AN INTRODUCTION TO MARKOV CHAIN MONTE CARLO METHOD

DADU CHEN

Abstract. This paper mainly follows from [1], notes created by students based on lectures from Professor Daniel Sanz-Alonso in the course STAT 31510 during Spring 2019. This paper introduces two algorithms in Markov Chain Monte Carlo (MCMC) method for random samplings, with Metropolis Adjusted Langevin Algorithm (MALA) on the basis of Metropolis Hastings (MH) Algorithm. An introduction about Monte Carlo integration and stochastic differential equations (SDEs) is also given along the way.

Contents

1. Preliminaries 1
2. Monte Carlo Integration 4
3. Markov Chain Monte Carlo (MCMC) 5
4. Stochastic Differential Equations (SDEs) 9
5. Metropolis Adjusted Langevin Algorithm (MALA) 13
Acknowledgments 15
References 15

1. Preliminaries

We assume basic familiarity of measure theory based probability, Markov chain, martingales, Brownian motion, and Real analysis. We now quickly review some key concepts and important properties of Markov chains. Note that we are not proving these properties as they are not the major points considered in this paper.

Definition 1.1. Markov chain. A collection $\mathbb{X} = \{X_n, n = 0, 1, 2, ...\}$ of random variables taking values in a state-space E is called a *Markov chain* if

$$\mathbb{P}(X_{n=1} \in A_{n+1} \mid X_n \in A_n, \ldots, X_0 \in A_0) = \mathbb{P}(X_{n+1} \in A_{n+1} \mid X_n \in A_n)$$

for all $n \geq 0$, and all measurable $A_{n+1}, A_n, \ldots, A_0 \subset E$.

The most classic way of interpreting the Markov property is that the future state only depends on the current state but not on any past states.

Definition 1.2. Time homogeneous Markov chain. A Markov chain $\mathbb{X}$ is called *time homogeneous* if its transition probabilities

$$P(x, A) = \mathbb{P}(X_{n+1} \in A \mid X_n = x)$$

do not depend on n. $P(x, A)$ is called the transition kernel or *Markov kernel* of $\mathbb{X}$.

In other words, a Markov chain is time homogeneous if its transition probabilities do not depend on time. Throughout this paper we will only work with time homogeneous Markov chain, and we will refer to the Markov kernel by $p(x, y)$.

Definition 1.3. $n-$**th step transition probability density**. The n-th step transition probability densities $p^n(x, y)$ are defined by

$$\mathbb{P}(X_n \in A \mid X_0 = x) = P^n(x, A) = \int_A p^n(x, y)dy.$$

A nice property of time homogeneous Markov chain is that we can easily compute the probability density function (PDF) given an initial probability distribution and its transition probability as we can see from the below theorem.

Theorem 1.4. *If $X_0 \sim \pi_0$ and $\mathbb{X}$ has n-th step transition probability densities $p^n(x, y)$, then the PDF of X_n is given by*

$$\pi_n(x) = \int_E \pi_0(y)p^n(y, x)dy.$$

If the state space E is finite, then π_n is a probability vector and $\pi_n = \pi_0 P^n$, where P is the transition probability matrix.

Moreover, it is very important to notice the large-time behavior of time-homogeneous Markov chain, i.e. the behavior of P^n for large n. It turns out that as n goes to infinity, the transition probability approaches to an equilibrium if certain conditions are satisfied. This gives rise to the following definition.

Definition 1.5. Invariant probability distribution. A Markov kernel $p(x, y)$ satisfies the *general balance equation* with respect to π if

$$\pi(x) = \int_E \pi(y)p(y, x)dy. \tag{1.6}$$

.

We then say that π is an invariant distribution of the Markov kernel $p(x, y)$. This is also called a stationary probability distribution.

Note that if E is discrete, the general balance equation is simply $\pi = \pi P$.

It is, however, not always guaranteed that such invariant probability distribution exists for every Markov chain. We then come up with the following definition.

Definition 1.7. Ergodic Markov chain. A Markov chain $\mathbb{X}$ is called *ergodic* if there exists a distribution π such that

$$\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \pi(A)$$

for all $A \subset E$ and initial distributions π_0. We say that π is the limit distribution of $\mathbb{X}$.

Theorem 1.8. *Let $\mathbb{X}$ be an ergodic Markov chain with Markov kernel $p(x, y)$ and limit distribution π. Then π is an invariant distribution for $p(x, y)$. In particular, if $\mathbb{X}$ is initialized at statistical equilibrium ($X_0 \sim \pi$) then $X_n \sim \pi$ for all $n \geq 0$.*

Given a probability distribution π, it is often difficult to find a Markov kernel for which π satisfies Equation (1.6). Hence we often find a Markov kernel based on another condition called the detailed balance, which is actually a stronger condition.

Definition 1.9. Detaled balance. We say that a transition density $p(x, y)$ satisfies *detailed balance* with respect to π if, for all $x, y \in E$,

$$\pi(x)p(x, y) = \pi(y)p(y, x).$$

Theorem 1.10. *Let $p(x, y)$ be a Markov kernel that satisfies detailed balance with respect to a distribution π. Then π is an invariant distribution for $p(x, y)$.*

The following concepts and results from measure theory are also important for later proofs. Similarly, we omit the proofs of these results.

Definition 1.11. Stationary sequence. Random variables $X_0, X_1, \ldots$ are said to be a stationary sequence if for every k, the shifted sequence $X_{k+n}, n > 0$ has the same distribution, i.e., the joint probability distribution of the sequence is invariant over time.

