Outline of Applied Stochastic Analysis

Tiejun Li
School of Mathematical Sciences, Peking University,
Beijing 100871, P.R. China
Email: tieli@pku.edu.cn

Lecture 1 Introduction *

Tiejun Li

1 Stochastics: why, what and where

There are still debates on whether the world is deterministic or stochastic. We take a practical point of view on this problem. The reason why we utilize stochastics is as below:

1. The problem itself is stochastic (quantum mechanics).
2. Even the problem is deterministic in nature, the degrees of freedom is too huge to be handled in a deterministic manner (statistical mechanics).
3. The considered problem is in deterministic form, but we utilize its equivalent stochastic form to do computing (Monte Carlo methods).

The course will be composed of three parts:

1. Monte Carlo methods.
2. SDEs and their simulations.
3. Applications.

The main application area of Monte Carlo method:

Statistical Physics, Statistical inference, Mathematical finance, Data Science.

2 Monte Carlo concepts

Example 1. *(Buffon test)*

1. *Parallel lines with distance a in the plane;*
2. *Tossing a needle of length l ($l < a$) randomly;*
3. *Intersection probability?*

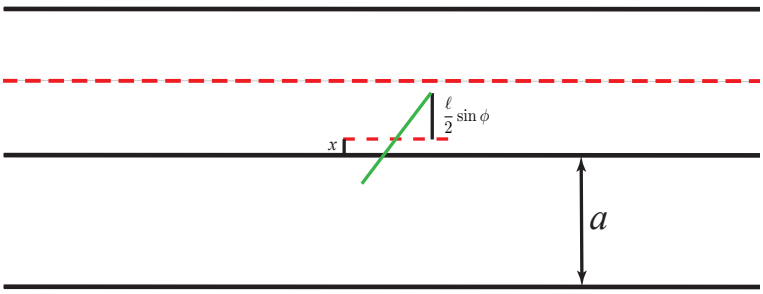


Figure 1: Schematics for Buffon's needle problem.

Solution. This is a geometric probability problem. The admissible set is

$$\Omega := \{0 \le x \le \frac{a}{2}, 0 \le \phi \le \pi\}.$$

The set of intersection is

$$G = \{x \le \frac{l}{2}\sin\phi\},$$

then the probability of intersection

$$P = \frac{meas(G)}{meas(\Omega)} = \Big(\int_0^\pi \frac{l}{2}\sin\phi d\phi\Big)\Big/\Big(\frac{a\pi}{2}\Big) = \frac{2l}{a\pi},$$

thus

$$\pi = \frac{2l}{aP}.$$

Another choice (taking into account more symmetry):

$$\Omega := \{0 \le x \le \frac{a}{2}, 0 \le \phi \le \frac{\pi}{2}\}.$$

and

$$G = \{x \le \frac{l}{2}\sin\phi\},$$

we also have

$$P = \frac{2l}{a\pi},$$

Example 2. *(Monte Carlo integration) Numerically solve*

$$I(f) = \int_0^1 f(x)dx.$$

1. Midpoint rule:

$$I_N^{(1)}(f) = h\sum_{i=1}^N f(x_i), \qquad h = \frac{1}{N},\ x_i = (i+\frac{1}{2})h$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Accuracy: $O(h^2)$.

2. Monte Carlo:

$$I_N^{(2)}(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1]$$

one has $\mathbb{E}I_N^{(2)}(f) = I(f)$, and the mean square error

$$\begin{aligned}\mathbb{E}|e_N|^2 = \mathbb{E}(I_N^{(2)}(f) - I(f))^2 &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(f(X_i) - I(f))\right)^2 \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}(f(X_i) - I(f))(f(X_j) - I(f)) \\ &= \frac{1}{N}\mathbb{E}(f(X_i) - I(f))^2 = \frac{1}{N}\mathrm{Var}\,(f),\end{aligned}$$

One obtains $e_N \sim \sqrt{\frac{\mathrm{Var}\,(f)}{N}} \sim O(h^{\frac{1}{2}})$ — half order convergence. (How to generate X_i?)

The above derivations are independent of dimensions.

3. High dimensional case:

Ensemble average in statistical mechanics

$$\langle A\rangle = \frac{1}{Z}\int_{R^{6N}} A(x)e^{-\beta H(x)}dx$$

where $Z = \int_{R^{6N}} e^{-\beta H(x)}dx$ is partition function, $\beta = (k_B T)^{-1}$, k_B is Boltzmann constant, T is the absolute temperature, $dx = dx_1\cdots dx_N dp_1\cdots dp_N$, N is the number of particles.

Deterministic quadrature: 10 segments in each direction, totally 10^{6N} nodes!

Monte Carlo method is the only viable approach!

4. Estimate of computational effort:

Dimension — d, ♯ of quadrature points — N

Midpoint rule $\sim O(N^{-\frac{d}{2}})$, Monte Carlo $\sim O(N^{-\frac{1}{2}})$.

If $d > 4$, Monte Carlo is better.

5. Brief summary:

The advantage of Monte Carlo:

- Half order convergence independent of dimensions;
- Parallel essentially;
- Versatile: If we can find a probabilistic interpretation of a problem, we can apply MC.

The disadvantage of Monte Carlo:

- Half order convergence (slow convergence);
- Noisy result.

3 Further applications

Example 3 (Randomized linear algebra)**.** *Compute the matrix product*

$$C = AB,$$

where $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, *and assume* $n \gg 1$.

When n is huge, which is possible in many applications in big data, the following randomized matrix multiplication was proposed:

Given any probability distribution $\{p_i\}$, *where* $p_i > 0$ *and* $\sum_{i=1}^{n} p_i = 1$, *randomly pick* K *columns with the* i_m*th column from* A, $L^{(m)}$ *and the* i_m*th row from* B, $R^{(m)}$ *according to* $\{p_i\}$. *Correspondingly define*

$$L^{(m)} = \frac{1}{\sqrt{Kp_{i_m}}} A_{\cdot,i_m}, \quad R^{(m)} = \frac{1}{\sqrt{Kp_{i_m}}} B_{i_m,\cdot}, \quad m = 1,\dots,K$$

then compute

$$C \approx \sum_{m=1}^{K} L^{(m)} R^{(m)}. \tag{1}$$

Does it work? Is it possible to generalize and improve it?

Example 4. *(Bayesian methods in statistical learning) Sampling the posterior distribution of the unknown parameters* θ.

In statistics, we have large amount of sampling data, and we want to extract the parameters from some type of probabilistic model. Suppose we have the *likelihood* function

$$L(\boldsymbol{\theta}|\boldsymbol{x}), \quad \boldsymbol{\theta} \in \Theta,$$

and the prior distribution of the parameter θ is $\pi(\boldsymbol{\theta})$, we would like to sample the posteriori distribution of θ

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto L(\boldsymbol{\theta}|\boldsymbol{x})\pi(\boldsymbol{\theta})$$

or compute the expectation of the parameters. Usually $\boldsymbol{\theta}$ is in a high dimensional space, and $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is only known up to a constant. We need the Monte Carlo sampling method here.

Example 5. *(Simulated annealing for optimization)* $\min_x H(x)$*, H(x) is an energy function.*

1. If $H(x)$ is convex, the problem is quite easy by steepest decent method

$$\frac{dx}{dt} = -\nabla H$$

2. If $H(x)$ is non-convex, the problem is complicate. The solution by steepest descent will fall into a local minimum generally.

3. Introduce thermal noise

$$\frac{dx}{dt} = -\nabla H + \epsilon \dot{w}$$

$\epsilon \sim$ temperature. Let $\epsilon \to 0$ with suitable speed, one can achieve the global minimum.

Example 6. *(Harmonic oscillator with random forcing) How to describe the noise mathematically? (Potential* $U(\boldsymbol{x}) = \frac{1}{2}k\boldsymbol{x}^2$*)*

1. Conservative harmonic oscillator

$$\begin{cases} \dot{\boldsymbol{x}} &= \boldsymbol{v} \\ m\dot{\boldsymbol{v}} &= -k\boldsymbol{x} \end{cases}$$

2. Frictional harmonic oscillator (frictional coefficient γ)

$$\begin{cases} \dot{\boldsymbol{x}} &= \boldsymbol{v} \\ m\dot{\boldsymbol{v}} &= -\gamma\boldsymbol{v} - k\boldsymbol{x} \end{cases}$$

3. White noise forcing (mesoscopic particles)

$$\begin{cases} \dot{\boldsymbol{x}} &= \boldsymbol{v} \\ m\dot{\boldsymbol{v}} &= -\gamma\boldsymbol{v} - k\boldsymbol{x} + \sqrt{2k_BT\gamma}\dot{\boldsymbol{w}} \end{cases}$$

$\dot{\boldsymbol{w}}$ is the temporal white noise. How to define $\boldsymbol{w}$?

Example 7. *(First exit time) Connection with PDEs.*

Solving the elliptic PDE

$$\begin{cases} \Delta u &= 0 & D \\ u &= f & \partial D \end{cases}$$

Traditional method: FEM, FD

$$u(\boldsymbol{x}) = \mathbb{E}\big(f(\boldsymbol{X}_{\tau_D})\big)$$

where $\boldsymbol{X}_{\tau_D}$ is the first exit point form ∂D of the Brownian motion starting at $\boldsymbol{x} \in D$.

One can compute the value of u at any point in Ω separately.

Example 8. *(Particle system) Macroscopic behavior from microscopic movements*

1. Deterministic case(without interaction): Liouville equation.

$$\frac{d\boldsymbol{x}_i}{dt} = \boldsymbol{b}(\boldsymbol{x}_i) \ \longrightarrow \ \psi_t + \nabla \cdot (\boldsymbol{b}\psi) = 0$$

2. Stochastic case(without interaction): Fokker-Planck equation.

$$\frac{d\boldsymbol{x}_i}{dt} = \boldsymbol{b}(\boldsymbol{x}_i) + \dot{\boldsymbol{w}}_i \ \longrightarrow \ \psi_t + \nabla \cdot (\boldsymbol{b}\psi) = \frac{1}{2}\Delta\psi$$

3. Stochastic case (with interaction): Mckean-Vlasov equation.

$$\frac{d\boldsymbol{x}_i}{dt} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{b}(\boldsymbol{x}_i - \boldsymbol{x}_j) + \dot{\boldsymbol{w}}_i \ \longrightarrow \ \psi_t + \nabla \cdot (\boldsymbol{U}\psi) = \frac{1}{2}\Delta\psi$$

where $\boldsymbol{U} = \int \boldsymbol{b}(\boldsymbol{x} - \boldsymbol{y})\psi(\boldsymbol{y})d\boldsymbol{y}$.

Example 9. *(Chemical reaction kinetics) Stochastic simulation algorithm.*

Traditional modeling of chemical reaction: reaction rate equation (RRE):

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{a}(\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{x}$ is the concentration of the reactants, $\boldsymbol{a}$ is the reaction rate. In biological reactions, the population of some species are very few. The concept concentration does not make any sense there. The reaction also shows the random character. How to model the chemical reaction kinetics?

Example 10. *(DLA model) Fractal growth of crystallization. (See Fig. 2)*

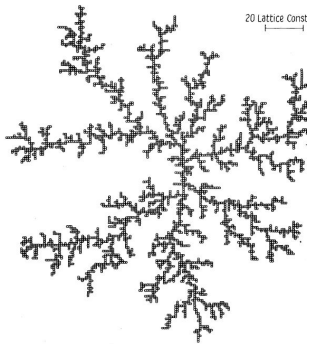


Figure 2: DLA model. Adapted from PRL 47(1981), 1400.

Example 11. *(Complex fluids) Such as the suspensions, colloids and liquid crystals, etc.*

How to describe the behavior of the fluids through describing the polymers?

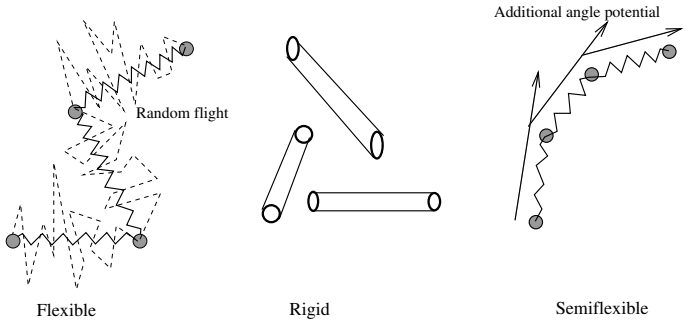


Figure 3: Schematics of flexible, semi-flexible and rigid polymers.

4 Course plan

The following topics will be covered in this course:

- Generation of pseudo random variables,
- Variance reduction methods,
- Simulated annealing and quasi-Monte Carlo,
- Large deviation principle,
- Metropolis algorithm (Markov chain Monte Carlo method),
- Multilevel sampling and kinetic MC,
- Wiener Process and its construction,
- Stochastic differential equations and Ito's formula,
- Fokker-Planck equation and diffusion process,
- Numerical solution of SDEs,
- Path integral methods and Girsanov transformation,
- Applications in material science(rare events),
- Applications in biology,
- Applications in networks,
- Applications in fluids.

We will have 2 numerical projects which will account for 15 pts. The homeworks will account for 15 pts, and the final exam will account for 70 pts.

5 Main references

- W.H. Press et al., Numerical Recipes: the Art of Scientific Computing, Cambridge university press, Cambridge, 1986.

- R.E. Caflish, Monte Carlo and Quasi-Monte Carlo methods, Acta Numerica, Vol. 7, 1-49, 1998.
- P. Glasserman, Monte Carlo methods in financial engineering, Springer-Verlag, New York, 2003.
- C.P. Robert and G. Casella, Monte Carlo Statistical methods, Springer-Verlag, New York, 2004.
- C.W. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry and Natural Sciences, Springer-Varlag, Berlin, New York, 1983.
- B. Oksendal, Stochastic Differential Equations: an Introduction with Applications, Springer-Verlag, Berlin Heidelberg New York, 2003(6th edition).
- I. Karatzas and S.E. Shreve, Brownian motion and stochastic calculus, Springer-Verlag, New York, 1991.
- P.E. Kloeden and E. Platen, Numerical Solution of Stochastic Differential Equations, Springer-Verlag, Berlin and Heidelberg, 1992.

6 Homeworks

- Discuss about the method (1). Why is it a valid method? How to characterize its accuracy?
- Show that the midpoint rule has second order convergence if $f \in C^2[0,1]$.
- Numerically testify the half order convergence of Monte Carlo integration for

$$I(f) = \int_0^1 \sin x dx = \mathbb{E} \sin X$$

where X is uniformly distributed in $[0,1]$.

Lecture 2 Random Variables *

Tiejun Li

1 A Crash Course on Basic Concepts

1.1 Discrete Examples

We will concentrate on the elementary and intuitive aspects of probability here. In the discrete case, the function $P(X)$ is called the probability mass function (pmf).

- Bernoulli distribution:
$$P(X) = \begin{cases} p, & X = 1, \\ q, & X = 0. \end{cases}$$
where $p > 0, q > 0, p + q = 1$. The mean and variance are
$$\mathbb{E}X = p, \mathrm{Var}(X) = pq.$$
If $p = q = \frac{1}{2}$, it is the well-known fair-coin tossing game.

- Binomial distribution $B(n, p)$:
n independent experiments of Bernoulli distribution X_k, $X := X_1 + \ldots + X_n$, then
$$P(X = k) = C_n^k p^k q^{n-k}.$$
The mean and variance are
$$\mathbb{E}X = np, \mathrm{Var}(X) = npq.$$

- Multinomial distribution $M(p_1, \ldots, p_r)$:
Multinomial distribution is a simple generalization of binomial distribution, in which each trial results in exactly one of some fixed number r possible outcomes with probability $p_1, p_2, \ldots, p_r$, where
$$\sum_{i=1}^{r} p_i = 1, \quad 0 \le p_i \le 1, \ i = 1, \ldots, r,$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

and we have n independent trials. Let the random variables X_i indicate the number of times the i-th outcome was observed over the n trials. $X = (X_1, \ldots, X_r)$ follows a multinomial distribution with parameters n and p, where $p = (p_1, \ldots, p_r)$.

The pmf of the multinomial distribution is:

$$P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}, \quad n = x_1 + \cdots + x_r.$$

The mean, variance and covariance are

$$\mathbb{E}(X_i) = np_i, \quad \mathrm{Var}(X_i) = np_i(1 - p_i), \quad \mathrm{Cov}(X_i, X_j) = -np_ip_j \ (i \neq j).$$

- Poisson distribution:

  The number X of radiated particles in a fixed time τ obeys

  $$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

  where λ is the average number of radiated particles each time. The mean and variance are

  $$\mathbb{E}X = \lambda, \mathrm{Var}(X) = \lambda.$$

  Poisson distribution may be viewed as the limit of binomial distribution (the law of rare events)

  $$C_n^k p^k q^{n-k} \longrightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad (n \to \infty, np = \lambda).$$

  Poisson distribution can also describe the spatial distribution of randomly scattered points. For example, Let A be a set in R^2. $X_A(\omega)$ be the number of points in A. If the points are uniformly distributed on the plane, and suppose the scattering density is λ (mean number of points per area), then X_A has Poisson distribution

  then

  $$\lambda = \text{ area of } A \times \text{ number of points/area.}$$

  X_A has Poisson distribution

  $$P(X_A = n) = \frac{(\lambda \cdot \mathrm{meas}(A))^n}{n!} e^{-\lambda \cdot \mathrm{meas}(A)}.$$

- Geometric probability.

  Probability = Ratio of areas

  Special case of continuous examples — uniform distribution.

Example 1. *Maxwell-Boltzmann, Bose-Einstein, Fermi-Dirac statistics.*

Suppose there are n particles and N bins, where $N > n$.

1. *Given n bins, what is the probability that each bin has one particle? (Boson)*

2. *What is the probability that there exist n bins such that each bin has exactly one particle? (Fermion, Pauli exclusion principle)*

In statistical physics the classical particles are distinguishable. If they satisfy the Pauli exclusion principle, then they are subject to Maxwell-Boltzmann statistis. The quantum particles are indistinguishable. If they satisfy the Pauli exclusion principle, then they are subject to Fermi-Dirac statistis (Fermions). If they do not satisfy the Pauli exclusion principle, then they are subject to Bose-Einstein statistis (Bosons). Distinguishable particles that are subject to the exclusion principle do not occur in physics.

The whole picture is as follows:

	Distinguishable balls (classical)	*Undistinguishable balls (quantum)*
Without exclusion	N^n *(Maxwell-Boltzmann)*	C^n_{N+n-1} *(Bose-Einstein)*
With exclusion	P^n_N	C^n_N *(Fermi-Dirac)*

1.2 Continuous Examples

In continuous case, the function $p(x)$ is called the probability density function (pdf).

- Uniform distribution $\mathcal{U}[0,1]$:

$$p(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

The mean and variance are

$$\mathbb{E}X = \frac{1}{2}, \mathrm{Var}(X) = \frac{1}{12}.$$

- Exponential distribution:($\lambda > 0$)

$$p(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

The mean and variance are

$$\mathbb{E}X = \frac{1}{\lambda}, \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Waiting time for continuous time Markov process also has exponential distribution, where λ is the rate of the process.

- Normal distribution(Gaussian distribution)($N(0,1)$):

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

or more generally $N(\mu, \sigma)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean (expectation), σ^2 is the variance.

High dimensional case ($N(\mu, \Sigma)$)

$$p(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} e^{-(\boldsymbol{X}-\mu)^T \Sigma^{-1} (\boldsymbol{X}-\mu)}$$

where μ is the mean, Σ is a symmetric positive definite matrix, which is the covariance matrix of $\boldsymbol{X}$. $\det \Sigma$ is the determinant of Σ. More general high dimensional normal distribution is defined with characteristic functions $g(t) = \exp\left(i\mu \cdot \boldsymbol{t} - \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}\right)$.

Remark 1. *In 1D case, the normal distribution $N(np, npq)$ may be viewed as the limit of the Binomial distribution $B(n,p)$ when n is large. This is the famous De Moivre-Laplace limit theorem. It is a special case of the central limit theorem (CLT). Notice that*

$$\frac{B(n,p) - np}{\sqrt{npq}} \longrightarrow N(0,1) \text{ as } n \to \infty.$$

Remark 2. *In 1D case, the normal distribution $N(\lambda, \lambda)$ may be viewed as the limit of the Poisson distribution $Poisson(\lambda)$ when λ is large. Notice the simple fact that the sum of two independent $Poisson(\lambda)$ and $Poisson(\mu)$ is $Poisson(\lambda + \mu)$ (why?), we can decompose $Poisson(\lambda)$ into the sum of n i.i.d. $Poisson(\lambda/n)$, we have*

$$\frac{Poisson(\lambda) - \lambda}{\sqrt{\lambda}} \longrightarrow N(0,1) \text{ when } \lambda \text{ is large.}$$

1.3 Probability Space

- σ-algebra $\mathcal{F}$

  $\mathcal{F}$ is a collection of subsets of Ω:

  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \backslash A \in \mathcal{F}$;
  3. If $A_1, A_2, \cdots, A_n, \cdots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

  Here $(\Omega, \mathcal{F})$ is called a measurable space.

- Probability measure P

  1. (Positive) $\forall A \in \mathcal{F}, P(A) \geq 0$;
  2. (Countably additive) If $A_1, A_2, \cdots \in \mathcal{F}$, and they are disjoint, then $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$;
  3. (Normalization) $P(\Omega) = 1$.

- Probability space — Triplet $(\Omega, \mathcal{F}, P)$

  1. Random variable: a measurable function $X : \Omega \to R$.
  2. Distribution(or law): a probability measure μ on R defined for any set $B \subset R$ by
  $$\mu(B) = \text{Prob}(X \in B) = P\{\omega \in \Omega : X(\omega) \in B\}.$$
  3. Probability density function(pdf): an integrable function $p(x)$ on R such that for any set $B \subset R$,
  $$\mu(B) = \int_B p(x)dx.$$
  4. Mean (expectation):
  $$\mathbb{E}f(X) = \int_\Omega f(X(\omega))P(d\omega) = \int_R f(x)d\mu(x) = \int_R f(x)p(x)dx.$$
  5. Variance:
  $$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$
  6. p-th moment: $\mathbb{E}|X|^p$.
  7. Covariance:
  $$\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$
  8. Independence:
  $$\mathbb{E}f(X)g(Y) = \mathbb{E}f(X)\mathbb{E}g(Y).$$
  for all continuous functions f and g.

1.4 Notions of Convergence

Probability space $(\Omega, \mathcal{F}, P)$, $\{X_n\}$ — a sequence of random variables, μ_n — the distirbution of X_n. X — another random variable with distribution μ.

Definition 1. *(Almost sure convergence) X_n converges to X almost surely as $n \to \infty$, $(X_n \to X,\ a.s.)$ if*
$$P\{\omega \in \Omega, \quad X_n(\omega) \to X(\omega)\} = 1$$

Definition 2. *(Convergence in probability) X_n converges to X in probability if for any $\epsilon > 0$,*

$$P\{\omega | X_n(\omega) - X(\omega)| > \epsilon\} \to 0$$

as $n \to +\infty$.

Definition 3. *(Convergence in distribution) X_n converges to X in distribution ($X_n \xrightarrow{d} X$) (i.e. $\mu_n \rightharpoonup \mu$ or $\mu_n \xrightarrow{d} \mu$, weak convergence), if for any bounded continuous function f*

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X)$$

Definition 4. *(Convergence in L^p) If $X_n, X \in L^p$, and*

$$\mathbb{E}|X_n - X|^p \to 0.$$

If $p = 1$, that is convergence in mean; if $p = 2$, that is convergence in mean square.

Relation:

$$\text{Almost sure convergence} \underset{\text{subsequence}}{\rightleftharpoons} \text{Converge in probability} \longrightarrow \text{Converge in distribution}$$

$$\Uparrow$$

$$L^p \text{ convergence}$$

1.5 Conditional Expectation

Let X and Y be two discrete random variables with joint probability

$$p(i, j) = \mathbb{P}(X = i, Y = j).$$

The *conditional probability* that $X = i$ given that $Y = j$ is given by

$$p(i|j) = \frac{p(i, j)}{\sum_i p(i, j)} = \frac{p(i, j)}{\mathbb{P}(Y = j)}$$

if $\sum_i p(i, j) > 0$ and conventionaly taken to be zero if $\sum_i p(i, j) = 0$. The natural definition of the *conditional expectation* of $f(X)$ given that $Y = j$ is

$$\mathbb{E}(f(X)|Y = j) = \sum_i f(i)p(i|j). \tag{1}$$

The axiomatic definition of the conditional expectation $Z = E(X|\mathcal{G})$ is defined with respect to a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ as follows.

Definition 5 (Conditional expectation)**.** *For any random variable X with $\mathbb{E}|X| < \infty$, the condition expectation Z of X given $\mathcal{G}$ is defined as*

(i) Z is a random variable which is measurable with respect to $\mathcal{G}$;

(ii) for any set $A \in \mathcal{G}$,

$$\int_A Z(\omega)\mathbb{P}(d\omega) = \int_A X(\omega)\mathbb{P}(d\omega).$$

The existence of $Z = E(X|\mathcal{G})$ comes from the Radon-Nikodym theorem by considering the measure μ on $\mathcal{G}$ defined by $\mu(A) = \int_A X(\omega)\mathbb{P}(d\omega)$ (see [3]). One can easily find that μ is absolutely continuous with respect to the measure $\mathbb{P}|_{\mathcal{G}}$, the probability measure confined in $\mathcal{G}$. Thus Z exists and is unique up to the almost sure equivalence in $\mathbb{P}|_{\mathcal{G}}$.

Theorem 1 (Properties of conditional expectation)**.** *Suppose X, Y are random variables with $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$, $a, b \in \mathbb{R}$. Then*

(i) $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$

(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$

(iii) $\mathbb{E}(X|\mathcal{G}) = X$*, if X is $\mathcal{G}$-measurable*

(iv) $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$*, if X is independent of $\mathcal{G}$*

(v) $\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$*, if Y is $\mathcal{G}$-measurable*

(vi) $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$ *for the sub-σ-algebras $\mathcal{G} \subset \mathcal{H}$.*

Lemma 1 (Conditional Jensen's inequality)**.** *Let X be a random variable such that $\mathbb{E}|X| < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function such that $\mathbb{E}|\phi(X)| < \infty$. Then*

$$\mathbb{E}(\phi(X)|\mathcal{G}) \geq \phi(\mathbb{E}(X|\mathcal{G})). \tag{2}$$

The readers may be referred to [4] for the details of the proof.

For the conditional expectation of a random variable X with respect to another random variable Y, it is natural to define it as

$$\mathbb{E}(X|Y) := \mathbb{E}(X|\mathcal{G}) \tag{3}$$

where $\mathcal{G}$ is the σ-algebra $Y^{-1}(\mathcal{B})$ generated by Y.

To realize the equivalence between the abstract definition (3) and (1) when Y only takes finitely discrete values, we suppose the following decomposition

$$\Omega = \bigcup_{j=1}^{n} \Omega_j$$

and $\Omega_j = \{\omega : Y(\omega) = j\}$. Then the σ-algebra $\mathcal{G}$ is simply the sets of all possible unions of Ω_j. The measurability of conditional expectation $\mathbb{E}(X|Y)$ with respect to $\mathcal{G}$ means $E(X|Y)$

takes constant on each Ω_j, which exactly corresponds to $E(X|Y=j)$ as we will see. By definition, we have

$$\int_{\Omega_j} \mathbb{E}(X|Y)\mathbb{P}(d\omega) = \int_{\Omega_j} X(\omega)\mathbb{P}(d\omega) \tag{4}$$

which implies

$$\mathbb{E}(X|Y) = \frac{1}{\mathbb{P}(\Omega_j)} \int_{\Omega_j} X(\omega)\mathbb{P}(d\omega). \tag{5}$$

This is exactly $\mathbb{E}(X|Y=j)$ in (1) when $f(X)=X$ and X also takes discrete values.

The conditional expectation has the following important property as the optimal approximation in L^2 norm among all of the Y-measurable functions.

Proposition 1. *Let $g(Y)$ be any measurable function of Y, then*

$$\mathbb{E}(X - \mathbb{E}(X|Y))^2 \leq \mathbb{E}(X - g(Y))^2. \tag{6}$$

Proof. We have

$$\begin{aligned}\mathbb{E}(X-g(Y))^2 = \mathbb{E}(X-E(X|Y))^2 + \mathbb{E}(E(X|Y)-g(Y))^2 \\ + 2\mathbb{E}\Big[(X-E(X|Y)(E(X|Y)-g(Y))\Big].\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\Big[(X-\mathbb{E}(X|Y)(\mathbb{E}(X|Y)-g(Y))\Big] \\ =&\mathbb{E}\Big[\mathbb{E}\big[(X-\mathbb{E}(X|Y)(E(X|Y)-g(Y))|Y\big]\Big] \\ =&\mathbb{E}\Big[(\mathbb{E}(X|Y)-\mathbb{E}(X|Y))(E(X|Y)-g(Y))\Big] = 0\end{aligned}$$

by properties (ii),(iii) and (v) in Theorem 1. The proof is done. □

2 Characteristic Function

The *characteristic function* of a random variable X or its distribution μ is defined as

$$f(\xi) = \mathbb{E}e^{i\xi X} = \int_{\mathbb{R}} e^{i\xi x}\mu(dx). \tag{7}$$

Proposition 2. *The characteristic function has the following properties:*

1. $\forall \xi \in \mathbb{R}$, $|f(\xi)| \leq 1$, $f(\xi) = \overline{f(-\xi)}$, $f(0) = 1$;
2. *f is uniformly continuous on $\mathbb{R}$;*
3. *$f^{(n)}(0) = i^n\mathbb{E}X^n$ provided $\mathbb{E}|X|^n < \infty$.*

Proof. The proof of statements 1 and 3 are straightforward. The second statement is valid by

$$\begin{aligned}|f(\xi_1) - f(\xi_2)| &= |\mathbb{E}(e^{i\xi_1 X} - e^{i\xi_2 X})| = |\mathbb{E}(e^{i\xi_1 X}(1 - e^{i(\xi_2 - \xi_1)X}))| \\ &\leq \mathbb{E}|1 - e^{i(\xi_2 - \xi_1)X}|.\end{aligned}$$

Dominated convergence theorem concludes the proof. □

Example 2. *The characteristic functions of some typical distributions are as below.*

1. *Bernoulli distribution:* $f(\xi) = q + pe^{i\xi}$.
2. *Binomial distribution* $B(n, p)$*:* $f(\xi) = (q + pe^{i\xi})^n$.
3. *Poisson distribtion* $\mathcal{P}(\lambda)$*:* $f(\xi) = e^{\lambda(e^{i\xi} - 1)}$.
4. *Exponential distribution* $\mathcal{E}xp(\lambda)$*:* $f(\xi) = (1 - \lambda^{-1} i\xi)^{-1}$.
5. *Normal distribution* $N(\mu, \sigma^2)$*:* $f(\xi) = \exp\left(i\mu\xi - \frac{\sigma^2\xi^2}{2}\right)$.

The following important theorem gives an explicit characterization of the weak convergence of probability measures based on their characteristic functions, which is a key in proving central limit theorem later.

Theorem 2 (Lévy's continuity theorem)**.** *Let* $\{\mu_n\}_{n\in\mathbb{N}}$ *be a sequence of probability measures, and* $\{f_n\}_{n\in\mathbb{N}}$ *be their corresponding characteristic functions. Assume that*

1. f_n *converges everywhere on* $\mathbb{R}$ *to a limiting function* f*.*
2. f *is continuous at* $\xi = 0$*.*

Then there exists a probability distribution μ *such that* $\mu_u \xrightarrow{d} \mu$*. Moreover* f *is the characteristic function of* μ*.*

Conversely, if $\mu_n \xrightarrow{d} \mu$*, where* μ *is some probability distribution then* f_n *converges to* f *uniformly in every finite interval, where* f *is the characteristic function of* μ*.*

For a proof, see [4].

As in Fourier transforms, one can also define the inverse transform

$$\rho(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\xi x} f(\xi) d\xi.$$

An interesting question arises as to when this gives the density of a probability measure. To answer this we define

Definition 6. *A function f is called positive semi-definite if for any finite set of values $\{\xi_1, \ldots, \xi_n\}$, $n \in \mathbb{N}$, the matrix $(f(\xi_i - \xi_j))_{i,j=1}^n$ is positive semi-definite, i.e.*

$$\sum_{i,j} f(\xi_i - \xi_j) v_i \bar{v}_j \geq 0, \tag{8}$$

for any $v_1, \ldots, v_n \in \mathbb{C}$.

Theorem 3 (Bochner's Theorem)**.** *A function f is the characteristic function of a probability measure if and only if it is a positive semi-definite and continuous at 0 with $f(0) = 1$.*

Proof. We only gives the necessity part. Suppose f is a characteristic function, then

$$\sum_{i,j=1}^n f(\xi_i - \xi_j) v_i \bar{v}_j = \int_{\mathbb{R}} \Big| \sum_{i=1}^n v_i e^{i\xi_i x} \Big|^2 \mu(dx) \geq 0. \tag{9}$$

The sufficiency part is difficult and the readers may be referred to [4]. □

3 Generating function

For discrete R.V. taking integer values, the generating function has the central importance

$$G(x) = \sum_{k=0}^{\infty} P(k) x^k.$$

One immediately has the formula:

$$P(k) = \frac{1}{k!} G^{(k)}(x)\Big|_{x=0}.$$

Definition 7. *Define the convolution of two sequences $\{a_k\}$, $\{b_k\}$ as $\{c_k\} = \{a_k\} * \{b_k\}$, the components are defined as*

$$c_k = \sum_{j=0}^{k} a_j b_{k-j}.$$

Theorem 4. *Consider two independent R.V. X and Y with PMF*

$$P(X = j) = a_j, \quad P(Y = k) = b_k$$

and $\{c_k\} = \{a_k\} * \{b_k\}$. Suppose the generating functions are $A(x)$, $B(x)$ and $C(x)$, respectively, then the generating function of $X + Y$ is $C(x)$.

Some generating functions:

- Bernoulli distribution: $G(x) = q + px$.
- Binomial distribution: $G(x) = (q + px)^n$.
- Poisson distribution: $G(x) = e^{-\lambda + \lambda x}$.

4 Moment Generating Function and Cumulants

The moment generating function of a random variable X is defined for all values of t by

$$M(t) = \mathbb{E}e^{tX} = \begin{cases} \sum_x p(x)e^{tx}, & X \text{ is discrete-valued} \\ \int_{\mathbb{R}} p(x)e^{tx}dx, & X \text{ is continuous} \end{cases} \tag{10}$$

provided that e^{tX} is integrable. It is obvious $M(0) = 1$.

Once $M(t)$ can be defined, one can show $M(t) \in C^\infty$ in its domain and its relation to the nth moments

$$M^{(n)}(t) = \mathbb{E}(X^n e^{tX}) \text{ and } \mu_n := \mathbb{E}X^n = M^{(n)}(0),\ n \in \mathbb{N}. \tag{11}$$

This gives

$$M(t) = \sum_{n=0}^{\infty} \mu_n \frac{t^n}{n!}, \tag{12}$$

which tells why $M(t)$ is called the moment generating function.

Theorem 5. *Denote $M_X(t), M_Y(t)$ and $M_{X+Y}(t)$ the moment generating functions of random variables X, Y and $X+Y$, respectively. If X, Y are independent, we have*

$$M_{X+Y}(t) = M_X(t)M_Y(t). \tag{13}$$

The proof is straightforward.

The following moment generating functions of typical random variables can be obtained by direct calculations.

(a) Binomial distribution: $M(t) = (pe^t + 1 - p)^n$.

(b) Poisson distribution: $M(t) = \exp[\lambda(e^t - 1)]$.

(c) Exponential distribution: $M(t) = \lambda/(\lambda - t)$ for $t < \lambda$.

(d) Normal distribution $N(\mu, \sigma^2)$: $M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$.

The cumulant generating function $K(t)$ is defined based on $M(t)$ by

$$K(t) = \ln M(t) = \ln \mathbb{E}e^{tX} = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}. \tag{14}$$

With such definition, we have the cumulants $\kappa_0 = 0$ and

$$\kappa_n = K^{(n)}(0), \quad n \in \mathbb{N}. \tag{15}$$

The moment generating function is not so powerful as the characteristic function since the integrable condition is usually too strong for many random variables. Under similar consideration, we can also define another type of cumulant generating function $H(t)$ as

$$H(t) = \ln \mathbb{E}e^{itX} = \sum_{n=1}^{\infty} \kappa_n \frac{(it)^n}{n!}.$$

All of the definitions above can be extended to random vectors without difficulty. In this circumstance, we have

$$M(\boldsymbol{t}) = \mathbb{E}e^{\boldsymbol{t}\cdot\boldsymbol{X}}, \quad \boldsymbol{t} \in \mathbb{R}^d$$

and correspondingly the moments

$$\mu_{\boldsymbol{k}} = \mathbb{E}(X_1^{k_1} \cdots X_d^{k_d}) = \frac{\partial^{|\boldsymbol{k}|} M}{\partial t_1^{k_1} \cdots \partial t_d^{k_d}}(\boldsymbol{0}), \quad \boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d,$$

where $|\boldsymbol{k}| := \sum_{j=1}^d k_j$ is the order of multi-index $\boldsymbol{k}$. The relation between $M(\boldsymbol{t})$ and $\mu_{\boldsymbol{k}}$ is simply

$$M(\boldsymbol{t}) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_d=0}^{\infty} \mu_{\boldsymbol{k}} \frac{t_1^{k_1} \cdots t_d^{k_d}}{k_1! \cdots k_d!}. \tag{16}$$

The $K(\boldsymbol{t}), H(\boldsymbol{t})$ can be defined similarly, and the corresponding cumulants are defined by

$$\kappa_{\boldsymbol{k}} = \frac{\partial^{|\boldsymbol{k}|} K}{\partial t_1^{k_1} \cdots \partial t_d^{k_d}}(\boldsymbol{0}), \quad \boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d,$$

and

$$K(\boldsymbol{t}) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_d=0}^{\infty} \kappa_{\boldsymbol{k}} \frac{t_1^{k_1} \cdots t_d^{k_d}}{k_1! \cdots k_d!}.$$

It is straightforward to verify the relations

$$\mu_X = \kappa_X, \quad \mu_{XY} = \kappa_{XY} + \mu_X \mu_Y,$$
$$\mu_{XYZ} = \kappa_{XYZ} + \mu_X \kappa_{YZ} + \mu_Y \kappa_{XZ} + \mu_Z \kappa_{XY} + \mu_X \mu_Y \mu_Z,$$

and so on. The general relation between μ and κ for scalar X is left as an exercise.

For the multi-variate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ we obtain

$$M(\boldsymbol{t}) = \exp\left(\boldsymbol{\mu}\cdot\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}\right), \quad K(\boldsymbol{t}) = \boldsymbol{\mu}\cdot\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}. \tag{17}$$

Note that only the cumulants κ_n with order $n \leq 2$ survive for Gaussian distributions. This property can be utilized to prove the useful Wick's theorem (see Exercise 6).

The moment and cumulant generating functions have explicit meaning in statistical physics, in which

$$Z(\beta) = \mathbb{E}e^{-\beta E}, \quad F(\beta) = -\beta^{-1} \ln Z(\beta)$$

are called *partition function* and *Helmholtz free energy*, respectively. Here $\beta = (k_B T)^{-1}$ is the inverse temperature, which is just a physical constant. They can be connected to M and K by

$$Z(\beta) = M_X(-\beta), \quad F(\beta) = -\beta^{-1} K_X(-\beta)$$

if X is taken as E, the energy of the system.

