda(z)\pi(z) = 0.$$

Substituting and simplifying this equation yields

$$-\pi(x)p(v|x)[v \cdot \nabla_x \log \pi(x) + v \cdot \nabla_x \log p(v|x) + \lambda(z)] + \int \lambda(x, v')q(v|x, v')\pi(x)p(v'|x)dv' = 0.$$

If we let v be independent of x for the invariant distribution, we have $\nabla_x \log p(v|x) = 0$. This means, we just need to satisfy

$$-\pi(x)p(v|x)[v \cdot \nabla_x \log \pi(x) + \lambda(z)] + \int \lambda(x, v')q(v|x, v')\pi(x)p(v'|x)dv' = 0,$$

which, after rearrangement, is

$$p_v(v)\lambda(x, v) - \int \lambda(x, v')q(v|x, v')p_v(v')dv' = -p_v(v)v \cdot \nabla_x \log \pi(x)$$

where $p_v(\cdot)$ is a distribution for the velocity v. If we integrate this over v, we would have

$$\int p_v(v)\lambda(x, v)dv - \int\int \lambda(x, v')q(v|x, v')p_v(v')dv'dv = \int -p_v(v)v \cdot \nabla_x \log \pi(x)dv,$$

which simplifies to

$$\mathbb{E}(V) \cdot \nabla_x \log \pi(x) = 0$$

by exchanging v and v' in the double integral term on the left. This equality holds for all π, so we must have $\mathbb{E}(V) = 0$, i.e. the average of all possible velocities must be zero.

In fact, we would commonly design the algorithm such that the set of possible velocities is symmetrical, and we have a flip operator F_x which is an involution, i.e. $F_x(F_x(v)) = v$ for all v. Thus, for any v and $v' = F_x(v)$, the invariant condition becomes

$$p_v(v)\lambda(x, v) - \lambda(x, v')p_v(v') = -p_v(v)v \cdot \nabla_x \log \pi(x)$$

and as $p_v(v) = p_v(v')$, we simply have

$$\lambda(x, v) - \lambda(x, v') = -v \cdot \nabla_x \log \pi(x)$$

for all x. As an implication of this relationship, we can add any constant to our rates λ without breaking this equality, but normally we would only use the smallest possible rate.

If we swap v and v' above, we would get

$$\lambda(x, v') - \lambda(x, v) = -v' \cdot \nabla_x \log \pi(x) = -F_x(v) \cdot \nabla_x \log \pi(x) = v \cdot \nabla_x \log \pi(x),$$

so F_x must satisfies $F_x(v) \cdot \nabla_x \log \pi(x) = -v \cdot \nabla_x \log \pi(x)$. As a result of this relationship, we only need to know π proportionally to obtain F_x due to $\nabla \log$.

3.2.2 Zig-Zag Algorithm

The **Zig-Zag** algorithm was first proposed in Bierkens and Roberts (2017) as a limiting form of a discrete-time MCMC algorithm. The set of velocities considered for Zig-Zag is, assuming the target distribution is d-dimensional, of the form $v = \sum_{i=1}^{d} \theta_i e_i$ where $\theta_i \in \{-1, +1\}$ and $\{e_i\}_{i=1}^{d}$ is a set of orthonormal basis of $\mathbb{R}^d$. There are, therefore, 2^d choices of v and the invariant of v is the uniform distribution over all these options. This choice of velocities enforces the algorithm to move in a zig-zag fashion, thus it is called the Zig-Zag algorithm.

The jump of Zig-Zag will occur coordinate-wise, and each coordinate will have its own rate. So, the flip of the i-th coordinate is $F^{(i)}$ which changes θ_i to $-\theta_i$. The overall jump rate λ will also be the sum of the coordinate-wise jump rates λ_i.

With these given structures, we can deduce that in order for the algorithm to admit π as its invariant, we need to have

$$\sum_{i=1}^{d} \{\lambda_i(x, v) - \lambda_i(x, F_i(v))\} = -\sum_{i=1}^{d} \theta_i \frac{\partial \nabla \log \pi(x)}{\partial x^{(i)}}.$$

One way for this condition to hold is to have

$$\lambda_i(x, v) - \lambda_i(x, F_i(v)) = -\theta_i \frac{\partial \nabla \log \pi(x)}{\partial x^{(i)}}.$$

for each coordinate.

There has been ongoing work on both the theory of Zig-Zag and the implementation of Zig-Zag. For theory of Zig-Zag, there are Bierkens et al. (2019), Vasdekis and Roberts (2022), and Bierkens and Duncan (2017). For the implementation of Zig-Zag, there is Corbella et al. (2022).

3.2.3 Bouncy Particle Sampler

The **Bouncy Particle Sampler** (BPS) is another main PDMP driven MCMC algorithm proposed in Bouchard-Côté et al. (2018). The flip operator F_x is defined to be

$$F_x(v) = v - 2\frac{v \cdot \nabla \log \pi(x)}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x).$$

Essentially, the flip will point v towards (or away from) the direction of the gradient of the target distribution π. This gives the algorithm a 'bouncy' trajectory, which gives the algorithm its name. The jump rate λ is chosen to be the smallest possible one such that the invariance condition is satisfied. For some target distributions, this algorithm will produce a reducible Markov chain, for which we need to introduce an additional refresh step from a Poisson process

with a constant rate. This issue is highlighted and remedied in the paper Bouchard-Côté et al. (2018).