Definition 1.12. Measure-preserving transformation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A measurable map $\varphi : \Omega \to \Omega$ is said to be measure preserving if $\mathbb{P}(\varphi^{-1}A) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$.

Remark 1.13. Let φ^n be the nth iterate of φ defined inductively by $\varphi^n = \varphi(\varphi^{n-1})$ for $n \geq 1$, where $\varphi^0(\omega) = \omega$. If $X \in \mathcal{F}$, then $X_m(\omega) = X(\varphi^m(\omega))$ defines a stationary sequence.

Remark 1.14. A set $A \in \mathcal{F}$ is said to be invariant if $\varphi^{-1}A = A$. We use $\mathcal{I}$ to denote the collection of invariant events. Note that $\mathcal{I}$ is a σ-field, and $X \in \mathcal{I}$ if and only if X is invariant, i.e., $X \circ \varphi = X$ *a.s.* [1]

Remark 1.15. A measure-preserving transformation on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be ergodic if $\mathcal{I}$ is trivial, i.e., for every $A \in \mathcal{I}, \mathbb{P}(A) \in \{0, 1\}$. For an ergodic Markov chain, $\mathcal{I}$ is trivial.

Theorem 1.16. *(Dominated Convergence Theorem)*[2]*. Suppose that f_n are measurable real-valued functions and $f_n(x) \to f(x)$ for each x. Suppose there exists a non-negative integrable function g such that $|f_n(x)| \leq g(x)$ for all x. Then*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

The conclusion still holds if $f_n(x) \to f(x)$ almost everywhere.

Theorem 1.17. *(Bounded convergence theorem). Let E be a set with $\mu(E) < \infty$. Suppose f_n vanishes on $E^c, |f_n(x)| \leq M$, and $f_n \to f$ in measure. Then,*

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu.$$

Finally, the following definitions are fundamental to our study in stochastic calculus in later sections. Notations and symbols of these terms will be the same throughout the paper.

Definition 1.18. Filtration. If $X_1, X_2, \ldots$ is a sequence of random variables, then the associated (discrete time) filtration is the collection $\{\mathcal{F}_n\}$ where $\mathcal{F}_n$ denotes the information in $X_1, \ldots, X_n$.

[1] *a.s.* stands for almost surely, which is the same thing as the definition of almost everywhere in measure theory.

[2] A detailed version of the proof is at [2], pg. 62-63.

Definition 1.19. Brownian Motion. A stochastic process[3] B_t is called a (one-dimensional) Brownian motion with drift m and variance σ^2 starting at the origin if it satisfies the following.

1. $B_0 = 0$.
2. For $s < t$, the distribution of $B_t - B_s$ is normal with mean $m(t-s)$ and variance $\sigma^2(t-s)$.
3. If $s < t$, the random variable $B_t - B_s$ is independent of the values B_r for $r \leq s$.
4. With probability 1, the function $t \to B_t$ is a continuous function of t.

If $m = 0, \sigma^2 = 1$, then B_t is called a standard Brownian motion.

We will encounter Brownian motion when we introduce stochastic differential equation (SDE).

2. Monte Carlo Integration

Monte Carlo method is usually used to approximate some distribution by random sampling. To first give an idea of it, we consider a simple but classic implementation: to find, or approximate, the value of π. The procedure is simple. We first consider a circle inscribed in a square, and then uniformly generate random points over the region of the square. Then, we count S, total number of points generated, and C, number of points inside the circle. The ratio $\frac{C}{S}$ will give an approximation to the value $\frac{\pi}{4}$, which gives the desired result. We will see later that the procedure of "generating random points over some region", in mathematical world, is just the repeated random sampling over a distribution.

There are undeniably various applications of Monte Carlo method in mathematics and other fields such as engineering, computer science, and finance. In this paper, we are particularly interested in Monte Carlo integration, an application that gives numerical approximation of a definite integral that is usually difficult to solve by normal computation.

Remark 2.1. Normally, Monte Carlo methods can compute integrals of the form

$$\mathcal{I}_f[h] = \int_E h(x)f(x)dx \equiv \mathbb{E}_{X\sim f}[h(X)], \tag{2.2}$$

where f is a PDF supported on E and $h : E \to \mathbb{R}$. The first equality is supported by the following result in measure theory.

Theorem 2.3. *Suppose X is a random variable with distribution μ_X, and f is a Borel measurable function. Then,*

$$\mathbb{E}[g(x)] = \int_{\mathbb{R}} g(x)d\mu_X.$$

Note that in measure theory, the expectation of $g(x)$ is defined by the Lebesgue integral of $g(x)$ with respect to a probability measure $\mathbb{P}$. If X has a PDF f, then we have

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

One way to approximate $\mathcal{I}_f[h]$ is to sample from f, and this is known as the *classical* Monte Carlo, which we will be using in later Markov Chain Monte Carlo (MCMC).

[3]A collection of random variables indexed by time.

Algorithm 1 Classical Monte Carlo Integration

Input: Target distribution f and test function h.

1: Sample $X^{(1)}, \dots, X^{(N)} \overset{i.i.d}{\sim} f$.

Output: Monte Carlo estimator $\mathcal{I}_f^{MC}[h] := \frac{1}{N}\sum_{n=1}^{N} h(X^{(n)}) \approx \mathcal{I}_f[h]$.