5 Borel-Cantelli Lemma

Let $\{A_n\}$ be a sequence of events, $A_n \in \mathcal{F}$. Define

$$\begin{aligned} \limsup_{n\to\infty}(A_n) &= \{\omega \in \Omega, \quad \omega \in A_n \text{ infinitely often (i.o.)}\} \\ &= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \end{aligned}$$

Lemma 2. *(First Borel-Cantelli Lemma) If* $\sum_{n=1}^{\infty} P(A_n) < +\infty$*, then* $P(\limsup_{n\to\infty} A_n) = P\{\omega : \omega \in A_n, i.o.\} = 0$*.*

Proof. $P\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} \leq P\{\bigcup_{k=n}^{\infty} A_k\} \leq \sum_{k=n}^{\infty} P(A_k)$ for any n, but the last term goes to 0, as $n \to \infty$.

As an example of the application of this result, we prove

Lemma 3. *Let* $\{X_n\}$ *be a sequence of identically distributed (not necessarily independent) random variables, such that* $\mathbb{E}|X_n| < +\infty$*. Then*

$$\lim_{n\to\infty} \frac{X_n}{n} = 0 \qquad a.s.$$

The proof of this relies on another useful fact.

Lemma 4. *(Chebyshev Inequality) Let* X *be a random variable such that* $\mathbb{E}|X|^k < +\infty$*, for some integer* k*. Then*

$$P\{|X| > \lambda\} \leq \frac{1}{\lambda^k} \mathbb{E}|X|^k$$

for any positive constant λ*.*

Proof. For any $\lambda > 0$,

$$\begin{aligned} \mathbb{E}|X|^k &= \int_{-\infty}^{\infty} |x|^k d\mu \geq \int_{|X|\geq\lambda} |X|^k d\mu \\ &\geq \lambda^k \int_{|X|\geq\lambda} d\mu = \lambda^k P\{|X| \geq \lambda\}. \end{aligned}$$

Proof of Lemma 3. For any $\epsilon > 0$, define

$$\begin{aligned} A_n &= \{\omega \in \Omega : \left|\frac{X_n(\omega)}{n}\right| > \epsilon\} \\ \sum_n P(A_n) &= \sum_n P\{|X_n| > n\epsilon\} \\ &= \sum_n \sum_{k=n} P\{k\epsilon < |X_n| < (k+1)\epsilon\} \\ &= \sum_k k P\{k\epsilon < |X_n| < (k+1)\epsilon\} \\ &\leq \frac{1}{\epsilon}\mathbb{E}|X| < +\infty \end{aligned}$$

Therefore if we define

$$B_\epsilon = \{\omega \in \Omega, \qquad \omega \in A_n \text{ i.o.}\}$$

then $P(B_\epsilon) = 0$. Let $B = \bigcup_{n=1}^{\infty} B_{\frac{1}{n}}$. Then $P(B) = 0$, and

$$\lim_{n\to\infty} \frac{X_n(\omega)}{n} = 0, \quad \text{if } \omega \notin B.$$

Lemma 5. *(Second Borel-Cantelli Lemma) If $\sum_{n=1}^{\infty} P(A_n) = +\infty$, and A_n are mutually independent, then*

$$P\{\omega \in \Omega, \quad \omega \in A_n \text{ i.o.}\} = 1$$

6 Homeworks

- HW1. Prove the second Borel-Cantelli Lemma.
- HW2. Prove that if $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$ and X is independent of Y, then $X + Y \sim \mathcal{P}(\lambda + \mu)$.
- HW3. Suppose $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$ are two independent Poisson random variables and the sum $X + Y = N$ is fixed. Then the conditional distribution of X (or Y) is a Binomial distribution with parameter $n = N$ and $p = \lambda/(\lambda + \mu)$ (or $p = \mu/(\lambda + \mu)$).
- HW4. Prove the following statements:

  1. (Memoryless property of exponential distribution) Suppose $X \sim \mathcal{E}(\lambda)$, prove that

  $$\text{Prob}(X > s + t | X > s) = \text{Prob}(X > t) \quad \text{for all } s, t > 0.$$

  2. Let X be a random variable such that

  $$\text{Prob}(X > s + t) = \text{Prob}(X > s)\text{Prob}(X > t) \quad \text{for all } s, t > 0,$$

  prove that there exists $\lambda > 0$ such that $X \sim E(\lambda)$.

- HW5. (**Wick's theorem**) For multi-variate Gaussian random variables $(X_1, X_2, \ldots, X_n)$ with mean 0, utilize (17) and (16) to prove

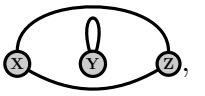
$$\mathbb{E}(X_1 X_2 \cdots X_k) = \begin{cases} \sum\prod \mathbb{E}(X_i X_j), & k \text{ is even}, \\ 0, & k \text{ is odd}, \end{cases}$$

where the notation $\sum\prod$ means summing of products over all possible partitions of $X_1, \ldots, X_k$ into pairs, e.g. for (X,Y,Z) is jointly Gaussian we obtain

$$\begin{aligned} \mathbb{E}(X^2Y^2Z^2) =& (\mathbb{E}X^2)(\mathbb{E}Y^2)(\mathbb{E}Z^2) + 2(\mathbb{E}YZ)^2\mathbb{E}X^2 + 2(\mathbb{E}XY)^2\mathbb{E}Z^2 + 2(\mathbb{E}XZ)^2\mathbb{E}Y^2 \\ &+ 8(\mathbb{E}XY)(\mathbb{E}YZ)(\mathbb{E}XZ). \end{aligned} \tag{18}$$

Each term in (18) can be schematically mapped to some graph as below

$(\mathbb{E}XY)^2\mathbb{E}Z^2 \longmapsto$  , $(\mathbb{E}YZ)^2\mathbb{E}X^2 \longmapsto$  ,

$(\mathbb{E}XZ)^2\mathbb{E}Y^2 \longmapsto$  , $(\mathbb{E}X^2)(\mathbb{E}Y^2)(\mathbb{E}Z^2) \longmapsto$  ,

$(\mathbb{E}XY)(\mathbb{E}YZ)(\mathbb{E}XZ) \longmapsto$  .

And the coefficient of each term is the combinatorial number for generating the corresponding schematic combinations. This is essentially the so-called Feynman diagrams.

- HW6. Suppose that the events A_n are mutually independent with $\text{Prob}(\cup_n A_n) = 1$ and $\text{Prob}(A_n) < 1$ for each n. Prove that $\text{Prob}\{A_n \ i.o.\} = 1$.

- HW7. Numerically investigate the limit process

$$\text{Binomial} \longrightarrow \text{Poisson} \longrightarrow \text{Normal distribution}$$

with MATLAB. Find the suitable parameter regime that the limit holds.

References

[1] Renguan Wang, An Introduction to Probability Theory, Peking University Press, 1998. (in chinese)

[2] B. Kisacanin, Mathematical problems and proofs: combinatorics, number theory, and geometry, Kluwer Academic Pub., New York, 2002.

[3] Billingsley. *Probability and measure*. John Wiley and Sons, New York, 1979.

[4] K.L. Chung. *A course in probability theory*. Academic Press, third edition, 2001.

Lecture 3 Generation of RVs *

Tiejun Li

1 Basic MC method

The MC method for integration is as follows:

$$I(f) = \int f(x)p(x)dx \quad \longrightarrow \quad I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \quad X_i \sim p(x) \;\; i.i.d.$$

From the WLLN, $I_N(f) \to I(f)$ in probability.

Problem: How to generate the random variables X_i? $(i = 1, \ldots, N)$

2 Generation of RVs

The first step to apply Monte Carlo method is to generate random variables. In computer simulations the random variables are replaced with pseudo-random variables for the reason of repeatability. We will show in the continued texts that the arbitrary distribution can be generated based on the uniform distributions. Let us start with generating the uniform distribution $\mathcal{U}[0, 1]$. We recommend [4] for the codes to be used in practice.

2.1 Uniform distribution

The most commonly used pseudo-random number generator (PRNG) for $\mathcal{U}[0, 1]$ is based on the linear congruential generator (LCG) and its different kinds of variants. It has the following simple form

$$X_{n+1} = aX_n + b(\text{mod } m) \tag{2.1}$$

where a, b and m are chosen natural numbers beforehand, and X_0 is the seed. The obtained sequence X_n/m is the desired pseudo-random number satisfies $\mathcal{U}[0, 1]$. The *period* for a typical sequence produced by the above recursion formula is defined as the length of the repeating cycle. It is proven in [1] that

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Theorem 2.1. *The period of a LCG is m if and only if*

(i) b and m are relatively prime;

(ii) every prime factor of m divides $a-1$;

(iii) if $m \big| 4$, then $(a-1) \big| 4$.

To achieve the goal of full period, a good choice in computer implementation is $m = 2^k$, $a = 4c+1$, and b is odd.

The LCG is also discussed when $b = 0$. In 1969, Lewis, Goodman and Miller proposed the following pseudo-random number generator

$$X_{n+1} = aX_n \pmod{m},$$

with $a = 7^5 = 16807, m = 2^{31} - 1$. This generator has passed all new theoretical tests in subsequent years, and resulted in a lot of successful use. They called it "Minimal standard generator" against which other generators should be judged. It is implemented in the function `ran0()` in the book *Numerical Recipes* [4]. The period of `ran0()` is about 2.1×10^9. With shuffling algorithm by combining sequences with different periods, a more powerful pseudo-random number generator `ran2()` with period about 2.3×10^{18} is constructed. The authors claim that they will pay $1,000 for the first person who may convince them by finding a statistical test that this generator fails in a nontrivial way!

More general LCG generators take the following form

$$X_{n+1} = a_0 X_n + a_1 X_{n-1} + \cdots + a_j X_{n-j} + b \pmod{m}.$$

These generators are characterized by the period τ, which in the best case can not proceed $m^{j+1} - 1$. The length of τ depends on the choice of a_j, b and m.

One important fact about the LCG is that it shows very poor distributions of s-tuples, i.e. the vectors $(X_n, X_{n+1}, \ldots, X_{n+s-1})$. In [2], Marsaglia proved the important fact

Theorem 2.2. *The s-tuples $(X_n, X_{n+1}, \ldots, X_{n+s-1})$ obtained via (2.1) lie on a maximum of $(s!m)^{\frac{1}{s}}$ equidistant parallel hyperplanes within the s-dimensional hypercube $(0,1)^s$.*

When s is large, the deviation with respect to the uniform distribution is apparent. Though the LCG has this drawback, it is still the most widely used pseudo random number generator in practice. The nonlinear generators are also discussed to overcome this limitation. Some very recent mathematical softwares adopt the so-called *Mersenne Twister* generator, which avoids the linear congruential steps and has the period up to $2^{19937} - 1$ [3].

We remark here that since the generation of RVs are essential for the success of the algorithm, one must use the reliable RV generators from available well-accepted codes or libraries!

2.2 Statistical testing

It is very difficult to distinguish whether a given sequence is generated from deterministic methods or stochastic methods. The practical way to handle this issue is to judge whether it can pass the corresponding statistical testing if the sequence is assumed to be random. That is the principle under which the pseudo-random number generator works. Below we show some of the statistical testing strategies for uniform distribution $\mathcal{U}[0,1]$. One may be referred to the book [1] or the document

https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-22r1a.pdf

maintained by the National Institute of Standards and Technology for more details on the empirical tests for PRNG.

- Equi-distribution test: The interval $(0,1)$ is divided into K subintervals. The number N_j of points falling into the j-th interval is then determined from a sample $\{X_1, \ldots, X_N\}$. A χ^2-test is performed where the expected number in each subinterval is N/K.

- Serial test: Consider the s-vector

$$\boldsymbol{X}_n = (X_n, X_{n+1}, \ldots, X_{n+s-1})$$

  in s-dimensional space ($s > 2$). The s-hypercube is divided into r^s equi-partitions and the frequency of the samples falling in each sub-partition is measured. Similarly a χ^2-test is applied to the sample sequences.

- Run test: Consider a short sequence $X_{n-1} > X_n < X_{n+1} > X_{n+2} < X_{n+3}$. We have a run-up of length 1 followed by two run-ups of length 2 since it has 3 increasing subsequences $X_{n-1}|X_n, X_{n+1}\,|X_{n+2}, X_{n+3}$. For a sequence of pseudo-random numbers, we can count the number of run-ups of length 1, length 2, ... and denote them by R_1, R_2, etc. It can be shown that $\{R_k\}$ is normally distributed in large sample size. Various statistical tests can be used to test such distributions.

2.3 Inverse Transformation Method

The general random variables $Y \in \mathbb{R}$ can be generated from $\mathcal{U}[0,1]$ in principle based on the following well-known proposition.

Proposition 2.3 (Inverse Transformation Method)**.** *Suppose the distribution function of Y is $F(y)$, i.e. $\mathbb{P}(Y \leq y) = F(y)$, which is strictly increasing. $X_i \sim \mathcal{U}[0,1]$, then $Y_i := F^{-1}(X_i)$ is the desired random variables.*

The geometric interpretation of the above proposition is clear from the following figure. When there are two sharp peaks at $Y = y_1$ and y_2 in the pdf of Y, the corresponding distribution function of Y will exhibit two sharp increase near y_1 and y_2. Thus the projection of $F(y)$ onto the vertical segment $[0, 1]$ has large portions near $F(y_1)$ and $F(y_2)$. The inverse transformation from $\mathcal{U}[0, 1]$ gives the desired distribution.

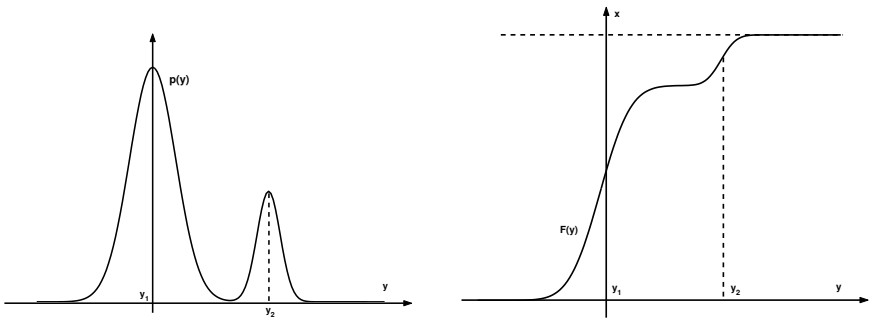


Figure 1: Left panel: The pdf of Y. Right panel: The distribution function $F(y)$.

Proof: If $F(y)$ is strictly increasing, we have the following simple proof for any $y \in \mathbb{R}$

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F^{-1}(X) \leq y) = \mathbb{P}(X \leq F(y)) = F(y).$$

When there are atoms in the distribution of Y or some parts have zero probability, the distribution function $F(y)$ is only non-decreasing and right continuous. In this case we should define the generalized inverse F^- of F as

$$F^-(u) = \inf\{x : F(x) \geq u\}.$$

With this definition, we have for any $u \in [0, 1]$ and for any $x \in F^-([0, 1])$ (the real domain), the generalizied inverse satisfies

$$F(F^-(u)) \geq u \quad \text{and} \quad F^-(F(x)) \leq x.$$

Thus

$$\{(u, x) : F^-(u) \leq x\} = \{(u, x) : F(x) \geq u\}$$

and

$$\mathbb{P}(F^-(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Some straightforward applications of the inverse transformation method are as follows.

- Generation of $\mathcal{U}[a, b]$:

The distribution function

$$F(y) = \frac{y-a}{b-a}, \qquad y \in [a,b],$$

then $F^{-1}(x) = (b-a)x + a$, so we can take $X_i \sim \mathcal{U}[0,1]$, $Y_i = (b-a)X_i + a$.

- Exponential distribution:

  The distribution function

  $$F(y) = 1 - e^{-\lambda y}$$

  then $F^{-1}(x) = -\ln(1-x)/\lambda,\ x \in [0,1]$, so we can take

  $$Y_i = -\frac{1}{\lambda}\ln X_i, \qquad (i = 1, 2, \ldots)$$

  where $X_i \sim \mathcal{U}[0,1]$ since $1 - X_i \sim \mathcal{U}[0,1]$ also.

Now let us investigate the possibility of generating $N(0,1)$ via inverse transformation method. We have

$$F(x) = \int_{-\infty}^{x} p(y)dy = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\Big(\frac{x}{\sqrt{2}}\Big).$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the error function. So $F^{-1}(x) = \sqrt{2}\mathrm{erf}^{-1}(2x-1)$. It is difficult to implement with this formula since it involves the solution of transcendental equations !

2.4 Box-Muller method for Gaussian RVs

A nice idea to generate Gaussian RVs is by the following Box-Muller method. The basic approach is though measure transformation on a lifted high dimensional space. Consider a two dimensional Gaussian distributed vector with independent components. With polar coordinates $x = r\cos\theta$, $y = r\sin\theta$, we have

$$\frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}dx_1dx_2 = \Big(\frac{1}{2\pi}d\theta\Big)\cdot(e^{-\frac{r^2}{2}}rdr).$$

So we transform the generation of a 2D Gaussian into the generation of Θ and R. Here the measure $1/2\pi d\theta$ corresponds to $\mathcal{U}[0,2\pi]$ in θ space, and $e^{-\frac{r^2}{2}}rdr$ corresponds to the distribution in r-direction with $F(r) = \int_0^r e^{-\frac{s^2}{2}}sds = 1 - e^{-\frac{r^2}{2}}$. $F^{-1}(r)$ is easy to be obtained and we obtain the method to generate Gaussian RV

$$Z_i = R_i\cos\Theta_i,$$

where $R_i = \sqrt{-2\ln X_i}$, $\Theta_i = 2\pi Y_i$ and $X_i, Y_i \sim \mathcal{U}[0,1] i.i.d.$.

Remark 2.4. *Another approximately generating Gaussian random variable is by central limit theorem*

$$X_n = \sqrt{12/N}\Big(\sum_{k=1}^{N} \xi_k - \frac{N}{2}\Big)$$

where $\xi_k \sim \mathcal{U}([0,1])$ *i.i.d.. The CLT asserts that* $N = 12$ *is sufficiently large for many purposes.*

2.5 Composition of random variables

Some distributions can be obtained by the composition of simple random variables instead of the direct application of the previous principles. Here are some examples.

- Sampling the hat pdf.

  Suppose the pdf is

  $$f(z) = \begin{cases} z, & 0 < z < 1, \\ 2 - z, & 1 \le z < 2. \end{cases}$$

  It is interesting to observe that Z has the same distribution with $X + Y$, where X and Y are i.i.d. with distribution $\mathcal{U}[0,1]$. This suggests that sampling Z can be obtained by the summation of two uniform RVs ξ_1 and ξ_2.

- Sampling a random variable raised to a power.

  Let $X_1, \ldots, X_n$ be drawn i.i.d. from the CDF $F_1(x_1), \ldots, F_n(x_n)$. If we set Z to be the largest number among X_i, i.e.

  $$Z = \max\{X_1, \ldots, X_n\}. \tag{2.2}$$

  Then the CDF of Z will be $F(z) = \prod_{i=1}^{n} F_i(z)$. Suppose we want to generate $Z \sim p(z) = nz^{n-1}$, where $z \in [0,1]$. Then $F(z) = z^n$ and we can take X_i are $\mathcal{U}[0,1]$ RVs in (2.2).

- Sampling the mixture models.

  Suppose the pdf

  $$f(x) = \sum_{i=1}^{n} \alpha_i g_i(x), \quad \alpha_i \ge 0,\ g_i(x) \ge 0.$$

  We can rewrite it as

  $$f(x) = \sum_{i=1}^{n} \beta_i h_i(x), \quad \beta_i = \alpha_i \int g_i(x)dx, \quad h_i(x) = \frac{g_i(x)}{\int g_i(x)dx},$$

  so we have the relation

  $$\int h_i(x)dx = 1, \quad \sum_{i=1}^{n} \beta_i = 1.$$

The sampling of X can be obtained by first sample the index I according to the distribution $\{\beta_i\}_{i=1}^n$, and then sample X according to the pdf $h_I(x)$. The rationale behind this is simply by the definition of conditional probability.

2.6 Acceptance-Rejection method

Though the inverse transformation method gives one approach to generate arbitrary RVs in principle, we have found that it encounters difficulty in implementations if there is no closed form inverse of the CDF. Next we present acceptance-rejection method, which is another general methodology to sample arbitrary RVs.

The aim is to generate RV with density $0 \le p(x) \le d$, $a \le x \le b$. The idea is to lift the state space into a higher dimensional space as shown in Figure 2. Suppose we can sample a uniformly distributed two dimensional random variable (X, Y) in the shaded domain A, where

$$A := \{(x, y) \ : \ x \in [a, b], y \in [0, p(x)]\}.$$

The pdf is $\chi_A(x, y)$ and its X-marginal distribution

$$p_X(x) = \int_0^{p(x)} \chi_A(x, y)dy = \int_0^{p(x)} 1dy = p(x),$$

which is exactly the desired distribution. The uniform distribution in domain A can be easily obtained by the inheritance from the uniform distribution in $[a, b] \times [0, d]$. This naturally leads to the *Acceptance-Rejection* algorithm

Algorithm 2.5 (Acceptance-Rejection method)**.** *Generate $X \sim p(x)$.*

Step1. Generate $X_i \sim \mathcal{U}[a, b]$.

Step2. Generate a decision-making RV $Y_i \sim \mathcal{U}[0, d]$.

Step3. If $0 \le Y_i < p(X_i)$, accept; otherwise, reject.

Step4. Back to Step1.

For the unbounded random variables, we should introduce more general *comparison functions*. We draw a curve $f(x)$ which lies everywhere above the original distribution density function $p(x)$. This $f(x)$ is called comparison function. Suppose we can generate the uniform distribution in the two dimensional domain covered by $f(x)$, we can apply the acceptance-rejection strategy to reduce it to the uniform distribution in the domain covered by $p(x)$. Then the X-marginal distribution assures us the correct sampling. Now let us consider the generation of uniform RVs with the support covered by $f(x)$ in 2D.

Suppose we have

$$\int_{-\infty}^{\infty} f(x)dx = A$$

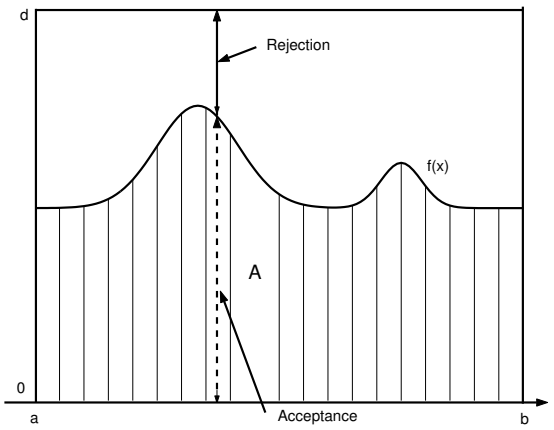


Figure 2: Schematics of the Acceptance-Rejection method

and we have the concrete form for $F^{-1}(x)$, where

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Then we consider the decomposition of uniform measure in $x \in (-\infty, \infty), y \in [0, f(x)]$

$$\frac{1}{A} f(x)dx \frac{1}{f(x)} dy.$$

This introduces a strategy for generating 2D uniform distribution by conditional sampling.

Algorithm 2.6 (Acceptance-Rejection method with general comparison function)**.** *Generate the unbounded $X \sim p(x)$.*

Step1. Generate $X_i = F^{-1}(AZ_i)$, where $Z_i \sim \mathcal{U}([0,1])$;

Step2. Generate decision-making RV $Y_i \sim \mathcal{U}[0, f(X_i)]$;

Step3. If $0 \leq Y_i < p(X_i)$, accept; otherwise, reject;

Step4. Back to Step1.

For bell-shaped random variables, the commonly used comparison function is the Cauchy distribution (or Lorentzian function) because of the slow decay when y is large

$$p(y) = \frac{1}{\pi(1+y^2)}.$$

One can check the first and second moments of the Cauchy distribution are both infinity though the principal integral of $p(y)$ is 0 because of symmetry. Since the standard deviation typically characterize the width of the "shoulder" near the center, the infinite second moment gives the reason why it is the usual candidate of comparison functions. Its inverse indefinite

integral is jus the tangent function. The comparison function is often chosen as the rescaled Cauchy function

$$f(x) = \frac{c_0}{1 + (x - x_0)^2/a_0^2} = c_0 p\Big(\frac{x - x_0}{a_0}\Big).$$

One can adjust the values of x_0, a_0 and c_0 such that it is everywhere greater than $p(x)$.

For the discrete random variables such as the Poisson and binomial distribution. one can extend it into a continuous distribution. With Poisson distribution as an example, we can extend it to $\mathbb{R}$ as

$$q(m) = \frac{x^{[m]} e^{-x}}{[m]!},$$

where $[m]$ represents the largest integer less than m. When x is large enough, we can take Cauchy function as the comparison function.

3 Homeworks

HW1. Familiarize the following functions in MATLAB.

mean, median, min, max, cov, hist

HW2. How many ways can you give to sample $\mathcal{U}(\mathbb{S}^2)$, the uniform distribution on the sphere surface $\mathbb{S}^2$. Implement them and make a comparison.

HW3. Derive the overall rejection probability of the Algorithm 2.6.

HW4. (Envelope Acceptance-Rejection) To generate the R.V. $X \sim p(x)$, we suppose that there exist bounds $g_l(x) \le p(x) \le M g_m(x)$, where $g_m(x)$ is a pdf, M is a positive constant and $g_l(x) \ge 0$ is a very simple function. Prove that the following algorithm

Step1: Generate $X \sim g_m(x)$, $U \sim \mathcal{U}[0, 1]$;

Step2: Accept X if $U \le g_l(X)/(M g_m(X))$;

Step3: Otherwise accept X if if $U \le p(X)/(M g_m(X))$.

generates X correctly and state its advantage compared with Algorithm 2.6.

References

[1] D.E. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley, third edition edition, 1998.

[2] Marsaglia. Random numbers fall mainly in the planes. *Proc. Nat. Acad. Sci. USA*, 61:25, 1968.

[3] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Mod. Comp. Simul.*, 8:3–30, 1998.

[4] W.T. Vetterling W.H. Press, S.A. Teukolsky and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, New York, Port Chester, Melbourne and Sydney, Second edition edition, 1995.

Lecture 4 Variance Reduction *

Tiejun Li

1 Necessity

The standard MC for computing $I(f) = \int_0^1 f(x)dx$ is

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1].$$

The mean square error

$$\mathbb{E}|e_N|^2 = \mathbb{E}(I_N(f) - I(f))^2 = \frac{1}{N}\mathrm{Var}\,(f), \tag{1.1}$$

where

$$\mathrm{var}(f) = \int_0^1 (f(x) - I(f))^2 dx.$$

If $\mathrm{var}(f) \gg 1$, the accuracy will be very poor!

2 Variance reduction.

We see from (1.1) that there are two factors that affect the error of Monte Carlo method: the sampling size N and the variance of f. N is clearly limited by the computational cost we are willing to afford. But the variance can be manipulated in order to reduce the size of the error.

The essence of variance reduction: to utilize some prior information about the integrand and try to extract the part which can be efficiently and accurately estimated through other ways.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

2.1 Importance Sampling

Consider for example the numerical evaluation of $\int_{-10}^{10} e^{-\frac{1}{2}x^2} dx$. Straightforward application of (1.1) would give

$$\int_{-10}^{10} e^{-\frac{1}{2}x^2} dx \approx \frac{20}{N} \sum_{i=1}^{N} e^{-\frac{1}{2}X_i^2},$$

where $\{X_i\}_{i=1}^N$ are i.i.d. random variables that are uniformly distributed on $[-10, 10]$. However notice that the integrand $e^{-\frac{1}{2}x^2}$ is an extremely non-uniform function, whose value is very small (and hence will have little contribution to the integral) everywhere except a small neighborhood of $x = 0$, most of the samples will be wasted in the region where the integrand is small. In other words, the uniform distribution ignores the importance of the integrand and thus the numerical quadrature is inefficient. The *importance sampling* embodies this idea by utilizing special distributions, which is schematically shown in Figure 1.

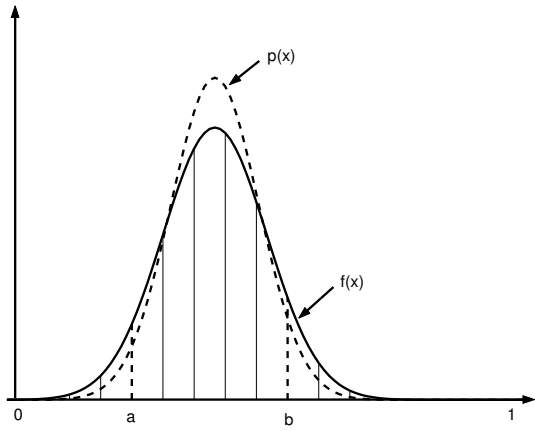


Figure 1: Schematics of importance sampling

Now if instead the $\{X_i\}$'s are distributed, differentially, say with density function $p(x)$, then we can use the fact that

$$\begin{aligned} \int f(x)dx &= \int \frac{f(x)}{p(x)} p(x)dx = \mathbb{E}\Big(\frac{f}{p}(X)\Big) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)} \end{aligned}$$

and approximate $\int f(x)dx$ by

$$I_N^p(f) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)},$$

where $X \sim g(x)$. The error can be estimated in the same way as before, and we get

$$\mathbb{E}(I(f) - I_N^p(f))^2 = \frac{1}{N}\mathrm{var}\Big(\frac{f}{p}\Big) = \frac{1}{N}\Big(\int \frac{f^2(x)}{p(x)}dx - I^2(f)\Big).$$

The Cauchy-Schwartz inequality shows

$$\Big(\int \frac{f(x)}{\sqrt{p(x)}}\sqrt{p(x)}dx\Big)^2 \le \int \frac{f^2}{p}dx \int p(x)dx$$

and the equality holds iff $p(x) = cf(x)$. Now we get an ideal importance function

$$p(x) = Z^{-1}f(x)$$

if f is nonnegative, where Z is the normalization factor $Z = \int f(x)dx$. In this case

$$I(f) = I_N^p(f).$$

This is not a miracle since all the necessary work has gone into computing Z which was our original task.

Though the perfect importance function is a mission impossible, it shows the direction toward which the sampled distribution should be constructed. For the example discussed earlier, we can pick $p(x)$ that behaves as $e^{-\frac{1}{2}x^2}$ and at the same time can be sampled with a reasonable cost.

Now let us discuss a slight variant of above direct implementation of the importance sampling [2]. Suppose we are interested in evaluating

$$I = \int f(x)\pi(x)dx,$$

we can proceed as the following steps.

- Draw $X_1, \ldots, X_n$ i.i.d. from a distribution $g(x)$.
- Calculate the importance weight

$$w_j = \frac{\pi(X_j)}{g(X_j)}, \quad \text{for } j = 1, 2, \ldots, n.$$

- Approximate the expectation by

$$\hat{I} = \frac{\sum_{i=1}^n w_i f(X_i)}{\sum_{i=1}^n w_i}. \tag{2.1}$$

Note that the expectation of $\hat{I}$ is not I, but we have by SLLN

$$\tilde{I} = \frac{1}{n}\sum_{i=1}^{n} w_i f(X_i) \to \mu \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} w_i \to 1$$

as $n \to \infty$. In this sense we call $\hat{I}$ a *biased estimator*. A major advantage of using (2.1) instead of $\tilde{I}$ is that in using the former, we *need only the ratio* $\pi(x)/g(x)$ *up to a multiplicative constant*, which is a usual case in statistics; whereas in the latter, the ratio needs to be known explicitly.

Example 2.1 (Toy example for importance sampling)**.** *Suppose we want to compute*

$$I = \int\int_{\mathcal{X}} f(x,y)dxdy,$$

where $\mathcal{X} = [-1,1] \times [-1,1]$ *and*

$$f(x,y) = 0.5\exp\Big(-90(x-0.5)^2 - 45(y+0.1)^4\Big) + \exp\Big(-45(x+0.4)^2 - 60(y-0.5)^2\Big).$$

The integrand resembles some renormalized Gaussian mixture distribution except the power 4 appearing in the first part for y variable. So the first step is to choose a suitable "Gaussian" to approximate the first part suitably. Here we take the trial distribution

$$g(x,y) \propto 0.5\exp\Big(-90(x-0.5)^2 - 10(y+0.1)^2\Big) + \exp\Big(-45(x+0.4)^2 - 60(y-0.5)^2\Big).$$

The reason that we take the number 10 before the y variable is as follows. Suppose we approximate $\exp(-10) \approx 0$, then from $45y^4 = 10$ we have the support radius for y is approximately $r = (10/45)^{1/4}$. With $kr^2 = 10$ we have $k = \sqrt{450} \gtrsim \mathcal{O}(10)$. A conservative choice may be $k = 10$. With the constraint $(x,y) \in \mathcal{X}$, it corresponds to a truncated mixture of Gaussian distribution

$$0.46N\left[\begin{pmatrix} 0.5 \\ -0.1 \end{pmatrix}, \begin{pmatrix} \frac{1}{180} & 0 \\ 0 & \frac{1}{20} \end{pmatrix}\right] + 0.54N\left[\begin{pmatrix} -0.4 \\ 0.5 \end{pmatrix}, \begin{pmatrix} \frac{1}{90} & 0 \\ 0 & \frac{1}{120} \end{pmatrix}\right].$$

We can sample $\boldsymbol{X}_n$ from this Gaussian mixture and compute the importance weight as

$$w_i = \frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)} \cdot \mathbf{1}_{\mathcal{X}}(\boldsymbol{X}_i).$$

One particular interesting specification of the importance sampling is the cross-entropy method [4]. Now suppose we want to compute

$$I(f) = \int f(x)\pi(x)dx$$

We assume $f \geq 0$. Then the perfect importance function will be $\mu(x) \propto f(x)\pi(x)$ but unable to sample in general. To compute a good sample average, one can assume a parameterized pdf with the form $\mu_u(x) = \mu(x; u)$ where u are the prescribed parameters. We choose u to minimize the cross-entropy (or Kullback-Leibler "distance", or relative entropy)

$$\min_u D(\mu||\mu_u) = \int \mu(x) \ln \frac{\mu(x)}{\mu_u(x)} dx.$$

Note the order matters here and it is important for the following derivations.

We have

$$\begin{aligned} D(\mu||\mu_u) &= \int \mu(x) \ln \mu(x) dx - \int \mu(x) \ln \mu_u(x) dx \\ &= \int \mu(x) \ln \mu(x) dx - \frac{1}{I(f)} \int f(x)\pi(x) \ln \mu_u(x) dx \end{aligned}$$

So minimizing cross-entropy is equivalent to maximize $F(x) = \int f(x)\pi(x) \ln \mu_u(x) dx$. The extremal point satisfies

$$\nabla F(x) = \int \frac{f(x)\pi(x)}{\mu_u(x)} \nabla_u \mu(x; u) dx = 0.$$

Solving this equation we obtain u^*, thus have a good candidate importance distribution μ_{u^*}.

The above argument is very useful for estimating the *rare events* such as the the small probability $p = \mathbb{P}(X \geq \gamma) = \mathbb{E}1_{\{X \geq \gamma\}}$. We have the relative error

$$\frac{\sqrt{\mathrm{V}ar(1_{\{X \geq \gamma\}})}}{I} = \sqrt{\frac{1-p}{p}} \gg 1 \quad \text{when} \quad p \ll 1.$$

One should introduce a multileveled version of the cross-entropy method to relax this issue with a step-by-step version.

2.2 Control Variates

Consider another form of $I(f)$

$$I(f) = \int f(\boldsymbol{x}) d\boldsymbol{x} = \int (f(\boldsymbol{x}) - g(\boldsymbol{x})) d\boldsymbol{x} + \int g(\boldsymbol{x}) d\boldsymbol{x}.$$

The idea of *control variates* is quite simple. If $g(\boldsymbol{x})$ is very similar to $f(\boldsymbol{x})$, and $I(g)$ is known or can be obtained in a highly accurate manner, then $\mathrm{var}(f - g) < \mathrm{var}(f)$, we will obtain a variance reduced estimator of $I(f)$. Similarly, an ideal control variates will be f itself, but we don't know $I(f)$! This is similar to the importance sampling. Though the perfect control variates is not practical, it tells us the direction toward which the approximate control variates should be constructed.

Another form of control variates is as follows. Suppose we have an unbiased estimator

$$U = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

for the integral $I(f)$, and we have another statistic V with known expectation $\mathbb{E}V = \mu$. Define a new static

$$\tilde{U} = U + c(V - \mu)$$

where c is to be determined. It is obvious that $\tilde{U}$ is also an unbiased estimator of $I(f)$. We have

$$\mathrm{Var}(\tilde{U}) = \mathrm{Var}(U) + c^2\mathrm{Var}(V) + 2c\mathrm{Cov}(U, V)$$

The optimal parameter for the minimization of the variance is

$$c^* = -\mathrm{Cov}(U, V)/\mathrm{Var}(V).$$

In this case

$$\mathrm{Var}(\tilde{U}^*) = (1 - \rho_{U,V}^2)\mathrm{Var}(U)$$

where $\rho_{U,V}$ is the correlation coefficient between U and V. So the more the introduced estimator V correlates with U, the more accurate the result will be. The constant c^* is usually computed from simulations in practice, e.g.

$$C_N^* = -\frac{\sum_{i=1}^{N}(U_i - \bar{U})(V_i - \bar{V})}{\sum_{i=1}^{N}(V_i - \bar{V})^2}.$$

There are also nonlinear version of control variates like

$$\bar{X} \cdot \frac{\mathbb{E}Y}{\bar{Y}} \quad \text{or} \quad \bar{X}\exp(\bar{Y} - \mathbb{E}Y)$$

in estimating $\mathbb{E}X$ through $\bar{X}$.

Example 2.2 (Toy example for control variates)**.** *Consider the following integral*

$$I(f) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}(1 + r)^{-1} e^{-\frac{x^2}{2}} dx,$$

where $r = e^{\sigma x},\ \sigma \gg 1$.

Notice that

$$(1 + r)^{-1} \approx h(x) = \begin{cases} 1, & x \le 0, \\ 0, & x > 0, \end{cases}$$

we have

$$I(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \big((1 + r)^{-1} - h(x)\big) e^{-\frac{x^2}{2}} dx + \frac{1}{2}.$$

Here $h(x)$ plays the role of control variates. Applying standard normal distribution can reduce the variance more.

2.3 Rao-Blackwellization

This method reflects a basic principle in Monte Carlo computation: *One should carry out analytical computation as much as possible.* Indeed this principle is also embodied in the idea of control variates. Suppose we have n independent samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn from pdf $\pi(\boldsymbol{x})$ and we are interested in evaluating $I = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}$. A straightforward estimator is

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i).$$

Suppose that $\boldsymbol{x}$ can be decomposed into two parts $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ and the conditional expectation $\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{x}^{(2)})$ can be obtained analytically or in a highly accurate manner. We can define another unbiased estimator of I as

$$\tilde{I} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}_i^{(2)}).$$

If the computational effort for obtaining the two estimates are similar, then $\tilde{I}$ should be preferred because of the variance identity [3]

$$\mathrm{var}(f(\boldsymbol{X})) = \mathrm{var}(\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)})) + \mathbb{E}(\mathrm{var}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)})), \tag{2.2}$$

which implies that

$$\mathrm{var}(\hat{I}) = \frac{\mathrm{var}(f(\boldsymbol{X}))}{m} \geq \frac{\mathrm{var}(\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)}))}{m} = \mathrm{var}(\tilde{I}).$$

The above procedure is called Rao-Blackwellization. The readers may be referred to [2] for more details.

2.4 Antithetic Variables

Proposition 2.3. *Suppose $X \sim \mathcal{U}[0,1]$, and $f(x)$ is monotone, then*

$$\mathrm{Cov}(f(X), f(1-X)) \leq 0.$$

Define

$$I_N(f) = \frac{1}{2N}\sum_{i=1}^{N}(f(X_i) + f(1-X_i)), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1],$$

then

$$\mathbb{E}I_N = I(f), \quad \mathrm{var}(I_N) = \frac{1}{2N}(\mathrm{var}(f) + \mathrm{Cov}(f(X), f(1-X))) \leq \frac{1}{2N}\mathrm{var}(f).$$

The variance is reduced!