It can be shown that, if the dimension of the target distribution π is one, Zig-Zag and BPS are identical. The difference between the two algorithms really lies in their extension to higher dimensions. There have been a series of followed up work on BPS, e.g. Deligiannidis et al. (2021), and there have been ongoing comparisons between Zig-Zag and BPS from both theoretical (Bertazzi et al.; 2022) and empirical perspectives (Bertazzi and Bierkens; 2022).

Another thing to note about BPS is that it has a discrete-time counterpart, called Discrete BPS (Sherlock and Thiery; 2022). This comes after the introduction of BPS, which is the opposite of Zig-Zag.

Bibliography

Andrieu, C., De Freitas, N., Doucet, A. and Jordan, M. I. (2003). An introduction to mcmc for machine learning, *Machine Learning* **50**(1): 5–43.

Bakry, D., Gentil, I., Ledoux, M. et al. (2014). *Analysis and Geometry of Markov Diffusion Operators*, Vol. 103, Springer.

Bertazzi, A. and Bierkens, J. (2022). Adaptive schemes for piecewise deterministic monte carlo algorithms, *Bernoulli* **28**(4): 2404–2430.

Bertazzi, A., Bierkens, J. and Dobson, P. (2022). Approximations of piecewise deterministic markov processes and their convergence properties, *Stochastic Processes and their Applications* **154**: 91–153.

Bierkens, J. and Duncan, A. (2017). Limit theorems for the zig-zag process, *Advances in Applied Probability* **49**(3): 791–825.

Bierkens, J., Fearnhead, P. and Roberts, G. (2019). The zig-zag process and super-efficient sampling for bayesian analysis of big data, *The Annals of Statistics* **47**(3): 1288–1320.

Bierkens, J. and Roberts, G. (2017). A piecewise deterministic scaling limit of lifted metropolis–hastings in the curie–weiss model, *The Annals of Applied Probability* **27**(2): 846–882.

Bouchard-Côté, A., Vollmer, S. J. and Doucet, A. (2018). The bouncy particle sampler: A non-reversible rejection-free markov chain monte carlo method, *Journal of the American Statistical Association* **113**(522): 855–867.

Conway, J. B. (2019). *A Course in Functional Analysis*, Vol. 96, Springer.

Corbella, A., Spencer, S. E. and Roberts, G. O. (2022). Automatic zig-zag sampling in practice, *Statistics and Computing* **32**(6): 1–16.

Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(3): 651–676.

Davis, M. H. (2018). *Markov Models and Optimization*, Routledge.

Deligiannidis, G., Paulin, D., Bouchard-Côté, A. and Doucet, A. (2021). Randomized hamiltonian monte carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates, *The Annals of Applied Probability* **31**(6): 2612–2662.

Diaconis, P. (2009). The markov chain monte carlo revolution, *Bulletin of the American Mathematical Society* **46**(2): 179–205.

Diaconis, P., Holmes, S. and Neal, R. M. (2000). Analysis of a nonreversible markov chain sampler, *Annals of Applied Probability* pp. 726–752.

Durmus, A. and Moulines, E. (2019). High-dimensional bayesian inference via the unadjusted langevin algorithm, *Bernoulli* **25**(4A): 2854–2882.

E, W., Li, T. and Vanden-Eijnden, E. (2021). *Applied Stochastic Analysis*, Vol. 199, American Mathematical Soc.

Faulkner, M. F. and Livingstone, S. (2022). Sampling algorithms in statistical physics: a guide for statistics and machine learning, *arXiv preprint arXiv:2208.04751* .

Fearnhead, P., Bierkens, J., Pollock, M. and Roberts, G. O. (2018). Piecewise deterministic markov processes for continuous-time monte carlo, *Statistical Science* **33**(3): 386–412.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications, *Biometrika* **57**(1): 97–109.

Karatzas, I., Shreve, S. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, Vol. 113, Springer Science & Business Media.

Law, K., Stuart, A. and Zygalakis, K. (2015). Data assimilation, *Cham, Switzerland: Springer* **214**: 52.

Lewis, P. W. and Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning, *Naval Research Logistics Quarterly* **26**(3): 403–413.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov Chains and Stochastic Stability*, Springer Science & Business Media.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics, *Handbook of Markov Chain Monte Carlo* **2**(11): 2.

Norris, J. R. (1998). *Markov chains 2nd ed.*, Cambridge university press.

Øksendal, B. (2013). *Stochastic Differential Equations: an introduction with applications*, Springer Science & Business Media.

Risken, H. (1996). Fokker-planck equation, *The Fokker-Planck Equation*, Springer, pp. 63–95.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Vol. 2, Springer.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms, *Probability Surveys* **1**: 20–71.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations, *Bernoulli* pp. 341–363.

Sauer, T. (2011). *Numerical Analysis*, Addison-Wesley Publishing Company.

Sherlock, C. (2018). Reversible markov chains: variational representations and ordering, *arXiv preprint arXiv:1809.01903* .

Sherlock, C. and Thiery, A. (2022). A discrete bouncy particle sampler, *Biometrika* **109**(2): 335–349.

Vasdekis, G. and Roberts, G. O. (2022). A note on the polynomial ergodicity of the one-dimensional zig-zag process, *Journal of Applied Probability* **59**(3): 895–903.