Observe that this procedure is exactly the same as the classic example of approximating the value of π as shown above. Using this method, one of the main reason we directly sample from f is because it is easy to sample from. In the case of approximating π, it is convenient to sample from a square because the area of a square is easily determined.

Also, it is worth noticing that in Classical Monte Carlo Integration, the target distribution needs to be uniformly distributed. Hence, this method faces large obstacles when we want to study any distribution, including those that are non-uniform. This gives rise to another method that combines Monte Carlo together with Markov Chain.

3. Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) method is another way of approximating integrals such as Equation (2.2). We will see later that combining Markov chain and Monte Carlo with several procedures helps to implicitly construct a Markov kernel that can be sampled from. We will show that f is exactly its invariant distribution and repeated random samplings over this Markov kernel can be used to approximate

$$\mathcal{I}_f[h] \approx \frac{1}{N}\sum_{n=1}^{N} h(X^{(n)}).$$

We begin by introducing the Metropolis Hastings algorithm, one of the MCMC method that gives a flexible way to construct a nice Markov kernel.

Algorithm 2 Metropolis Hastings Algorithm

Input: Target f, initial distribution π_0, and proposal Markov kernel $q(x, y)$.

Initial draw: Sample $X^{(0)} \sim \pi_0$.

Subsequent samples: For $n = 0, \dots, N-1$ do:

1: Sample $Y^* \sim q(X^{(n)}, \cdot)$.

2: Update $X^{(n+1)} = \begin{cases} Y^* \text{ w.p. } a(X^{(n)}, Y^*), \\ X^{(n)} \text{ w.p. } 1 - a(X^{(n)}, Y^*). \end{cases}$

Output: Samples $X^{(1)}, \dots, X^{(N)}$ and approximation $\mathcal{I}_f[h] \approx \frac{1}{N}\sum_{n=1}^{N} h(X^{(n)})$.

Note that in Metropolis Hastings, the initial proposal Markov kernel $q(x, y)$ is not strictly determined, although we do hope that its stationary distribution is close to f.

In this algorithm, we draw samples iteratively based on an accept/reject mechanism as shown in the second procedure. As we will show later that, by specific selection of $a(x, y)$, the probability of accepting a proposed move from x to y, we indeed turns $q(x, y)$ into another Markov kernel for which f is invariant.

Observe that in classical Monte Carlo method, the samples drawn from the distribution are independent, which is a necessary condition in Algorithm 1. However, in MCMC the samples are clearly not independent as they satisfy the Markov property. Luckily, the Metropolis Hastings Algorithm still holds based on the following theorem, which is known as the Law of Large Numbers.

Theorem 3.1. *Let $h : E \to \mathbb{R}$ and let $\mathbb{X} = \{X_n\}_{n=0}^{\infty}$ be an ergodic Markov chain with stationary distribution π. Then $\pi-$almost surely*

$$\frac{1}{N}\sum_{n=1}^{N} h(X_n) \overset{N\to\infty}{\longrightarrow} \int_E h(x)\pi(x)dx \equiv \mathcal{I}_\pi[h].$$

This result directly follows from Birkhoff's Ergodic Theorem, which is stated as follows.

Theorem 3.2. *(Birkhoff's Ergodic Theorem). For any $X \in L^1$,*

$$\frac{1}{n}\sum_{m=0}^{n-1} X(\varphi^m\omega) \to E(X \mid \mathcal{I}) \text{ a.s. and in } L^1.$$

Since this theorem is vital to the implementation of many MCMC methods, it deserves a detailed proof. Many of the preliminaries have already been stated starting from Definition 1.12, while we still need the following lemma for preparation.

Lemma 3.3. *(Maximal ergodic lemma). Let $X_j(\omega) = X(\varphi^j\omega), S_k(\omega) = X_0(\omega) + \cdots + X_{k-1}(\omega)$, and $M_k(\omega) = \max(0, S_1(\omega), \ldots, S_k(\omega))$. Then $E(X; M_k > 0) \geq 0$.*

Proof. Let $X_j(\omega) = X(\varphi^j\omega), S_k(\omega) = X_0(\omega) + \cdots + X_{k-1}(\omega)$, and $M_k(\omega) = \max(0, S_1(\omega), \ldots, S_k(\omega))$.

For $j \leq k$, we have $M_k(\varphi\omega) \geq S_j(\varphi\omega)$, which implies that

$$X(\omega) \geq X(\omega) + S_j(\omega) - M_k(\varphi\omega) = S_{j+1}(\omega) - M_k(\varphi\omega) \text{ for } j = 1, = \ldots, k.$$

Since $S_1(\omega) = X_0(\omega) = X(\omega)$ and $M_k(\varphi\omega) \geq 0$, we have

$$X(\omega) \geq S_1(\omega) - M_k(\varphi\omega).$$

Therefore,

$$\begin{aligned} E(X(\omega); M_k > 0) &\geq \int_{M_k>0} [\max(S_1(\omega), \ldots, S_k(\omega)) - M_k(\varphi\omega)]dP \\ &= \int_{M_k>0} [M_k(\omega) - M_k(\varphi\omega)]dP \end{aligned}$$

Since $M_k(\omega) = 0$ and $M_k(\varphi\omega) \geq 0$ on $\{M_k > 0\}^c$, we have

$$\int_{M_k>0} [M_k(\omega) - M_k(\varphi\omega)]dP \geq \int [M_k(\omega) - M_k(\varphi\omega)]dP$$

Finally, since φ is measure preserving, $\int M_k(\omega) - M_k(\varphi\omega)dP = 0$. Hence, the theorem is proved. □

We now give a proof to Birkhoff's Ergodic Theorem. This part requires heavy knowledge on real analysis and many of the results will be directly used.