2.5 Stratified Sampling

Consider two numerical integration strategies:

1. Monte Carlo: $I_N^{(1)}(f) = \frac{1}{N}\sum_{i=1}^N f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1].$
2. Midpoint rule: $I_N^{(2)}(f) = \frac{1}{N}\sum_{i=1}^N f(Y_i), \qquad Y_i = \frac{1}{2N} + \frac{i-1}{N}.$

The error estimate:

$$|e_N^{(1)}| \sim \mathcal{O}(\frac{1}{\sqrt{N}}), \quad |e_N^{(2)}| \sim \mathcal{O}(\frac{1}{N^2}).$$

The comparison of accuracy:

$$\text{Uniform} > \text{Quasi-random} > \text{Random}.$$

To improve the accuracy, one applies

$$\text{Uniform} + \text{Adaptive} \to \text{Moving Mesh}.$$

$$\text{Random} + \text{Adaptive} \to \text{Importance Sampling}.$$

Idea: If we combine the uniform and random sample points, we obtain the stratified sampling, and the accuracy will be improved.

Strategy: Divide $\Omega = [0,1]$ into M equi-partitions

$$\Omega_k = \Big[\frac{k-1}{M}, \frac{k}{M}\Big], \qquad k = 1, 2, \ldots, M.$$

Sample $N_k = N/M$ points uniformly in Ω_k, denoted as $X_i^{(k)},\ i = 1, \ldots, N_k$. Define

$$\bar{f}(x) = \bar{f}_k = |\Omega_k|^{-1} \int_{\Omega_k} f(x)dx = \mathbb{E}f(X^{(k)}), \quad x \in \Omega_k$$

and

$$I_N = \frac{1}{N}\sum_{k=1}^M \sum_{i=1}^{N_k} f(X_i^{(k)}), \tag{2.3}$$

we have

$$\begin{aligned}
\mathbb{E}I_N &= \frac{1}{N}\sum_{k=1}^M (N_k \cdot \bar{f}_k) = I(f), \\
\text{Var}\,(I_N) &= \frac{1}{N^2}\sum_{i,k}\sum_{j,l} \mathbb{E}\Big[\big(f(X_i^{(k)}) - \bar{f}_k\big)\big(f(X_j^{(l)}) - \bar{f}_l\big)\Big] \\
&= \frac{1}{N^2}\sum_{k=1}^M \Big(N_k \cdot |\Omega_k|^{-1} \int_{\Omega_k} (f(x) - \bar{f}_k)^2 dx\Big) \\
&= \frac{1}{N}\int_\Omega (f(x) - \bar{f}(x))^2 dx.
\end{aligned}$$

Proposition 2.4. *Define*

$$\sigma_s = \Big(\int_\Omega (f(x) - \bar{f}(x))^2 dx\Big)^{\frac{1}{2}},$$

then

$$\sigma_s \le \sigma = \Big(\int_\Omega (f(x) - I(f))^2 dx\Big)^{\frac{1}{2}}.$$

Proof. The quadratic function of c

$$g(c) = \int_{\Omega_k} (f(x) - c)^2 dx$$

takes minimum at $c = \bar{f}_k$, so we have

$$\sigma_s^2 = \sum_k \int_{\Omega_k} (f(x) - \bar{f}_k)^2 dx \le \sum_k \int_{\Omega_k} (f(x) - I(f))^2 dx = \sigma^2.$$

The variance is reduced! □

The stratified sampling can be combined with importance sampling.

Let $\Omega = \bigcup_{k=1}^M \Omega_k$, take N_k points $\{X_i^{(k)}\}_{i=1}^{N_k}$ in Ω_k, $\sum_{k=1}^M N_k = N$. Assume $\{X_i^{(k)}\}_{i=1}^{N_k} \sim$ *i.i.d.* $p^{(k)}(x) = p(x)/\bar{p}_k$, $x \in \Omega_k$, and $\bar{p}_k = \int_{\Omega_k} p(x)dx$, then

$$I_N = \sum_{k=1}^M \frac{\bar{p}_k}{N_k} \sum_{i=1}^{N_k} f(X_i^{(k)}).$$

Define

$$\bar{f}(x) = \bar{f}_k = \mathbb{E} f(X^{(k)}) = \bar{p}_k^{-1} \int_{\Omega_k} f(x)p(x)dx, \quad x \in \Omega_k,$$

we have

$$\mathbb{E} I_N = \sum_{k=1}^M \int_{\Omega_k} f(x)p(x)dx = I(f),$$

$$\mathrm{Var}\,(I_N) = \sum_{k=1}^M \frac{\bar{p}_k}{N_k} \int_{\Omega_k} (f(x) - \bar{f}_k)^2 p(x)dx = \sum_{k=1}^M \frac{\bar{p}_k}{N_k} \sigma_k^2,$$

where $\sigma_k^2 \triangleq \int_{\Omega_k} (f(x) - \bar{f}_k)^2 p(x)dx$.

Proposition 2.5. *If the balance condition $\bar{p}_k/N_k = \frac{1}{N}$ is satisfied, the variance is reduced.*

In a nutshell, the stratified sampling can be described as

$$\begin{aligned}\mathbb{E}Y &= \sum_{k=1}^K \mathbb{P}(Y \in A_k)\mathbb{E}(Y|Y \in A_k) = \sum_{k=1}^K p_k \mathbb{E}(Y|Y \in A_k) \\ &= \sum_{k=1}^K \frac{n_k}{n} \mathbb{E}(Y|Y \in A_k) \approx \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} Y_{kj}\end{aligned}$$

where $Y_{kj} \sim Y|Y \in A_k$, and $n_k = np_k$ which is enforced in the partition.

When the stratified sampling is applied to the realistic high dimensional problems, rather than attempt to stratify all the dimensions, it is better to identify which variables (if any) carry most of the variation of the integrand and stratify these. Significant reduction in the variance can sometimes be achieved by stratifying a single dimension in a many-dimensional integral.

3 Homeworks

- HW1. Prove the Proposition 2.3.
- HW2. Prove the identity (2.2).
- HW3. Prove that the relative entropy has the property that $D(f||g) \geq 0$ and $D(f||g) = 0$ if and only if $f(x) = g(x)$ for distribution density $f(x)$ and $g(x)$. Here we assume $g(x) > 0$ and take the convention that $0 \ln 0 = 0$.

References

[1] R.E. Caflish, Monte Carlo and Quasi-Monte Carlo methods, Acta Numerica, Vol. 7, 1-49, 1998.

[2] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer.

[3] R. Durrett, Probability: Theory and Examples, Wiley and Sons.

[4] P.-T. De Boer et al, A tutorial on the cross-entropy method, Ann. Oper. Res. 134 (2005), 19-67.

Lecture 5 Limit theorems *

Tiejun Li

1 Law of Large Numbers

Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independently and identically distributed (abbreviated as i.i.d. in the later text) random variables. Let $\eta = \mathbb{E}X_1$ and S_n the partial sum of X_j from 1 to n. The well-known law of large numbers validates the intuitive characterization of the mathematical expectation: it is the limit of empirical average when the sample size n goes to infinity. It is also the theoretical basis of the Monte Carlo methods.

Theorem 1.1 (Weak law of large numbers (WLLN)). *For i.i.d. random variables $\{X_j\}_{j=1}^{\infty}$ with $\mathbb{E}|X_j| < \infty$, we have*

$$\frac{S_n}{n} \to \eta \qquad \textit{in probability.}$$

Proving the result under the stated assumption is quite involved. We will give a proof of the WLLN under the stronger assumption that $\mathbb{E}|X_j|^2 < \infty$.

Proof. Without loss of generality, we can assume $\eta = 0$. Using Chebyshev's inequality, we have

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \le \frac{1}{\epsilon^2}\mathbb{E}\left|\frac{S_n}{n}\right|^2$$

for any $\epsilon > 0$. Using independence, we have

$$\mathbb{E}|S_n|^2 = \sum_{i,j=1}^{n} \mathbb{E}(X_i X_j) = n\mathbb{E}|X_1|^2.$$

Hence

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \le \frac{1}{n\epsilon^2}\mathbb{E}|X_1|^2 \to 0,$$

as $n \to \infty$. □

Theorem 1.2 (Strong law of large numbers (SLLN)). *For i.i.d. random variables $\{X_j\}_{j=1}^{\infty}$ we have*

$$\frac{S_n}{n} \to \eta \quad \textit{a.s.}$$

if and only if $\mathbb{E}|X_j| < \infty$.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Proof. We will only give a proof of SLLN here under the stronger assumption that $\mathbb{E}|X_j|^4 < \infty$. The proof in the most general condition may be referred to [2].

Without loss of generality, we can assume $\eta = 0$. Using Chebyshev's inequality, we obtain

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \frac{1}{\epsilon^4}\mathbb{E}\left|\frac{S_n}{n}\right|^4.$$

Using independence, we get

$$\mathbb{E}|S_n|^4 = \sum_{i,j,k,l=1}^{n} \mathbb{E}(X_i X_j X_k X_l) = n\mathbb{E}|X_j|^4 + 3n(n-1)(\mathbb{E}|X_j|^2)^2.$$

We have $(\mathbb{E}|X_j|^2)^2 \leq \mathbb{E}|X_j|^4 < \infty$ by Hölder inequality. Hence

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \frac{C}{n^2\epsilon^4}.$$

Since the series $1/n^2$ is summable we get

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon, \text{ i.o.}\right\} = 0$$

by Borel-Cantelli lemma. This implies that

$$\frac{S_n}{n} \to 0 \qquad \text{a.s.}$$

and we are done. □

Example 1.3 (Cauchy distribution)**.** *The following example shows that the law of large numbers does not hold if the assumed condition is not satisfied. Consider the i.i.d. random variables $\{X_j\}_{j=1}^{\infty}$ with Cauchy distribution having probability density function*

$$\frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}. \tag{1.1}$$

We have $\mathbb{E}X_j = 0$ by symmetry and $\mathbb{E}|X_j| = \infty, \mathbb{E}|X_j|^2 = \infty$. In this case, we can prove S_n/n always has the same distribution as X_1. Thus the weak and strong law of large numbers are both violated.

2 Central Limit Theorem

The following central limit theorem explains why the normal or normal-like distributions are so widely observed in the nature.

Theorem 2.1 (Lindeberg-Lévy central limit theorem (CLT))**.** *Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of i.i.d. random variables. Assume that $\mathbb{E}X_j^2 < \infty$ and let $\sigma^2 = \text{var}(X_j)$. Then*

$$\frac{S_n - n\eta}{\sqrt{n\sigma^2}} \to N(0,1)$$

in the sense of distribution.

Proof. Assume without loss of generality $\eta = 0$ and $\sigma = 1$, otherwise we can shift and rescale X_j. Let f be the characteristic function of X_1 and let g_n be the characteristic function of $S_n/\sqrt{n}$. Then

$$g_n(\xi) = \mathbb{E}e^{i\xi S_n/\sqrt{n}} = \prod_{j=1}^{n} \mathbb{E}e^{i\xi X_j/\sqrt{n\sigma^2}} = \prod_{j=1}^{n} f\left(\frac{\xi}{\sqrt{n}}\right) = f^n\left(\frac{\xi}{\sqrt{n}}\right).$$

Using Taylor expansion and the properties of characteristic functions we obtain

$$\begin{aligned} f\left(\frac{\xi}{\sqrt{n}}\right) &= f(0) + \frac{\xi}{\sqrt{n}} f'(0) + \frac{1}{2}\left(\frac{\xi}{\sqrt{n}}\right)^2 f''(0) + o\left(\frac{1}{n}\right) \\ &= 1 - \frac{\xi^2}{2n} + o\left(\frac{1}{n}\right) \end{aligned}$$

Hence

$$g_n(\xi) = f\left(\xi/\sqrt{n}\right)^n = \left(1 - \frac{\xi^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-\frac{1}{2}\xi^2} \qquad \text{as } n \to \infty$$

for every $\xi \in \mathbb{R}^1$. This completes the proof by using Levy's continuity theorem. □

The central limit theorem is the theoretical basis for the assumption that additive noise can be modeled by Gaussian noises. It also gives an estimate for the rate of convergence in the law of large numbers. Since by CLT we have

$$\frac{S_n}{n} - \eta \sim \frac{\sigma}{\sqrt{n}}.$$

The rate of convergence of S_n/n to η is $O(n^{-\frac{1}{2}})$. This is the reason why most Monte Carlo methods has a rate of convergence of $O(n^{-\frac{1}{2}})$ where n is the sample size.

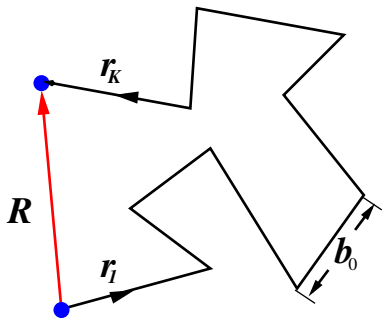


Figure 1: Schematics of the freely jointed chain.

Application in polymer physics. The central limit theorem is fundamental to understand the end-to-end statistics for a polymer. The simplest model for flexible polymers is called the freely jointed chain, in which a polymer consists of K units, each of length b_0 and able to point in any direction independently of each other (Figure 1). Denote the bond vectors as $\boldsymbol{r}_k$ $(k = 1, \ldots, K)$, which has i.i.d. distribution density

$$p(\boldsymbol{r}) = \frac{1}{4\pi b_0^2}\delta(\boldsymbol{r} - b_0).$$

The end-to-end vector

$$\boldsymbol{R} = \sum_{k=1}^{K} \boldsymbol{r}_k.$$

From the central limit theorem, we have asymptotically

$$\boldsymbol{R} \sim N(0, Kb_0^2\boldsymbol{I}), \quad K \gg 1.$$

Note that this Gaussian type approximation as $K \gg 1$ is independent of the choice of the bond vector distribution. This model is called *Gaussian chain* in polymer physics.

Remark 2.2 (Stable laws)**.** *Theorem 2.1 requires that the variance of X_j be finite. For variables with unbounded variances one can show the following. If there exists $\{a_n\}$ and $\{b_n\}$ such that*

$$\mathbb{P}\{a_n(S_n - b_n) \le x\} \to G(x) \quad \text{as } n \to \infty,$$

then the distribution $G(x)$ is stable. For details see [3].

3 Laplace asymptotics

Laplace method is the basis of large deviation theory. It is widely used in many fields of applied mathematics. We will only introduce the one-dimensional version of Laplace asymptotics in this section. For more details, see [1].

Let us consider the Laplace integral

$$F(t) = \int_{\mathbb{R}} e^{th(x)}dx, \quad t \gg 1$$

where $h(x) \in C^2(\mathbb{R})$, $h(0) = 0$ is the only global maximum such that

$$h(x) \le -b \quad \text{if} \quad |x| \ge c$$

for positive reals b, c. Suppose $h(x) \to -\infty$ fast enough as $x \to \infty$ to ensure the convergence of F for $t = 1$ and assume $h^{''}(0) < 0$, then the Laplace Lemma holds.

Lemma 3.1. *(Laplace method) As $t \to \infty$, to leading order*

$$F(t) \sim \sqrt{2\pi}(-th^{''}(0))^{-\frac{1}{2}}.$$

Proof. If $h(x) = h''(0)x^2/2, h^{''}(0) < 0$, then

$$\int_{\mathbb{R}} e^{th(x)}dx = \sqrt{2\pi}(-th^{''}(0))^{-\frac{1}{2}}.$$

In general, for any $\epsilon > 0$, there exists $\delta > 0$ such that for any $|x| \le \delta$,

$$|h(x) - \frac{h^{''}(0)}{2}x^2| \le \epsilon x^2.$$

It follows that

$$\int_{[-\delta,\delta]} \exp\Big(\frac{tx^2}{2}(h''(0) - 2\epsilon)\Big)dx \le \int_{[-\delta,\delta]} \exp\Big(th(x)\Big)dx \le \int_{[-\delta,\delta]} \exp\Big(\frac{tx^2}{2}(h''(0) + 2\epsilon)\Big)dx.$$

For this $\delta > 0$, there exists $\eta > 0$ by assumptions such that

$$h(x) \le -\eta \quad \text{if} \quad |x| \ge \delta,$$

thus

$$\int_{|x|\ge\delta} \exp\Big(th(x)\Big)dx \le e^{-(t-1)\eta} \int_{\mathbb{R}} e^{h(x)}dx \sim \mathcal{O}(e^{-\alpha t}),\ \alpha > 0, \quad \text{for} \quad t > 1.$$

First consider the upper bound, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\Big(th(x)\Big)dx \\
\le &\int_{\mathbb{R}} \exp\Big(\frac{tx^2}{2}(h''(0)+2\epsilon)\Big)dx - \int_{|x|\ge\delta} \exp\Big(\frac{tx^2}{2}(h''(0)+2\epsilon)\Big)dx + \mathcal{O}(e^{-\alpha t}) \\
= &\sqrt{2\pi}\Big[t(-h^{''}(0)-2\epsilon)\Big]^{-\frac{1}{2}} + \mathcal{O}(e^{-\beta t})
\end{aligned}$$

where $\beta > 0$. In fact, we ask $\epsilon < -h''(0)/2$ here.

The proof of lower bound is similar. By the arbitrary smallness of ϵ, we have

$$\lim_{t\to\infty} F(t)/\sqrt{2\pi}(-th^{''}(0))^{-\frac{1}{2}} = 1,$$

which completes the proof. □

The result is easily extended to the case where $h(0) \neq 0$. The term $e^{th(0)}$ will appear in the leading order and another commonly used form ignoring the prefactor is the so-called saddle point approximation

$$\lim_{t\to\infty} \frac{1}{t} \log F(t) = \sup_{x\in\mathbb{R}} h(x),$$

which is the typical form in large deviation theory and widely used in physics literature.

4 Cramér's Theorem for Large Deviations

Let $\{X_j\}_{j=1}^n$ be a sequence of i.i.d. random variables and let $\eta = \mathbb{E}X_j$. The laws of large numbers says that for any $\epsilon > 0$, with probability close to 1, $|S_n/n - \eta| < \epsilon$ for large enough n; conversely if $y \neq \eta$, then the probability that S_n/n is close to y goes to zero as $n \to \infty$. Events of this type, i.e. $\{|S_n/n - y| < \epsilon\}$, are called *large deviation events* compared with the *small deviation events* from the mean like the set $\{|S_n/n - \eta| \le c/\sqrt{n}\}$ in central limit theorem.

To estimate the precise rate at which $\mathbb{P}\{|S_n/n - y| < \epsilon\}$ goes to zero, we assume here that the distribution μ of the X_j's have finite exponential moments. Let us define the moment generating function

$$M(\lambda) = \mathbb{E}e^{\lambda X_j} = \int_{\mathbb{R}} e^{\lambda x} d\mu(x) < \infty, \quad \lambda \in \mathbb{R},$$

the cumulant generating function

$$\Lambda(\lambda) = \log M(\lambda) \tag{4.1}$$

and the Legendre-Fenchel transform of $\Lambda(\lambda)$ as

$$I(x) = \sup_{\lambda}\{x\lambda - \Lambda(\lambda)\}. \tag{4.2}$$

Then we have the large deviation type theorem for the i.i.d. sums.

Theorem 4.1 (Cramér's Theorem)**.** *The distribution of the empirical average μ_n defined by*

$$\mu_n(\Gamma) = \mathbb{P}\{S_n/n \in \Gamma\}$$

satisfies the large deviation principle:

(i) For any closed set $F \in \mathcal{B}$

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(F) \leq - \inf_{x\in F} I(x).$$

(ii) For any open set $G \in \mathcal{B}$

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x\in G} I(x).$$

$I(x)$ is called the rate function.

For the so-called I-continuity set Γ, i.e. $\inf_{x\in\Gamma^\circ} I(x) = \inf_{x\in\bar{\Gamma}} I(x)$, this theorem suggests that roughly

$$\mu_n(\Gamma) \asymp \exp\left(-n \inf_{x\in\Gamma} I(x)\right).$$

Here we use the notation "$\asymp$" instead of "$\approx$" since the equivalence is in the logarithmic scale. Before the proof, we need some results on the Legendre-Fenchel transform and some elementary properties of $I(x)$.

Lemma 4.2. *Suppose $f(x) : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is a lower semicontinuous convex function. The conjugate function $F(y)$ of $f(x)$ (Legendre-Fenchel transform) defined as*

$$F(y) = \sup_x\{(x, y) - f(x)\}$$

has the following properties:

(i) F is also a lower semicontinuous convex function.

(ii) Fenchel inequality holds

$$(x, y) \leq f(x) + F(y).$$

(iii) The conjugacy relation holds:

$$f(x) = \sup_y\{(x, y) - F(y)\}.$$

where we utilize the rule

$$\begin{aligned}
&\alpha + \infty = \infty,\ \alpha - \infty = -\infty \quad \textit{for } \alpha \textit{ finite}\\
&\alpha \cdot \infty = \infty,\ \alpha \cdot (-\infty) = -\infty, \quad \textit{for } \alpha > 0\\
&0 \cdot \infty = 0 \cdot (-\infty) = 0, \quad \inf\emptyset = \infty, \sup\emptyset = -\infty
\end{aligned}$$

The readers may be referred to [5, 6] for proof details.

Heuristic derivation of the rate function. Now we apply the Laplace asymptotics to explain heuristically why the rate function takes the interesting form in (4.2). Suppose the Cramér's theorem is already correct, then roughly we have

$$\mu(dx) \propto \exp(-nI(x))dx$$

and thus by Laplace asymptotics

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}^{\mu_n}(\exp(n\Phi(x))) := \lim_{n\to\infty}\frac{1}{n}\log\int_{\mathbb{R}}\exp(n\Phi(x))\mu_n(dx) = \sup_x\{\Phi(x)-I(x)\}. \tag{4.3}$$

Now we take $\Phi(x)=\lambda x$ then

$$\mathbb{E}^{\mu_n}(\exp(n\lambda x)) = \mathbb{E}\exp(\lambda\sum_{j=1}^{n}X_j) = [\mathbb{E}\exp(\lambda X_j]^n = \Big(M(\lambda)\Big)^n.$$

The equation (4.3) leads to

$$\Lambda(\lambda) = \sup_x\{\lambda x - I(x)\}.$$

By the conjugacy relation of Legendre-Fenchel transform, we obtain the rate function $I(x)$ as shown in (4.2).

Lemma 4.3. *The rate function $I(x)$ has the following properties:*

(i) $I(x)$ is convex and lower semicontinuous.

(ii) $I(x)$ is non-negative and $I(\eta)=0$.

(iii) $I(x)$ is non-decreasing in $[\eta,\infty)$ and non-increasing in $(-\infty,\eta]$.

(iv) If $x>\eta$, $I(x)=\sup_{\lambda>0}\{\lambda x-\Lambda(\lambda)\}$; If $x<\eta$, $I(x)=\sup_{\lambda<0}\{\lambda x-\Lambda(\lambda)\}$.

Proof. (i) The convexity of $\Lambda(\lambda)$ follows by Hölder's inequality. For any $0\le\theta\le 1$,

$$\begin{aligned}\Lambda(\theta\lambda_1+(1-\theta)\lambda_2) &= \log\mathbb{E}\Big(\exp(\theta\lambda_1X_j)\exp((1-\theta)\lambda_2X_j)\Big)\\ &\le \log\Big((\mathbb{E}\exp\left(\lambda_1X_j\right))^{\theta}\big(\mathbb{E}\exp(\lambda_2X_j)\big)^{(1-\theta)}\Big)\\ &= \theta\Lambda(\lambda_1)+(1-\theta)\Lambda(\lambda_2)\end{aligned}$$

Thus $\Lambda(\lambda)$ is a convex function. The rest is a direct application of Lemma 4.2.

(ii) Taking $\lambda=0$, we obtain $x\cdot 0-\Lambda(0)=0$. Thus $I(x)\ge 0$. On the other hand, we have

$$\Lambda(\lambda)=\log\mathbb{E}\exp(\lambda X_j)\ge\log\exp(\lambda\eta)=\lambda\eta$$

by Jensen's inequality. This gives $I(\eta)\le 0$. Combing with $I(x)\ge 0$ we get the result.

(iii) From the convexity of $I(x)$ and it achieves minimum at $x=\eta$, we immediately obtain the desired monotone property in $(-\infty,\eta]$ and $[\eta,\infty)$.

(iv) If $x>\eta$, then when $\lambda\le 0$

$$\lambda x-\Lambda(\lambda)\le\lambda\eta-\Lambda(\lambda)\le 0,$$

Thus the supremum is only achieved when $\lambda>0$ by the non-negativity of $I(x)$. Similar proof can be applied to the case $x<\eta$. □

Proof of Theorem 4.1. Without loss of generality, we assume $\eta = 0$.

(i) *Upper bound.* Suppose $x > 0$, $J_x := [x, \infty)$. For $\lambda > 0$,

$$\begin{aligned}\mu_n(J_x) = \int_x^\infty \mu_n(dy) &\le e^{-\lambda x} \int_x^\infty e^{\lambda y} \mu_n(dy) \\ &\le e^{-\lambda x} \int_{-\infty}^\infty e^{\lambda y} \mu_n(dy) = e^{-\lambda x} \Big[M(\frac{\lambda}{n}) \Big]^n.\end{aligned}$$

Taking $n\lambda$ instead of λ in the above equation, we obtain

$$\frac{1}{n} \log \mu_n(J_x) \le -(\lambda x - \Lambda(\lambda))$$

and accordingly

$$\frac{1}{n} \log \mu_n(J_x) \le -\sup_{\lambda > 0} \{\lambda x - \Lambda(\lambda)\} = -I(x).$$

If $x < 0$, we can define $\tilde{J}_x = (-\infty, x]$. Similarly as above we get

$$\frac{1}{n} \log \mu_n(\tilde{J}_x) \le -I(x).$$

For a closed set $F \in \mathcal{B}$, if $0 \in F$, $\inf_{x \in F} I(x) = 0$. Then the upper bound holds obviously. Otherwise, let (x_1, x_2) is the maximal interval satisfying the condition $(x_1, x_2) \cap F = \emptyset$ and $0 \in (x_1, x_2)$. So $x_1, x_2 \in F$, $F \subset \tilde{J}_{x_1} \cup J_{x_2}$. From monotonicity of $I(x)$ in $(-\infty, 0]$ and $[0, \infty)$, we obtain

$$\begin{aligned}\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(F) &\le \max \left(\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(\tilde{J}_{x_1}), \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(J_{x_2}) \right) \\ &\le -\min(I(x_1), I(x_2)) = -\inf_{x \in F} I(x).\end{aligned}$$

(ii) *Lower bound.* For any nonempty open set G, it is sufficient to prove that for any $x \in G$

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(G) \ge -I(x).$$

Now fix x and assume $I(x) < \infty$.

Case 1. If the supremum

$$I(x) = \sup_\lambda \{\lambda x - \Lambda(\lambda)\}$$

can not be achieved, then $x \neq 0$. Suppose $x > 0$ and there exists $\lambda_n \to \infty$ such that

$$I(x) = \lim_{n\to\infty} (\lambda_n x - \Lambda(\lambda_n)).$$

We have

$$\int_{-\infty}^{x-0} \exp(\lambda_n (y - x) \mu(dy) \to 0 \text{ as } n \to \infty.$$

by dominated convergence theorem. On the other hand

$$\lim_{n\to\infty} \int_x^\infty \exp(\lambda_n (y - x) \mu(dy) = \lim_{n\to\infty} \int_{-\infty}^\infty \exp(\lambda_n (y - x) \mu(dy) = \exp(-I(x)) < \infty.$$

Thus $\mu((x, \infty)) = 0$ and

$$\exp(-I(x)) = \lim_{n\to\infty} \int_x^\infty \exp(\lambda_n (y - x) \mu(dy) = \mu(\{x\}).$$

We have

$$\mu_n(G) \geq \mu_n(\{x\}) \geq (\mu(\{x\}))^n = \exp(-nI(x))$$

and thus

$$\frac{1}{n}\log \mu_n(G) \geq -I(x).$$

Similar proof can be applied to the case $x < a$.

Case 2. Suppose that the supremum is attained at λ_0 such that

$$I(x) = \lambda_0 x - \Lambda(\lambda_0).$$

Then $x = \Lambda'(\lambda_0) = M'(\lambda_0)/M(\lambda_0)$. Define a new probability measure as

$$\tilde{\mu}(dy) = \frac{1}{M(\lambda_0)}\exp(\lambda_0 y)\mu(dy).$$

It has the expectation

$$\int_{\mathbb{R}} y\tilde{\mu}(dy) = \frac{1}{M(\lambda_0)}\int_{\mathbb{R}} y\exp(\lambda_0 y)\mu(dy) = \frac{M'(\lambda_0)}{M(\lambda_0)} = x.$$

If $x \geq 0$, then $\lambda_0 \geq 0$. For sufficiently small $\delta > 0$, we have $(x-\delta, x+\delta) \subset G$,

$$\begin{aligned}
\mu_n(G) &\geq \mu_n(x-\delta, x+\delta) \\
&= \int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \mu(dy_1)\cdots\mu(dy_n) \\
&\geq \exp(-n\lambda_0(x+\delta)) \int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \exp(\lambda_0 y_1)\cdots\exp(\lambda_0 y_n)\mu(dy_1)\cdots\mu(dy_n) \\
&= \exp(-n\lambda_0(x+\delta))M(\lambda_0)^n \int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \tilde{\mu}(dy_1)\cdots\tilde{\mu}(dy_n).
\end{aligned}$$

By the WLLN, we have

$$\int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \tilde{\mu}(dy_1)\cdots\tilde{\mu}(dy_n) \to 1 \text{ as } n \to \infty.$$

Thus

$$\varliminf_{n\to\infty} \frac{1}{n}\log \mu_n(G) \geq -\lambda_0(x+\delta) + \Lambda(\lambda_0) = -I(x) - \lambda_0\delta \text{ for all } 0 < \delta \ll 1.$$

Similar proof can be applied to the case $x < a$. □

Example 4.4 (Cramér's theorem applied to the Bernoulli distribution with parameter p $(0 < p < 1)$)**.** *We have* $\Lambda(\lambda) = \ln(pe^{\lambda} + q)$ *where* $q = 1 - p$*. The rate function*

$$I(x) = \begin{cases} x\log\frac{x}{p} + (1-x)\ln\frac{1-x}{q}, & x \in [0,1], \\ \infty, & \textit{otherwise}. \end{cases} \tag{4.4}$$

Here we take the convention $0\log 0 = 0$*. It is obvious that* $I(x) \geq 0$*, and* $I(x)$ *achieves its global minimum* 0 *at* $x^* = p$*.* $I(x)$ *has important background in information theory. It is called relative entropy, or Kullback-Leibler distance between two distributions* μ *and* ν *defined as follows*

$$D(\mu||\nu) = \sum_{i=1}^{r} \mu_i \log\frac{\mu_i}{\nu_i}, \tag{4.5}$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_r)$, $\nu = (\nu_1, \nu_2, \ldots, \nu_r)$. *In the previous case, we have* $r = 2$, $\mu = (x, 1-x)$ *and* $\nu = (p, q)$, *the underlying Bernoulli distribution.*

Connections with statistical mechanics. There are intimate relations between the large deviation theory and equilibrium statistical mechanics [4]. Now let us only consider the simplest case here. For the Bernoulli trials with parameter p, we can obtain the rate function as (see Exercise 5)

$$I(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{q}, \ x \in [0,1]$$

which is also called the relative entropy. When $p = 1/2$ we have

$$I(x) = x \ln x + (1-x) \ln(1-x) + \ln 2, \ x \in [0,1].$$

In this case, the rate function is exactly the negative Shannon entropy up to a constant $\ln 2$. Below we will show that it has direct connection to Boltzmann entropy in statistical mechanics.

Consider a system with n independent spins being up or down with equal probability $1/2$. If it is up, we label it as 1, and 0 otherwise. We define the set of microstates as

$$\Omega = \{\omega : \omega = (s_1, s_2, \ldots, s_n), s_i = 1 \text{ or } 0\}.$$

For each microstate ω, we define its mean energy as

$$h_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

In thermodynamics, the entropy is a function of the macrostate energy. In statistical mechanics, Boltzmann gives a clear mathematical definition of the entropy

$$S = k_B \ln W \tag{4.6}$$

in the micro-canonical ensemble (the number of spins n and total energy $h_n = E$ are fixed in this set-up), where k_B is the Boltzmann constant, W is the number of the microstates corresponding to the fixed energy E. Actually this formula is carved in Boltzmann's tombstone. From large deviation theory we have

$$I(E) = \lim_{n \to \infty} -\frac{1}{n} \ln \mathbb{P}(h_n \in [E, E + dE]),$$

where dE is an infinitesimal quantity and

$$\begin{aligned} I(E) &= \lim_{n \to \infty} -\frac{1}{n} \ln \frac{W(h_n \in [E, E + dE])}{2^n} \\ &= \ln 2 - \frac{1}{k_B} \lim_{n \to \infty} \frac{1}{n} S_n(E). \end{aligned}$$

Taking the normalization of S in (4.6) with $1/n$ in the $n \to \infty$ limit, we obtain

$$k_B I(E) = k_B \ln 2 - S(E),$$

where $S(E)$ is the Boltzmann entropy in statistical mechanics. So we have that the rate function is the negative entropy (with factor $1/k_B$) up to an additive constant. This is a general statement.

In the canonical ensemble in statistical mechanics (the number of spins n and the temperature T are fixed in this set-up), let us investigate the physical meaning of Λ. The logarithmic moment generating function of $H_n(\omega) = nh_n(\omega)$ with normalization $1/n$ is

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E} e^{\lambda H_n},$$

where we take H_n instead of a single R.V. s_i since it admits more general interpretation. Take $\lambda = -\beta = -(k_B T)^{-1}$, we have

$$\Lambda(-\beta) = \lim_{n\to\infty} \frac{1}{n} \ln(\sum_{\omega} e^{-\beta H_n(\omega)}) - \ln 2.$$

Define the partition function

$$Z_n(\beta) = \sum_{\omega} e^{-\beta H_n(\omega)}$$

and free energy

$$F_n(\beta) = -\beta^{-1} \ln Z_n(\beta),$$

we have

$$\Lambda(-\beta) = -\beta \lim_{n\to\infty} \frac{1}{n} F_n(\beta) - \ln 2 = -\beta F(\beta) - \ln 2.$$

Thus the free energy $F(\beta)$ is the negative logarithmic moment generating function up to a constant.

According to the large deviation theory we have

$$-\beta F(\beta) - \ln 2 = \sup_E \{-\beta E - \ln 2 + k_B^{-1} S(E)\},$$

i.e.

$$F(\beta) = \inf_E \{E - TS(E)\}.$$

The infimum is achieved at the critical point E^* such that

$$\frac{\partial S(E)}{\partial E}\Big|_{E=E^*} = \frac{1}{T},$$

which is exactly a thermodynamic relation between S and T. Here E^* is essentially the internal energy U.

Exercises

1. Denote X_j the i.i.d. $\mathcal{U}[0,1]$ random variables. Prove that

$$\lim_{n\to\infty} \frac{n}{X_1^{-1} + \cdots + X_n^{-1}}, \quad \lim_{n\to\infty} \sqrt[n]{X_1 X_2 \cdots X_n}, \quad \lim_{n\to\infty} \sqrt{\frac{X_1^2 + \cdots X_n^2}{n}}$$

   exit almost surely and find their values.

2. The central limit of *i.i.d.* random variables as the Gaussian distribution can be understood from the following viewpoint. Denote $X_1, X_2, \ldots$ the *i.i.d.* random variables with mean 0. Suppose

$$Z_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} X \quad \text{and} \quad Z_{2n} = \frac{X_1 + \cdots + X_{2n}}{\sqrt{2n}} \xrightarrow{d} X. \tag{4.7}$$

   Denote the characteristic function of X is $f(\xi)$.

(a) Prove that $f(\xi) = f^2(\xi/\sqrt{2})$.

(b) Prove that $f(\xi)$ is the characteristic function of a Gaussian random variable under the condition $f \in C^2(\mathbb{R})$.

(c) Investigate the situation if the scaling $1/\sqrt{n}$ in (4.7) is replaced with $1/n$. Prove that X corresponds to the Cauchy-Lorentz distribution under the symmetry condition $f(\xi) = f(-\xi)$ or $f(\xi) \equiv 1$.

(d) If the scaling $1/\sqrt{n}$ is replaced with $1/n^\alpha$, what can we infer about the characteristic function $f(\xi)$ if we assume $f(\xi) = f(-\xi)$? What is the correct range of α?

3. Prove the assertion in the Example 1.3.

4. (Single-side Laplace lemma) Suppose that $h(x)$ attains the only maximum at $x = 0$, $h^{'} \in C^1(0, +\infty)$, $h^{'}(0) < 0$, $h(x) < h(0)$ for $x > 0$. $h(x) \to -\infty$ as $x \to \infty$, and $\int_0^\infty e^{h(x)}dx$ converges. To the leading order

$$\int_0^\infty e^{th(x)}dx \sim (-th^{'}(0))^{-1}e^{th(0)}$$

as $t \to \infty$.

5. Compute $I(x)$ for $N(\mu, \sigma^2)$ and the exponential distribution with parameter $\lambda > 0$.

References

[1] C.M. Bender and S.A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory.* Springer-Verlag, New York, Berlin and Heidelberg, 1999.

[2] K.L. Chung. *A course in probability theory.* Academic Press, third edition, 2001.

[3] R. Durrett. *Probability: Theory and Examples.* Cambridge University Press, Cambridge, fourth edition edition, 2010.

[4] R.S. Ellis. *Entropy, Large deviations, and statistical mechanics.* Springer-Verlag, 1st edition edition, 1985.

[5] R.T. Rockafellar. *Convex analysis.* Princeton University Press, Princeton, 1970.

[6] R.T. Rockafellar and R. J-B Wets. *Variational Analysis.* Springer-Verlag, Berlin and Heidelberg, 2009.

Lecture 6 Markov Chains *

Tiejun Li

Markov process is one of the most important stochastic processes in application. Roughly speaking, A Markov process is independent of the past, knowing the present state. In this lecture, we only consider the finite state Markov chain. The readers may be refereed to [2] for further information.

1 Markov Chains

Example 1. *(1D Random Walk) Let ξ_i are i.i.d. random variables such that $\xi_i = \pm 1$ with probability $\frac{1}{2}$, and let*

$$X_n = \xi_1 + \xi_2 + \ldots + \xi_n$$

$\{X_n\}$ represents a unconstrained unbiased random walk on $\mathbb{Z}$, the set of integers. Given $X_n = i$, we have

$$\begin{aligned} P\{X_{n+1} = i \pm 1 |\ X_n = i\} &= \frac{1}{2}, \\ P\{X_{n+1} = \textit{anything else} |\ X_n = i\} &= 0. \end{aligned}$$

We see that the distribution of X_{n+1} depends only on the value of X_n.

The result above can be restated as the *Markov property*

$$P\{X_{n+1} = i_{n+1} |\ \{X_m = i_m\}_{m=1}^n\} = P\{X_{n+1} = i_{n+1} |\ X_n = i_n\},$$

and the sequence $\{X_n\}_{n=1}^\infty$ is called a realization of a Markov process.