Proof. Let $X_j(\omega) = X(\varphi^j\omega), S_k(\omega) = X_0(\omega) + \cdots + X_{k-1}(\omega)$, and $M_k(\omega) = max(0, S_1(\omega), \ldots, S_k(\omega))$. By Remark 1.14, $E(X \mid \mathcal{I})$ is invariant under φ. Then, by letting $X' = X - E(X \mid \mathcal{I})$ we can assume without lost of generality that $E(X \mid \mathcal{I}) = 0$.

Let $\overline{X} = \limsup S_n/n$. Take arbitrary $\epsilon > 0$, and let $D = \{\omega : \overline{X} > \epsilon\}$. Note that $\overline{X}(\varphi\omega) = \overline{X}(\omega)$, which implies $D \in \mathcal{I}$ by Remark 1.14. We now want to show that $P(D) = 0$.

Let

$$X^*(\omega) = (X(\omega) - \epsilon)1_D(\omega), S_n^*(\omega) = X_n^*(\omega) + \cdots + X^*(\varphi^{n-1}\omega),$$

$$M_n^*(\omega) = \max(0, S_1^*(\omega), \ldots, S_n^*(\omega)), F_n = \{M_n^* > 0\},$$

$$F = \bigcup_n F_n = \left\{\sup_{k\geq 1} S_k^*/k > 0\right\}.$$

Note that since $X^*(\omega) = (X(\omega) - \epsilon)1_D(\omega)$ and $D = \{\limsup S_k/k > \epsilon\}$, it follows that

$$F = \left\{\sup_{k\geq 1} S_k/k > 0\right\} \cap D = D.$$

Observe that since $E|X^*| \leq E|X| + \epsilon < \infty$, by Theorem 1.16 we have $E(X^*; F_n) \to E(X^*; F)$. Also, by Lemma 3.3 which we just proved, $E(X^*; F_n) \geq 0$. Therefore, we have that $E(X^*; F) \geq 0$.

Now, since $F = D \in \mathcal{I}$, this implies that

$$0 \leq E(X^*; D) = E(X - \epsilon; D) = E(E(X \mid \mathcal{I}); D) - \epsilon P(D) = -\epsilon P(D).$$

Hence, $P(D) = 0 = P(\limsup S_n/n > \epsilon)$. Since $\epsilon > 0$ is arbitrary, it follows that $\limsup S_n/n \leq 0$. Similarly, we can apply the last result to $-X$ and thus we show that $S_n/n \to 0$ *a.s.*

Let $X'_M(\omega) = X(\omega)1_{(|X(\omega)|\leq M)}$ and $X''_M(\omega) = X(\omega) - X'_M(\omega)$. At this stage we have shown that

$$\frac{1}{n}\sum_{m=0}^{n-1} X'_M(\varphi^m\omega) \to E(X'_M \mid \mathcal{I}) \text{ a.s.}$$

We now show that this convergence occurs in L^1. Note that since X'_M is bounded, by Theorem 1.17 we have that

$$E\left|\frac{1}{n}\sum_{m=0}^{n-1} X'_M(\varphi^m\omega) - E(X'_M \mid \mathcal{I})\right| \to 0.$$

Also, observe that

$$E\left|\frac{1}{n}\sum_{m=0}^{n-1} X''_M(\varphi^m\omega)\right| \leq \frac{1}{n}\sum_{m=0}^{n-1} E|X''_M(\varphi^m\omega)| = E|X''_M|$$

and that

$$E|E(X''_M \mid \mathcal{I})| \leq E(E(|X''_M| \mid \mathcal{I})) = E|X''_M|.$$

Hence, we have that

$$E\left|\frac{1}{n}\sum_{m=0}^{n-1} X''_M(\varphi^m\omega) - E(X''_M \mid \mathcal{I})\right| \leq 2E|X''_M|$$
$$\limsup_{n\to\infty} E\left|\frac{1}{n}\sum_{m=0}^{n-1} X''_M(\varphi^m\omega) - E(X''_M \mid \mathcal{I})\right| \leq 2E|X''_M|.$$

Finally, by Theorem 1.16, $E|X''_M| \to 0$ as $M \to \infty$. This completes the proof. □

Theorem 3.1 is an application of Birkhoff's Ergodic Theorem as $\mathcal{I}$ becomes trivial(Remark 1.15) and applying the theorem to $h(X_n(\omega))$ gives

$$\frac{1}{n}\sum_{n=1}^{n} f(X_n) \to \int_E f(x)\pi(x)dx \text{ a.s. and in } L^1.$$

As stated before, it is necessary to decide $a(x, y)$, the probability of accepting a proposed move from x to y. This is given by

$$a(x, y) := \min\{1, \frac{f(y)q(y, x)}{f(x)q(x, y)}\}.$$

The approximation $\mathcal{I}_f[h] \approx \frac{1}{N}\sum_{n=1}^{N} h(X^{(n)})$ relies on the construction of Markov kernel by this algorithm. The following theorem gives the construction and a proof of discrete case is provided.