Example 2 (Ehrenfest's diffusion model)**.** *An urn contains a mixture of red and black balls. At each time $1, 2, \ldots$ a ball is picked at random from the urn and replaced by a ball of the other colour. The total number of balls in the urn is therefore a constant N, say. Let the state X_n of the system at time n be the number of black balls in the urn.*

As will be stated below, the one-step transition matrix can be given as

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{N} & 0 & \frac{N-1}{N} & 0 & 0 & 0 \\ 0 & \frac{2}{N} & 0 & \frac{N-2}{N} & 0 & 0 \\ 0 & 0 & \ddots & 0 & \ddots & 0 \\ 0 & \ldots & 0 & \frac{N-1}{N} & 0 & \frac{1}{N} \\ 0 & \ldots & 0 & 0 & 1 & 0 \end{pmatrix} \tag{1}$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Example 3. *(Finite state Markov chain) Suppose a Markov chain only takes a finite set of possible values, without loss of generality, we let the state space be* $\{1, 2, \ldots, N\}$. *Define the* **transition probabilities**

$$p_{jk}^{(n)} = P\{X_{n+1} = k | X_n = j\}$$

This uses the Markov property that the distribution of X_{n+1} *depends only on the value of* X_n.

Proposition 1. *(Chapman-Kolmogorov equation)*

$$P(X_n = j | X_0 = i) = \sum_k P(X_n = j | X_m = k) P(X_m = k | X_0 = i), \quad 1 \leq m \leq n-1.$$

Definition 1. *(Time-stationary, or time homogeneous) A Markov chain is called stationary if* p_{jk}^n *is independent of* n. *From now on we will discuss only stationary Markov chains and let* $P = (p_{jk})_{j,k=1}^N$. P *is called the transition probability matrix(TPM).*

Markov property implies that

$$P\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\} = (\mu_0)_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n}$$

where $(\mu_0)_{i_0}$ is defined by the intial distribution $(\mu_0)_{i_0} = P\{X_0 = i_0\}$.

From this we get

$$\begin{aligned} P\{X_n = i_n | X_0 = i_0\} &= \sum_{i_1, \ldots, i_{n-1}} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n} \\ &= (P^n)_{i_0 i_n} \end{aligned}$$

The last quantity denotes the (i_0, i_n)-th entry of the matrix P^n.

P is also called a *stochastic matrix*, in the sense that

$$p_{ij} \geq 0, \quad \sum_{j=1}^N p_{ij} = 1.$$

Given the initial distribution of the Markov chain μ_0, the distribution of X_n is then given by

$$\mu_n = \mu_0 P^n$$

Example 4. μ_n *satisfies the recurrence relation* $\mu_n = \mu_{n-1} P$. *This equation can also be rewritten as*

$$(\mu_n)_i = (\mu_{n-1})_i (1 - \sum_{j \neq i} p_{ij}) + \sum_{j \neq i} (\mu_{n-1})_j p_{ji}.$$

The interpretation is clear.

The following two questions are of special interest.

- Is there an invariant distribution? π is called an invariant distribution if

  $$\pi = \pi P$$

  This is equivalent to say that there exists a nonnegative left eigenvector of P with eigenvalue equal to 1. Notice that 1 is always an eigenvalue of P since it always has the right eigenvector $(1, \ldots, 1)^T$.

- When is the invariant distribution unique?

To answer these questions, it is useful to recall some general results on nonnegative matrices.

Definition 2. *(Reducibility) If there exists a permutation matrix Q such that*

$$QPQ^T = \begin{pmatrix} A_1 & B \\ 0 & A_2 \end{pmatrix}$$

then P is called reducible. Otherwise P is called irreducible.

Example 5. *(Graph representation of Markov chains) Any Markov chain can be sketched by their graph representation as in Figure 1. The arrows and real numbers show the transition probability of the Markov*

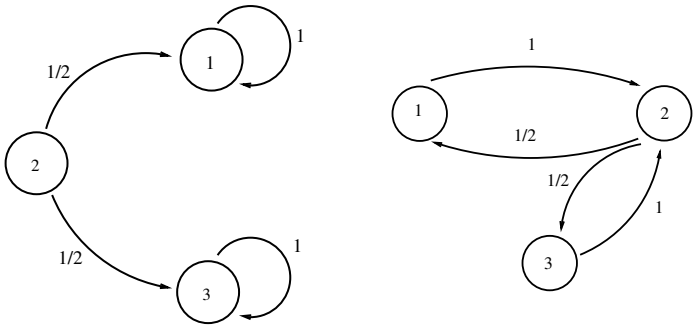


Figure 1: Graph representation of Markov chains. Left panel: chain 1, right panel: chain 2.

chain. The TPM corresponds to left panel is

$$P = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix},$$

It's quite clear that P is a reducible matrix, and it has two invariant distributions $\pi_1 = (1, 0, 0)$ and $\pi_2 = (0, 0, 1)$.

The TPM corresponds to the right panel is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}.$$

It's a irreducible matrix, and the only invariant distribution is $\pi = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

The following theorem is a key answer for invariant distribution of a Markov chain

Theorem 1. *(Perron-Frobenius) Let A be an irreducible nonnegative matrix, and let $\rho(A)$ be its spectral radius: $\rho(A) = \max_\lambda |\lambda|$, where λ is an eigenvalue of A. Then,*

1. *There exists a positive right eigenvector x of A, such that*

$$Ax = \rho(A)x$$

$x = (x_1, \ldots, x_N)^T, x_i > 0.$

2. *$\lambda = \rho(A)$ is an eigenvalue of multiplicity 1.*

Coming back to Markov chains, we obtain as a consequence of the Perron-Frobenius Theorem that

- If P is irreducible, then there exists exactly one invariant distribution.
- If P is reducible, then there are some cases that we can decompose the state space into ergodic components for the Markov chain. On each component there exists a unique in variant distribution. Arbitrary convex combinations of these invariant distributions on each component are invariant distributions for the whole chain. However in this case, the invariant distribution for the whole chain is clearly not unique. One typical example may be as follows:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0.3 & 0 & 0.4 & 0.3 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}.$$

  In this case, states $1, 2$ and $4, 5$ form two closed irreducible sub-chains, but P is reducible. There are infinite many invariant distributions. But reducibility itself is not a sufficient condition for the non-uniqueness of the invariant distribution, e.g.

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 & 0 \\ 0.3 & 0 & 0.4 & 0.3 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}.$$

  Though the invariant distribution has some zero components which are related to the transience of the states, it is unique.

Irreducibility is equivalent to the property that all nodes on the chain *communicate*, i.e. given any pair (i, j) we have

$$p_{ik_1} p_{k_1 k_2} \cdots p_{k_s j} > 0,$$

for some $(k_1, k_2, \ldots, k_s)$ (if there is only transition from $i \to k_1 \to \cdots \to k_j \to j$, we say that j is accessible from i).

The following theorem gives the asymptotic states of a Markov chain

Theorem 2. *Assume that for any pairs (i, j), there exists an s such that $(P^s_{i,j}) > 0$ (irreducible). Then*

1. *There exists a unique invariant distribution π. π is strictly positive.*
2. *For any μ_0,*

$$\pi_n = \mu_0 \bar{P}_n \to \pi \qquad \textit{exponentially fast as } n \to \infty,$$

   where

$$\bar{P}_n = \frac{1}{n} \sum_{j=1}^{n} P^j.$$

Remark 1. *A stronger assumption is "primitive" which says that there exist an natural number s, such that*

$$(P^s)_{ij} > 0, \quad \textit{for all } i, j$$

and a stronger convergence theorem $\mu_n = \mu_0 P^n \to \pi$ can be obtained. A critical example is that

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which is called a periodic chain. Actually we have primitive $\Leftrightarrow$ irreducible + aperiodic for finite Markov chains.

Theorem 3. *Assume that the Markov chain is primitive. Then for any initial distribution $\boldsymbol{\mu}_0$*

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \boldsymbol{P}^n \to \boldsymbol{\pi} \quad \textit{exponentially fast as } n \to \infty,$$

where $\boldsymbol{\pi}$ is the unique invariant distribution.

Proof. Given two distributions, $\boldsymbol{\mu}_0$ and $\tilde{\boldsymbol{\mu}}_0$, we define the *total variation distance* by

$$d(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\mu}}_0) = \frac{1}{2} \sum_{i \in S} |\mu_{0,i} - \tilde{\mu}_{0,i}|.$$

Since

$$0 = \sum_{i \in S} (\mu_{0,i} - \tilde{\mu}_{0,i}) = \sum_{i \in S} (\mu_{0,i} - \tilde{\mu}_{0,i})^+ - \sum_{i \in S} (\mu_{0,i} - \tilde{\mu}_{0,i})^- ,$$

where $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0)$. We also have

$$\begin{aligned} d(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\mu}}_0) &= \frac{1}{2} \sum_{i \in S} (\mu_{0,i} - \tilde{\mu}_{0,i})^+ + \frac{1}{2} \sum_{i \in S} (\mu_{0,i} - \tilde{\mu}_{0,i})^- \\ &= \sum_{i \in S} (\mu_{0,i} - \tilde{\mu}_{0,i})^+ \leq 1. \end{aligned}$$

Let $\boldsymbol{\mu}_s = \boldsymbol{\mu}_0 \boldsymbol{P}^s$, $\tilde{\boldsymbol{\mu}}_s = \tilde{\boldsymbol{\mu}}_0 \boldsymbol{P}^s$ and consider $d(\boldsymbol{\mu}_s, \tilde{\boldsymbol{\mu}}_s)$. We have

$$\begin{aligned} d(\boldsymbol{\mu}_s, \tilde{\boldsymbol{\mu}}_s) &= \sum_{i \in S} \left[\sum_{j \in S} \Big(\mu_{0,j} (\boldsymbol{P}^s)_{ji} - \tilde{\mu}_{0,j} (\boldsymbol{P}^s)_{ji} \Big) \right]^+ \\ &\leq \sum_{j \in S} \Big(\mu_{0,j} - \tilde{\mu}_{0,j} \Big)^+ \sum_{i \in B_+} (\boldsymbol{P}^s)_{ji}, \end{aligned}$$

where B_+ is the subset of indices where $\sum_{j \in S} (\mu_{0,j} - \tilde{\mu}_{0,j}) (\boldsymbol{P}^s)_{ji} > 0$. We note that B_+ cannot contain all the elements of S, otherwise one must have $(\boldsymbol{\mu}_0 \boldsymbol{P}^s)_i > (\tilde{\boldsymbol{\mu}}_0 \boldsymbol{P}^s)_i$ for all i, and

$$\sum_{i \in S} (\boldsymbol{\mu}_0 \boldsymbol{P}^s)_i > \sum_{i \in S} (\tilde{\boldsymbol{\mu}}_0 \boldsymbol{P}^s)_i,$$

which is impossible since both sides sum to 1. Therefore at least one element is missing in B_+. By assumption, there exists an $s > 0$ and $\alpha \in (0, 1)$ such that $(\boldsymbol{P}^s)_{ij} \geq \alpha$ for all pairs (i, j). Hence $\sum_{i \in B_+} (\boldsymbol{P}^s)_{ji} \leq (1 - \alpha) < 1$. Therefore

$$d(\boldsymbol{\mu}_s, \tilde{\boldsymbol{\mu}}_s) \leq d(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\mu}}_0)(1 - \alpha),$$

i.e. the Markov chain is contractive after every s steps. Similarly for any $m \geq 0$

$$d(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n+m}) \leq d(\boldsymbol{\mu}_{n-sk}, \boldsymbol{\mu}_{n+m-sk})(1-\alpha)^k \leq (1-\alpha)^k,$$

where k is the largest integer such that $n - sk \geq 0$. If n is sufficiently large the right hand side can be made arbitrarily small. Therefore the sequence $\{\boldsymbol{\mu}_n\}_{n=0}^{\infty}$ is a Cauchy sequence. Hence it has to converge to a limit π, which satisfies

$$\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\mu}_0 \boldsymbol{P}^{n+1} = \lim_{n\to\infty} (\boldsymbol{\mu}_0 \boldsymbol{P}^n)\boldsymbol{P} = \boldsymbol{\pi}\boldsymbol{P}.$$

Such a $\boldsymbol{\pi}$ satisfying such a property is also unique. For if there were two such distributions, $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$, then $d(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = d(\boldsymbol{\pi}^{(1)}\boldsymbol{P}^s, \boldsymbol{\pi}^{(2)}\boldsymbol{P}^s) < d(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)})$. This implies $d(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = 0$, i.e $\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(2)}$. □

Remark 2. *We do not discuss the convergence speed here. But in fact it is exponential, which depends on the spectral gap of the transition probability matrix P. The readers may be referred to [3, 4].*

Theorem 4 (Ergodic theorem)**.** *let X_n be an irreducible, positive recurrent Markov chain with invariant distribution $\pi(x)$, and f be a bounded function, then*

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \to \langle f \rangle_\pi, \quad a.s.$$

1.1 Time Reversal

Theorem 5. *Assume that the Markov chain $\{X_n\}_{n\geq 0}$ admits a unique invariant distribution $\boldsymbol{\pi}$ and is also initially distributed according to $\boldsymbol{\pi}$. Denote by $\boldsymbol{P}$ its transition probability matrix. Define a new Markov chain $\{Y_n\}_{0\leq n\leq N}$ by $Y_n = X_{N-n}$ where $N \in \mathbb{N}$ is fixed. Then $\{Y_n\}_{0\leq n\leq N}$ is also an Markov chain with invariant distribution $\boldsymbol{\pi}$. Its transition probability matrix $\hat{\boldsymbol{P}}$ is given by*

$$\hat{p}_{ij} = \frac{\pi_j}{\pi_i} p_{ji}. \tag{2}$$

Proof. It is straightforward to check that $\hat{\boldsymbol{P}}$ is a stochastic matrix with an invariant distribution $\boldsymbol{\pi}$. To prove that $\{Y_n\}$ is Markov with transition probability matrix $\hat{\boldsymbol{P}}$, it is enough to observe that

$$\begin{aligned}\mathbb{P}(Y_0 = i_0, Y_1 = i_1, \ldots, Y_N = i_N) &= \mathbb{P}(X_N = i_0, X_{N-1} = i_1, \ldots, X_0 = i_N) \\ &= \pi_{i_N} p_{i_N i_{N-1}} \cdots p_{i_1 i_0} = \pi_{i_0} \hat{p}_{i_0 i_1} \cdots \hat{p}_{i_{N-1} i_N}\end{aligned}$$

for any $i_0, i_1, \ldots, i_N$. □

A particularly important class of Markov chains are those that satisfy the condition of *detailed balance*

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{3}$$

In this case, we have $\hat{p}_{ij} = p_{ij}$. We call the chain *reversible*. The reversible chain can be equipped with variational structure and has nice spectral properties. Define the matrix

$$\boldsymbol{L} = \boldsymbol{P} - \boldsymbol{I}$$

and correspondingly its action on any function f

$$(\boldsymbol{L}f)(i) = \sum_{j \in S} p_{ij}(f(j) - f(i)).$$

Let $L^2_{\boldsymbol{\pi}}$ be the space of square summable functions f endowed with the $\boldsymbol{\pi}$-weighted scalar product

$$(f, g)_{\boldsymbol{\pi}} = \sum_{i \in S} \pi_i f(i) g(i). \tag{4}$$

Denote the *Dirichlet form* or energy of a function f by

$$D(f) = \sum_{i,j \in S} \pi_i p_{ij}(f(j) - f(i))^2.$$

One can show that $D(f) = (f, -\boldsymbol{L}f)_{\boldsymbol{\pi}}$. These formulations are particularly useful in potential theory for Markov chains.

1.2 Hitting time distribution

Example 6. *(Hitting time distribution of a Markov chain) Consider TPM of a 4-state Markov chain(1,2,3,4):*

$$P = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Define the first hitting time $n_* = \inf\{n |\ X_n = 3 \text{ or } 4\}$ *and the hitting time probability* $q(m) = \text{Prob}\{n_* = m\}$, *an interesting question is to ask how to obtain* $q(m)$. *The idea is to modify the chain to a 3-state chain*

$$\tilde{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 0 & 1 \end{bmatrix}.$$

then

$$1 - (\mu_n)_3 = \sum_{m=n+1}^{\infty} q(m),$$

hence

$$q(n) = (\mu_n)_3 - (\mu_{n-1})_3 = \mu_0 \cdot (\tilde{P}^n - \tilde{P}^{n-1}) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

2 Continuous time Markov chains

2.1 Poisson Process

Definition 3. *(Poisson Process) Let* $X(t)$ *be the number of calls received up to time* t, *and assume the follows:*

1. $X(0) = 0$;

2. *$X(t)$ has independent increments, i.e. for any $0 \le t_1 < t_2 < \cdots < t_n$,*

$$X(t_2) - X(t_1), X(t_3) - X(t_2), \ldots, X(t_n) - X(t_{n-1})$$

are independent;

3. *for any $t \ge 0, s \ge 0$, we have the distribution of the increment $X(t+s) - X(t)$ is independent of t (time-homogeneous);*

4. *for any $t \ge 0, h > 0$, we have*

$$\begin{aligned} P\{X(t+h) = X(t) + 1|\ X(t)\} &= \lambda h + o(h), \\ P\{X(t+h) = X(t)|X(t)\} &= 1 - \lambda h + o(h), \\ P\{X(t+h) \ge X(t) + 2\} &= o(h), \end{aligned}$$

where λ is called the rate.

Then $X(t)$ is called a Poisson process.

Let $p_m(t) = P\{X(t) = m\}$, then

$$p_0(t+h) = p_0(t)p_0(h) = p_0(t)(1 - \lambda h) + o(h).$$

This gives

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) + o(1).$$

As $h \to 0$, we obtain

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t), \quad p_0(0) = 1.$$

The solution is given by

$$p_0(t) = e^{-\lambda t}.$$

For $m > 0$, we have

$$p_m(t+h) = p_m(t)p_0(h) + p_{m-1}(t)p_1(h) + \sum_{i=2}^{m} p_{m-i}(t)p_i(h).$$

From the definition of Poisson process, we get

$$p_m(t+h) = p_m(t)(1 - \lambda h) + p_{m-1}(t)\lambda h + o(h).$$

Taking the limit as $h \to 0$, we get

$$\frac{dp_m(t)}{dt} = -\lambda p_m(t) + \lambda p_{m-1}(t)$$

Using the fact $p_m(0) = 0 (m > 0)$, we get

$$p_m(t) = \frac{(\lambda t)^m}{m!} e^{-\lambda t}$$

by induction method. This means that for any fixed t, the distribution of $X(t)$ is Poisson with parameter λt.

The waiting times can be obtained in the following way. Define

$$\mu_t = P\{\text{Waiting time} \geq t\},$$

then $\mu_0 = 1$, and it obeys $\mu_t - \mu_{t+h} = \mu_t \lambda h + o(h)$, thus $\mu_t' = -\lambda \mu_t$, we get

$$\mu_t = e^{-\lambda t}.$$

i.e. The waiting times are i.i.d. exponentially distributed with rate λ.

2.2 Q-Process

Now let us turn to general continuous time Markov chains. We will restrict only on finite state space case in this text. We define

$$p_{ij}(t) = \text{Prob}\{X(t+s) = j | X(s) = i\}.$$

Here we also assumed the stationarity of the Markov chain, i.e. the right hand side is independent of s. By definition we have

$$p_{ij}(t) \geq 0, \qquad \sum_{j=1}^{N} p_{ij}(t) = 1.$$

In addition we require that

$$p_{ii}(h) = 1 - \lambda_i h + o(h), \quad \lambda_i > 0, \tag{5}$$

$$p_{ij}(h) = \lambda_{ij} h + o(h), \quad j \neq i. \tag{6}$$

(5) is a statement about the regularity in time of the Markov chain; together with the obvious constraint that $p_{jj}(0) = 1$. (6) states that if the process is in state j at time t and a change occurs between t and $t+h$, the process must have jumped to some state $i \neq j$; λ_{ij} is the rate of switching from state i to state j.

From the non-negativity and normalization condition of the probability, we have

$$\lambda_{ij} \geq 0, \qquad \sum_{j=1, j\neq i}^{N} \lambda_{ij} = \lambda_i. \tag{7}$$

The Markov property of the process requires the Chapman-Kolmogorov equation

$$p_{ij}(t+s) = \sum_{k=1}^{N} p_{ik}(t) p_{kj}(s). \tag{8}$$

Using matrix notation $P(t) = (p_{ij}(t))$, we can express the Chapman-Kolmogorov relation as

$$P(t+s) = P(t)P(s) = P(s)P(t).$$

Similarly, if we define

$$Q = \lim_{h \to 0+} h^{-1}(P(h) - I), \tag{9}$$

and denote $Q = (q_{ij})$, (5), (6) and (7) can be stated as

$$q_{ii} = -\lambda_i, \qquad q_{ij} = \lambda_{ij} \quad (i \neq j), \qquad \sum_{j=1}^{N} q_{ij} = 0.$$

Q is called the *generator* of the Markov chain.

Since

$$\frac{P(t+h) - P(t)}{h} = \frac{P(h) - I}{h} P(t)$$

as $s \to 0+$, we get

$$\frac{dP(t)}{dt} = QP(t) = P(t)Q \tag{10}$$

The solution of this equation is given by

$$P(t) = e^{Qt} P(0) = e^{Qt},$$

since $P(0) = I$.

Next we discuss how the distribution of the Markov chain evolves in time. Let $\nu(t)$ be the distribution of $X(t)$. Then

$$\begin{aligned} \nu_j(t+dt) &= \sum_{i \neq j} \nu_i(t) p_{ij}(dt) + \nu_j(t) p_{jj}(dt) \\ &= \sum_{i \neq j} \nu_i(t) q_{ij} dt + \nu_j(t)(1 + q_{jj} dt) + o(dt) \end{aligned}$$

for infinitesimal dt. This gives

$$\frac{d\nu(t)}{dt} = \nu(t) Q, \tag{11}$$

which is called the forward Kolmogorov equation for the distribution. Its solution can be given as

$$\nu_j(t) = \sum_{i=1}^{N} \nu_i(0) p_{ij}(t),$$

or, in matrix notation,

$$\nu(t) = \nu(0) e^{Qt}.$$

Similar as the Poisson process, we can consider the waiting time distribution for each state j,

$$\mu_j(t) = \text{Prob}\{\tau \geq t | X(0) = j\}.$$

The same procedure as previous section leads to

$$\frac{d\mu_j(t)}{dt} = q_{jj} \mu_j(t), \qquad \mu_j(0) = 1.$$

Thus the waiting time at state j is exponentially distributed with rate $-q_{jj} = \sum_{k \neq j} q_{jk}$. From the memoryless property of exponential distribution, the waiting time can be counted from any starting point.

It is interesting to investigate the probability

$$\begin{aligned} p(\theta, j | 0, i) d\theta := \text{Prob}\{&\text{The jump time } \tau \text{ is in } [\theta, \theta + d\theta) \\ &\text{and } X(\tau) = j \text{ given } X(0) = i\}. \end{aligned}$$

We have

$$\begin{aligned} p(\theta, j|0, i)d\theta =& \text{Prob}\{\text{No jump occurs in } [0, \theta) \text{ given } X(0) = i\} \\ & \times \text{Prob}\{\text{One jump occurs from } i \text{ to } j \text{ in } [\theta, \theta + d\theta)\} \\ =& \mu_i(\theta) q_{ij} d\theta = \exp(q_{ii}\theta) q_{ij} d\theta. \end{aligned} \tag{12}$$

Thus we obtain the marginal probability

$$\text{Prob}(X(\tau) = j | X(0) = i) = p(j|0, i) = -\frac{q_{ij}}{q_{ii}} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$$

where τ is the waiting time. These results are particularly useful for the numerical simulation of the trajectories of the Q-process.

Define the *jump times* of $(X_t)_{t \geq 0}$

$$J_0 = 0, \quad J_{n+1} = \inf\{t : t \geq J_n, X_t \neq X_{J_n}\}, \quad n \in \mathbb{N}$$

where we take the convention $\inf \emptyset = \infty$, and *holding times*

$$H_n = \begin{cases} J_n - J_{n-1}, & \text{if } J_{n-1} < \infty, \\ \infty, & \text{otherwise.} \end{cases}$$

for $n = 1, 2, \ldots$. We define $X_\infty = X_{J_n}$ if $J_{n+1} = \infty$. Define the *jump chain* induced by X_t

$$Y_n = X_{J_n}, \quad n \in \mathbb{N}.$$

From Strong Markov property and the derivation of $p(\theta, j|0, i)$, we know that the holding times $H_1, H_2, \ldots$ are independent exponential random variables with parameters $q_{Y_0}, q_{Y_1}, \ldots$, respectively, and the jump chain Y_n is a Markov chain with $\tilde{\boldsymbol{Q}}$ as the transition probability matrix, where $\tilde{\boldsymbol{Q}} = (\tilde{q}_{ij})$ defined as

$$\tilde{q}_{ij} = \begin{cases} q_{ij}/q_i, & \text{if } i \neq j \text{ and } q_i > 0, \\ 0, & \text{if } i \neq j \text{ and } q_i = 0, \end{cases} \tag{13}$$

$$\tilde{q}_{ii} = \begin{cases} 0, & \text{if } q_i > 0, \\ 1, & \text{if } q_i = 0. \end{cases} \tag{14}$$

It is called the *jump matrix*, and the corresponding Markov chain is called the *embedded chain* or *jump chain* of the original Q-process.

It is natural to consider the invariant distribution for the Q-processes as in the discrete time Markov chains. From the forward Kolmogorov equation (11), the invariant distribution must satisfy

$$\pi Q = 0, \quad \pi \cdot \mathbf{1}^T = 1.$$

But to ensure the convergence $\nu(t) \to \pi$, we need the following theorem on the finite state space.

Theorem 6 (Convergence to equilibrium)**.** *Suppose the matrix Q is irreducible with invariant distribution π, then for all states i, j we have*

$$p_{ij}(t) \to \pi_j \text{ as } t \to \infty.$$

Note that we do NOT need the primitive condition since in the continuous time case if $q_{ij} > 0$ we have

$$p_{ij}(t) \geq \mathbb{P}_i(J_1 \leq t, Y_1 = j, H_2 > t) = \int_0^t e^{-q_i u} q_{ij} du \cdot e^{-q_j t} = \frac{q_{ij}}{q_i}(1 - e^{-q_i t})e^{-q_j t} > 0.$$

Similarly we also have the ergodic theorem

Theorem 7 (Ergodic theorem)**.** *Suppose the matrix Q is irreducible with invariant distribution π, then for any bounded function f we have*

$$\frac{1}{t}\int_0^t f(X(s))ds \to \langle f \rangle_\pi, \qquad a.s.$$

We should remark that the irreducibility condition is not enough to establish the above ergodic theorems in the countable state space case. We need the so-called positive recurrent condition in both theorems.

3 Homeworks

- HW1. Discuss the invariant distribution of the Ehrenfest's model.
- HW2. Rederive the distribution of Poisson process through characteristic function method.
- HW3. Let f be a function defined on the state space, and let

  $$h_i(t) = \mathbb{E}^i f(X(t)),$$

  where $\mathbb{E}^i$ means the expectation with respect to initial state i. Derive an equation for $h(t)$.
- HW4. Consider the following binomial process: we repeatedly throw an unfair coin with parameter p (say, the proability that the HEAD appears) with time unit τ. If the HEAD appears, we denote it as a jump. Then we let $p, \tau \to 0$ and consider the limiting process. In which regime you can intuitively get the Poisson process with parameter λ?
- HW5. For the Poisson process, if the condition 3 is removed, and the rate λ depends on t. That is, λ is replaced with $\lambda(t)$ in conidtion 4, then what about $p_m(t)$ and the waiting time distribution μ_s conditioned at the current time t?

References

[1] K.L. Chung, A Course in Probability Theory, Academic Press, New York, 1974.

[2] R. Durret, Probability: theory and examples (3rd edition), Thomson Learning, 2005.

[3] C.P. Robert and G. Casella, Monte Carlo statistical methods, Second edition, Springer-Verlag, 2004.

[4] G. Winkler, Image analysis, random fields and dynamic Monte Carlo methods, Springer-Verlag, Berlin and Heidelberg, 1995.

Lecture 7 Metropolis Algorithm *

Tiejun Li

1 Introduction

Metropolis algorithm is one of the Top10 algorithms in 20th century [1]. The aim is to compute the thermodynamic quantities through ensemble average, such as

$$\langle A \rangle = \sum_{\sigma} A(\sigma) P(\sigma) \quad (\text{or } \int_S A(\sigma) P(d\sigma)).$$

where σ represents all possible configurations, and

$$P(\sigma) = \frac{1}{Z_\beta} e^{-\beta H(\sigma)} \qquad \beta = (k_B T)^{-1}.$$

$H(\sigma)$ is the energy function $Z_\beta = \sum_\sigma e^{-\beta H(\sigma)}$ is the partition function, $\beta = (k_B T)^{-1}$.

Intuitively, we could approximate $\langle A \rangle$ by

$$\langle A \rangle \approx \langle A \rangle_N := \frac{1}{N} \sum_{i=1}^{N} A(\sigma_i), \tag{1}$$

where $\sigma_i \sim P(\sigma)$ *i.i.d.*

But the problem is **how to generate** σ_i!

2 Ising Model

Example 1. *(1D Ising model) The magnetization of a ferromagnet may be described essentially by a spin model as shown in Figure 1.*

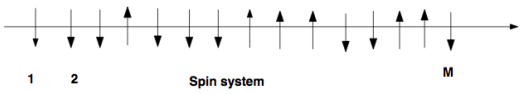


Figure 1: Sketch of 1D Ising model

The macroscopic magnetization $\bar{H}$ can be obtained from ensemble average of all the possible microscopic spin configurations in statistical physics. Consider one microscopic sate with M sites $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_M)$, $\sigma_i = +1$ or -1. If the i-th spin is $\uparrow$, then $\sigma_i = +1$; If the i-th spin is $\downarrow$, then $\sigma_i = -1$.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Physically the following quantities are of interest:

1. *Hamiltonian:*
$$H(\sigma) = -J \sum_{<ij>} \sigma_i \sigma_j - h \sum_i \sigma_i,$$
where $\sigma_i = \pm 1$, $< ij >$ means to take sum w.r.t all neighboring spins $|i-j| = 1$, and h is the strength of external magnetic field.

2. *Free energy:*
$$F_M = -\beta^{-1} \ln Z_M, \qquad f_M = \frac{1}{M} F_M$$
where $Z_M = \sum_\sigma \exp\{-\beta H(\sigma)\}$ is the partition function.

3. *Internal energy:*
$$U_M = \langle H(\sigma) \rangle = \sum_\sigma H(\sigma) \frac{\exp\{-\beta H(\sigma)\}}{Z_M} = -\frac{\partial \ln Z_M}{\partial \beta}, \tag{2}$$
Correspondingly define the internal energy per site: $u_M = \frac{1}{M} U_M$.

4. *Specific heat:*
$$C_M = \frac{\partial U_M}{\partial T} = k_B \beta^2 \{\langle H^2(\sigma) \rangle - \langle H(\sigma) \rangle^2\} = k_B \beta^2 \frac{\partial^2 \ln Z_M}{\partial \beta^2}. \tag{3}$$
Correspondingly define the specific heat per site: $c_M = \frac{1}{M} C_M$.

5. *Magnetization:*
$$G_M = \left\langle \sum_i \sigma_i \right\rangle, \qquad g_M = \frac{1}{M} G_M$$

6. *Magnetic susceptibility:*
$$\Phi_M = \frac{\partial G_M}{\partial h} = \beta \mathrm{Var}(G_M), \qquad \phi_M = \frac{1}{M} \Phi_M.$$

For 2D Ising model, if $h = 0$ and $M \to \infty$, the second order phase transition exists for the internal energy u when the temperature varies. If one takes $J = 1$, and the periodic boundary condition is assumed, the critical temperature can be obtained exactly

$$k_B T_c = (\beta_c)^{-1} = \frac{2}{\ln(1 + \sqrt{2})}.$$

And there exists so called spontaneous magnetization for g_M as $M \to \infty$, we have

$$\begin{aligned}
0 < \beta < \beta_c, \quad & g(\beta, h) \to g(\beta, 0) = 0 \ \text{as } h \to 0+, \\
& g(\beta, h) \to g(\beta, 0) = 0 \ \text{as } h \to 0-, \\
\beta > \beta_c, \quad & g(\beta, h) \to g(\beta, +) > 0 \qquad\qquad \text{as } h \to 0+, \\
& g(\beta, h) \to g(\beta, -) = -g(\beta, +) < 0 \ \text{as } h \to 0-.
\end{aligned}$$

For refs, see [2, 3].

How to compute the approximate $k_B T_c$ with computer?

3 Metropolis Algorithm

1. Basic idea:

   The central problem for standard MC(equation (1)) is that it is difficult to generate $\sigma_i \sim P(\sigma)$ $i.i.d.$. To overcome this difficulty, Metropolis algorithm takes an "**iteration**" procedure to produce these random variables. This is very similar with finding a root of a nonlinear equation $f(x) = 0$. There isn't direct method to do this, but if one uses the iteration $x_{k+1} = g(x_k)$ such that $x^* = g(x^*)$ is one root of the equation $f(x) = 0$, one approximate value of x^* will be obtained.

   Metropolis algorithm is also called Markov Chain Monte Carlo method. It sets up a Markov chain by defining a suitable transition probability matrix P, such that the probability density $\frac{1}{Z_M} e^{-\beta H(\sigma)}$ is the only equilibrium distribution of this Markov Chain. If we define π the final Gibbs distribution, then we need

   $$\pi P = \pi.$$

   The single step transition $\nu_{n+1} = \nu_n P$ satisfies similar contraction mapping property as the fixed point iteration. One could obtain the Gibbs distribution from any initial state ν_0. That's the essence of Metropolis algorithm!

2. Physical interpretation(a heuristic observation):

   In statistical physics, the equations (2) and (3) are called ensemble average, which means that the macroscopic state is an average result of all possible microscopic states. Suppose we image large amount of gas molecules in a container, and they are in equilibrium, we can measure the temperature, pressure etc. In the viewpoint of statistical physics, this system corresponds lots of microscopic systems. This is shown schematically in Figure 2.

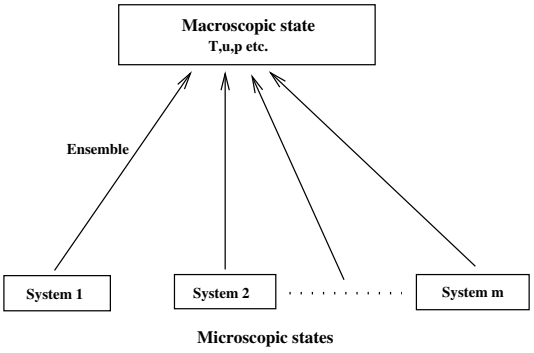


Figure 2: Sketch of ensemble

Though system $1, 2$ and system m are independent evolving systems, the probability distribution anytime for all the m systems keeps the same. Metropolis algorithm takes the following viewpoint: since the macroscopic quantity T, p, ρ is invariant all the time, and the systems are in a dynamic

equilibrium, this means

$$\text{Ensemble average} = \text{Time average},$$

i.e.

$$\langle H(\sigma)\rangle \approx \frac{1}{N}\sum_{i=1}^{N} H(\sigma_i) \tag{4}$$

If we can find an appropriate collision rule for the molecules to sample the state sequence $\{\sigma_i\}_{i=1}^{N}$, we can simulate this process in computer, and find an approximate value of $\langle H(\sigma)\rangle$. This collision rule is the setup of Markov Chain stated as before.

3. Mathematical description:

From the **equilibrium condition** $\pi P = \pi$, where

$$\pi(\sigma) = \frac{1}{Z_M}\exp\{-\beta H(\sigma)\},$$

the choice of matrix P is infinite. One must add more constraints to reduce the degrees of freedom.

Detailed Balance Condition(DBC):

$$\pi(\sigma)P(\sigma \to \sigma') = \pi(\sigma')P(\sigma' \to \sigma).$$

Furthermore one has

$$\frac{P(\sigma \to \sigma')}{P(\sigma' \to \sigma)} = \frac{\pi(\sigma')}{\pi(\sigma)} = e^{-\beta\Delta H},$$

where $\Delta H = H(\sigma') - H(\sigma)$.

Intuitively, if $\Delta H > 0$, set $P(\sigma \to \sigma') = 1$; else set $P(\sigma \to \sigma') = e^{-\beta\Delta H}$. This choice satisfies the DBC. But **how about** $P(\sigma \to \sigma') = 1$**?**

We lose a proposal process!

In general, set

$$P(\sigma \to \sigma') = Q(\sigma \to \sigma')A(\sigma \to \sigma'), \quad \sigma' \neq \sigma$$

and

$$P(\sigma \to \sigma') = 1 - \sum_{\tau \neq \sigma} P(\sigma \to \tau), \quad \sigma' = \sigma$$

where $Q(\sigma \to \sigma')$ is the choosing probability corresponds to **Proposal Matrix**, and $A(\sigma \to \sigma')$ is the acceptance probability corresponds to **Decision Matrix**. In many cases, we take $Q(\sigma \to \sigma') = Q(\sigma' \to \sigma)$, which is a symmetric matrix, and the following two decision strategies:

A. Metropolis algorithm:

$$A(\sigma \to \sigma') = \min(1, e^{-\beta\Delta H}).$$

B. Glauber dynamics:

$$A(\sigma \to \sigma') = (1 + e^{\beta\Delta H})^{-1}.$$

Remark 1. *The proposal matrix Q can also be unsymmetric, then the reversibility is realized in the decision step, which is called Metropolis-Hastings algorithm in statistical learning theory.*

Example 2. *(2D Ising model - $M \times M = M^2$ sites)*

Step1 Generate a proposal σ' from current state σ;

A. Equi-probability proposal:

$$Q(\sigma \to \sigma') = \begin{cases} \frac{1}{N_t - 1} & \sigma \neq \sigma' \\ 0 & \sigma = \sigma' \end{cases}$$

where $N_t = 2^{M^2}$ is the total number of microscopic states.

B. Single flip proposal(usually used):

$$Q(\sigma \to \sigma') = \begin{cases} \frac{1}{M^2} & \sigma \neq \sigma' \text{ only at one site} \\ 0 & \text{otherwise} \end{cases}$$

Step2 Decide to accept or reject the proposal:

Compute $\Delta H(\sigma) = H(\sigma') - H(\sigma)$. If $\Delta H < 0$, accept; else, accept with probability $e^{-\beta \Delta H}$.

4. Algorithm:

Algorithm 1. *(Metropolis Algorithm)*

Step1 Generate the proposal sate σ' from state σ_n according to some strategy;

Step2 Define $\Delta H(\sigma) = H(\sigma') - H(\sigma)$, compute

$$\begin{aligned} A(\sigma, \sigma') &= \min\left\{1, \frac{\pi(\sigma')}{\pi(\sigma)}\right\} \\ &= \begin{cases} 1 & H(\sigma') \leq H(\sigma), \\ \exp(-\beta \Delta H(\sigma)) & \text{otherwise}; \end{cases} \end{aligned}$$

Step3 Generate R.V. $r \sim \mathcal{U}[0,1]$;

Step4 If $r \leq A(\sigma, \sigma')$, then $\sigma_{n+1} = \sigma'$; else, $\sigma_{n+1} = \sigma_n$, turn to Step1.

5. Theorems:

The proof of the theorems presented below can be found in [4].

Definition 1. *The total variation of a vector is defined as*

$$\|v\| = \sum_i |v_i|.$$

This is the commonly known L^1-norm in matrix theory.

Definition 2. *If there is a natural number τ such that the transition probability matrix P satisfies $P^\tau > 0$(where $A > 0$ means each element of A is greater than 0), this Markov chain is called primitive.*

Theorem 1. *(Ergodic theorem for time-homogeneous Markov Chain)*

There is only one invariant distribution π for a primitive Markov chain, and for any initial distribution ν, one has

$$\lim_{n\to+\infty} \|\nu P^n - \pi\| = 0.$$

Theorem 2. *(WLLN for time-homogeneous Markov Chain)*

Suppose $\{\xi_i\}_{i\geq 1}$ is a realization of a primitive Markov chain, then one has

$$\text{Prob}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) - \langle f\rangle_\pi\right| > \epsilon\right) \leq \frac{C}{n\epsilon^2},$$

where $\langle f\rangle_\pi$ means the expectation of f w.r.t. the probability π. C depends on R.V. f and Markov chain P.

6. The improved Metropolis-type algorithm in different extensions such as the Swendsen-Wang algorithm, parallel tempering etc. can be referred to [5].

4 Homeworks

- HW1. You want to compute
$$\frac{\int_{-10}^{10} e^{-x^2/a}dx}{\int_{-10}^{10} e^{-x^2/b}dx},$$
where $a,\ b$ are positive constants. Describe a Metropolis Monte Carlo algorithm which does the job; use matlab to test it when $a = 10, b = 12$.

- HW2. Check the detailed balance condition for Metropolis and Glauber dynamics.

- HW3. Check the Markov chain setup by Metropolis and Glauber dynamics for Ising model are both primitive.

References

[1] I. Beichl and F. Sullivan, The Metropolis algorithm, Computing in Science and Engineering 2(2000), 65-69.

[2] L. Onsager, Crystal Statistics. I. A Two Dimensional Model with an Order-Disorder Transition, Phys. Rev. 65(1944), 117-149.

[3] C.N. Yang, The Spotaneous Magnetization of a Two-Dimensional Ising Model, Phys. Rev. 85(1952), 808-816.

[4] G. Winkler, Image Analysis, Random Fields and Dynamic Monte Carlo Methods, Springer-Verlag, Berlin and New York, 1995.

[5] J.S. Liu, Monte Carlo strategies in scientific computing, Springer-Verlag, 2005.

Lecture 8 Multilevel sampling and KMC *

Tiejun Li

1 Swendsen-Wang algorithm

For the numerical computation of the Ising model, a commonly used approach is the Gibbs sampling to flip single site at each step. However, this single-site update algorithm slows down rapidly once the temperature is approaching or below the critical value T_0, the so-called "critical slowing down". Swendsen and Wang [6] introduced a powerful clustering algorithm which together with an implementation modification by Wolff [8], amost completely eliminates the critical slowing down. Below explanation to Swendsen-Wang algorithm is from data augmentation viewpoint by Higdon [7].

We have the Gibbs distribution for Ising model

$$\begin{aligned}\pi(\boldsymbol{x}) &\propto \exp\Big\{\beta J \sum_{<i,j>} x_i x_j\Big\} \\ &\propto \prod_{<i,j>} \exp\Big\{\beta J(1+x_i x_j)\Big\}.\end{aligned}$$

Note that $1+x_ix_j$ is equal to either 0 or 2. Hence if we introduce an auxiliary variable $\boldsymbol{u}$ on each edge such that

$$\pi(\boldsymbol{x},\boldsymbol{u}) \propto \prod_{<i,j>} I\Big[0 \le u_{ij} \le \exp\{\beta J(1+x_i x_j)\}\Big].$$

Then the marginal distribution of $\boldsymbol{x}$ is the Gibbs distribution. And under this joint distribution, the conditional distribution $\boldsymbol{u}|\boldsymbol{x}$ is a product of uniform distributions with ranges depending on two neighboring spins. Conversely, the conditional distribution $\boldsymbol{x}|\boldsymbol{u}$ is: if $u_{ij} > 1$, then $x_i = x_j$; otherwise there is no constraint on x_i's. Thus $\boldsymbol{u}$ affects $\boldsymbol{x}$ only through the event $I[u_{ij} > 1]$. Based on the configuration $\boldsymbol{u}$, we cluster those lattice sites according to whether they have a mutual bond ($u_{ij} > 1$). We formulate the following algorithm

Algorithm 1. *Swendsen-Wang algorithm:*

- *Step 1. For a given configuration of the spins, form the bond variable by giving every edge of the lattice $< i, j >$, between two "like spins" ($x_i = x_j$) a bond value of 1 with probability $1 - \exp(-2\beta J)$, and a bond value of 0 otherwise.*

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

- *Step 2. Conditional on the bond variable* $\boldsymbol{u}$*, update the spin variable* $\boldsymbol{x}$ *by drawing from* $\pi(\boldsymbol{x}|\boldsymbol{u})$*, which is uniform on all compatible spin configurations; that is, clusters are produced by connecting neighboring sites with bond value 1. Each cluster is the flipped with probability 0.5.*

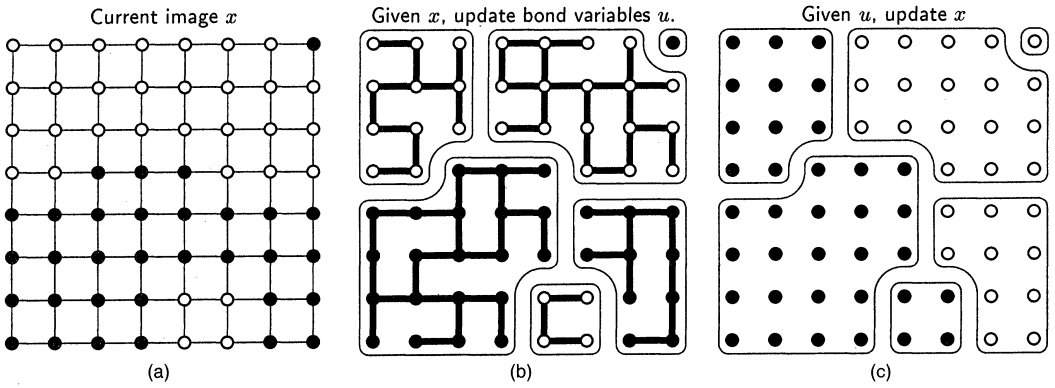


Figure 2. The Swendsen–Wang Algorithm for the Ising Model on the 8 × 8 Lattice. (a) Initial image x and Markov random field graph for π(x). (b) Given the current image x, the bond variables u are generated uniformly over the interval (0, $e^{\beta I[x_i=x_j]}$). If $u_{ij} > 1$ (marked by the thick lines), x_i is constrained to equal x_j. These constraints partition the image into clusters of like-colored sites. Clusters induced by this realization of u|x are outlined. The Markov random field graph for x|u differs from that of x, marginally; the auxiliary vector u strengthens the dependence between some neighboring sites, while completely removing it from others. (c) Given the bond variables u, x is now a coarse image of independent clusters. Because there is no external field in this example, each cluster is recolored black or white with probability .5.

2 The modification by Wolff

Wolff introduced a modification for the Swendsen-Wang algorithm, which, although both conceptually and operationally simple, significantly outperforms the SW algorithm.

Algorithm 2. *Wolff's algorithm:*

- *Step 1. For a given configuration* $\boldsymbol{x}$*, one randomly picks a site, say* x_i*, and grow recursively from it a "bonded set"* C *as follows:*
  - *Check all the unchecked neighboring sites of a current set* $C^{(old)}$*; add a bond between a neighboring site and* $C^{(old)}$ *the same way as in the Swendsen-Wang algorithm.*
  - *Add those newly bonded neighboring sites to* $C^{(old)}$ *so as to form a new set* $C^{(new)}$*.*
  - *Stop the recursion when there is no unchecked neighbor to add; name the final set* C*.*
- *Flip all the spins corresponding to the sites in set* C *to their opposites.*

The only difference between Wolff's algorithm and SW is that in each iteration, only one cluster is constructed and all spins in that cluster are changed to their opposite value. This algorithm actually offers a new insight which is different from the one based on the data augmentation. Suppose all of the states in cluster C has spin $+1$ and it has $m+n$ neighboring links among which m are linked with $+1$ spins and n with -1 spins. We have the acceptance probability

$$A_{\text{old}\to\text{new}} = \min\left\{1, \frac{Q_{\text{new}\to\text{old}}\pi_{\text{new}}}{Q_{\text{old}\to\text{new}}\pi_{\text{old}}}\right\}$$

and

$$\frac{\pi_{\text{new}}}{\pi_{\text{old}}} = \frac{e^{\beta J(n-m)}}{e^{\beta J(m-n)}} = e^{2\beta J(n-m)}, \quad \frac{Q_{\text{new}\to\text{old}}}{Q_{\text{old}\to\text{new}}} = \frac{e^{-2\beta Jn}}{e^{-2\beta Jm}} = e^{2\beta J(m-n)}$$

Thus these two probability cancel each other and the proposed change is accepted with probability one.

3 Simulated tempering

To sample the distribution

$$p(x) \propto \exp\Big(-\frac{U(x)}{T}\Big)$$

or compute the ensemble average with the type

$$\langle H\rangle = \int H(x)\frac{1}{Z}\exp\Big(-\frac{U(x)}{T}\Big)dx,$$

one usually apply the Metropolis-Hastings MCMC algorithm. But when the temperature T is very low, that is, we have many high peaks in the pdf $p(x)$, which may cause the acceptance probability small thus decrease the mixing.

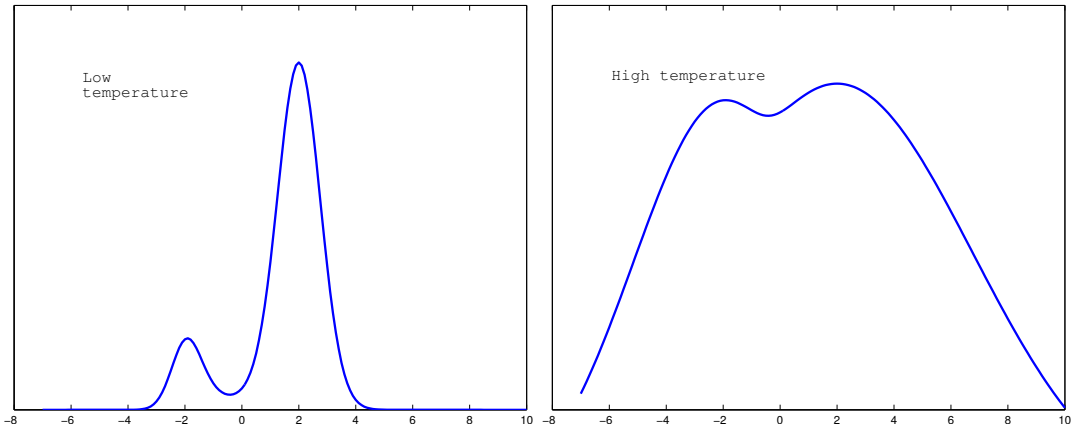


Figure 1: Sketch of the Gibbs distribution at low and high temperature.

In order to let a MCMC scheme move more freely in the sate space, Marinari and Parasi [4] and Geyer and Thompson [3] proposed a data augmentation strategy to increase the mixing, which is called *simulated tempering*. Algorithmically, their basic idea is to extend the state space $x \in \mathcal{X}$ into $(x, i) \in \mathcal{X} \times I$ and perform conditional sampling in this extended space. Physically, to approach the low temperature case, they consider the pdf at the heated temperature, which can give high acceptance ratio for traversing the state space $\mathcal{X}$, and then jump in the different ensembles.

Mathematically, they let

$$I = \{1, 2, \ldots, L\}, \quad T_1 < T_2 < \ldots < T_L$$

and $T_1 = T$, $T_L = T_{\text{high}}$. Then they ask the stationary distribution in the extended space as

$$\pi_{\text{st}}(x, i) \propto \pi_i \exp\Big(-\frac{U(x)}{T_i}\Big),$$

where π_i is called pseudopriors which is set up *a priori*.

From this form, we know the conditional distribution

$$f(x|i) \propto \exp\Big(-\frac{U(x)}{T_i}\Big)$$

which is the standard Gibbs distribution. The marginal distribution

$$f(i) \propto \int \pi_i \exp\Big(-\frac{U(x)}{T_i}\Big)dx = \pi_i Z_i.$$

To make the transition in different ensembles more uniformly, the best choice for the parameter $\pi_i \propto 1/Z_i$. But in the computations, it is not feasible and only updated with the time.

To do the conditional sampling in the extended space, we list a mixture-type transition kernel here.

Algorithm 3 (Simulated tempering)**.** *Mixture type of the simulated tempering.*

- *Step 1. With the current state $(x_n, i_n) = (x, i)$, we draw $u \sim \mathcal{U}[0,1]$.*
- *Step 2. If $u < \alpha_0$, we let $i_{n+1} = i$ and let x_{n+1} be drawn from a MCMC transition $T_i(x, x_{n+1})$ that leaves $f(x|i)$ invariant (this is also Metropolis-Hastings strategy).*
- *Step 3. If $u > \alpha_0$, we let $x_{n+1} = x$ and propose a level transtion $i \to j$, from a transition function $\alpha(i,j)$, and let $i_{n+1} = j$ with probability*

$$\min\Big(1, \frac{\pi_{st}(x,j)\alpha(j,i)}{\pi_{st}(x,i)\alpha(i,j)}\Big).$$

Otherwise let $i_{n+1} = i$.

A commonly used strategy for $\alpha(i,j)$ is the random walk proposal with reflecting barrier, that is,

$$\alpha(i, i \pm 1) = 1/2, \quad i = 2, \ldots, L-1$$

and $\alpha(1,2) = \alpha(L, L-1) = 1$.

The idea of simulated tempering is further generalized by Liu and Sabatti [5] into the so called "simulated sintering" scheme.

4 Parallel tempering

The pararellel tempering is first proposed by Geyer [2] in 1991. Instead of augmenting $\mathcal{X}$ into $\mathcal{X} \times I$, Geyer suggested directly dealing with the product space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L$, where the $\mathcal{X}_i$ are identical copies of $\mathcal{X}$, suppose

$$(x_1, \ldots, x_L) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_L,$$

we define the stationary distribution

$$\pi_{st}(x_1, \ldots, x_L) = \prod_{i \in I} \pi_i(x_i)$$

where $\pi_i(x_i) = 1/Z_i \exp(-U(x_i)/T_i)$ the Gibbs distribution at $T = T_i$. The parallel tempering is run on all of the $\mathcal{X}_i$. An "index swapping" operation is conducted in place of the temperature transition. The algorithm is defined as follows:

Algorithm 4 (Parallel tempering algorithm)**.** *Mixture type transition kernel.*

- *Step 1: Let the current state be* $(x_1^{(n)}, \ldots, x_L^{(n)})$*. Draw* $u \sim \mathcal{U}[0,1]$*.*
- *Step 2: If* $u \leq \alpha_0$*, we conduct the parallel step. That is, we update each* $x_i^{(n)}$ *to* $x_i^{(n+1)}$ *via their respective MCMCM scheme.*
- *Step 3: If* $u > \alpha_0$*, we conduct the swapping setp. That is, we randomly choose a neighboring pair, say* i *and* $i+1$*, and propose "swapping"* $x_i^{(n)}$ *and* $x_{i+1}^{(n)}$*. Accept this swap with probability*

$$\min\left\{1, \frac{\pi_i(x_{i+1}^{(n)})\pi_{i+1}(x_i^{(n)})}{\pi_i(x_i^{(n)})\pi_{i+1}(x_{i+1}^{(n)})}\right\}.$$

  In computations, $T_1 < T_2 < \ldots < T_L$*, and it is very important to choose a proper number of temperature levels. A rough guideline is to choose* T_i *such that*

$$\Big(\frac{1}{T_i} - \frac{1}{T_{i+1}}\Big)|\Delta U| \approx -\log p_a,$$

  where $|\Delta U|$ *is the typical energy difference (e.g., the mean energy change under the target distribution) and* p_a *is the lower bound for the acceptance rate.*

Remark 1. *The rationale behind the choice of temperature* T_i *is to make the acceptance probability is relatively large since*

$$\frac{\pi_i(x_{i+1}^{(n)})\pi_{i+1}(x_i^{(n)})}{\pi_i(x_i^{(n)})\pi_{i+1}(x_{i+1}^{(n)})} \sim \exp\left(-\left(\frac{1}{T_i} - \frac{1}{T_{i+1}}\right)\Delta U\right).$$

5 Kinetic Monte Carlo

Kinetic Monte Carlo is also called BKL algorithm [9]. It is widely used in simulating crystal growth.

- Drawbacks of standard MC:

  At the metastable state σ_m, suppose the proposal state is σ', then

$$r = e^{-\beta\Delta H}, \quad \Delta H = H(\sigma') - H(\sigma_m).$$

  If $r \ll 1$, rejection occurred very often! The sample sequence will be like

$$\sigma_m, \sigma_m, \ldots, \sigma_m, \sigma_{new} \ldots.$$

  That's very inefficient!

  KMC aims to setup a **rejection free** algorithm.

- Generation of new state:

  Consider 2D Ising model: (ten-fold way in BKL algorithm)

  For a given state σ, there are 10 kinds of flips(single flip proposal):

Class	Spin	Number of spins up (nearest neighbors)
1	$\uparrow$	4
2	$\uparrow$	3
3	$\uparrow$	2
4	$\uparrow$	1
5	$\uparrow$	0
6	$\downarrow$	4
7	$\downarrow$	3
8	$\downarrow$	2
9	$\downarrow$	1
10	$\downarrow$	0

Table 1: Classification of spins in the 10-fold way

  There are 10 kinds of flipping probability $P_j = \min(1, \exp(-\beta \Delta H_j)),\ j = 1, \ldots, 10$. Suppose there are n_j sites at j class $j = 1, \ldots, 10$. Define

$$Q_i = \sum_{j=1}^{i} n_j P_j, \quad i = 1, \ldots, 10,$$

  then the BKL algorithm is as follows:

  Algorithm 5. *(BKL Algorithm)*

  Step1 Generate $R \sim \mathcal{U}[0, Q_{10})$*;*

  Step2 Identify $Q_{i-1} \leq R < Q_i, \quad (Q_0 = 0)$*;*

  Step3 Randomly choose one site to flip in class i*.*

- Time increment between two flips:

  Suppose on the average there is one attempted flip per lattice site in time τ (physical time), (# of sites $= N = M^2$) then

$$\frac{Q_{10}}{N} : \text{ Probability of flipping for a spin (only one) on a given attempt.}$$

  Note that the above procedure has homogenized the successful flipping probability to each site. We have the successful flipping probability for one site in unit time

$$\frac{Q_{10}}{N} \Big/ \frac{\tau}{N} = \frac{Q_{10}}{\tau} : \text{ Flip one spin unit time.}$$

  Define $P(\Delta t)$ is the probability that no flip occurs before time Δt has elapsed since the previous flip, then

$$P(\Delta t) - P(\Delta t + dt) = P(\Delta t) \cdot \frac{Q_{10}}{\tau} dt,$$

so one has

$$P(\Delta t) = \exp(-\frac{Q_{10}\Delta t}{\tau}), \quad P(0) = 1.$$

i.e. the time increment

$$\Delta t = -\frac{\tau}{Q_{10}} \ln R, \quad R \sim \mathcal{U}[0,1],\ 0 \leq R \leq 1.$$

Remark 2. *Essence: The Markov chain in Metropolis algorithm is some skeleton of a continuous time Q-process with Q-matrix*

$$q_{ij} = 1_{Q_{ij}} A_{ij} \tag{1}$$

in KMC, where $1_{Q_{ij}}$ *is defined as* $1_{Q_{ij}} = 1$ *if* $Q_{ij} > 0$ *and* $1_{Q_{ij}} = 0$ *otherwise.* A_{ij} *is the acceptance probability* P *shown above.*

Remark 3. *If one applies KMC to compute the ensemble average, the time increment occurs as a weight for different states.*

Remark 4. *KMC can simulate the non-equilibrium process such as crystal growth, but the connection between the process and the real physics is not clear!*

6 Homeworks

1. Write down the transition kernel of simulated tempering method (transition probability matrix in the case of discrete state Markov chain).

2. Write down the transition kernel of parallel tempering method (transition probability matrix in the case of discrete state Markov chain).

References

[1] J. Liu, Monte Carlo strategies in scientific computing, Springer-Verlag, 2001.

[2] C.J. Geyer, Markov chain Monte Carlo maximum likelihood, in E. Keramigas (ed.), COmputing Science and statistics: The 23rd symposium on the interface, Interface Foundation, Fairfax, 156-163, 1991.

[3] C.J. Geyer and E.A. Thompson, Annealing Markov Chain Monte Carlo with applications to ancestral inference, J. Amer. Stat. Assoc. 431 (1995), 909-919.

[4] E. Marinari and G. Parisi, Simulated tempering: a new Monte Carlo scheme, Europhys. Lett. 19 (1992), 451-458.

[5] J. Liu and C. Sabatti, Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions, in J. Bernardo, J. Berger, A. Dawid and A. Smith (eds.), Bayesian statistics, 6, 386-413, Oxford University Press, New York, 1998.

[6] R.H. Swendsen and J.S. Wang, Nonuniversal critical dynamics in Monte Carlo simulations, Phys. Rev. Lett. 58 (1987), 86-88.

[7] D.M. Higdon, Auxiliary variable methods for Markov Chain Monte Carlo with applications, J. Amer. Stat. Assoc. 93 (1998), 585-595.

[8] U. Wolff, Collective Monte Carlo updating for spin systems, Phys. Rev. Lett. 62 (1989), 361-364.

[9] A.B. Bortz, M.H. Kalos and J.L. Lebowitz, J. Comp. Phys. 17(1975), 10-18.

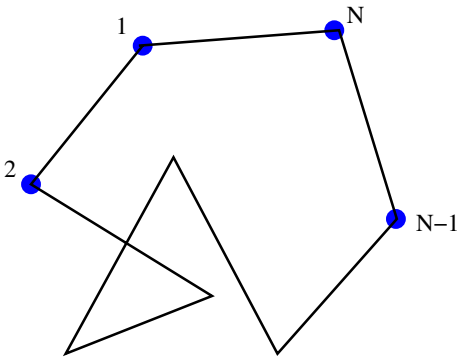
Lecture 9 Simulated Annealing and QMC *

Tiejun Li

1 Simulated Annealing

We already have very efficient algorithms for traditional convex programming. But how about the non-convex programming problems, such as the following combinatorial optimization problem?

Example 1. *(Traveling Salesman Problem) Suppose there are N cities and there exists one path ($l_{ij} = l_{ji}$) for each two. Try to find a minimal path passing all the cities such that each city is passed and only passed one time.*



$$\min_{x\in X} H(x) = \sum_{i=1}^{N} l_{x_i x_{i+1}}, \quad x_{N+1} := x_1.$$

$X = \{(x_1, \ldots, x_N),\ x_1, \ldots, x_N$ is a permutation of $1, 2, \ldots, N\}$

Figure 1: Traveling Salesman Problem

The number of all the possible paths is $\frac{N!}{2}$. It is a typical combinatorial explosion problem(NP problem). This number increases exponentially with N, and there isn't any rules for the function $H(x)$. The traditional algorithms are inapplicable here.

Example 2. *(Image smoothing problem) Suppose there are J pixels for an image, and there are 256 colors for each pixel. Any image can be represented as one element in*

$$X = \{(x_1, \ldots, x_J) \ : \ x_i \in \{0, 1, \ldots, 255\}\}.$$

The smoothness of an image is defined as

$$H(x) = \alpha \sum_{<s,t>} (x_s - x_t)^2, \qquad \alpha > 0,$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

where $< s,t >$ means the neighboring pixels in the lattice among $x = (x_1, \ldots, x_J)$. Then define the comparison function for images x and y where y is the reference image

$$H(x|y) = \alpha \sum_{<s,t>} (x_s - x_t)^2 + \frac{1}{2\sigma^2} \sum_s (x_s - y_s)^2.$$

An image recovering problem for polluted y may be proposed as minimizing the following function:

$$\min_{x \in X} H(x|y).$$

The number of all the possible states is 256^J! Traditional algorithms are still inapplicable here!

Simulated annealing algorithm is one of the framework to handle this kind of non-convex global optimization problem from stochastics recent years [3, 5]. While the effectivity is still under discussion.

1.1 Basic framework

For optimization problem

$$\min_{x \in X} H(x),$$

Define the global minimizers of $H(x)$

$$M = \{x_0 \ : \ H(x_0) = \min_{x \in X} H(x)\},$$

and introduce the parameter $\beta > 0$, define

$$\Pi^\beta(x) = \frac{1}{Z_\beta} e^{-\beta H(x)}, \qquad Z_\beta = \sum_{x \in X} \exp(-\beta H(x)),$$

then $\Pi^\beta(x)$ is a probability distribution on X.

Theorem 1. *$\Pi^\beta(x)$ has the property*

$$\lim_{\beta \to +\infty} \Pi^\beta(x) = \begin{cases} \frac{1}{|M|} & \text{if } x \in M, \\ 0 & \text{else.} \end{cases}$$

and if β is sufficiently large, then $\Pi^\beta(x)$ is monotonely increasing as a function of β for any $x \in M$, and $\Pi^\beta(x)$ is monotonely decreasing as a function of β for any $x \notin M$.

Proof. Rewrite

$$\Pi^\beta(x) = \frac{e^{-\beta(H(x)-m)}}{\sum_{z:H(z)=m} e^{-\beta(H(z)-m)} + \sum_{z:H(z)>m} e^{-\beta(H(z)-m)}}$$
$$\overset{\beta \to +\infty}{\longrightarrow} \begin{cases} \frac{1}{|M|}, & x \in M, \\ 0, & x \notin M, \end{cases}$$

where $m = \min_x H(x)$.

If $x \in M$, we have

$$\Pi^\beta(x) = \frac{1}{|M| + \sum_{z:H(z)>m} e^{-\beta(H(z)-m)}},$$

then $\Pi^\beta(x)$ monotonely increases with β increasing.

If $x \notin M$, we have

$$\begin{aligned}\frac{\partial \Pi^\beta(x)}{\partial \beta} &= \frac{1}{\tilde{Z}_\beta^2}\Big(e^{-\beta(H(x)-m)}(m-H(x))\tilde{Z}_\beta - e^{-\beta(H(x)-m)}\sum_{z\in X} e^{-\beta(H(z)-m)}(m-H(z))\Big)\\ &= \frac{1}{\tilde{Z}_\beta^2}\Big(e^{-\beta(H(x)-m)}\big[(m-H(x))\tilde{Z}_\beta - \sum_{z\in X} e^{-\beta(H(z)-m)}(m-H(z))\big]\Big),\end{aligned}$$

where

$$\tilde{Z}_\beta \triangleq \sum_{z\in X}\exp(-\beta(H(z)-m)).$$

Pay attention that

$$\lim_{\beta\to+\infty}\big[(m-H(x))\tilde{Z}_\beta - \sum_{z\in X} e^{-\beta(H(z)-m)}(m-H(z))\big] = |M|(m-H(x)) < 0,$$

The proof is completed.

Remark 1. *The construction of $\Pi^\beta(x)$ opens a new way to optimize $H(x)$ via stochastics. Theorem 1 shows that if we can generate the random sequence with distribution $\Pi^\beta(x)$, then the random numbers will finally jump among the minimizers when $\beta = +\infty$. This procedure is called* **annealing**. *It corresponds to the physical crystallization. In physics, β corresponds to $1/T$, where T is temperature. Global energy minimization means a perfect crystal without defects. The observed crystals with defects in nature can be understood as the local minimum state. In order to obtain a perfect crystal, one may image the following process: The crystals will take the form of liquids in the high temperature, then one decreases the temperature very slowly until the perfect crystal forms at the zero temperature. This is the basic idea of simulated annealing.*

The random number generation with distribution $\Pi^\beta(x)$ can be created by Metropolis algorithm.

1.2 Theoretical results

Assuming the Metropolis sampler for simulated annealing is

$$P^\beta(\sigma,\sigma') = \begin{cases} G(x,y)\frac{\pi^\beta(y)}{\pi^\beta(x)}, & \pi^\beta(y) < \pi^\beta(x) \text{ and } x \neq y,\\ G(x,y), & \pi^\beta(y) \geq \pi^\beta(x) \text{ and } x \neq y,\\ 1-\sum_{z\neq x} P^\beta(x,z) & x = y.\end{cases}$$

where $G(x,y)$ is the proposal matrix. It is symmetric as before.

In order to state the fundamental theorem of simulated annealing, we define the follows.

Definition 1. *(Neighborhood system) The neighborhood system of x is defined as $N(x) = \{y \in X | x \neq y, G(x,y) > 0\}$.*

Definition 2. *Given x and y, if there exists sequence $x = u_0, u_1, \ldots, u_{\sigma(x,y)} = y$ such that $u_{j+1} \in N(u_j)$ for any $j = 0, 1, \ldots, \sigma(x,y)-1$, then we say that the states x and y communicate, where $\sigma(x,y)$ is the length of the shortest path along which x and y communicate.*

Definition 3. *The maximal local increase of energy is defined as*

$$\Delta = \max\{H(y) - H(x) : x \in X, y \in N(x)\}.$$

Theorem 2. *(Fundamental theorem of simulated annealing) Suppose that X is a finite set, $H(x)$ is a nonconstant function, $G(x, y)$ is a symmetric irreducible proposal matrix,*

$$\tau = \max\{\sigma(x, y) : x, y \in X\}.$$

If the annealing procedure is chosen such that $\beta(n) \leq \frac{1}{\tau\Delta} \ln n$, then for any initial distribution ν, we have

$$\lim_{n \to +\infty} \|\nu P^{\beta(1)} \cdots P^{\beta(n)} - \Pi^{\infty}\| = 0.$$

The proof of the theorem may be refereed to [5].

Remark 2. *Theorem 2 shows that the annealing rate must be slow enough such that $\beta(n) \leq \frac{1}{\tau\Delta} \ln n$. It is a very very slow rate because $n \geq \exp(\tau\Delta\beta(n))$, we need $n \sim \exp(N_0)$ if $\beta(n) = N_0 \gg 1$. This means high accuracy needs exponential computing time, which is impossible for realistic computation. We should take more rapid annealing rates such as $\beta(n) \sim p^{-n}$ $(p \lesssim 1)$ or others. Of course, it has no theoretical foundations. The implementation details may be refereed to [2].*

2 Quasi-Monte Carlo Method

The standard MC is of $O(\frac{\sigma}{\sqrt{N}})$. In order to improve the accuracy, one has two choices

- Take N very large — Huge computational effort;
- Variance reduction techniques.

In the follows we will introduce the QMC to replace the pseudo-random sequence with quasi-random sequence. It improves the convergence rate to $O((\ln N)^k N^{-1})$, where k depends on the space dimension. Finally we will find that QMC is essentially a deterministic method which is very similar with MC. The main contents may be refereed to [1].

2.1 Discrepancy

The concept of discrepancy is an estimate of the uniformity of the points. For N points $\{x_n\}_{n=1}^N$ belonging to the unit d-cube $I^d = [0, 1]^d$, define

$$R_N(J) = \frac{1}{N} \#\{x_n \in J\} - m(J) \tag{1}$$

for any set $J \subset I^d$, where $\#\{x_n \in J\}$ means the number of the points in set J, and $m(J)$ is the measure of J. Intuitively $R_N(J)$ is the difference between the exact volume and the random sampling estimate of the volume.

Definition 4. *Define the whole set of rectangles in I^d as*

$$E = \{J(x, y) \ : \ (0, 0, \ldots, 0) \leq x \leq y \leq (1, 1, \ldots, 1)\},$$

where $x \leq y$ means $x_i \leq y_i$,$i = 1, \ldots, d$, $J(x, y)$ means the set of rectangles with the lower left node x and the upper right node y. Define

$$E^* = \{J(0, y) \ : \ (0, 0, \ldots, 0) \leq y \leq (1, 1, \ldots, 1)\}.$$

Definition 5. *The L^∞-discrepancy of a sequence $\{x_n\}_{n=1}^N$ is defined as*

$$D_N = \sup_{J \in E} |R_N(J)|;$$

and the L^2-discrepancy

$$T_N = \Big(\int_{(x,y) \in I^{2d}, x \leq y} R_N(J(x, y))^2 dxdy \Big)^{\frac{1}{2}}.$$

The L^p-discrepancy can be defined similarly. Specially we define the discrepancy

$$D_N^* = \sup_{J \in E^*} |R_N(J)|,$$
$$T_N^* = \left(\int_{I^d} R_N(J(0, x))^2 dx \right)^{\frac{1}{2}}.$$

2.2 Total variation

In 1D case, the total variation of a function is defined as the sum of the jumps:

$$V[f] = \sup_{\tau} \sum_i |f(x_{i+1}) - f(x_i)|,$$

where τ is taken to all the possible partitions of the domain. If f is differentiable, then

$$V[f] = \int_0^1 |df| = \int_0^1 |f'(x)| dx.$$

The total variation of function f in unit d-cube $[0, 1]^d$ is defined as

$$V[f] = \int_{I^d} \Big| \frac{\partial^d f}{\partial x_1 \cdots \partial x_d} \Big| dx_1 \cdots dx_d + \sum_{i=1}^d V[f_1^{(i)}],$$

where $f_1^{(i)}$ is the restriction of f on the boundary $x_i = 1$. It is a recursive definition of total variation.

Theorem 3. *(Koksma-Hlawka) For any sequence $\{x_n\}_{n=1}^N \subset I^d$ and the function f with bounded variation in I^d, the integration error $\mathcal{E}$ obeys the following inequality*

$$\mathcal{E}[f] \leq V[f] D_N^*,$$

where $\mathcal{E}[f] \triangleq |I[f] - I_N[f]| = | \int_{I^d} f(x) dx - \frac{1}{N} \sum_{i=1}^N f(x_i)|$.

Proof. We only present the intuitive proof here.

For the function $f(x)$ which takes value 0 on the boundary of I^d, define

$$R(x) = R_N(J(0,x)),$$

then

$$dR(x) = \{\frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i) - 1\}dx,$$

where $dR = \frac{\partial^d R}{\partial x_1 \cdots \partial x_d}$, $dx = dx_1 \cdots dx_d$. $\delta(x-x_i)$ is the Dirac-δ function centered at x_i, then we have

$$\begin{aligned}
\mathcal{E}[f] &= \Big|\int_{I^d} f(x)dx - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\Big| \\
&= \Big|\int_{I^d} \{1 - \frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i)\}f(x)dx\Big| \\
&= \Big|\int_{I^d} R(x)df(x)\Big| \\
&\leq (\sup_x R(x))\int_{I^d}|df(x)| = D_N^* V[f].
\end{aligned}$$

Koksma-Hlawka theorem shows that the discretization error can be described by the total variation $V[f]$ and the discrepancy for the sample points. QMC gives some special quasi random sequences which have good discrepancy properties. It is a pure number theoretic result.

2.3 Quasi Monte Carlo integration

Definition 6. *A sequence $\{x_n\}_{n=1}^N \subset I^d$ is called quasi-random if*

$$D_N \leq C(\ln N)^k N^{-1},$$

in which c and k are constants that are independent of N, but may depend on the dimension d.

What follows are some typical quasi random sequences:

- Van der Corput sequence($d = 1$):

  The generation of sequence $\{x_i\}_{i=1}^N$ is composed of two steps:

  Step1. Write out n in base 2:

  $$n = \big(a_m a_{m-1} \cdots a_1 a_0\big)_2,$$

  where $(\cdot)_2$ means in base 2, $a_i \in \{0,1\}$ is the i-th bit of n;

  Step2. Generate x_n in base 2

  $$x_n = \Big(0.\, a_0 a_1 \cdots a_m\Big)_2.$$

- Halton sequence($d > 1$):

  Denote $x_n = (x_n^1, x_n^2, \ldots, x_n^d)$, where the k-th component x_n^k is obtained by two steps.

Step1. Write out n in base p_k. (where p_k is the k-th prime number, e.g. $p_1 = 2, p_2 = 3$)

$$n = \left(a^k_{m_k} a^k_{m_k-1} \cdots a^k_1 a^k_0\right)_{p_k};$$

Step2. Generate x^k_n in base p_k:

$$x^k_n = \left(0.\, a^k_0 a^k_1 \cdots a^k_{m_k}\right)_{p_k}.$$

The number theorists has proved

$$D_N(\text{Halton}) \leq C_d (\ln N)^d N^{-1}.$$

Some other quasi random number sequences such as Sobol sequence, Faure sequence etc. may be refereed to [4].

2.4 Limitations of QMC

QMC has the following limitations:

- QMC are designed for integration and are not directly applicable to simulations. This is because of the correlations between the points of a quasi-random sequence.
- Because the theoretical basis of QMC is from Koksma-Hlawka theorem, and the generation style of quasi-random numbers is very special, it is commonly applied to the integral in rectangle with the form $\int_{I^d} f(x)dx$. For the powerful Metropolis algorithm in statistical physics, how to design the corresponding QMC version is an open problem.
- QMC is found to lose its effectiveness when the dimension of the integral becomes large. This can be anticipated from the bound $(\ln N)^d N^{-1}$ on discrepancy. For large dimension d, this bound is dominated by the $(\ln N)^d$ term unless $N > 2^d$;
- QMS is found to lose its effectiveness if the integrand f is not smooth. The factor $V[f]$ in the Koksma-Hlawka inequality is an indicator of this dependence.

All in all, QMC is suitable for the integration in which the space dimension is not so big, the integrand f is relatively smooth. Though it has better convergence rate than Monte Carlo method, its applicability is limited.

3 Homeworks

HW1. If we apply the argument in the simulated annealing to the continuous space case with smooth energy $V(x)$ and isolated minimizers, what can we say about the limit as $\beta \to \infty$?

References

[1] R.E. Caflisch, Monte Carlo and Quasi-Monte Carlo methods, Acta Numerica, Vol. 7, 1-49, 1998.

[2] Lishan Kang, Non-Numerical Parallel Algorithms: Simulated Annealing Algorithms, Science Press, Beijing, 1994.(In Chinese)

[3] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by Simulated Annealing, Science 220(1983), 671-680.

[4] W.H. Press et al., Numerical Recipes: the Art of Scientific Computing, Cambridge university press, Cambridge, 1986.

[5] G. Winkler, Image Analysis, Random Fields and Dynamic Monte Carlo Methods, Springer-Verlag, Berlin and New York, 1995.

Lecture 10 Random Walk and Brownian motion *

Tiejun Li

1 1D Symmetric Random Walk

Example 1. *(1D Random Walk) Suppose a particle suffers displacements along a straight line from the origin, denote its position $X_n \in \mathbb{Z}$. Let ξ_i are i.i.d. random moves such that $\xi_i = \pm 1$ with probability $\frac{1}{2}$, and let*

$$X_n = \xi_1 + \xi_2 + \ldots + \xi_n \quad (i.e.\ \ X_0 = 0)$$

$\{X_n\}$ is called a unconstrained symmetric random walk on $\mathbb{Z}$. Given $X_n = i$, we have

$$\begin{aligned} P\{X_{n+1} = i \pm 1 |\ X_n = i\} &= \frac{1}{2}, \\ P\{X_{n+1} = \textit{anything else} |\ X_n = i\} &= 0. \end{aligned}$$

It is a typical example of the simplest Markov chains.

After taking N steps, the particle could be at any of the points

$$-N, -N+2, \ldots, \ldots, N-2, N.$$

1.1 Distribution of X_N

One basic question is the probability $W(m, N) = \text{Prob}\{X_N = m\}$ that the particle arrives at the point m after suffering N displacements.

It is not difficult to find that $W(m, N)$ obeys binomial distribution

$$W(m, N) = \frac{N!}{(\frac{N+m}{2})!(\frac{N-m}{2})!}\Big(\frac{1}{2}\Big)^N,$$

and it is easy to note that m can be odd or even only according as N is odd or even.

The expectation position and mean square deviation are

$$\mathbb{E}X_N = 0, \quad \mathbb{E}X_N^2 = N,$$

then the root mean square displacement is $\sqrt{N}$.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Definition 1. *(Diffusion coefficient) The 1D diffusion coefficient D is defined as*

$$D = \frac{\langle (X_N - X_0)^2 \rangle}{2N}.$$

It is assumed $\mathbb{E}X_N = X_0$ here. In general continuous case, it is defined as

$$D = \lim_{t\to\infty} \frac{\langle (X_t - X_0)^2 \rangle}{2dt},$$

where d is the space dimension.

For this simplest random walk, $D = \frac{1}{2}$.