Theorem 3.4. *The Metropolis Hastings Markov kernel is given by*

$$p_{MH}(x, y) = q(x, y)a(x, y) + \delta_x(y)r(x),$$

where

$$r(x) = \begin{cases} \sum_{y\in E} q(x, y)(1 - a(x, y)) \text{ if } E \text{ is discrete,} \\ \int_E q(x, y)(1 - a(x, y))dy \text{ if } E \text{ is continuous,} \end{cases}$$

and $\delta_x(y)$ denotes a Dirac measure at x.

Proof. (Discrete case.) In order to move from x to y, assuming $x \neq y$, then state y need to be proposed and accepted. Then, the probability of moving from x to y is given by $q(x, y)a(x, y)$.

In order to move from x to x, or x to y where $x = y$, the following two things may happen:

1. Propose x as new state and accept it. This happens with probability $q(x, x)a(x, x)$.
2. Propose any $y \in E$ and reject it. This happens with probability

$$r(x) = \sum_{y\in E} q(x, y)(1 - a(x, y)).$$

Therefore, Theorem 3.4 satisfies all condition where x moves to y. □

Observe that if $x \neq y$, then we have

$$p_{MH}(x, y) = q(x, y)a(x, y).$$

This helps to show the following theorem.

Theorem 3.5. *The Metropolis Hastings kernel p_{MH} satisfies detailed balance with respect to f.*

Proof. Take arbitrary $x, y \in E$. If $x = y$, then we simply have

$$f(x)p_{MH}(x, y) = f(y)p_{MH}(y, x).$$

Suppose that $x \neq y$, then

$$\begin{aligned} p_{MH}(x, y) &= q(x, y)a(x, y) \\ &= f(x)q(x, y) \min\{1, \frac{f(y)q(y, x)}{f(x)q(x, y)}\} \\ &= \min\{f(x)q(x, y), f(y)q(y, x)\}. \end{aligned}$$

Since the RHS is symmetric in x and y, then we again have

$$f(x)p_{MH}(x, y) = f(y)p_{MH}(y, x).$$

Therefore the result follows.

□

As a result, by Theorem 1.10 we have that f is an invariant distribution for p_{MH}, and thus we can apply Theorem 3.1 to obtain an approximation for $\mathcal{I}_f[h]$.

4. Stochastic Differential Equations (SDEs)

Before we get into Metropolis Adjusted Langevin Algorithm (MALA), it is important to understand some basic concepts in stochastic differential equations and ways to find their solution to these equations. This is clearly not a detailed introduction to stochastic differential equations as many important properties and proofs are omitted. Instead, we aim to quickly get a sense of what stochastic differential equations are and equip with us most of the knowledge and intuition we need for MALA. We will see later that the core of MALA relies on a nonlinear SDE which does not have an analytical solution. Hence, we will need to use numerical approximation via discretization.

Before we get to SDEs, it is reasonable to first look back at a general differential equation written in the following way

$$df(t) = C(f(t), t)dt$$

with initial condition $f(0) = x_0$. It is known that the solution to this equation will be given by

$$f(t) = x_0 + \int_0^t C(f(s), s)ds.$$

Stochastic differential equations look similar, except that there is another "white noise" that adds randomness to the system. The general equation of this form looks like the following:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t. \tag{4.1}$$

It is worth noticing that the white noise is created by the derivative of B_t, a standard Brownian motion. The equation is interpreted as stating that X_t is evolving like a Brownian motion with drift $\mu(t, X_t)$ and variance $\sigma(t, X_t)^2$ at time t.

Generally, the SDE will have the following solution

$$X_t = X_0 + \int_0^t \mu(X_s, s)ds + \int_0^t \sigma(X_s, s)dBs$$

where the first integral is easy to compute by ordinary integral method. The crucial step now is how we should make sense of the second term.

Naturally, we would try using previously known integration method like Riemann integral. However, recall that a standard Brownian motion has the following property:

Theorem 4.2. *With probability 1, the function $t \to B_t$ is nowhere differentiable.*

Although Brownian motion is continuous, it is nowhere differentiable and therefore not Riemann integrable[4]. One commonly used approach to defining the second term leads to the Ito integral.

The idea to construct the Ito integral is actually similar to both Riemann integral and Lebesgue integral.

Recall that when we define Riemann integral to compute integral of the form

$$\int_a^b f(t)dt,$$

a key idea is to partition the interval $[a, b]$ into different intervals with the convention that $a = t_0 < t_1 < \cdots < t_n = b$ and then approximate $f(t)$ by a step function

$$f_n(t) = f(s_j),\ t_{j-1} < t \leq t_j$$

for $s_j \in [t_{j-1}, t_j]$. We define

$$\int_a^b f_n(t)dt = \sum_{j=1}^n f(s_j)(t_j - t_{j-1})$$

and if the norm (the maximum length of the subintervals of $[a, b]$) goes to 0, the limit

$$\int_a^b f(t)dt = \lim_{n\to\infty} \int_a^b f_n(t)dt$$

exists and defines the integral.

Ito's integral is similar in the way that it first defines a *simple process* as an analogy to a step function in Riemann integral. As we can see below from the definition, a subtle difference is that the first endpoint $t_0 = 0$ and the last endpoint t_n goes to ∞ for a simple process.