Next we consider the case $N, m \gg 1$, and $m \ll N$ since we will rescale the process with the relation $x = ml, t = N\tau$ and $l \sim \mathcal{O}(\sqrt{\tau})$, $\tau \to 0$. Thus $m/N = x/t \cdot \tau/l \to 0$. So only the range $m \ll N$ matters. By Stirling's formula

$$\log n! = (n + \frac{1}{2}) \log n - n + \frac{1}{2} \log 2\pi + O(n^{-1}) \quad (n \to +\infty),$$

we have

$$\begin{aligned} \log W(m, N) \quad \approx \quad & (N + \frac{1}{2}) \log N - \frac{1}{2}(N + m + 1) \log \Big[\frac{N}{2}(1 + \frac{m}{N})\Big] \\ & -\frac{1}{2}(N - m + 1) \log \Big[\frac{N}{2}(1 - \frac{m}{N})\Big] - \frac{1}{2} \log 2\pi - N \log 2. \end{aligned}$$

Since $m \ll N$ we have Taylor series expansion for $x \ll 1$

$$\log(1 + x) = x - \frac{1}{2}x^2 + O(x^3),$$

thus

$$\log W(m, N) \approx -\frac{1}{2} \log N + \log 2 - \frac{1}{2} \log 2\pi - \frac{m^2}{2N} + O\Big(\Big(\frac{m}{N}\Big)^2\Big).$$

In other words, one obtains the the asymptotic formula

$$W(m, N) \approx \Big(\frac{2}{\pi N}\Big)^{\frac{1}{2}} \exp(-\frac{m^2}{2N}).$$

An interesting thing is to take the continuum limit of random walk. Now suppose we rescale the random walk with the spatial steplength l and the time spacing τ for each movement, we take the limit in the following sense when considering the point (x, t)

$$N, m \to \infty, \quad l, \tau \to 0, \text{ and } \quad N\tau = t, \ ml = x. \tag{1}$$

To make the continuum limit physically reasonable, we also ask to fix the diffusion coefficient

$$D = \frac{\langle (X_{N\tau} - X_0)^2 \rangle}{2N\tau} = \frac{l^2}{2\tau}$$

in the limit. Consider the intervals Δx which are large compared with the length l, we have the probability that $y \in (x - \Delta x/2, x + \Delta x/2)$ for the continuous probability density $W(x, t)$ satisfies

$$W(x, t)\Delta x \approx \int_{x-\Delta x/2}^{x+\Delta x/2} W(y, t)dy \approx \sum_{\substack{m' \in \{m, m\pm 2, m\pm 4, \ldots\} \\ m'l \in (x-\Delta x/2, x+\Delta x/2)}} W(m', N) \approx W(m, N)\frac{\Delta x}{2l} \quad (x = ml)$$

since m can take only even or odd values depending on whether N is even or odd. Combining the results above one has

$$W(x,t)\Delta x = \frac{1}{\sqrt{2\pi t\frac{l^2}{\tau}}}\exp(-\frac{x^2}{2t\frac{l^2}{\tau}})\Delta x,$$

thus the limiting probability density at time t

$$W(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp(-\frac{x^2}{4Dt}).$$

1.2 Random walk with reflecting and absorbing Barriers

Case 1: A reflecting barrier at $m = m_1$;

Suppose $m_1 > 0$ We now ask the probability $W(m, N; m_1)$ that the particle will arrive at $m(\leq m_1)$ after N steps.

This problem may be solved very efficiently in the $(m - N)$ plane in a neat way.

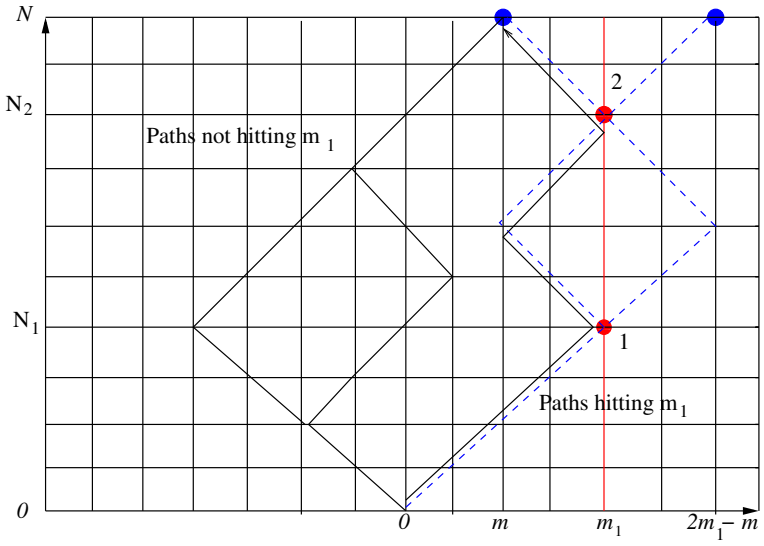


Figure 1: Schematics of reflection principle.

From Fig. 1, the actual sample paths are shown with solid lines (including the reflected path), and the paths crossing the barrier m_1 in the unrestricted random walk case are shown with dashed lines. These paths can be classified into two classes. One class only contains the paths not hitting m_1 and finally reaching m; the other class contains the paths hitting m_1 before time N and finally reaching m_1 or $2m_1 - m$. We have the following two assertions on these paths:

- In the unrestricted random walk, all of the sample paths have equal probability $1/2^N$;
- The probability of the reflected paths which hits m_1 is equal to the sum of the probability of the paths hitting m_1 and reaching m and the paths reaching $2m_1 - m$. A simple argument to prove this is to observe that the reflecting probability is 1 at the reflection point shown as points 1 and 2 in the

figure. From $1 = 1/2 + 1/2$, this decomposition actually decompose the paths into those go leftwards and rightwards with equal probability, which corresponds to all the paths just stated.

- The number of the paths hitting m_1 and hitting m finally is equal to that of the paths hitting $2m_1 - m$ finally. This can be understood because these paths have to cross m_1 and we can denote the final hitting time as N_2 as shown in the figure. So after N_2, the paths go leftwards or rightwards with mirror symmetry to hit m or $2m_1 - m$. Before N_1, the paths can be either branch.

These assertions are called the **reflection principle**, which is the basis of the following calculations for reflection and absorbing barrier problem.

So we have the following identity

$$W_r(m, N; m_1) = W(m, N) + W(2m_1 - m, N).$$

If we take large N limit we have

$$W_r(m, N; m_1) \approx \Big(\frac{2}{\pi N}\Big)^{\frac{1}{2}} \big[\exp(-\frac{m^2}{2N} + \exp(-\frac{(2m_1 - m)^2}{2N}\big],$$

then passing to the continuum limit we have

$$W_r(x, t; x_1) = \frac{1}{\sqrt{4\pi D t}} \big[\exp(-\frac{x^2}{4Dt}) + \exp(-\frac{(2x_1 - x)^2}{4Dt})\big], \tag{2}$$

and we may note in this case

$$\frac{\partial W_r}{\partial x}\Big|_{x=x_1} = 0.$$

with W is defined in (2).

Case 2: Absorbing wall at $m = m_1$;

Similarly as before we easily deduce that

$$W_a(m, N; m_1) = W(m, N) - W(2m_1 - m, N).$$

by reflection principle.

In the large N limit we have

$$W_a(m, N; m_1) \approx \Big(\frac{2}{\pi N}\Big)^{\frac{1}{2}} \big[\exp(-\frac{m^2}{2N}) - \exp(-\frac{(2m_1 - m)^2}{2N})\big],$$

and the continuum limit is

$$W_a(x, t; x_1) = \frac{1}{\sqrt{4\pi D t}} \big[\exp(-\frac{x^2}{4Dt}) - \exp(-\frac{(2x_1 - x)^2}{4Dt})\big], \tag{3}$$

and we may note in this case

$$W_a(x, t; x_1) = 0.$$

with W_a is defined in (3).

Define the first hitting probability $a(m_1, N) = \text{Prob}\{X_N = m_1, \text{ and } X_n < m_1, \forall n < N\}$ that taking N steps the particle will arrive at m_1 without ever hitting $m = m_1$ at any earlier step. Then we have

$$a(m_1, N) = \frac{1}{2} W_a(m_1 - 1, N - 1; m_1) = \frac{m_1}{N} W(m_1, N)$$

by the above formula and the relation $W(m,N)=\frac{N}{N+m}W(m-1,N-1)$. In the large N limit we have

$$a(m_1,N)\approx\frac{m_1}{N}\Big(\frac{2}{\pi N}\Big)^{\frac{1}{2}}\exp(-\frac{m_1^2}{2N}).$$

The continuous probability density $a(m_1,t)$ becomes

$$a(m_1,t)\Delta t\approx a(m_1,N)\frac{\Delta t}{2\tau}\quad(t=N\tau)$$

In the continuum limit one obtains

$$a(x_1,t)=\frac{x_1}{t}\frac{1}{\sqrt{4\pi Dt}}\exp(-\frac{x_1^2}{4Dt}).$$

We may note in this case

$$a(x_1,t)=-D\frac{\partial W}{\partial x}\Big|_{x=x_1}.$$

with W is defined in (3).

2 Arcsine law and the law of iterated logarithm

To simplify the notations, we define the first hitting time

$$\sigma_{2n}=\min\{1\le k\le 2n: S_k=0\}$$

and we define $\sigma_{2n}=+\infty$ if $S_k\neq 0$ for $1\le k\le 2n$. For $0\le k\le n$ we define

$$u_{2k}=\mathbb{P}(S_{2k}=0),\quad f_{2k}=\mathbb{P}(\sigma_{2n}=2k). \tag{4}$$

It is clear that $u_{2k}=C_{2k}^k\cdot 2^{-2k}$. From the reflection principle, we have

$$f_{2k}=2\frac{1}{2}\cdot\frac{1}{2k-1}W(1,2k-1)=\frac{1}{2k}u_{2(k-1)}=u_{2(k-1)}-u_{2k}. \tag{5}$$

Now define $P_{2k,2n}$ be the probability that during the interval $[0,2n]$ the particle spends $2k$ units of time on the positive side (We say that the particle is on the positive side in the interval $[m-1,m]$ if one, at least, of the value S_{m-1} and S_m is positive).

Lemma 1. *Let $u_0=1$ and $0\le k\le n$. Then*

$$P_{2k,2n}=u_{2k}\cdot u_{2n-2k}. \tag{6}$$

Proof. At first let us show that (6) holds for $k=0$. Suppose we have a path with $S_{2n}=0$ and $\min_{0\le k\le 2n}X_k=-m$, where $m>0$. Denote $l=\min\{k|X_k=-m\}$. We can map this path into a path only in the positive side. Take a reflection of the path $\{X_k\}_{0\le k\le l}$ with respect to the axis $t=l$ and denote the new path by $\{\tilde{X}_k\}_{0\le k\le l}$ such that $\tilde{X}_k=X_{l-k}$. Concatenate $\tilde{X}_0$ to the point $(2n,0)$ and translate the left endpoint of the new path into the origin. With such manipulation, we get a path on the positive side and the right endpoint is $(2n,2m)$. Conversely, for each path on the positive side with the right endpoint is $(2n,2m)$, we take $l=\max\{k|X_k=m\}$. We can cut the part beyond $t=k$, make a reflection with respect

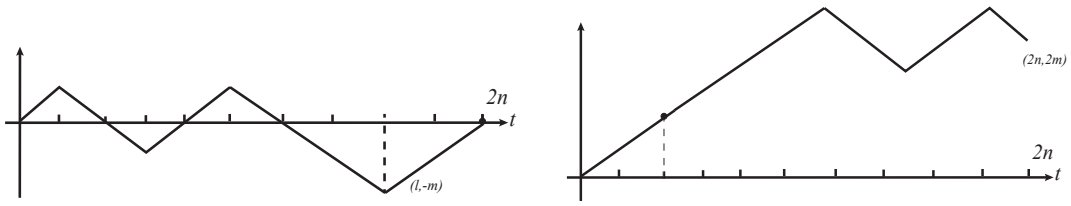


Figure 2: Schematics of construction from a path with $S_{2n} = 0$ to a new path on the positive side.

to $t = l$, concatenate it to the left endpoint of the rest part and translate the whole path into the origin, we then get a new path with $S_{2n} = 0$. A special case is illustrate in Figure 2. The case for $k = n$ is trivially true by symmetry and the case $k = 0$.

Then let us prove the following relation

$$u_{2k} = \sum_{r=1}^{k} f_{2r} \cdot u_{2(k-r)}. \tag{7}$$

Since $\{S_{2k} = 0\} \subset \{\sigma_{2n} \leq 2k\}$, we have

$$\{S_{2k} = 0\} = \{S_{2k} = 0\} \cap \{\sigma_{2n} \leq 2k\} = \sum_{r=1}^{k} \{S_{2k} = 0\} \cap \{\sigma_{2n} = 2r\} \tag{8}$$

Consequently

$$\begin{aligned} u_{2k} &= \mathbb{P}(S_{2k} = 0) = \sum_{r=1}^{k} \mathbb{P}(S_{2k} = 0, \sigma_{2n} = 2r) \\ &= \sum_{r=1}^{k} \mathbb{P}(S_{2k} = 0 | \sigma_{2n} = 2r) \mathbb{P}(\sigma_{2n} = 2r). \end{aligned} \tag{9}$$

But

$$\begin{aligned} \mathbb{P}(S_{2k} = 0 | \sigma_{2n} = 2r) &= \mathbb{P}(S_{2k} = 0 | S_1 \neq 0, \ldots, S_{2r-1} \neq 0, S_{2r} = 0) \\ &= \mathbb{P}(S_{2r} + (\xi_{2r+1} + \cdots + \xi_{2k}) = 0 | S_1 \neq 0, \ldots, S_{2r-1} \neq 0, S_{2r} = 0) \\ &= \mathbb{P}(S_{2r} + (\xi_{2r+1} + \cdots + \xi_{2k}) = 0 | S_{2r} = 0) \\ &= \mathbb{P}(\xi_{2r+1} + \cdots + \xi_{2k} = 0) = \mathbb{P}(S_{2(k-r)} = 0). \end{aligned} \tag{10}$$

Combing (9) and (10) we obtain (7). To prove (6), we apply the induction method. Now let $1 \leq k \leq n-1$. If the particle is on the positive side for exactly $2k$ instants, it must pass through zero. Let $2r$ be the time of first passage through zero. There are two possibilities: either $S_k \geq 0$, $k \leq 2r$, or $S_k \leq 0$, $k \leq 2r$.

The number of paths of the first kind is

$$(2^{2r} \cdot \frac{1}{2} f_{2r}) \cdot (2^{2(n-r)} \cdot P_{2(k-r),2(n-r)}) = \frac{1}{2} 2^{2n} f_{2r} P_{2(k-r),2(n-r)}.$$

The number of paths of the second kind is

$$\frac{1}{2} 2^{2n} f_{2r} P_{2k,2(n-r)}.$$

Consequently, for $1 \le k \le n-1$,

$$P_{2k,2n} = \frac{1}{2}\sum_{r=1}^{k} f_{2r} P_{2(k-r),2(n-r)} + \frac{1}{2}\sum_{r=1}^{k} f_{2r} P_{2k,2(n-r)}. \tag{11}$$

Suppose that $P_{2k,2m} = u_{2k} \cdot u_{2m-2k}$ holds for $m = k, k+1, \ldots, n-1$. Then by (7) and (11) we have (How is the induction applied here?)

$$\begin{aligned} P_{2k,2n} &= \frac{1}{2}u_{2n-2k}\sum_{r=1}^{k} f_{2r}u_{2k-2r} + \frac{1}{2}u_{2k}\sum_{r=1}^{k} f_{2r}u_{2n-2k-2r} \\ &= \frac{1}{2}u_{2n-2k}u_{2k} + \frac{1}{2}u_{2k}u_{2n-2k} = u_{2k}u_{2n-2k}. \end{aligned}$$

This completes the proof. □

Now let $\gamma(2n)$ be the number of time units that the particle spends on the positive axis in the interval $[0, 2n]$. Then when $x < 1$,

$$\mathbb{P}\Big\{\frac{1}{2} < \frac{\gamma(2n)}{2n} \le x\Big\} = \sum_{k,1/2<2k/2n\le x} P_{2k,2n}.$$

Since

$$u_{2k} \sim \frac{1}{\sqrt{\pi k}}$$

by Stirling's formula as $k \to \infty$, we have

$$P_{2k,2n} \sim \frac{1}{\pi\sqrt{k(n-k)}}$$

as $k, n-k \to \infty$.

Therefore

$$\sum_{\{k,1/2<2k/2n\le x\}} P_{2k,2n} - \sum_{k,1/2<2k/2n\le x} \frac{1}{\pi n}\cdot\Big[\frac{k}{n}\Big(1-\frac{k}{n}\Big)\Big]^{-\frac{1}{2}} \to 0, \quad n \to \infty,$$

Whence

$$\sum_{\{k,1/2<2k/2n\le x\}} P_{2k,2n} \to \frac{1}{\pi}\int_{\frac{1}{2}}^{x} \frac{dt}{\sqrt{t(1-t)}}, \quad n \to \infty.$$

From the symmetry,

$$\sum_{\{k,2k/2n\le 1/2\}} P_{2k,2n} \to \frac{1}{2}$$

and

$$\frac{1}{\pi}\int_{\frac{1}{2}}^{x} \frac{dt}{\sqrt{t(1-t)}} = \frac{2}{\pi}\arcsin\sqrt{x} - \frac{1}{2}.$$

Thus we have the following theorem:

Theorem 1 (Arcsine Law)**.** *The probability that the fraction of the time spent by the particle on the positive side is at most x tends to $\frac{2}{\pi}\arcsin\sqrt{x}$:*

$$\sum_{\{k,k/n\le x\}} P_{2k,2n} \to \frac{2}{\pi}\arcsin\sqrt{x}.$$

The following deep theorem is due to Hartman and Wintler (1941).

Theorem 2 (Law of Iterated Logarithm)**.** *Let $\xi_1, \ldots, \xi_n$ are i.i.d. R.V. with $\mathbb{E}\xi_i = 0, \mathrm{Var}\xi_i = \sigma^2 > 0$, Then*

$$\mathbb{P}\Big\{ \limsup \frac{S_n}{\sqrt{2\sigma^2 n \ln\ln n}} = 1 \Big\} = 1.$$

Remark 1. *Application of the above result to $-\xi_i$, one also obtains*

$$\mathbb{P}\Big\{ \liminf \frac{S_n}{\sqrt{2\sigma^2 n \ln\ln n}} = -1 \Big\} = 1.$$

3 Random Flights With Gaussian Displacements

In the general problem of random flights, the position $\boldsymbol{R}$ of the particle after N displacements is given by

$$\boldsymbol{R} = \sum_{i=1}^{N} \boldsymbol{r}_i,$$

where the $\boldsymbol{r}_i = (r_i^1, r_i^2, r_i^3)$'s denote the different displacements. Assume the probability that the ith displacement between $\boldsymbol{r}_i$ and $\boldsymbol{r}_i + d\boldsymbol{r}_i$ is given by

$$\tau_i(r_i^1, r_i^2, r_i^3) dr_i^1 dr_i^2 dr_i^3 = \tau_i d\boldsymbol{r}_i \quad (i = 1, \ldots, N).$$

Now we ask the probability $W_N(\boldsymbol{R})d\boldsymbol{R}$ that the position of the particle after N displacements lies in the interval $\boldsymbol{R}, \boldsymbol{R} + d\boldsymbol{R}$. The method presented in the following is originally devised by A.A. Markov.

It is a standard exercise to have

$$\widehat{W_N(\boldsymbol{R})} = \int W_N(\boldsymbol{R}) \exp(i\boldsymbol{\rho}\cdot\boldsymbol{R}) d\boldsymbol{R} = \prod_{i=1}^{N} \int \tau_i(\boldsymbol{r}_i) \exp(i\boldsymbol{\rho}\cdot\boldsymbol{r}_i) d\boldsymbol{r}_i = \prod_{i=1}^{N} \widehat{\tau_i(\boldsymbol{r}_i)}.$$

In the case of Gaussian distribution of random displacement $\boldsymbol{r}_i$, we have the pdf

$$\tau_i(\boldsymbol{r}_i) = \frac{1}{(2\pi l^2)^{\frac{3}{2}}} \exp(-\frac{|\boldsymbol{r}_i|^2}{2\ l^2}),$$

From the property of Fourier transform for Gaussian distribution, we have

$$W_N(\boldsymbol{R}) = \frac{1}{(2\pi N l^2)^{\frac{3}{2}}} \exp(-\frac{|\boldsymbol{R}|^2}{2Nl^2}).$$

Suppose the time spacing is τ each time and define the diffusion coefficient as before

$$D = \lim_{t\to 0} \frac{\langle (X_t - X_0)^2 \rangle}{2dt} = \frac{3Nl^2}{6N\tau} = \frac{l^2}{2\tau},$$

Then we have the continuum limit pdf for free Gaussian random flight

$$W(\boldsymbol{R}, t) = \frac{1}{(4\pi D t)^{\frac{3}{2}}} \exp\Big(-\frac{|\boldsymbol{R}|^2}{4Dt} \Big) \quad (t = N\tau).$$

Remark 2. *Similar results for $W_N(\boldsymbol{R})$ holds for other distributions which can be refereed in [1]. These results will be further clarified in next lecture on Brownian motion.*

Question 1. *How about more general reflecting and absorbing barrier problem in high dimensions?*

4 Einstein's work on the theory of Brownian motion

In 1905, A. Einstein published a seminal paper on the theory of Brownian motion (he also publishes two other seminal papers on Special Relativity and photoemission in this year). Two major points in Einstein's solution to Brownian motion are

(i) The motion is caused by the exceedingly frequent impacts on the pollen grain of the incessantly moving molecules of liquid in which it is suspended;

(ii) The motion of these molecules is so complicated that its effect on the pollen grain can only be described probabilistically in terms of exceedingly frequent statistically independent impacts.

His mathematical interpretation is as follows (1D version).

In a small time interval τ, the X-coordinates of an individual particle will increase by an amount Δ. There will be a certain "frequency law" for Δ

$$dn = n\phi(\Delta)d\Delta$$

where

$$\int_{-\infty}^{+\infty} \phi(\Delta)d\Delta = 1, \quad \phi(-\Delta) = \phi(\Delta),$$

and ϕ is only different from 0 for very small values of Δ.

Let $f(x,t)$ be the number of particles per unit volume, then

$$f(x,t+\tau)dx = \int_{-\infty}^{+\infty} f(x-\Delta,t)dx\phi(\Delta)d\Delta.$$

Since τ is small

$$f(x,t+\tau) = f(x,t) + \frac{\partial f}{\partial t}\tau,$$

furthermore

$$f(x-\Delta,t) = f(x,t) - \Delta\frac{\partial f}{\partial x} + \frac{\Delta^2}{2}\frac{\partial^2 f}{\partial x^2} + \cdots .$$

Thus

$$f(x,t) + \frac{\partial f}{\partial t}\tau = f\int_{-\infty}^{+\infty} \phi(\Delta)d\Delta + \frac{\partial f}{\partial x}\int_{-\infty}^{+\infty} \Delta\phi(\Delta)d\Delta + \frac{\partial^2 f}{\partial x^2}\int_{-\infty}^{+\infty} \frac{\Delta^2}{2}\phi(\Delta)d\Delta + \cdots .$$

Set

$$\frac{1}{\tau}\int_{-\infty}^{+\infty} \frac{\Delta^2}{2}\phi(\Delta)d\Delta = D$$

throwing h.o.t., we have

$$\frac{\partial f}{\partial t} = D\frac{\partial^2 f}{\partial x^2}.$$

His description contains very many of the major concepts which have been developed more and more generally and rigorously since then, such as

(i) Chapman-Kolmogorov equation;

(ii) Fokker-Planck equation;

(iii) Kramers-Moyal expansion;

etc.

5 Homeworks

- HW1. Prove that the continuum limit pdf $W(x,t)$ for free random walk satisfies the PDE

$$\begin{cases} \dfrac{\partial W}{\partial t} = D\dfrac{\partial^2 W}{\partial x^2}, \quad x \in R,\ t \geq 0 \\ W(x,t)\Big|_{t=0} = \delta(x). \end{cases}$$

- HW2. Prove that the continuum limit pdf $W(x,t)$ with reflecting barrier satisfies the PDE

$$\begin{cases} \dfrac{\partial W}{\partial t} = D\dfrac{\partial^2 W}{\partial x^2}, \quad x \leq x_1,\ t \geq 0 \\ W(x,t)\Big|_{t=0} = \delta(x), \\ \dfrac{\partial W}{\partial x}(x,t)\Big|_{x=x_1} = 0. \end{cases}$$

- HW3. Prove that the continuum limit pdf $W(x,t)$ with absorbing barrier satisfies the PDE

$$\begin{cases} \dfrac{\partial W}{\partial t} = D\dfrac{\partial^2 W}{\partial x^2}, \quad x \leq x_1,\ t \geq 0 \\ W(x,t)\Big|_{t=0} = \delta(x), \\ W(x,t)\Big|_{x=x_1} = 0. \end{cases}$$

- HW4 (optional). Give the detailed procedure for the induction in proving the Arcsine Law.

References

[1] S. Chandrasekhar, Stochstic problems and in physics and astronomy, Rev. Mod. Phys. 15(1943), 1-89.

[2] A.N. Shiryaev, Probability, Springer-Verlag, Berlin and Heidelberg, 1996.

Lecture 11 Stochastic Process and Brownian Motion *

Tiejun Li

1 Axiomatic Construction of Stochastic Process

Example 1.1. *Consider the independent fair coin tossing process described by the sequence*

$$X = (X_1, X_2, \ldots, X_n, \ldots) \in \{0,1\}^{\mathbb{N}},$$

where $X_n = 0$ or 1 if the nth output is 'Tail' (T) or 'Head' (H), respectively. Different trials are assumed to be independent and $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = 1/2$.

Notice that for this process the number of all possible outputs is uncountable. One can not define the probability of an event through summation of the probability of each atom as the case of discrete random variables. In fact, if we define $\Omega = \{H,T\}^{\mathbb{N}}$, the probability of an *atom* $\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \{H,T\}^{\mathbb{N}}$ is 0, i.e.

$$\mathbb{P}(X_1(\omega) = k_1, X_2(\omega) = k_2, \ldots, X_n(\omega) = k_n, \ldots) = \lim_{n\to\infty} \left(\frac{1}{2}\right)^n = 0, \quad k_j \in \{0,1\},\ j = 1,2,\ldots$$

and events like $\{X_n(\omega) = 1\}$ involve uncountably many atoms.

To set up a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for this process, it is natural to take $\Omega = \{H,T\}^{\mathbb{N}}$ and the σ-algebra $\mathcal{F}$ as the smallest σ-algebra containing all events of the form:

$$C = \left\{\omega | \omega \in \Omega, (\omega_j)_{j=1:m} \in C_m, C_m \subset \{H,T\}^m\right\} \tag{1.1}$$

for any $m \in \mathbb{N}$, i.e. the sets whose finite time projections are specified. These sets are called *cylinder sets*, which is meaningful from the experimental observation point of view. The probability measure $\mathbb{P}$ of an event of the form (1.1) is defined to be

$$\mathbb{P}(C) = \frac{1}{2^m}|C|.$$

Denote $\mathcal{C}$ the set of cylinder sets. One can easily show that $\mathcal{C}$ is an algebra which is only closed under finite union/intersection operation. To extend the probability measure $\mathbb{P}$ from $\mathcal{C}$ to $\mathcal{F}$, we need to verify that $\mathbb{P}$ is countably additive on $\mathcal{C}$.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Lemma 1.2. *If $A_n \downarrow A$ and $A_n \in \mathcal{C}$ is non-empty, then A is non-empty.*

With Lemma 1.2, we obtain if $A_n \downarrow \emptyset$, then $\mathbb{P}(A_n) \downarrow 0$, which is equivalent to the countable additivity. From the extension theorem of measures, this probability measure $\mathbb{P}$ is well-defined on $\mathcal{F}$.

Proof of Lemma 1.2. Denote

$$A_n = \{\omega | (\omega_1, \omega_2, \ldots, \omega_{m_n}) \in C_n\}$$

where $\omega_k \in \{H, T\}$. From the non-empty condition of A_n, there exists $\omega^n \in A_n$. Consider

$$\begin{matrix} \omega_1^1 & \omega_1^2 & \omega_1^3 & \cdots \\ \omega_2^1 & \omega_2^2 & \omega_2^3 & \cdots \\ \omega_3^1 & \omega_3^2 & \omega_3^3 & \cdots \\ \vdots & \vdots & \vdots & \ddots, \end{matrix}$$

there exist infinite superscripts n_k^1 such that $\omega_1^{n_k^1} = H$ or T always in the 1st row. Similar argument can be applied to the continued rows by an subsequence trick. Take the diagonal indices and define $n_k := n_k^k$ and $u_k := \omega_k^{n_k}$ for $k = 1, 2, \ldots$. Denote $u = (u_1, u_2, \ldots)$.

For any r, if $k \geq r$, one has $\omega_j^{n_k} = u_j$ for $1 \leq j \leq r$. For any n, if $k \geq n$, then $n_k \geq n$, and $\omega^{n_k} \in A_{n_k} \subset A_n$. So $(\omega_1^{n_k}, \omega_2^{n_k}, \ldots, \omega_{m_n}^{n_k}) \in C_n$. Take $k \geq m_n$. We get $\omega_j^{n_k} = u_j$ for $1 \leq j \leq m_n$, i.e. $u \in A_n$ for any n.

In summary, $u \in A$ and we are done. □

It is straightforward to check that for any cylinder set $F \in \{0, 1\}^{\mathbb{N}}$, the probability $\mathbb{P}(X(\omega) \in F)$ coincides with the definition we made in Example 1.1 for independent coin tossing process. We remark that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is not uniquely defined. Another natural way is to take $\Omega = \{0, 1\}^{\mathbb{N}}$, $\mathcal{F}$ the smallest σ-algebra containing all cylinder sets in Ω, and similar probability measure $\mathbb{P}$ on $\mathcal{F}$. With this choice we have

$$X_n(\omega) = \omega_n, \quad \omega \in \Omega = \{0, 1\}^{\mathbb{N}}$$

which is called a *coordinate process* in the sense that $X_n(\omega)$ is just the nth coordinate of ω.

In general, a stochastic process is a parameterized random variables $\{X_t\}_{t \in \mathbf{T}}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\mathbb{R}$, the parameter set $\mathbf{T}$ can be $\mathbb{N}$, $[0, +\infty)$ or some finite interval. For any fixed $t \in \mathbf{T}$, we have a random variable

$$X_t : \Omega \to \mathbb{R} \qquad \omega \rightarrowtail X_t(\omega).$$

For any fixed $\omega \in \Omega$, we have a real-valued measurable function on $\mathbf{T}$

$$X_\cdot(\omega) : \mathbf{T} \to \mathbb{R} \qquad t \rightarrowtail X_t(\omega),$$

which is called a trajectory or sample path of X. As a bi-variate function, a stochastic process can also be viewed as a measurable function from $\Omega \times \mathbf{T}$ to $\mathbb{R}$

$$(\omega, t) \rightarrowtail X(\omega, t) := X_t(\omega),$$

with the σ-algebra in $\Omega \times \mathbf{T}$ been chosen as $\mathcal{F} \times \mathcal{T}$, and $\mathcal{T}$ is the Borel σ-algebra on $\mathbf{T}$.

The largest probability space that one can take is the infinite product space $\Omega = \mathbb{R}^{\mathbf{T}}$, i.e. Ω is the space of all real-valued functions on $\mathbf{T}$. $\mathcal{F}$ can be taken as the infinite product σ-algebra $\mathcal{B}^{\mathbf{T}}$, which is the smallest σ-algebra containing all cylinder sets

$$C = \{\omega \in \mathbb{R}^{\mathbf{T}} | (\omega(t_1), \omega(t_2), \ldots, \omega(t_k)) \in A\}, \quad A \in \mathcal{B}^k, \ t_i \in \mathbf{T} \text{ for } i = 1, \ldots, k,$$

where $\mathcal{B}, \mathcal{B}^k$ is the Borel σ-algebra on $\mathbb{R}$ and $\mathbb{R}^k$, respectively. When $\mathbf{T} = \mathbb{N}$ and X_t only takes values in $\{0, 1\}$, we are back to the setting of Example 1.1.

Finite dimensional distributions are particularly interesting for a stochastic process, since they are the ones we can really observe. Let

$$\mu_{t_1,\ldots,t_k}(F_1 \times F_2 \times \cdots \times F_k) = \mathbb{P}[X_{t_1} \in F_1, \ldots, X_{t_k} \in F_k]$$

for all $F_1, F_2, \ldots, F_k \in \mathcal{B}$. $\mu_{t_1,\ldots,t_k}$ is called the finite dimensional distributions of $\{X_t\}_{t\in\mathbf{T}}$ at the time slice $(t_1, \ldots, t_k)$, where $t_i \in \mathbf{T}$ for $i = 1, 2, \ldots, k$.

The following theorem of Kolmogorov states that an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ can be established for a stochastic process X by knowing its all finite dimensional distributions with suitable consistency conditions.

Theorem 1.3 (Kolmogorov's extension theorem)**.** *Assume that a family of finite dimensional distributions $\{\mu_{t_1,\ldots,t_k}\}$ satisfy the following two consistency conditions for arbitrary sets of $t_1, t_2, \ldots, t_k \in T, k \in \mathbb{N}$:*

(i) For any permutation σ of $\{1, 2, \ldots, k\}$,

$$\mu_{t_{\sigma(1)},\ldots,t_{\sigma(k)}}(F_1 \times F_2 \times \cdots \times F_k) = \mu_{t_1,\ldots,t_k}(F_{\sigma^{-1}(1)} \times F_{\sigma^{-1}(2)} \times \cdots \times F_{\sigma-1(k)}).$$

(ii) For any $m \in \mathbb{N}$,

$$\mu_{t_1,\ldots,t_k}(F_1 \times F_2 \times \cdots \times F_k) = \mu_{t_1,\ldots,t_k,t_{k+1},\ldots,t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbb{R} \times \cdots \times \mathbb{R}).$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process $\{X_t\}_{t\in\mathbf{T}}$ such that

$$\mu_{t_1,\ldots,t_k}(F_1 \times F_2 \times \cdots \times F_m) = \mathbb{P}(X_{t_1} \in F_1, X_{t_2} \in F_2, \ldots, X_{t_m} \in F_m)$$

for any $t_1, t_2, \ldots, t_m \in \mathbf{T}$, $m \in \mathbb{N}$.

The proof the Kolmogorov extension theorem may be referred to [4, 5]. The advantage of the Kolmogorov theorem is that it is very general. The problem is that the probability space Ω is too big, so big that we can not say anything about features of paths on this space. Therefore the real challenge is to define probability measures on smaller spaces.

2 Filtration

The more we observe about a stochastic process, the more information we have at our disposal. This gives rise to a family of increasingly larger σ-algebras, which we call the *filtration* associated with the stochastic process. The filtration is the main conceptual difference between the random variables and and stochastic processes.

Definition 2.1 (Filtration)**.** *Given the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the filtration is a nondecreasing family of σ-algebras $\{\mathcal{F}_t\}_{t\geq 0}$ such that $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for any $0 \leq s < t$.*

A stochastic process $\{X_t\}$ is called $\mathcal{F}_t$-*adapted* if X_t is $\mathcal{F}_t$-measurable, i.e. $X_t^{-1}(B) \in \mathcal{F}_t$, for any $t \geq 0$ and $B \in \mathcal{B}$. Given a stochastic process $\{X_t\}$, one can define the filtration generated by this process by: $\mathcal{F}_t^X = \sigma(X_s, s \leq t)$, which is the smallest σ-algebra such that the $\{X_s\}_{s\leq t}$ are measurable. $\mathcal{F}_t^X$ is the smallest filtration such that the process $\{X_t\}$ is adapted. The filtration $\mathcal{F}_t^X$ can be thought of as the information supplied by the process up to time t. Taking again the independent coin tossing as the example and $\Omega = \{H, T\}^{\mathbb{N}}$. In this case, $\mathbf{T} = \mathbb{N}$ and the filtration is $\{\mathcal{F}_n^X\}_{n\geq 0}$. When $n = 0$, the σ-algebra is trivial

$$\mathcal{F}_0^X = \{\emptyset, \Omega\},$$

which means that we do not know any information about the output of the coin tossing. When $n = 1$, the σ-algebra is

$$\mathcal{F}_1^X = \{\emptyset, \Omega, \{H\}, \{T\}\}$$

since the first output gives either Head or Tail and we only know this information about the first output. When $n = 2$, we have

$$\mathcal{F}_2^X = \{\emptyset, \Omega, \{H\cdot\}, \{T\cdot\}, \{\cdot H\}, \{\cdot T\}, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \ldots\},$$

which contains all possible combinations of the outputs for the first two rounds of experiments. Sets like

$$\{HH\cdots T\} \quad \text{or} \quad \{HH\cdots H\}$$

are not contained in $\mathcal{F}_0^X$, $\mathcal{F}_1^X$ or $\mathcal{F}_2^X$ since the first two outputs can not tell such information. It is obvious that $\mathcal{F}_n^X$ becomes finer and finer as n increases.

3 Gaussian Process

In order to study Wiener process or Brownian motion (Brownian motion is also called Wiener process because its first rigorous mathematical foundation was established by N. Wiener in 1923 [11]), we will first introduce the Gaussian process on the continuous state space $\boldsymbol{R}$.

Definition 3.1. *A Gaussian process means that all of the finite dimensional distributions* $\mu_{t_1,\dots,t_k}$ *are Gaussian for any* $t_1, t_2, \dots, t_k \in T$.

We know that any Gaussian vector $X = (X_1, X_2, \dots, X_n)^T$ is completely determined by its first moment $m = \mathbb{E}X$ and second moment $K = \mathbb{E}(X-m)(X-m)^T$, where $m_i = \mathbb{E}X_i$ and $K_{ij} = \mathbb{E}(X_i - m_i)(X_j - m_j)^T$. The corresponding pdf is

$$p(\boldsymbol{x}) = \frac{1}{Z} e^{-\frac{1}{2}(\boldsymbol{x}-m)^T K^{-1}(\boldsymbol{x}-m)}$$

if K is invertible, where Z is a normalization constant. For the general case, we need to represent X via the characteristic function

$$\mathbb{E}e^{i\boldsymbol{\xi}\cdot X} = e^{i\boldsymbol{\xi}\cdot m - \frac{1}{2}\boldsymbol{\xi}^T K \boldsymbol{\xi}}.$$

From the above interpretation, a Gaussian process is uniquely determined by the mean function $m(t) = \mathbb{E}X_t$ and the covariance function $K(s,t) = \mathbb{E}(X_s - m(s))(X_t - m(t))$. We have $K(s,t) = K(t,s)$ by definition. If we consider the finite dimensional distribution at the time slice $(t_1, t_2, \dots, t_n)$, then $m(t)$ and $K(s,t)$ give the first moment

$$M = \big(m(t_1), m(t_2), \dots, m(t_n)\big)$$

and second moment

$$K = \begin{bmatrix} K(t_1,t_1) & K(t_1,t_2) & \cdots & K(t_1,t_n) \\ K(t_2,t_1) & K(t_2,t_2) & \cdots & K(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(t_n,t_1) & K(t_n,t_2) & \cdots & K(t_n,t_n) \end{bmatrix}.$$

It is straightforward to observe that for any $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ we have

$$\begin{aligned} \sum_{i,j} K(t_i,t_j) x_i x_j &= \sum_{i,j} \mathbb{E}(X_{t_i} - m(t_i))(X_{t_j} - m(t_j)) x_i x_j \\ &= \mathbb{E}\Big(\sum_i (X_{t_i} - m(t_i)) x_i \Big)^2 \geq 0. \end{aligned}$$

Thus we may view $m(t)$ as an infinite dimensional vector, and $K(s,t)$ as an infinite dimensional positive semi-definite matrix. From the characteristic function point of view, the Gaussian process X can be explained as a Gaussian random element in an infinite dimensional space $L^2(T)$ since we have at least formally in the current stage

$$\mathbb{E}e^{i(\xi,X)} = e^{i(\xi,m) - \frac{1}{2}(\xi, K\xi)},$$

where $(\xi, m) = \int_a^b \xi(t) m(t) dt$ is the inner-product in $L^2(T)$, and $(K\xi)(t) = \int_a^b K(t,s)\xi(s) ds$ is the action of the kernel function K on the function ξ. Based on the Kolmogorov's

extension theorem, we can construct a Gaussian process X from a given mean function $m(t)$ and covariance function $K(s,t)$.

The covariance function K is obviously symmetric, i.e. $K(t,s) = K(s,t)$, by definition. In addition, we have the semi-positivity of K in the following sense.

Theorem 3.2. *Assume the Gaussian process $(X_t)_{t\in[0,T]}$ possesses the regularity $X \in L^2_\omega L^2_t$ in the sense that $X \in L^2(\Omega; L^2[0,T])$, i.e.*

$$\mathbb{E}\int_0^T X_t^2 dt < \infty.$$

We have $m \in L^2_t$ and the operator

$$\mathcal{K}f(s) := \int_0^T K(s,t)f(t)dt, \quad s \in [0,T]$$

is a positive compact operator on L^2_t.

Proof. The mean function $m \in L^2_t$ is obvious since

$$\int_0^T m^2(t)dt = \int_0^T (\mathbb{E}X_t)^2 dt \le \int_0^T \mathbb{E}X_t^2 dt < \infty.$$

In addition, we have

$$\begin{aligned}\int_0^T\int_0^T K^2(s,t)dsdt &= \int_0^T\int_0^T \Big(\mathbb{E}(X_t - m(t))(X_s - m(s))\Big)^2 dsdt \\ &\le \int_0^T\int_0^T \mathbb{E}(X_t - m(t))^2\mathbb{E}(X_s - m(s))^2 dsdt \le \Big(\int_0^T \mathbb{E}X_t^2 dt\Big)^2,\end{aligned}$$

which means $K \in L^2([0,T]\times[0,T])$. Thus $\mathcal{K}$ is a compact operator on L^2_t (c.f. [6]).

It is easy to find that the adjoint operator of $\mathcal{K}$ is

$$\mathcal{K}^* f(s) := \int_0^T K(t,s)f(t)dt, \quad s \in [0,T].$$

From the symmetry of $K(s,t)$, we know that $\mathcal{K}$ is self-adjoint.

To show the positivity of $\mathcal{K}$, we have

$$\begin{aligned}(\mathcal{K}f, f) &= \int_0^T\int_0^T \mathbb{E}(X_t - m(t))(X_s - m(s))f(t)f(s)dsdt \\ &= \mathbb{E}\Big(\int_0^T (X_t - m(t))f(t)dt\Big)^2 \ge 0.\end{aligned}$$

The proof is completed. □

The following important closure property for a collection of Gaussian random variables will be used frequently in this chapter.

Theorem 3.3. *Suppose $X_1, X_2, \ldots$ are a sequence of Gaussian random variables and X_n converges to X in probability. Then X is also Gaussian.*

Proof. Let us denote

$$m_k = \mathbb{E}X_k, \quad \sigma_k^2 = \mathrm{var}X_k.$$

Then by dominated convergence theorem we have

$$e^{i\xi m_k - \frac{1}{2}\sigma_k^2\xi^2} = \mathbb{E}e^{i\xi X_k} \to \mathbb{E}e^{i\xi X} \quad \text{for any } \xi \in \boldsymbol{R}.$$

From the existence of the limit of the above equation, there are numbers m and σ^2 such that

$$m = \lim m_k, \quad \sigma^2 = \lim \sigma_k^2$$

and $\mathbb{E}e^{i\xi X} = e^{i\xi m - \frac{1}{2}\sigma^2\xi^2}$. The proof is completed. □

4 Wiener Process

The rigorous mathematical definition of the Brownian motion, or the Wiener Process, is defined as follows.

Definition 4.1. *(Brownian motion) The one dimensional Brownian motion W_t is defined as*

1. *It is a Gaussian process.*
2. *It has mean function $m(t) = 0$, and covariance function $K(s,t) = s \wedge t = \min(s,t)$.*
3. *With probability one, $t \mapsto W_t$ is continuous.*

The m-dimensional Brownian motion $\boldsymbol{W}_t$ has the form $\boldsymbol{W}_t = (W_t^1, W_t^2, \ldots, W_t^m)$, where each component W_t^j is a Brownian motion and they are independent each other. The Brownian motion (m-dimensional Brownian motion) is usually denoted as W_t or B_t ($\boldsymbol{W}_t$ or $\boldsymbol{B}_t$).

It is not difficult to prove that the above three conditions are equivalent to the following definition.

1′. For any $t_0 < t_1 < \cdots < t_n$, the random variables $W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_n} - W_{t_{n-1}}$ are independent.

2′. For any $s, t \geq 0$, $W_{s+t} - W_s \sim N(0, t)$.

3. With probability one, $t \mapsto W_t$ is continuous.

One straightforward implication of the second equivalent definition is that we can immediately write down the joint probability distribution density for $(W_{t_1}, W_{t_2}, \ldots, W_{t_n})$ $(t_1 < t_2 < \cdots < t_n)$ as

$$p_n(w_1, w_2, \ldots, w_n) = \frac{1}{\sqrt{2\pi t_1}} e^{-\frac{w_1^2}{2t_1}} \frac{1}{\sqrt{2\pi(t_2 - t_1)}} e^{-\frac{(w_2 - w_1)^2}{2(t_2 - t_1)}} \cdots \frac{1}{\sqrt{2\pi(t_n - t_{n-1})}} e^{-\frac{(w_n - w_{n-1})^2}{2(t_n - t_{n-1})}}.$$

More compactly

$$p_n(w_1, w_2, \ldots, w_n) = \frac{1}{Z_n} \exp(-I_n(w)),$$

where

$$I_n(w) = \frac{1}{2} \sum_{j=1}^{n} \Big(\frac{w_j - w_{j-1}}{t_j - t_{j-1}} \Big)^2 (t_j - t_{j-1}), \quad t_0 := 0, w_0 := 0,$$

$$Z_n = (2\pi)^{\frac{n}{2}} \big[t_1(t_2 - t_1) \cdots (t_n - t_{n-1}) \big]^{\frac{1}{2}}.$$

This also explicitly shows the stationarity and Markovianity of the Brownian motion with transition kernel function $p(x, t|y, s)$

$$\mathbb{P}(W_t \in B | W_s = y) = \int_B \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(x-y)^2}{2(t-s)}} dx = \int_B p(x, t|y, s) dx$$

where $s < t$ and B is a Borel set on $\mathbf{R}$. The transition probability density $p(x, t|y, s)$ satisfies the stationarity $p(x, t|y, s) = p(x - y, t - s|0, 0)$ and $p(x, t|0, 0)$ satisfies the PDE

$$\frac{\partial p}{\partial t} = \frac{1}{2} \frac{\partial^2 p}{\partial x^2}, \quad p(x, 0|0, 0) = \delta(x).$$

Now mathematically the first question is "Is there a process with these properties?". Though from Kolmogorov's extension theorem we can construct a probability space on $(\boldsymbol{R}^{[0,\infty)}, \mathcal{R}^{[0,\infty)})$ by the consistency of the finite dimensional distributions, it is not straightforward that the condition 3 in Definition 4.1 must be satisfied automatically. In fact, if we define the set

$$C = \{\omega | \omega \in \boldsymbol{R}^T, \omega \text{ is continuous on } T\}, \tag{4.1}$$

we will show that C is not a measurable set in $\mathcal{R}^T$! To understand this, one needs the following theorem

Theorem 4.2. *For any family of real functions $X_t : \Omega \to \boldsymbol{R}$, $t \in T$.*

(i) If $A \in \sigma\{X_t, t \in T\}$ and $\omega \in A$, and if $X_t(\omega') = X_t(\omega)$ for all $t \in T$, then we have $\omega' \in A$.

(ii) If $A \in \sigma\{X_t, t \in T\}$, then $A \in \sigma\{X_t, t \in S\}$ for some countable subset $S \subset T$.

The proof of this theorem may be referred to [2]. To apply the above theorem, we take $T = [0, \infty)$ and S a countable dense subset of T. We will have $C \in \mathcal{R}^S$ if $C \in \mathcal{R}^T$ by the second statement. From the first statement, C should contain all functions which have the same value with some $f \in C$ on S. This should contain lots of discontinuous functions. This contradicts with that C is the set of continuous functions.

To handle this issue, we need the concept "modification" of a process.

Definition 4.3 (Modification)**.** *Two processes X and X' defined on the same probability space are said to be modifications of each other if for each t,*

$$X_t = X'_t \quad a.s.$$

They are called indistinguishable if for almost all ω

$$X_t(\omega) = X'_t(\omega) \quad \text{for every } t.$$

It is clear that if X and X' are modifications of each other, they have the same finite dimensional distribution. If X and X' are modifications of each other and are almost surely continuous, they are indistinguishable.

Theorem 4.4 (Kolmogorov's continuity theorem)**.** *A real-valued process X for which there exist three strictly positive constants γ, β, C such that*

$$\mathbb{E}(|X_t - X_s|^\alpha) \leq C|t - s|^{1+\beta}$$

for any $s, t \geq 0$, then there is a modification $\tilde{X}$ of X which is almost-surely continuous.

For Brownian motion, one has $\alpha = 4, \beta = 1$, thus the condition of the above theorem is satisfied and the continuity of Brownian motion can be ensured in the sense of modifications.

5 Homeworks

- HW1. Let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of *i.i.d.* random variables taking values $+1$ with probability $\frac{2}{3}$ and -1 with probability $\frac{1}{3}$. Consider the (asymmetric) random walk on $\mathbb{Z}$ defined as

  $$S_n = \sum_{j=1}^{n} \xi_j.$$

  We wish to construct a stochastic process Z_t defined for $t \in [0, 1]$ by appropriate rescaling of S_n (similar to what we did to construct the Wiener process). That is we

want to show that there exists $\alpha \in \mathbb{R}$ such that the sequence of piecewise constant functions (here $\lfloor z \rfloor$ denotes the biggest integer smaller or equal to $z \in \mathbb{R}$)

$$Z_t^N = \frac{S_{\lfloor Nt \rfloor}}{N^\alpha}$$

converge as $N \to +\infty$ to some nontrivial Z_t. What is the α you need to chose for this to be the case? And what is Z_t?

- HW2. Let W_t be a Wiener process. Compute

  (a) $\mathbb{E}W_t^4$.

  (b) $\mathbb{E}(W_t - W_s + W_z)^2 \quad (t, s, z \in [0, 1])$.

- HW3. Let $X \in R^n$ be a n-dimensional Gaussian R.V. with mean zero and covariance matrix A (i.e. $\mathbb{E}X_i X_j = A_{ij}$). Suppose B is another strictly positive definite symmetric $n \times n$ matrix. Compute

$$\mathbb{E} \exp(-\frac{1}{2} X^T B X).$$

- HW4. Prove the equivalence of the two definitions for Brownian motion.

References

[1] R.J. Adler. *The Geometry of random fields*. John Wiley & Sons, New York, 1981.

[2] Billingsley. *Probability and measure*. John Wiley and Sons, New York, 1979.

[3] R.H. Cameron and W.T. Martin. The orthogonal development of nonlinear functionals in series of fourier-hermite functionals. *Ann. Math.*, pages 385–392, 1947.

[4] I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, Berlin, Heidelberg and New York, 1991.

[5] J. Lamperti. *Stochastic processes: A survey of the mathematical theory*. Springer-Verlag, New York, 1977.

[6] P.D. Lax. *Functional analysis*. John Wiley & Sons, 2002.

[7] P. Möters and Y. Peres. *Brownian motion*. Cambridge University Press, Cambridge, UK, 2010.

[8] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Springer-Verlag, Berlin and Heidelberg, 3rd edition edition, 2005.

[9] F. Riesz and B. Sz.-Nagy. *Functional analysis.* Blackie & Son limited, London and Glasgow, 1956.

[10] D.F. Walnut. *An introduction to wavelet analysis.* Birkhauser, Boston, Basel and Berlin, 2002.

[11] N. Wiener. Differential space. *J. Math. and Phys.*, 2:131–174, 1923.

[12] D. Xiu and G. Karniadakis. The wiener-askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comp.*, pages 619–644, 2002.

Lecture 12 Construction of BM and its properties *

Tiejun Li

1 Construction of Wiener Process

Below we will show three approaches to construct the Wiener process. Different forms play different roles in different circumstances.

A. Construction from invariance principle

The first construction from the invariance principle embodies the idea of taking continuum limit of symmetric random walk.

Theorem 1.1. *(Invariance Principle) Suppose* $\{\xi_i\}$ *are i.i.d.* $N(0,1)$ *random variables, define* $S_n = \sum_{i=1}^n \xi_i$ *and* X_t^n *as follows:*

$$X_t^n = \begin{cases} \dfrac{s_k}{\sqrt{n}}, & t = \dfrac{k}{n}, \\ (1-\theta)\dfrac{s_k}{\sqrt{n}} + \theta\dfrac{s_{k+1}}{\sqrt{n}}, & t \in \left(\dfrac{k}{n}, \dfrac{k+1}{n}\right), \quad \theta = nt - k, \end{cases}$$

then $X^n \in C[0,\infty)$ *and*

$$X^n \xrightarrow{d} W,$$

where $\xrightarrow{d}$ *is the weak convergence on the function space* $C[0,\infty)$ *to be defined below.*

Before stating the sketch of the proof, let us consider a special case by taking

$$\boldsymbol{P}(\xi_i) = \begin{cases} 1/2, & \xi_i = 1, \\ 1/2, & \xi_i = -1, \end{cases}$$

then $\mathbb{E}\xi_i = 0$, $\mathrm{var}\xi_i = 1$. The state of X_t^n at the time $t_k = k/n$ is nothing but the random walk considered before. The construction from invariance principle indicates that the standard Brownian motion is just the rescaled limit of the random walk with spatial scale $l = 1/\sqrt{n}$ and time scale $\tau = 1/n$. The relation $l^2/\tau = 1$ is exactly the regime considered before. This approximation is the most common one in computations.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Proposition 1.2. *The function space $C[0,\infty)$ equipped with the metric*

$$\rho(x,y)=\sum_{k=1}^{\infty}\frac{1}{2^k}\left(\|x-y\|_{L^\infty([0,k])}\wedge 1\right),\qquad x,y\in C[0,\infty)$$

is a complete, separable metric space.

It is also natural to define the σ-algebra $\mathcal{B}(C[0,\infty))$ in space $C[0,\infty)$ through finite dimensional cylinder sets

$$C=\{\omega\in C[0,\infty)|(\omega(t_1),\omega(t_2),\ldots,\omega(t_n))\in A\},\quad n\geq 1, A\in\mathcal{R}^n \tag{1.1}$$

One can show $\mathcal{B}(C[0,\infty))$ is equivalent to the Borel σ-algebra generated by the open sets in the metric space $(C[0,\infty),\rho)$.

Definition 1.3. *A family of probability measures $\{P_n\}_{n=1}^{\infty}$ on the metric space S with Borel σ-algebra $\mathcal{B}(S)$ is said to converge weakly to another probability measure P on the same space if and only if*

$$\lim_{n\to\infty}\int_S f(s)P_n(ds)=\lim_{n\to\infty}\int_S f(s)P(ds)$$

for every bounded, continuous real-valued function f on S.

Definition 1.4. *Let X_n be the random variables defined on the probability space $(\Omega_n,\mathcal{F}_n,P_n)$ and X be defined on another probability space $(\Omega,\mathcal{F},P)$. Both X_n and X take their values on the metric space S equipped with the Borel σ-algebra $\mathcal{B}(S)$. The random variables $\{X_n\}$ are said to converge weakly to X if the corresponding distribution $\mu_n=P_n\circ X_n^{-1}$ converges weakly to $\mu=P\circ X^{-1}$. It is usually denoted as*

$$X^n\xrightarrow{d}X.$$

The proof of the weak convergence in the invariance principle relies on the Prohorov's theorem on the weak compactness of the probability measures and the probabilistic type of Arzela-Ascoli compactness theorem in space $S=C[0,+\infty)$. It is quite involved so we will skip the detailed proof. The interested readers may be referred to [2].

The probability measure P_* as the weak convergence limit of $P_n\circ X_n^{-1}$ on the space $S=C[0,\infty)$ is called the *Wiener measure* and the probability space $(C[0,\infty),\mathcal{B}(C[0,\infty)),P_*)$ is called the *canonical probability space* for Wiener process, under which the coordinate mapping $W_t(\omega)=\omega(t)$ is a standard Brownian motion.

Heuristic Check. Now we give a heuristic check for the validity of invariance principle based on the central limit theorem for some discrete time. From the definition $S_n=\sum_{i=1}^n\xi_i$, where $\{\xi_i\}$ are i.i.d. $N(0,1)$ random variables, then by CLT

$$\frac{S_k}{\sqrt{n}}=\frac{\sqrt{k}}{\sqrt{n}}\cdot\frac{S_k}{\sqrt{k}}\xrightarrow{d}N(0,t),\ \text{as } k,n\to\infty \text{ and } t=\frac{k}{n}.$$

The limit X of X^n is then a Gaussian process formally with $X_0 = 0$ and

$$\begin{aligned}\mathbb{E}X_t X_s &\sim \mathbb{E}X_t^n X_s^n \\ &= \mathbb{E}X_{t\wedge s}^n (X_{t\vee s}^n - X_{t\wedge s}^n + X_{t\wedge s}^n) \\ &= \mathbb{E}(X_{t\wedge s}^n)^2 + \mathbb{E}X_{t\wedge s}^n (X_{t\vee s}^n - X_{t\wedge s}^n) \\ &\to t \wedge s. \quad \text{for } t = k/n, s = l/n \text{ and } k, l, n \to \infty.\end{aligned}$$

The last identity holds because of the independence between $X_{t\wedge s}^n$ and $X_{t\vee s}^n - X_{t\wedge s}^n$, and

$$\mathbb{E}(X_{t\wedge s}^n - X_{t\vee s}^n) = 0.$$

Heuristically the key point in the invariance principle is CLT when n, k is sufficiently large. This implies the condition $\xi_n \sim i.i.d.\ N(0,1)$ may be relaxed to ξ_n be $i.i.d.$ with mean 0 and variance 1. The distribution of ξ_n is not important. That is why the theorem is called "invanriance" principle.

A realization of Wiener process with finite N is shown in Fig. 1.

B. Construction from Karhunen-Loeve Expansion

The construction from Karhunen-Loeve expansion is based on the theory for Gaussian random fields. It can be easily extended to the case of Brownian bridge or high dimensional cases like the Brownian sheet etc [5].

Theorem 1.5. *(Karhunen-Loeve expansion) Let X_t ($t \in [0,1]$) be a Gaussian process with mean function $m(t) = 0$ and continuous covariance function $K(s,t)$. Consider the following eigenvalue problem*

$$\int_0^1 K(s,t)\phi_k(t)dt = \lambda_k \phi_k(s), \qquad k = 1, 2, \cdots$$

where $\int_0^1 \phi_k \phi_j dt = \delta_{kj}$. We have

$$X_t = \sum_{k=1}^{\infty} \alpha_k \sqrt{\lambda}_k \phi_k(t), \tag{1.2}$$

in the sense that the series

$$X_t^N = \sum_{k=1}^{N} \alpha_k \sqrt{\lambda}_k \phi_k(t) \to X_t \qquad \text{in } L_t^\infty L_P^2, \tag{1.3}$$

i.e. we have

$$\lim_{N\to\infty} \sup_{t\in[0,1]} \mathbb{E}|X_t^N - X_t|^2 = 0.$$

Here α_k are $i.i.d.\ N(0,1)$ random variables.

Proof. At first it is easy to find that the operator $K : L^2[0,1] \to L^2[0,1]$ defined as

$$(K\phi)(s) := \int_0^1 K(s,t)\phi(t)dt$$

through the covariance kernel function is nonnegative, self-adjoint and compact from the non-negativity, symmetry and continuity of $K(s,t)$ on $[0,1]^2$ [3]. From the theory of functional analysis, there are countable real eigenvalues, and 0 is the only possible accumulation point. For each nonzero eigenvalue, the eigensubspace is finite dimensional. This verifies the formal validity of the definition (1.2).

From Mercer's theorem which states that the convergence

$$\sum_{k=1}^{N} \lambda_k \phi_k(s)\phi_k(t) \to K(s,t), \quad s,t \in [0,1],\ N \to \infty$$

holds in absolute and uniform sense when K is continuous [6], we have for $N > M$

$$\mathbb{E}|X_t^N - X_t^M|^2 = \sum_{k=M+1}^{N} \lambda_k \phi_k^2(t) \to 0$$

in the absolute and uniform sense when $N, M \to \infty$. This implies X_t^N is a Cauchy sequence in the Banach space $L_t^\infty L_P^2$, thus the limit X_t exists and is unique in this space. For each fixed t, the mean square convergence of the Gaussian random vector $(X_{t_1}^N, X_{t_2}^N, \ldots, X_{t_m}^N)$ to $(X_{t_1}, X_{t_2}, \ldots, X_{t_m})$ implies the convergence in probability for any $t_1, t_2, \ldots, t_m \in [0,1]$. Application of the Theorem **??** ensures that the limit X_t is indeed a Gaussian process. It is not difficult to prove that

$$\mathbb{E}X_t = \lim_{N\to\infty} \mathbb{E}X_t^N = 0,$$

$$\mathbb{E}X_s X_t = \lim_{N\to\infty} \mathbb{E}X_s^N X_t^N = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t) = K(s,t)$$

by the convergence of X^N to X in $L_t^\infty L_P^2$. The proof is completed.

As an application of Karhunen-Loeve expansion to Brownian motion, one can obtain the eigensystem $\{\lambda_k, \phi_k(t)\}$ as follows. We have

$$\int_0^1 (s \wedge t)\phi_k(t)dt = \lambda_k \phi_k(s)$$

and thus

$$\int_0^s t\phi_k(t)dt + \int_s^1 s\phi_k(t)dt = \lambda_k \phi_k(s). \tag{1.4}$$

Taking differentiation with respect to s we obtain

$$\lambda_k \phi_k'(s) = s\phi_k(s) + \int_s^1 s\phi_k(t)dt - s\phi_k(s) = \int_s^1 s\phi_k(t)dt. \tag{1.5}$$

Differentiating once again gives a Sturm-Liouville problem

$$\lambda_k \phi_k''(s) = -\phi_k(s).$$

This naturally suggests $\lambda_k \neq 0$. Take $s = 0$ in (1.4), we obtain $\phi_k(0) = 0$; take $s = 1$ in (1.5), we have $\phi_k'(1) = 0$.

Solving this boundary value problem we obtain

$$\lambda_k = \left((k - \frac{1}{2})\pi\right)^{-2}, \quad \phi_k(s) = \sqrt{2}\sin\left((k - \frac{1}{2})\pi s\right), \quad k = 1, 2, \ldots.$$

Thus we get another representation of Brownian motion

$$W_t = \sum_{k=1}^{\infty} \alpha_k \frac{\sqrt{2}}{(k - \frac{1}{2})\pi} \sin\left((k - \frac{1}{2})\pi t\right). \tag{1.6}$$

It is easy to find that $W_0 = 0$ with this representation. To understand why it is almost surely continuous, we need the following theorem.

Theorem 1.6. *For the Karhunen-Loeve expansion to the Gaussian random field X_t with the same condition as in Theorem 1.5, if additionally*

$$\int_0^1 (-\ln u)^{1/2} dp(u) < \infty, \tag{1.7}$$

where

$$p(u) := \max\{\sigma(s,t) : |s - t| \leq |u|\}$$

and

$$\sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k (\phi_k(s) - \phi_k(t))^2 = K(s,s) + K(t,t) - 2K(s,t),$$

then X_t^N converges to X_t uniformly for $t \in [0,1]$ with probability one, and thus X has continuous trajectory almost surely.

The proof of this theorem may be referred to [1]. For the Wiener process, $\sigma(s,t) = t \vee s - t \wedge s$ and $p(u) = |u|$, so the condition (1.7) is satisfied and we have the continuity of the constructed W_t almost surely.

A realization with cutoff $N = 1000$ is shown in Fig. 1.

C. Construction from Haar basis

The construction below based on the Haar basis is originated from P. Lévy's interpolation method for Brownian motion. At first let us define the mother function

$$\psi(t) = \begin{cases} 1, & t \in [0, 1/2), \\ -1, & t \in [1/2, 1), \\ 0, & \text{otherwise.} \end{cases}$$

The multilevel Haar functions $\{H_k^{(n)}\}$ are defined as $H_0^{(0)}(t) = 1$ and

$$H_k^{(n)}(t) = 2^{\frac{n-1}{2}}\psi(2^{n-1}t - k), \quad n \geq 1,\ k \in I_n := \{0, 1, \ldots, 2^{n-1} - 1\}$$

for $t \in [0, 1]$, where n is the level and we take the convention that $I_0 = \{0\}$. It is a standard result that the Haar system $\{H_k^{(n)}\}$ for $n \in \mathbb{N}$ and $k \in I_n$ forms an orthonormal basis in $L^2[0, 1]$ [7]. We have the following theorem.

Theorem 1.7. *Let the random variables $\{\alpha_k^{(n)}\}$ i.i.d. $N(0, 1)$. Then*

$$W_t^N = \sum_{n=0}^{N}\sum_{k\in I_n} \alpha_k^{(n)} \int_0^t H_k^{(n)}(s)ds \longrightarrow W_t, \quad N \to \infty,$$

uniformly in $t \in [0, 1]$ in the almost sure sense.

A direct check on the finite terms approximation shows

$$\mathbb{E}W_t^N = \sum_{n=0}^{N}\sum_{k\in I_n} \mathbb{E}\alpha_k^{(n)} \int_0^t H_k^{(n)}(s)ds = 0,$$

and

$$\begin{aligned}
\mathbb{E}W_t^N W_s^N &= \sum_{n,m=0}^{N}\sum_{k\in I_n, l\in I_m} \mathbb{E}(\alpha_k^{(n)}\alpha_l^{(m)}) \int_0^t H_k^{(n)}(\tau)d\tau \int_0^s H_l^{(m)}(\tau)d\tau \\
&= \sum_{n=0}^{N}\sum_{k\in I_n} \int_0^t H_k^{(n)}(\tau)d\tau \int_0^s H_k^{(n)}(\tau)d\tau \\
&= \sum_{n=0}^{N}\sum_{k\in I_n} \int_0^1 H_k^{(n)}(\tau)\chi_{[0,t]}(\tau)d\tau \int_0^1 H_k^{(n)}(\tau)\chi_{[0,s]}(\tau)d\tau \\
&\to \int_0^1 \chi_{[0,t]}\chi_{[0,s]}(\tau)d\tau = t \wedge s. \qquad (1.8)
\end{aligned}$$

where $\chi_{[0,t]}(\tau)$ is the indicator function on $[0, t]$. Here the last convergence in the above equations is due to Parseval's identity because $\{H_k^{(n)}\}$ is an orthonormal basis. Below we give the rigorous proof.

Proof. At first, we show W_t^N uniformly converges to some continuous function W_t in the almost sure sense. We have the following tail estimate for any Gaussian distributed random variable $\xi \sim N(0, 1)$.

$$\mathbb{P}(|\xi| > x) = \sqrt{\frac{2}{\pi}}\int_x^\infty e^{-\frac{y^2}{2}}dy \leq \sqrt{\frac{2}{\pi}}\int_x^\infty \frac{y}{x}e^{-\frac{y^2}{2}}dy = \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{x^2}{2}}}{x}, \quad x > 0.$$

Define $a_n = \max_{k\in I_n} \alpha_k^{(n)}$, then we obtain

$$\mathbb{P}(a_n > n) = \mathbb{P}\left(\bigcup_{k\in I_n} \alpha_k^{(n)} > n\right) \le 2^{n-1}\sqrt{\frac{2}{\pi}}\frac{e^{-\frac{n^2}{2}}}{n}, \quad n \ge 1.$$

From $\sum_{n=1}^{\infty} 2^{n-1}\sqrt{\frac{2}{\pi}}\frac{e^{-\frac{n^2}{2}}}{n} < \infty$, the Borel-Cantelli lemma implies that there exists a set $\tilde{\Omega}$ with $\mathbb{P}(\tilde{\Omega}) = 1$ such that for any $\omega \in \tilde{\Omega}$ there is a $N(\omega)$ satisfying $a_m(\omega) \le m$ for any $m \ge N(\omega)$. In this case

$$\left|\sum_{m=N(\omega)}^{\infty}\sum_{k\in I_m} \alpha_k^{(m)} \int_0^t H_k^{(m)}(s)ds\right| \le \sum_{m=N(\omega)}^{\infty} m \sum_{k\in I_m}\int_0^t H_k^{(m)}(s)ds \le \sum_{m=N(\omega)}^{\infty} m2^{-\frac{m+1}{2}} < \infty,$$

which shows the uniform convergence of W_t^N to a continuous function W_t in the almost sure sense.

Now we prove W_t is indeed the standard Brownian motion. From the uniform convergence of W_t^N with respect to t in a almost sure sense, the limit W_t is indeed a Gaussian process from Theorem **??**. From the initial condition $W_0 = 0$ and the covariance function relation (1.8), we obtain a new representation of the Wiener process W_t.

A realization with finite cutoff is shown in Fig. 1.

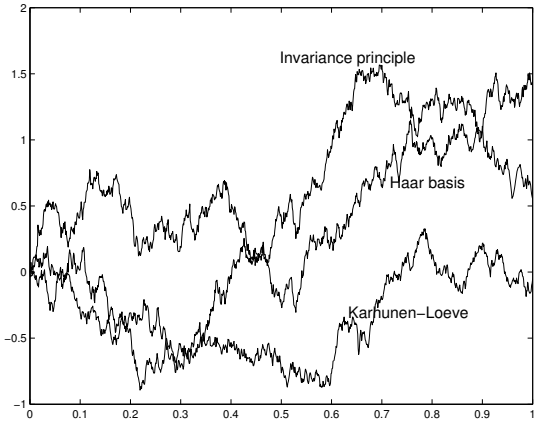


Figure 1: Numerical constructions of Brownian motion

Example 1.8. *Compute the expectation*

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big).$$

Solution. Note that it is not straightforward to compute this expectation since the integrand involves the whole Wiener path, i.e. a Wiener functional. From the Karhunen-Loeve

expansion,

$$\begin{aligned}\int_0^1 W_t^2 dt &= \int_0^1 \sum_{k,l} \sqrt{\lambda_k \lambda_l}\alpha_k\alpha_l\phi_k(t)\phi_l(t)dt \\ &= \sum_k \int_0^1 \lambda_k \alpha_k^2 \phi_k^2(t)dt = \sum_k \lambda_k \alpha_k^2.\end{aligned}$$

Then

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) = \mathbb{E}\Big(\prod_k \exp(-\frac{1}{2}\lambda_k\alpha_k^2)\Big) = \prod_k \mathbb{E}\exp(-\frac{1}{2}\lambda_k\alpha_k^2).$$

From the identity

$$\mathbb{E}\exp(-\frac{1}{2}\lambda_k\alpha_k^2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot e^{-\frac{1}{2}\lambda_k x^2} dx = \sqrt{\frac{1}{1+\lambda_k}}$$

we obtain

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) = \prod_k \sqrt{\frac{1}{1+\lambda_k}} := M,$$

where

$$M^{-2} = \prod_{k=1}^{\infty}\Big(1 + \frac{4}{(2k-1)^2\pi^2}\Big).$$

From the identities for infinite product series we have

$$\cosh(x) = \prod_{n=1}^{\infty}\Big(1 + \frac{4x^2}{(2n-1)^2\pi^2}\Big),$$

where $\cosh(x) = (e^x + e^{-x})/2$. Thus

$$M = (\cosh(1))^{-\frac{1}{2}} = \sqrt{\frac{2e}{1+e^2}}.$$

2 Properties of Wiener path

In this section, we investigate some basic properties and the regularity of the Wiener path.

Theorem 2.1 (Basic properties)**.** *Suppose W_t is a standard Brownian motion, then*

1. *Time-homogeneity: For any $s > 0$, $W_{t+s} - W_s$, $t \geq 0$, is a Brownian motion;*
2. *Symmetry: The process $-W_t$, $t \geq 0$, is a Brownian motion;*

3. *Scaling: For every $c > 0$, the process cW_{t/c^2}, $t \geq 0$, is a Brownian motion;*

4. *Time-inversion: The process X defined by $X_0 = 0$, $X_t = tW_{1/t}$ for $t > 0$, is a Brownian motion.*

The proof of these properties are left as exercise. Specially, the scaling property 3 has important implication for the dimensional analysis involving Brownian motion, which states

$$W_{kt} \sim \sqrt{k} W_t, \quad \dot{W}_{kt} \sim \frac{1}{\sqrt{k}} \dot{W}_t, \tag{2.1}$$

where $\dot{W}_t$ means the formal derivative of W_t with respect to t as discussed later. Note that for a standard smooth function $f(t)$ with the change of variable $t = k\tau$, we have the relation

$$\frac{df}{dt}(k\tau) = \frac{1}{k} \frac{df}{d\tau}(k\tau), \tag{2.2}$$

instead of (2.1).

Now let us investigate the regularity of the Brownian motion. The total variation of a specific path of the process X on $[a, b]$ is defined as

$$V(X(\omega); [a, b]) = \sup_{\Delta} \sum_k |X_{t_{k+1}}(\omega) - X_{t_k}(\omega)|,$$

where $\Delta = \cup_k [t_k, t_{k+1}]$ is any fixed subdivision of $[a, b]$. The discrete quadratic variation of X on $[0, t]$ with subdivision Δ is defined as

$$Q_t^{\Delta} = \sum_k |X_{t_{k+1}}(\omega) - X_{t_k}(\omega)|^2.$$

If for any t and any sequence Δ_n of subdivisions of $[0, t]$ such that $|\Delta_n|$ goes to zero, there exists a finite process $\langle X, X \rangle$ such that

$$Q_t^{\Delta_n} \to \langle X, X \rangle_t \quad \text{in Probability as } n \to \infty,$$

then $\langle X, X \rangle$ is called the *quadratic variation process* of X. It is obvious that $\langle X, X \rangle$ is increasing. The definition can be straightforwardly extended to the case on the interval $[a, b]$ as

$$Q_{[a,b]}^{\Delta_n} \to \langle X, X \rangle_b - \langle X, X \rangle_a \quad \text{as } n \to \infty.$$

Proposition 2.2. *For any t and subdivision Δ of $[0, t]$, we have for Wiener process W*

$$\mathbb{E}(Q_t^{\Delta} - t)^2 = 2 \sum_k (t_{k+1} - t_k)^2, \tag{2.3}$$

thus we get

$$Q_t^{\Delta} \longrightarrow t \text{ in } L^2(\mathbb{P}) \text{ as } |\Delta| \to 0$$

and $\langle W, W \rangle_t = t$ a.s.

The proof of Proposition 2.2 is straightforward and left as an exercise. This result is sometimes formally stated as $dW_t^2 = dt$.

Theorem 2.3 (Unbounded variation of the Wiener path)**.** *The Wiener paths are a.s. of infinite variations on any interval.*

Proof. Suppose the probability space is $(\Omega, \mathcal{F}, \mathbb{P})$. Based on (2.3) and the subsequence argument, there is a set $\Omega_0 \subset \Omega$ such that $\mathbb{P}(\Omega_0) = 1$, and there exits a subsequence of the subdivisions, still denoted as Δ_n, such that for any rational pair $p < q$,

$$Q_{[p,q]}^{\Delta_n} \to q - p, \quad \text{on } \Omega_0.$$

Now for any rational interval $[p, q]$, we have

$$q - p \leftarrow \sum_k (W_{t_{k+1}} - W_{t_k})^2 \leq \sup_k |W_{t_{k+1}} - W_{t_k}| \cdot V(W(\omega), [p, q]).$$

From the uniformly continuity of W on $[p, q]$, $\sup_k |W_{t_{k+1}} - W_{t_k}| \to 0$, thus we complete the proof.

The following result shows the Brownian motion has very curious smoothness.

Theorem 2.4 (Smoothness of the Wiener path)**.** *Consider the Wiener process on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define Ω_α the set of functions that are Hölder continuous with exponent α $(0 < \alpha < 1)$*

$$\Omega_\alpha = \left\{ f \in C[0,1], \sup_{0 \leq s,t \leq 1} \frac{|f(t) - f(s)|}{|t - s|^\alpha} < \infty \right\}.$$

Then if $0 \leq \alpha < \frac{1}{2}$, $\mathbb{P}(W_t \in \Omega_\alpha) = 1$; if $\alpha \geq \frac{1}{2}$, $\mathbb{P}(W_t \in \Omega_\alpha) = 0$.

The proof of Theorem 2.4 relies on the following generalized Kolmogorov's continuity theorem, which can be referred to [5].

Theorem 2.5. *Let X_t $(t \in [0,1]^d)$ be a Banach-valued process for which there exist three strictly positive constants γ, c, ε such that*

$$\mathbb{E}(|X_t - X_s|^\gamma) \leq c|t - s|^{d+\varepsilon},$$

then there is a modification $\tilde{X}$ of X such that

$$\mathbb{E}\Big(\sup_{s \neq t} (|\tilde{X}_t - \tilde{X}_s| / |t - s|^\alpha) \Big)^\gamma < \infty$$

for every $\alpha \in [0, \varepsilon/\gamma)$. In particular, the paths of $\tilde{X}$ are Hölder continuous of order α.

Proof of Theorem 2.4. When $\alpha < 1/2$, according to the generalized Kolmogorov continuity theorem and the following identity for 1D Gaussian R.V.

$$\mathbb{E}|B_t|^{2p} = Ct^p$$

for any $p \in \mathbb{N}$, we have $\epsilon/\gamma = (p-1)/2p = 1/2 - 1/2p$. Thus for $\alpha < 1/2$, $\mathbb{P}(W_t \in \Omega_\alpha) = 1$.

When $\alpha > 1/2$, if there exists rational interval $[p, q]$ such that $|W_t - W_s| \leq c|t-s|^\alpha$ for any $p \leq s, t \leq q$ then by Proposition 2.2

$$q-p \leftarrow \sum_k (W_{t_{k+1}} - W_{t_k})^2 \leq c^2 \sum_k |t_{k+1} - t_k|^{2\alpha-1} |t_{k+1} - t_k| \leq c^2 (q-p) \sup_k |t_{k+1} - t_k|^{2\alpha-1} \to 0,$$

which is a contradiction.

For the critical case $\alpha = 1/2$, one should apply the deep theorem on Lévy's modulus of continuity. The readers may be referred to [5].

From Fig. 1 we may observe the Brownian path is always fluctuating and it is a very noisy curve. Theorem 2.3 and 2.4 tell us that each trajectory is continuous and nowhere differentiable and it has unbounded variation in any finite interval! All of these results show the Brownian motion is a very subtle and strange mathematical object.

The following theorem due to A. Khinchin, characterizes the local behavior of W_t when t goes to zero.

Theorem 2.6 (Local law of the iterated logarithm)**.** *For the standard Brownian motion, we have*

$$\mathbb{P}\Big(\limsup_{t \to 0} \frac{W_t}{\sqrt{-2t \ln \ln t}} = 1 \Big) = 1.$$

Correspondingly

$$\mathbb{P}\Big(\liminf_{t \to 0} \frac{W_t}{\sqrt{-2t \ln \ln t}} = -1 \Big) = 1.$$

For the long time behavior of the Brownian motion, we have the following type of strong law of large numbers.

Theorem 2.7 (Strong Law of Large Numbers)**.** *For the standard Brownian motion, we have*

$$\lim_{t \to \infty} \frac{W_t}{t} = 0, \quad a.s.$$

The readers may be referred to [2, 4, 5] for more properties of Brownian motion.