Definition 4.3. Simple process. A process A_t is a simple process if there exist times

$$0 = t_0 < t_1 < \cdots < t_n < \infty$$

and random vvariables $Y_j, j = 0, 1, \ldots, n$ that are $\mathcal{F}_{t_j}$-measurable such that

$$A_t = Y_j,\ t_j \leq t < t_{j+1}.$$

[4]Since Brownian motion is nowhere differentiable, it has unbounded variation. This implies that we cannot directly use Riemann-Stieltjes integral either. More can be seen at [5].

Here we set $t_{n+1} = \infty$. Since Y_j is $\mathcal{F}_{t_j}$-measurable, A_t is $\mathcal{F}_t$-measurable. We also assume that $\mathbb{E}[Y_j^2] < \infty$ for each j.

Just like what we did in Riemann integral, we now define the Ito integral of a simple process.

Definition 4.4. If A_t is a simple process we define

$$Z_t = \int_0^t A_s dB_s$$

by

$$Z_{t_j} = \sum_{i=0}^{j-1} Y_i[B_{t_{i+1}} - B_{t_i}],$$

and more generally,

$$Z_t = Z_{t_j} + Y_j[B_{t_{i+1}} - B_{t_i}] \text{ if } t_j \leq t \leq t_{j+1}$$

$$\int_r^t A_s dB_s = Z_t - Z_r.$$

Recall that in the construction of Lebesgue integral, we first define the Lebesgue integral of a simple function and then use this definition to define the Lebesgue integral of a general measurable function. Similarly, in Ito integral we first define a simple process and then move to the construction of integration of continuous processes. Before giving the definition, we need the following theorem.

Theorem 4.5. *Suppose A_t is a process with continuous paths, adapted to the filtration $\{\mathcal{F}\}$ [5]. Suppose also that there exists $C < \infty$ such that with probability one $|A_t| \leq C$ for all t. Then there exists a sequence of simple processes $A_t^{(n)}$ such that for all t,*

$$\lim_{n\to\infty} \int_0^t \mathbb{E}[|A_s - A_s^{(n)}|^2] ds = 0. \tag{4.6}$$

Moreover, for all $n, t, |A_t^{(n)}| < C$.

With the help of Theorem 4.5, we are able to find a sequence of simple processes $A_s^{(n)}$ satisfying Equation (4.6). It turns out that the existence of

$$Z_t = \lim_{n\to\infty} \int_0^t A_s^{(n)} dBs$$

can be shown [6] and thus we can define

$$\int_0^t A_s dB_s = Z_t.$$

So far, we have given a mathematical definition to Ito integral and have made solving a SDE like (4.1) meaningful. However, solving SDEs using the definition can be inefficient and thus we require stronger tools. Hence, we are now introducing the Ito's formula, a general approach to solve many linear SDEs. Its role can be viewed as the fundamental theorem of stochastic calculus and is vital to many theorems in stochastic calculus. The Ito's formula presented below is one of the many forms of Ito's formula and is sufficient for our later need.

[5] Note that A_t is adapted to the filtration $\{\mathcal{F}\}$ if A_t is $\mathcal{F}_t$-measurable for each t.

[6] For a detailed description on this fact, one can look for [6], pg. 90-91.

Theorem 4.7. *(Ito's Formula). Suppose X_t satisfies*

$$dX_t = R_t dt + A_t dB_t$$

and suppose $f(x,t)$ is a function that is C^1 in t and C^2 in x. Then,

$$df(X_t,t)[\partial_t f(X_t,t) + R_t \partial_x f(X_t,t) + \frac{A_t^2}{2}\partial_{xx} f(X_t,t)]dt + A_t \partial_x f(X_t,t) dB_t.$$

We will not focus on how Ito's formula contribute to the derivation of many analytical solutions of some common linear SDEs since the SDE we will encounter in later section is hardly solvable. Instead, we will use Ito's formula in our derivation of the Fokker-Planck-Kolmogorov (FPK) equation which we will see very soon. Before that, we introduce the concept of *diffusion* and discuss its connection to Markov chain.

Definition 4.8. Diffusion. We say that X_t is a diffusion if it is a solution to an SDE of Equation (4.1). Here, $\mu(x,t)$ and $\sigma(x,t)$ are functions. It is called time-homogeneous if the functions do not depend on t, namely

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t.$$

Note that a diffusion is a Markov chain since at time t, the state of the system has all the information about the past to predict the future. In other words, present is only what matters. This leads us to further discussion of its ergodic property when we consider specific SDEs in later section.

For nonlinear SDE, in most cases there does not exist a general solution to it by using traditional approach in linear SDEs. Guessing a solution by empirical knowledge and using the Ito formula to check the solution is also not efficient. It is, in fact, quite common that one expects a numerical approximation to a hardly solvable nonlinear SDE.

Before we move to MALA, a final preparation is the following Fokker-Planck-Kolmogorov (FPK) equation.

Theorem 4.9. *The probability density p(x, t) of the diffusion of the SDE in Equation (4.1) solves the partial differential equation (PDE):*

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)],$$

where $D(x_t,t) = \frac{\sigma^2(X_t,t)}{2}$.

This PDE is called the Fokker-Planck-Kolmogorov(FPK) equation.

We now give a proof through a direct computation approach where less analysis is involved.

Proof. We first apply Theorem 4.7 where $R_t = \mu(X_t,t)$ and $A_t = \sigma(X_t,t)$ for an arbitrary twice differentiable function $f(x)$ and we get

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\mu(X_t,t) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\sigma^2(X_t,t).$$

By taking the expectation of both sides we get

$$\frac{d\mathbb{E}[f]}{dt} = \mathbb{E}[\frac{\partial f}{\partial x}\mu(X_t,t)] + \frac{1}{2}\mathbb{E}[\frac{\partial^2 f}{\partial x^2}\sigma^2(X_t,t)] \tag{4.10}$$

Using integration by parts, the first term of the right-hand side yields

$$\begin{aligned}\mathbb{E}[\frac{\partial f}{\partial x}\mu(X_t,t)] &= \int \frac{\partial f}{\partial x}\mu(X_t,t)p(x,t)dx \\ &= -\int f(x)\frac{\partial}{\partial x}[\mu(X_t,t)p(x,t)]dx \end{aligned} \tag{4.11}$$

where $p(x,t)$ is the probability density of the solution to Equation (4.1).

Through integration by parts again, the second term becomes

$$\begin{aligned}\frac{1}{2}\mathbb{E}[\frac{\partial^2 f}{\partial x^2}\sigma^2(X_t,t)] &= \frac{1}{2}\int \frac{\partial^2 f}{\partial x^2}\sigma^2(X_t,t)p(x,t)dx \\ &= \frac{1}{2}\int (\frac{\partial f}{\partial x})\frac{\partial}{\partial x}[\sigma^2(X_t,t)p(x,t)]dx \\ &= \frac{1}{2}\int f(x)\frac{\partial^2}{\partial x^2}[\sigma^2(X_t,t)p(x,t)]dx. \end{aligned} \tag{4.12}$$

Note that we also have

$$\begin{aligned}\frac{d\mathbb{E}[f]}{dt} &= \frac{d}{dt}\int f(x)p(x,t)dx \\ &= \int f(x)\frac{\partial}{\partial t}p(x,t)dx. \end{aligned} \tag{4.13}$$

Equation (4.10), (4.11), (4.12), and (4.13) together give the following equation

$$\begin{aligned}\int f(x)\frac{\partial}{\partial t}p(x,t)dx = &-\int f(x)\frac{\partial}{\partial x}[\mu(X_t,t)p(x,t)]dx + \\ &\frac{1}{2}\int f(x)\frac{\partial^2}{\partial x^2}[\sigma^2(X_t,t)p(x,t)]dx\end{aligned}$$

which by rearrangement yields

$$\int f(x)\{\frac{\partial}{\partial t}p(x,t) + \frac{\partial}{\partial x}[\mu(x,t)p(x,t)] - \frac{1}{2}\frac{\partial^2}{\partial x^2}[\sigma^2(X_t,t)]p(x,t)\}dx = 0.$$

Since we are considering arbitrary $f(x)$ at the beginning, we have

$$\frac{\partial}{\partial t}p(x,t) + \frac{\partial}{\partial x}[\mu(x,t)p(x,t)] - \frac{1}{2}\frac{\partial^2}{\partial x^2}[\sigma^2(X_t,t)]p(x,t) = 0,$$

which gives the FPK equation. □

5. Metropolis Adjusted Langevin Algorithm (MALA)

Metropolis Adjusted Langevin Algorithm (MALA) is another MCMC method to generate random samples from a target distribution. Similar to traditional Metropolis Hastings, MALA provides a way to compute integrals like Equation (2.2).

A simple but direct motivation to the study of MALA is that we can decide the choice of our initial Markov kernel by inputting a step-size ϵ which we will see later, rather than aiming to find a nice Markov kernel by empirical insight as in traditional Metropolis Hastings. Moreover, MALA is actually more efficient than traditional Metropolis Hastings when running simulations on computers because it has a higher acceptance rate during the accept/reject mechanism as we will see below[7].

[7]One can look at [10] for more on the computation complexity of MALA.

Algorithm 3 Metropolis Adjusted Langevin Algorithm (MALA)

Input: Target f, initial distribution π_0, and step-size $\epsilon > 0$.

1: Define the special Markov kernel:

$$q_{Lgv}(x, \cdot) := \mathcal{N}(x + \epsilon \nabla \log f(x), 2\epsilon I_d).$$

2: Run Metropolis Hastings with inputs f, π_0 and $q_{Lgv}(x, y)$.

Output: Samples $X^{(1)}, \ldots, X^{(N)}$ and approximation $\mathcal{I}_f[h] \approx \frac{1}{N} \sum_{n=1}^{N} h(X^{(n)})$.

In MALA, the proposals are made according to the special Markov kernel defined above, and the accept/reject mechanism is based on the same acceptance probability

$$a(x, y) := min\{1, \frac{f(y)q_{Lgv}(y, x)}{f(x)q_{Lgv}(x, y)}\}$$

as in traditional Metropolis Hastings Algorithm. The rest of the paper aim to discuss why MALA works.

The phrase "Langevin" in MALA comes from Langevin equation, a common type of SDE in the form of

$$dX_t = \mu(X_t, t)dt + \sigma dB_t.$$

We are particularly interested in the following Langevin equation

$$dX(t) = -\nabla V(X(t))dt + \sqrt{2}dW, \ 0 \leq t < \infty \tag{5.1}$$

because its diffusion in $\mathbb{R}^d$ is ergodic and has $\pi(x) = e^{-V(x)}$ as its invariant distribution. For this particular SDE, we shall now show that the diffusion has stationary distribution $\pi(x) = e^{-V(x)}$. We will use the Fokker-Planck-Kolmogorov (FPK) equation in Theorem 4.9 to achieve this.

Proof. We only need to put $p(x, t) = e^{-V(x)}$ into FPK equation and check if the equilibrium holds. Given Equation (5.1), $\mu(X_t, t) = -\nabla V(X_t) = -V'(X_t)$ and $\sigma(X_t, t) = \sqrt{2}$.

Observe that

$$\begin{aligned}
\frac{\partial}{\partial t}p(x, t) &= -\frac{\partial}{\partial x}[\mu(x, t)p(x, t)] + \frac{\partial^2}{\partial x^2}[\frac{\sigma^2(X_t, t)p(x, t)}{2}] \\
&= -\frac{\partial}{\partial x}[-V'(x)e^{-V(x)}] + \frac{\partial^2}{\partial x^2}e^{-V(x)} \\
&= -\frac{\partial}{\partial x}[-V'(x)e^{-V(x)}] + \frac{\partial}{\partial x}[-V'(x)e^{-V(x)}] \\
&= 0
\end{aligned}$$

□

Interestingly, we can think of choosing $V(x) = -\log f(x)$ so that the solution of Equation (5.1) has f as its invariant distribution. This gives a very natural way to think of applying this equation into MCMC method. As mentioned above for several times, we wish to approximate the solution of the SDE by discretization.

One of such method is the Euler-Muruvama method, which is actually very similar to Euler's method in approximating ordinary differential equations.

Theorem 5.2. *(Euler-Muruyama method). In a given time $t \in (0, T]$, we can approximate the solution of Equation (4.1) by letting*

$$X^{(0)} := X_0,$$
$$X^{(k+1)} := X^{(k)} + \mu(X^{(k)}, t)\epsilon + \sigma(X^{(k)}, t)\Delta B_{i+1},$$

where

$$\epsilon = t_{i+1} - t_i,$$
$$\Delta B_{i+1} = B_{t_{i+1}} - B_{t_i}.$$

The random increment ΔB_i is computed as $\Delta B_i = z_i\sqrt{\Delta t_i}$ where z_i is chosen from $\mathcal{N}(0, I_d)$. Note that this gives us a Markov chain.

Applying Euler-Muruyama method to Equation (5.1) with $V(x) = -\log f(x)$ gives the approximation of $X(t_n)$ with $X^{(n)}$ where

$$X^{(n+1)} = X^{(n)} + \epsilon\nabla \log(f(X^{(n)})) + \sqrt{2\epsilon}Z_i. \tag{5.3}$$

A problem of such discretization is that the ergodic property of the original SDE is lost, namely f is no longer the invariant distribution of the output samples $\{X^{(n)}\}_{n\geq 0}$.[8] By this reason, we use the Markov kernel

$$q_{Lgv}(x, \cdot) := \mathcal{N}(x + \epsilon\nabla \log f(x), 2\epsilon I_d)$$

implicitly defined by Equation (5.3) that plays its role in keeping the ergodic property.

Acknowledgments

I would like to thank my mentor Mark Olson for guiding me through MCMC method and providing insightful suggestions about many of the reading materials and writing techniques. I would also like to thank Professor Peter May for organizing the Math REU program and Professor Gregory Lawler for providing great lectures on analysis and probability. Besides, I would like to thank all the students[9] who participated in creating [1], which provides a clear and detailed guidance to the world of MCMC. Finally, I would like to give special thanks to my family, who support and encourage me all the time.

References

[1] D. Sanz-Alonso and students (2022). Monte Carlo Simulation. STAT 31511.
[2] Richard F. Bass (2022). Real Analysis for Graduate Student. Version 4.3. https://bass.math.uconn.edu/v43.pdf
[3] Rick Durrett (2019). Probability: Theory and Examples. Version 5.
[4] J. M. Hammersley and D. C. Handscomb (1964). Monte Carlo Methods. http://www.cs.fsu.edu/ mascagni/Hammersley-Handscomb.pdf
[5] Christopher Heil. Functions of Bounded Variation. https://heil.math.gatech.edu/6337/fall07/section3.5a.pdf

[8] In general, Euler-Maruyama method does not maintain the ergodic property of Langevin equations even when the original SDE is geometrically ergodic. For more on the conditions to test ergodicity of SDEs with time discretization, one can look at [7].

[9] They are Weilin Chen, Fuheng Cui, Zhen Dai, Marlin Figgins, Kim Liu, Yuwei Luo, Shinpei Nakamura Sakai, David Noursi, Nick Rittler, Matthew Shin, Zihao Wang, Wen Yuan Yen, Joey Yoo, and Yanfei Zhou.

[6] Gregory F. Lawler. Stochastic Calculus: An Introduction with Applications. http://www.math.uchicago.edu/ lawler/finbook.pdf
[7] J. Mattingly, A.M. Stuart and D.J. Higham, Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. Stochastic Processes and their Applications 101(2) (Oct 2002) 185232. Doi: 10.1016/S03044149(02)001503
[8] Timothy Sauer. Computational solution of stochastic differential equations. WIREs Comput Stat 2013. doi: 10.1002/wics.1272
[9] Simo Srkk and Arno Solin (2019). Applied Stochastic Differential Equations. Cambridge University Press. https://users.aalto.fi/ asolin/sde-book/sde-book.pdf
[10] Rong Tang and Yun Yang (2022). On the Computational Complexity of Metropolis-Adjusted Langevin Algorithms for Bayesian Posterior Sampling. https://doi.org/10.48550/arXiv.2206.06491.