3 Homeworks

- HW1. (**Scaling invariance of Wiener Process**) Let W_t be a Wiener process. Show that

$$X_t = \begin{cases} 0 & \text{if } t = 0 \\ tW_{1/t} & \text{if } t \in (0, 1] \end{cases}$$

$$Y_t = \frac{1}{\sqrt{c}} W_{ct}, \quad t > 0, c > 0$$
$$Z_t = W(T) - W(T-t), \quad 0 < t \leq T.$$

are all Wiener processes in the sense that they have the same finite dimensional distributions.

- HW2. Prove Proposition 2.2 and if we set the points $t_k = k2^{-n}t, k = 0, 1, \ldots, 2^n$ and consider the discrete quadrative variation of Brownian motion in $[0, t]$, prove the following sharpening of the Proposition 2.2.

$$\lim_{n\to\infty} Y_N(t, \omega) \to t, \quad \text{a.s.}$$

- HW3. Prove that $C[0, \infty)$ is a complete, separable metric space with the metric defined as

$$d(x, y) = \sum_{n=1}^{\infty} \frac{1}{2^n} \Big(\|x - y\|_{L^\infty[0,n]} \wedge 1 \Big)$$

- HW4. Prove that that given $0 \leq s < t$, $W_s = x$, $W_t = y$, then the conditional distribution

$$W_{\frac{s+t}{2}} | W_s = x, W_t = y \sim N\Big(\frac{x+y}{2}, \frac{t-s}{4} \Big).$$

References

[1] R.J. Adler. *The Geometry of random fields*. John Wiley & Sons, New York, 1981.

[2] I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, Berlin, Heidelberg and New York, 1991.

[3] P.D. Lax. *Functional analysis*. John Wiley & Sons, 2002.

[4] P. Möters and Y. Peres. *Brownian motion*. Cambridge University Press, Cambridge, UK, 2010.

[5] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Springer-Verlag, Berlin and Heidelberg, 3rd edition edition, 2005.

[6] F. Riesz and B. Sz.-Nagy. *Functional analysis*. Blackie & Son limited, London and Glasgow, 1956.

[7] D.F. Walnut. *An introduction to wavelet analysis*. Birkhauser, Boston, Basel and Berlin, 2002.

Lecture 13 SDE and Ito's formula *

Tiejun Li

1 White noise

In physics literature, the physicists usually use the stochastic differential equations (SDEs) like

$$\dot{X}_t = b(X_t, t) + \sigma(X_t, t)\dot{W}_t, \quad X|_{t=0} = X_0, \tag{1.1}$$

where $\dot{W}_t$ is called the *temporal Gaussian white noise*, which is the formal derivative of the Brownian motion W_t with respect to time. Its formal definition is that it is a Gaussian process with mean and covariance functions as

$$m(t) = \mathbb{E}(\dot{W}_t) = 0, \quad K(s,t) = \mathbb{E}(\dot{W}_s\dot{W}_t) = \delta(t-s).$$

It can be formally understood as

$$m(t) = \frac{d}{dt}\mathbb{E}(W_t) = 0, \quad K(s,t) = \frac{\partial^2}{\partial s \partial t}\mathbb{E}(W_s W_t) = \frac{\partial^2}{\partial s \partial t}(s \wedge t) = \delta(t-s).$$

The name *white noise* comes from its power spectral density (PSD) $S(\omega)$ defined as the Fourier transform of its autocorrelation function $R(t) = \mathbb{E}(\dot{W}_0\dot{W}_t) = \delta(t)$, thus $S(\omega) = \widehat{(\delta(t))} = 1$ which corresponds to a flat constant at all frequencies ω. We call it white as an analogy to the frequency spectrum of white light. If the frequency spectrum of the noise is not flat, it is called colored noise. From practical point view, the white noise is not physical since it requires infinite energy

$$E = \int_{-\infty}^{\infty} S(\omega)d\omega = \infty.$$

From the regularity theory of the Brownian motion, the function $\dot{W}$ is meaningless since W_t has less than half order smoothness. In fact, it is not a traditional function but a distribution [1]. However, the rigorous mathematical foundation of the white noise calculus can be also established [2]. But we will only introduce the Itô's classical way to establish the well-posedness of the stochastic differential equations.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Mathematically, the SDEs (1.1) are often denoted as

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t, \tag{1.2}$$

to avoid the ambiguity of the white noise, where W_t is the standard Wiener process. X_t may be viewed as a process induced by W_t. If there is no term $\sigma(X_t, t)dW_t$, it is a deterministic ODEs. The effect of $b(X_t, t)$ is to drive the mean position of the system, while the effect of $\sigma(X_t, t)dW_t$ is to diffuse around the mean position which we will see later. To make sense of (1.2), one natural way is to define X_t through its integral form

$$X_t = X_0 + \int_0^t b(X_s, s)ds + \int_0^t \sigma(X_s, s)dW_s. \tag{1.3}$$

We will show the first mathematical issue is how to define the integral $\int_0^t \sigma(X_s, s)dW_s$ involving Brownian motion.

2 Itô integral

First suppose X_t is continuous with respect to time t. For a fixed sample ω, we borrow the idea for defining the Riemann-Stieljes integral to make the definition

$$\int_0^t \sigma(X_s, s)dW_s = \lim_{|\Delta|\to 0} \sum_j \sigma(X_j, t_j^*)\Big(W_{t_{j+1}} - W_{t_j}\Big),$$

where Δ is a subdivision of $[0, t]$, X_j is the function value $X_{t_j^*}$ and t_j^* is chosen from the interval $[t_j, t_{j+1}]$. One critical issue about the above definition is that it depends on the choice of t_j^* when we are handling W_t, which has unbounded variation in any interval almost surely.

To have a sense on this, consider the Riemann-Stieltjes integral to $\int_a^b f(t)dg(t)$, where f and g are all assumed continuous. So

$$\int_a^b f(t)dg(t) \approx \sum_j f_j\Big(g(t_{j+1}) - g(t_j)\Big). \tag{2.1}$$

If one takes another value for f_j in $[t_j, t_{j+1}]$ under the same subdivision, then

$$\int_a^b f(t)dg(t) \approx \sum_j \tilde{f}_j\Big(g(t_{j+1}) - g(t_j)\Big).$$

If $g(t)$ has bounded total variation, we subtract the right hand side of the above two definitions and obtain

$$\begin{aligned}\Big|\sum_j (f_j - \tilde{f}_j)\Big(g(t_{j+1}) - g(t_j)\Big)\Big| &\le \max_j |f_j - \tilde{f}_j| \sum_j \Big|g(t_{j+1}) - g(t_j)\Big| \\ &\le \max_j |f_j - \tilde{f}_j| V(g; [a, b]) \to 0\end{aligned}$$

as $|\Delta| \to 0$ by the uniform continuity of f on $[a, b]$. Thus we get a well-defined definition which is independent of the choice of reference point in the approximation. If $g(t) = W_t(\omega)$, let us see what will happen in this case.

Example 2.1. *Different choices for the stochastic integral* $\int_0^T W_t dW_t$.

Choice 1: Leftmost endpoint integral.

$$\int_0^T W_t dW_t \approx \sum_j W_{t_j}(W_{t_{j+1}} - W_{t_j}) := I_N^L.$$

Choice 2: Rightmost endpoint integral.

$$\int_0^T W_t dW_t \approx \sum_j W_{t_{j+1}}(W_{t_{j+1}} - W_{t_j}) := I_N^R.$$

Choice 3: Midpoint integral.

$$\int_0^T W_t dW_t \approx \sum_j W_{t_{j+\frac{1}{2}}}(W_{t_{j+1}} - W_{t_j}) := I_N^M.$$

Without looking into the exact pathwise result for the three choices, we have the following identities from the statistical average sense.

$$\begin{aligned}
\mathbb{E}(I_N^L) &= \sum_j \mathbb{E}W_{t_j}\mathbb{E}(W_{t_{j+1}} - W_{t_j}) = 0, \\
\mathbb{E}(I_N^R) &= \sum_j \Big[\mathbb{E}(W_{t_{j+1}} - W_{t_j})^2 + \mathbb{E}W_{t_j}\mathbb{E}(W_{t_{j+1}} - W_{t_j})\Big] = \sum_j \Delta t_j = T, \\
\mathbb{E}(I_N^M) &= \mathbb{E}\Big[\sum_j W_{t_{j+\frac{1}{2}}}(W_{t_{j+1}} - W_{t_{j+\frac{1}{2}}}) + \sum_j W_{t_{j+\frac{1}{2}}}(W_{t_{j+\frac{1}{2}}} - W_{t_j})\Big] \\
&= \sum_j \mathbb{E}(W_{t_{j+\frac{1}{2}}} - W_{t_j})^2 = \sum_j (t_{j+\frac{1}{2}} - t_j) = \frac{T}{2}.
\end{aligned}$$

The reason is that the Brownian motion has unbounded variations for any finite interval. The example above also shows that we should take special attention to stochastic integrals.

One important remark on the definition of stochastic integrals like (2.1) is that it can not be defined for arbitrary continuous functions f, otherwise the function g must have bounded variations on compacts [9]. To overcome this issue, one rescue is to restrict the integrands to be a special class of functions, the adapted processes. That is the key point of the well-known Itô integral to be introduced below.

The first stochastic integral which is studied rigorously in the history is Itô's leftmost endpoint integral [4], which is named Itô integral from then on. It turns out that the different choices of the reference point correspond to different consistent definitions of stochastic

integrals under suitable conditions, but they can be connected by some simple transformation rules (See [9], Theorem 30 in Chapter 5). To understand Itô's definition for stochastic integral, we take the filtration generated by standard Wiener process as $\mathcal{F}_t^W$ (we also assume all of the sets of measure zero has been contained in $\mathcal{F}_t^W$). The construction of Itô integral takes the leftmost endpoint approximation

$$\int_0^T f(t,\omega)dW_t \approx \sum_j f_{t_j}(W_{t_{j+1}} - W_{t_j}).$$

Mathematically, to understand Itô integral, we need the concept *simple function* which takes the form

$$f(t,\omega) = \sum_{j=1}^{n} e_j(\omega)\chi_{[t_j,t_{j+1})}(t), \tag{2.2}$$

where $e_j(\omega)$ is $\mathcal{F}_{t_j}^W$-measurable and $\chi_{[t_j,t_{j+1})}(t)$ is the indicator function on $[t_j, t_{j+1})$. It is natural to define

$$\int_0^T f(t,\omega)dW_t = \sum_j e_j(\omega)(W_{t_{j+1}} - W_{t_j}) \tag{2.3}$$

for this choice of simple functions.

Lemma 2.2. *For any $S \le T$, the stochastic integral for the simple functions satisfies*

$$(1) \qquad \mathbb{E}\left(\int_S^T f(t,\omega)dW_t\right) = 0, \tag{2.4}$$

$$(2) \qquad \textit{(Itô isometry)} \quad \mathbb{E}\Big(\int_S^T f(t,\omega)dW_t\Big)^2 = \mathbb{E}\left(\int_S^T f^2(t,\omega)dt\right). \tag{2.5}$$

Proof. The first property is straightforward by the independence between $\Delta W_j := W_{t_{j+1}} - W_{t_j}$ and $e_j(\omega)$ and $\Delta W_j \sim N(0, t_{j+1} - t_j)$. For the second property we have

$$\begin{aligned}
\mathbb{E}\left(\int_S^T f(t,\omega)dW_t\right)^2 &= \mathbb{E}\Big(\sum_j e_j\Delta W_j\Big)^2 = \mathbb{E}\Big(\sum_{j,k} e_j e_k \Delta W_j \Delta W_k\Big) \\
&= \mathbb{E}\Big(\sum_j e_j^2 \Delta W_j^2 + 2\sum_{j<k} e_j e_k \Delta W_j \Delta W_k\Big) \\
&= \sum_j \mathbb{E}e_j^2 \cdot \mathbb{E}\Delta W_j^2 + \sum_{j<k} \mathbb{E}(f_j f_k \Delta W_j)\cdot \mathbb{E}(\Delta W_k) \\
&= \sum_j \mathbb{E}e_j^2 \Delta t_j = \mathbb{E}\left(\int_S^T f^2(t,\omega)dt\right).
\end{aligned}$$

where the last third identity holds because of the independence between ΔW_k and $e_j e_k \Delta W_j$ for $j < k$.

Now for $f(t,\omega)$ which belongs to the class of functions $\mathcal{V}[S,T]$ defined as

(i) f is $\mathcal{B}([0,\infty)) \times \mathcal{F}$-measurable as a function from (t,ω) to $\mathbb{R}$,

(ii) $f(t,\omega)$ is $\mathcal{F}_t^W$-adapted,

(iii) $f \in L_P^2 L_t^2$, that is $\mathbb{E}\left(\int_S^T f^2(t,\omega)dt\right) < \infty$,

we have the approximation property through simple functions $\phi_n(t,\omega)$

$$\mathbb{E}\left(\int_S^T (f(t,\omega) - \phi_n(t,\omega))^2 dt\right) \to 0, \tag{2.6}$$

i.e. $\phi_n \to f$ in $L_P^2 L_t^2$ (c.f. [5,8]). With this setup, we can define the Itô integral as

$$\int_S^T f(t,\omega)dW_t = \lim_{n\to\infty} \int_S^T \phi_n(t,\omega)dW_t \qquad \text{in } \ L_P^2. \tag{2.7}$$

From (2.5), $\int_S^T \phi_n(t,\omega)dW_t$ is in L_P^2 for any simple function $\phi_n(t,\omega)$. Furthermore we have

$$\mathbb{E}\Big(\int_S^T \phi_n dW_t - \int_S^T \phi_m dW_t\Big)^2 = \mathbb{E}\left(\int_S^T (\phi_n - \phi_m)^2 dt\right). \tag{2.8}$$

From (2.6), the approximation sequence $\{\phi_n\}$ is a Cauchy sequence in $L_P^2(\Omega; L_t^2[S,T])$. This implies $\{\int_S^T \phi_n dW_t\}$ is also a Cauchy sequence in L_P^2. From the completeness of $L_P^2(\Omega)$, it has a unique limit and we define it as

$$\int_S^T f(t,\omega)dW_t$$

in the definition (2.7). The independence on the choice of the approximating sequence $\{\phi_n\}$ is left as an exercise.

As a natural extension of Lemma 2.2, we have

Theorem 2.3. *For $f \in \mathcal{V}[S,T]$, the Itô integral satisfies*

$$(1) \qquad \mathbb{E}\left(\int_S^T f(t,\omega)dW_t\right) = 0, \tag{2.9}$$

$$(2) \qquad \textit{(Itô isometry)} \quad \mathbb{E}\Big(\int_S^T f(t,\omega)dW_t\Big)^2 = \mathbb{E}\left(\int_S^T f^2(t,\omega)dt\right). \tag{2.10}$$

Proof. Based on Lemma 2.2, we have

$$\begin{aligned}\left|\mathbb{E}\left(\int_S^T f(t,\omega)dW_t\right)\right| &= \left|\mathbb{E}\left(\int_S^T f(t,\omega)dW_t - \int_S^T \phi_n(t,\omega)dW_t\right)\right| \\ &\le \mathbb{E}\left(\int_S^T f(t,\omega)dW_t - \int_S^T \phi_n(t,\omega)dW_t\right)^2 \to 0\end{aligned}$$

by Hölder's inequality and the definition (2.7).

It is a standard result that if $X_n \to X$ in a Hilbert space H, then $|X_n| \to |X|$ and thus $|X_n|^2 \to |X|^2$, where $|\cdot|$ is the correspoding norm in Hilbert space H. So we have

$$\mathbb{E}\left(\int_S^T \phi_n(t,\omega)dW_t\right)^2 \to \mathbb{E}\left(\int_S^T f(t,\omega)dW_t\right)^2 \qquad \text{in } L_P^2(\Omega)$$

and

$$\mathbb{E}\left(\int_S^T \phi_n^2(t,\omega)dt\right) \to \mathbb{E}\left(\int_S^T f^2(t,\omega)dt\right) \qquad \text{in } L_P^2(\Omega; L_t^2[S,T])$$

From the Itô isometry for simple functions, we obtain (2.10) immediately.

The following properties can be proved for the Itô integral easily.

Proposition 2.4. *For $f, g \in \mathcal{V}[S,T]$ and $U \in [S,T]$, we have*

(i) $\int_S^T f dW_t = \int_S^U f dW_t + \int_U^T f dW_t$ *a.s..*

(ii) $\int_S^T (f+cg)dW_t = \int_S^T f dW_t + c\int_S^T g dW_t$ *(c is a constant) a.s..*

(iii) $\int_S^T f dW_t$ *is $\mathcal{F}_t^W$-measurable.*

Furthermore, we have the regularity of the path of the process defined via Itô integral, whose proof may be referred to [5, 8, 10].

Lemma 2.5. *For $f \in \mathcal{V}[0,T]$, $X_t := \int_0^t f(s,\omega)dW_s$ has continuous trajectories in the almost sure sense.*

We remark that the class of functions $\mathcal{V}[0,T]$ to make sense of the Itô integral and keep the above properties can be weakened by replacing the conditions (ii) and (iii) in $\mathcal{V}[0,T]$ as

(ii)' f is $\mathcal{F}_t$-adapted, where $\{\mathcal{F}_t\}$ is a filtration such that W_t is a $\mathcal{F}_t$-martingale.

(iii)' $\int_0^T f^2(s,\omega)ds < \infty$ almost surely.

The readers may be referred to [5, 8, 10] for more details. With this weaker setup, one can define the multi-dimensional Ito integral

$$\int_0^T \boldsymbol{\sigma}(t,\omega)\cdot d\boldsymbol{W}_t,$$

where $\boldsymbol{W}_t$ is an m-dimensional Wiener process, and $\boldsymbol{\sigma} \in \mathbb{R}^{n\times m}$ is $\mathcal{F}_t^{\boldsymbol{W}}$-adapted. To compute their expectation, We have the similar property as the Ito isometry

$$\mathbb{E}\left(\int_S^T \sigma(t,\omega)dW_t^j\right) = 0, \qquad \mathbb{E}\Big(\int_S^T \sigma(t,\omega)dW_t^j\Big)^2 = \mathbb{E}\left(\int_S^T \sigma^2(t,\omega)dt\right), \quad \forall j.$$

and especially the cross product expectation

$$\mathbb{E}\left(\int_S^T \sigma_1(t,\omega)dW_t^i \cdot \int_S^T \sigma_2(t,\omega)dW_t^j\right) = 0, \qquad \forall i \neq j,$$
$$\mathbb{E}\Big(\int_S^T \sigma_1(t,\omega)dW_t^j \int_S^T \sigma_2(t,\omega)dW_t^j\Big) = \mathbb{E}\left(\int_S^T \sigma_1(t,\omega)\sigma_2(t,\omega)dt\right), \quad \forall j.$$

Example 2.6. *With Itô integral we have*

$$\int_0^t W_s dW_s = \frac{W_t^2}{2} - \frac{t}{2}. \tag{2.11}$$

Proof. From the definition of Itô integral

$$\begin{aligned}\int_0^t W_s dW_s &\approx \sum_j W_{t_j}(W_{t_{j+1}} - W_{t_j}) = \sum_j \frac{2W_{t_j}W_{t_{j+1}} - 2W_{t_j}^2}{2} \\ &= \sum_j \frac{W_{t_{j+1}}^2 - W_{t_j}^2}{2} - \sum_j \frac{W_{t_{j+1}}^2 - 2W_{t_{j+1}}W_{t_j} + W_{t_j}^2}{2} \\ &= \frac{W_t^2}{2} - \frac{1}{2}\sum_j (W_{t_{j+1}} - W_{t_j})^2 \to \frac{W_t^2}{2} - \frac{t}{2},\end{aligned}$$

where the last limit is due to the fact $\langle W, W\rangle_t = t$ in Proposition **??**.

3 Itô's formula

Let's take the differential form of the identity (2.11), then we have

$$dW_t^2 = 2W_t dW_t + dt.$$

Note that it is different from the traditional Newton-Leibnitz calculus which suggests $dW_t^2 = 2W_t dW_t$ with chain rule. This exactly manifests the specialty of Itô calculus to be introduced in this section. To further understand the previous specific example, we consider a more general situation.

Proposition 3.1. *For any bounded and continuous function $f(t,\omega)$ in t,*

$$\sum_j f(t_j^*,\omega)(W_{t_{j+1}} - W_{t_j})^2 \to \int_0^t f(s,\omega)ds, \quad \textit{for any } \ t_j^* \in [t_j, t_{j+1}]$$

in probability when the subdivision size goes to zero.

Proof. Straightforward calculation shows

$$\begin{aligned}\mathbb{E}\left(\sum_j f(t_j)\Delta W_{t_j}^2 - \sum_j f(t_j)\Delta t_j\right)^2 &= \mathbb{E}\left(\sum_{j,k} f(t_j)f(t_k)(\Delta W_{t_j}^2 - \Delta t_j)(\Delta W_{t_k}^2 - \Delta t_k)\right)\\ &= \mathbb{E}\left(\sum_j f^2(t_j)\cdot \mathbb{E}\left((\Delta W_{t_j}^2 - \Delta t_j)^2|\mathcal{F}_{t_j}\right)\right)\\ &= 2\sum_j \mathbb{E}f^2(t_j)\Delta t_j^2 \to 0.\end{aligned}$$

At the same time, we have

$$|\sum_j (f(t_j^*) - f(t_j))\Delta W_{t_j}^2| \le \sup_j |f(t_j^*) - f(t_j)| \cdot \sum_j \Delta W_{t_j}^2.$$

The first term on the right hand side goes to zero almost surely because of the uniform continuity of f on $[0,t]$, and the second term converges to the quadratic variation of W_t in probability. Combining the results above leads to the desired conclusion.

It is exactly this reason that we simply denoted it as

$$dW_t^2 = dt$$

for calculations. The Itô's formula to be introduced below gives this a rigorous foundation.

Now let us consider the Itô process defined as

$$X_t = X_0 + \int_0^t b(s,\omega)ds + \int_0^t \sigma(s,\omega)dW_s,$$

which is usually denoted as

$$dX_t = b(t,\omega)dt + \sigma(t,\omega)dW_t, \quad X_t|_{t=0} = X_0 \tag{3.1}$$

for functions

$$\sigma \in \mathcal{W}[0,T], \quad b \text{ is } \mathcal{F}_t\text{-adapted and } \int_0^T |b(t,\omega)|dt < \infty \text{ a.s.}$$

We have the following important result, whose rigorous proof can be referred to [3, 5].

Theorem 3.2 (1D Itô's formula)**.** *If X_t is an Itô process as in Equation* (3.1)*, $Y_t = f(X_t)$ where f is a twice differentiable function. Then Y_t is also an Itô process and*

$$dY_t = f'(X_t)dX_t + \frac{1}{2}f''(X_t)(dX_t)^2,$$

where the rule of simplification is $dt^2 = 0$, $dtdW_t = dW_t dt = 0$ and $(dW_t)^2 = dt$, i.e.

$$(dX_t)^2 = (bdt + \sigma dW_t)^2 = b^2dt^2 + 2b\sigma dtdW_t + \sigma^2(dW_t)^2 = \sigma^2 dt.$$

Thus finally

$$dY_t = \left(b(t,\omega)f'(X_t) + \frac{1}{2}\sigma^2(t,\omega)f''(X_t)\right)dt + \sigma(t,\omega)f'(X_t)dW_t.$$

Sketch of Proof. We will only consider the situation that f, f' and f'' are bounded and continuous here. At first, if b and σ are simple functions, we have

$$Y_t - Y_0 = \sum_j (f(X_{t_{j+1}}) - f(X_{t_j})) = \sum_j \left(f'(X_{t_j})\Delta X_{t_j} + \frac{1}{2} f''(X_{t_j})\Delta X_{t_j}^2 + R_j \right),$$

where $\Delta X_{t_j} = X_{t_{j+1}} - X_{t_j}$ and $R_j = o(|\Delta X_{t_j}|^2)$. Without loss of generality we assume the discontinuity of the step functions are embedded in the current subdivision grid points. We obtain

$$\begin{aligned}\sum_j f'(X_{t_j})\Delta X_{t_j} &= \sum_j f'(X_{t_j})b(t_j)\Delta t_j + \sum_j f'(X_{t_j})\sigma(t_j)\Delta W_{t_j} \\ &\to \int_0^t b(s)f'(X_s)ds + \int_0^t \sigma(s)f'(X_s)dW_s\end{aligned}$$

and

$$\sum_j f''(X_{t_j})\Delta X_{t_j}^2 = \sum_j f''(X_{t_j}) \left(b^2(t_j)\Delta t_j^2 + 2b(t_j)\sigma(t_j)\Delta t_j \Delta W_{t_j} + \sigma^2(t_j)\Delta W_{t_j}^2 \right).$$

We have

$$|\sum_j f''(X_{t_j})b^2(t_j)\Delta t_j^2| \le K\sum_j \Delta t_j^2 \le KT \sup_j \Delta t_j \to 0,$$

$$|\sum_j f''(X_{t_j})b(t_j)\sigma(t_j)\Delta t_j \Delta W_{t_j}| \le K\sum_j |\Delta t_j \Delta W_{t_j}| \le KT \sup_j |\Delta W_{t_j}| \to 0$$

as the subdivision size goes to zero, where K is the bound of b, σ and f''. From Proposition 3.1, we get

$$\sum_j f''(X_{t_j})\sigma^2(t_j)\Delta W_{t_j}^2 \to \int_0^t \sigma^2(s)f''(X_s)ds \qquad \text{in} \quad L_P^2.$$

The general situation can be done by taking approximation through simple functions.

The above result can be generalized to multidimensional case as

Theorem 3.3 (Multidimensional Ito formula)**.** *If $d\boldsymbol{X}_t = \boldsymbol{b}(t,\omega)dt + \boldsymbol{\sigma}(t,\omega)\cdot d\boldsymbol{W}_t$, where $\boldsymbol{X}_t \in \mathbb{R}^n$, $\boldsymbol{\sigma} \in \mathbb{R}^{n\times m}$, $\boldsymbol{W} \in \mathbb{R}^m$. Define $Y_t = f(\boldsymbol{X}_t)$, where f is a twice differentiable function. Then*

$$dY_t = \nabla f(\boldsymbol{X}_t)\cdot d\boldsymbol{X}_t + \frac{1}{2}(d\boldsymbol{X}_t)^T \cdot \nabla^2 f(\boldsymbol{X}_t)\cdot (d\boldsymbol{X}_t),$$

where the rule of simplification is $dt^2 = 0$, $dtdW_t^i = dW_t^i dt = dW_t^i dW_t^j = 0$ $(i \ne j)$, $(dW_t^i)^2 = dt$. That is

$$\begin{aligned}(d\boldsymbol{X}_t)^T \cdot \nabla^2 f(\boldsymbol{X}_t)\cdot (d\boldsymbol{X}_t) &= \sum_{l,k,i,j} dW_t^l \sigma_{il}\partial_{ij}^2 f \sigma_{jk} dW_t^k \\ &= \sum_{k,i,j} \sigma_{ik}\sigma_{jk}\partial_{ij}^2 f dt = \boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f dt,\end{aligned}$$

where $\boldsymbol{A} : \boldsymbol{B} = \sum_{ij} a_{ij} b_{ji}$ *is the twice contraction for second order tensors. Finally*

$$dY_t = (\boldsymbol{b} \cdot \nabla f + \frac{1}{2} \boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f)dt + \nabla f \cdot \boldsymbol{\sigma} \cdot d\boldsymbol{W}_t.$$

Example 3.4. *Integration by part*

$$\int_0^t s dW_s = tW_t - \int_0^t W_s ds. \tag{3.2}$$

Proof. Define $f(x, y) = xy$, $X_t = t$, $Y_t = W_t$, then from multidimensional Itô's formula

$$df(X_t, Y_t) = X_t dY_t + Y_t dX_t + dX_t dY_t.$$

With the rule $dt dW_t = 0$, we obtain $d(tW_t) = t dW_t + W_t dt$ and the result follows.

Example 3.5. *Iterated Itô integrals*

$$\int_0^t dW_{t_1} \int_0^{t_1} dW_{t_2} \dots \int_0^{t_{n-1}} dW_{t_n} = \frac{1}{n!} t^{\frac{n}{2}} h_n \left(\frac{W_t}{\sqrt{t}} \right), \tag{3.3}$$

where $h_n(x)$ *is the n-th order Hermite polynomial*

$$h_n(x) = (-1)^n e^{\frac{1}{2}x^2} \frac{d^n}{dx^n} \left(e^{-\frac{1}{2}x^2} \right).$$

Proof. It is easy to verify that

$$\int_0^t W_s dW_s = \frac{t}{2!} h_2 \left(\frac{W_t}{\sqrt{t}} \right),$$

where $h_2(x) = x^2 - 1$ is the second order Hermite polynomial. In the same fashion, we have

$$\int_0^t \Big(\int_0^s W_u dW_u \Big) dW_s = \frac{1}{2} \int_0^t (W_s^2 - s) dW_s.$$

Using Itô's formula, we have

$$\int_0^t W_s^2 dW_s = \frac{1}{3} W_t^3 - \int_0^t W_s ds.$$

Hence, using (3.2) we obtain

$$\int_0^t \Big(\int_0^s W_u dW_u) dW_s = \frac{1}{6} W_t^3 - \frac{1}{2} t W_t = \frac{1}{3!} t^{\frac{3}{2}} h_3 \left(\frac{W_t}{\sqrt{t}} \right),$$

where $h_3(x) = x^3 - 3x$ is the third order Hermite polynomial. The general case is left as an exercise.

4 SDE

4.1 Wellposed-ness

With the help of Itô's integral, we can establish the classical well-posedness result for the stochastic differential equations

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t) \cdot d\boldsymbol{W}_t, \tag{4.1}$$

through Picard-type iterations.

Theorem 4.1. *Let $\boldsymbol{X} \in \mathbb{R}^n, \boldsymbol{W} \in \mathbb{R}^m$. Suppose the coefficients $\boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{\sigma} \in \mathbb{R}^{n\times m}$ satisfy global Lipschitz and linear growth conditions as*

$$|\boldsymbol{b}(\boldsymbol{x}, t) - \boldsymbol{b}(\boldsymbol{y}, t)| + |\boldsymbol{\sigma}(\boldsymbol{x}, t) - \boldsymbol{\sigma}(\boldsymbol{y}, t)| \leq K|\boldsymbol{x} - \boldsymbol{y}|, \tag{4.2}$$

$$|\boldsymbol{b}(\boldsymbol{x}, t)|^2 + |\boldsymbol{\sigma}(\boldsymbol{x}, t)|^2 \leq K(1 + |\boldsymbol{x}|^2) \tag{4.3}$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, t \in [0, T]$, where K is a positive constant and $|\cdot|$ means the Frobenius norm, that is

$$|\boldsymbol{b}|^2 := \sum_i b_i^2, \quad |\boldsymbol{\sigma}|^2 := \sum_{i,j} \sigma_{ij}^2.$$

Assume the initial value $\boldsymbol{X}_0 = \xi$ is a random variable which is independent of $\mathcal{F}_\infty^{\boldsymbol{W}}$ and satisfies $\mathbb{E}|\xi|^2 < \infty$. Then (4.1) has a unique t-continuous solution $\boldsymbol{X}_t \in \mathcal{V}[0, T]$.

The proof can be referred to [5].

4.2 Diffusion process

The SDEs driven by Wiener processes is the typical Markov process which is also called the *diffusion processes* in stochastic analysis. Mathematically, the diffusion process is defined for a Markov process $\{\boldsymbol{X}_t\}$ with continuous trajectory and its transition density $p(\boldsymbol{x}, t|\boldsymbol{y}, s)$ $(t \geq s)$ satisfies the following conditions for any $\epsilon > 0$:

$$\lim_{t\to s} \frac{1}{t - s} \int_{|\boldsymbol{x}-\boldsymbol{y}|<\epsilon} (\boldsymbol{x} - \boldsymbol{y}) p(\boldsymbol{x}, t|\boldsymbol{y}, s) d\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{y}, s) + O(\epsilon), \tag{4.4}$$

$$\lim_{t\to s} \frac{1}{t - s} \int_{|\boldsymbol{x}-\boldsymbol{y}|<\epsilon} (\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})^T p(\boldsymbol{x}, t|\boldsymbol{y}, s) d\boldsymbol{x} = \boldsymbol{a}(\boldsymbol{y}, s) + O(\epsilon), \tag{4.5}$$

where $\boldsymbol{b}(\boldsymbol{y}, s)$ is called the drift of the considered diffusion process and $\boldsymbol{a}(\boldsymbol{y}, s)$ is called the diffusion matrix at time s and position $\boldsymbol{y}$. The conditions (4.4) and (4.5) can also be represented as

$$\lim_{t\to s} \frac{1}{t - s} \mathbb{E}^{\boldsymbol{y},s}(\boldsymbol{X}_t - \boldsymbol{y}) = \boldsymbol{b}(\boldsymbol{y}, s), \tag{4.6}$$

$$\lim_{t\to s}\frac{1}{t-s}\mathbb{E}^{\boldsymbol{y},s}(\boldsymbol{X}_t-\boldsymbol{y})(\boldsymbol{X}_t-\boldsymbol{y})=\boldsymbol{a}(\boldsymbol{y},s). \tag{4.7}$$

It is easy to find that the diffusion matrix $\boldsymbol{a}=\boldsymbol{\sigma}\boldsymbol{\sigma}^T$ in (4.1).

4.3 Simple SDEs

Example 4.2 (Ornstein-Uhlenbeck process)**.**

$$dX_t=-\gamma X_t dt+\sigma dW_t. \tag{4.8}$$

The Ornstein-Uhlenbeck process (OU process) has fundamental importance in statistical physics since it serves as the simplest model for many complex diffusion dynamics.

Solution. The equation above is equivalent to

$$dX_t+\gamma X_t dt=\sigma dW_t. \tag{4.9}$$

By applying Ito's formula to $e^{\gamma t}X_t$, we get

$$d(e^{\gamma t}X_t)=\gamma e^{\gamma t}X_t dt+e^{\gamma t}dX_t.$$

Integrating from 0 to t we have

$$e^{\gamma t}X_t-X_0=\int_0^t(\gamma e^{\gamma s}X_s ds+e^{\gamma s}dX_s).$$

Timing $e^{\gamma t}$ to both sides of (4.9) and taking integration, we get

$$e^{\gamma t}X_t-X_0=\int_0^t\sigma e^{\gamma s}dW_s.$$

Thus the solution

$$X_t=e^{-\gamma t}X_0+\sigma\int_0^t e^{-\gamma(t-s)}dW_s.$$

If we define $Q_t:=\int_0^t e^{-\gamma(t-s)}dW_s$, then it is not difficult to know that Q_t is a Gaussion process with

$$\mathbb{E}Q_t=0,\qquad \mathbb{E}Q_t^2=\int_0^t\mathbb{E}e^{-2\gamma(t-s)}ds=\frac{1}{2\gamma}(1-e^{-2\gamma t}).$$

From this result we can observe that X_t is also a Gaussian process if X_0 is Gaussian, and the limit behavior of X_t is

$$X_t\xrightarrow{d}N\left(0,\frac{\sigma^2}{2\gamma}\right),\qquad (t\to+\infty).$$

This equation is called the SDE with additive noise since the coefficient of dW_t term is just a constant.

Example 4.3 (Geometric Brownian motion)**.**

$$dN_t = rN_t dt + \alpha N_t dW_t, \quad r, \alpha > 0. \tag{4.10}$$

This model has strong background in mathematical finance, in which N_t represents the asset price, r is the interest rate and α is called the volatility.

Solution. Divide N_t to both sides we have $dN_t/N_t = rdt + \alpha dW_t$. In deterministic calculus $1/N_t dN_t = d(\log N_t)$, so we apply Ito's formula to $\log N_t$, then

$$\begin{aligned} d(\log N_t) &= \frac{1}{N_t} dN_t - \frac{1}{2N_t^2}(dN_t)^2 \\ &= \frac{1}{N_t} dN_t - \frac{1}{2N_t^2}\alpha^2 N_t^2 dt \\ &= \frac{1}{N_t} dN_t - \frac{\alpha^2}{2} dt. \end{aligned}$$

Substitute the equation of dN_t we get

$$d(\log N_t) = (r - \frac{\alpha^2}{2})dt + \alpha dW_t.$$

Integrate from 0 to t to both sides

$$\begin{aligned} \log N_t - \log N_0 &= (r - \frac{\alpha^2}{2})t + \alpha W_t, \\ N_t &= N_0 \exp\left\{(r - \frac{\alpha^2}{2})t + \alpha W_t\right\}. \end{aligned}$$

This equation is called the SDE with multiplicative noise since the coefficient of dW_t term depends on N_t.

4.4 Brownian motion: revisited

Example 4.4 (Langevin equation)**.** *Mathematically a mesoscopic particle obeys the following well-known Langevin equation by Newton's Second Law*

$$\begin{cases} d\boldsymbol{X}_t &= \boldsymbol{V}_t dt, \\ m d\boldsymbol{V}_t &= \Big(-\gamma \boldsymbol{V}_t - \nabla V(\boldsymbol{X}_t)\Big) dt + \sqrt{2\sigma} d\boldsymbol{W}_t, \end{cases}$$

where γ is frictional coefficient, $V(\boldsymbol{X})$ is external potential, $\boldsymbol{W}_t$ is standard Wiener process, and σ is the strength of fluctuating force.

This example is used to show that the strength of fluctuating force must be related to the frictional coefficient in a physical setup. In principle the fluctuating force must be independent of external potential. In the case that the external force is zero, we have

$$m d\boldsymbol{V}_t = -\gamma \boldsymbol{V}_t dt + \sqrt{2\sigma} d\boldsymbol{W}_t.$$

This is exactly an Ornstein-Uhlenbeck process for $\boldsymbol{V}_t$. In the limit $t \to \infty$, we have

$$\langle \frac{1}{2} m \boldsymbol{V}^2 \rangle = \frac{3\sigma}{2\gamma}.$$

From equilibrium thermodynamics, the average kinetic energy must obey the rule

$$\langle \frac{1}{2} m \boldsymbol{V}^2 \rangle = \frac{3 k_B T}{2}.$$

Thus we obtain the well-known *fluctuation-dissipation relation*:

$$\sigma = k_B T \gamma.$$

It can be proved that in this case the diffusion coefficient

$$D := \lim_{t \to \infty} \frac{\langle (\boldsymbol{X}_t - \boldsymbol{X}_0)^2 \rangle}{6t} = \frac{k_B T}{\gamma} \tag{4.11}$$

which is called *Einstein's relation*.

For more general forms of fluctuation-dissipation relation, the readers may be referred to [7].

Example 4.5 (Brownian dynamics)**.** *In the high γ case, the velocity $\boldsymbol{V}_t$ will always stay at an equilibrium Gaussian distribution, which means formally we can take $d\boldsymbol{V}_t = 0$. Then the Langevin equation is approximated by*

$$d\boldsymbol{X}_t = -\frac{1}{\gamma} \nabla V(\boldsymbol{X}_t) dt + \sqrt{\frac{2 k_B T}{\gamma}} d\boldsymbol{W}_t,$$

which is called Brownian dynamics or Smoluchowski approximation. A mathematically rigorous derivation of Brownian dynamics from Langevin equations may be referred to [6] and the references therein.

5 Stratonovich integral

Another very important definition of the stochastic integral is the so-called Stratonovich (or Fisk-Stratonovich) integral which is defined as the limit of the following approximation

$$\int_0^T f(t, \omega) \circ dW_t \approx \sum_j \frac{f(t_j) + f(t_{j+1})}{2} (W_{t_{j+1}} - W_{t_j}).$$

Note that we use the special notation $\circ$ for stochastic integral to distinguish the Ito and Stratonovich integrals. As one can follow the similar way as in the definition for the Ito

integral, we can also establish a consistent stochastic calculus based on the Stratonovich integral. It turns out that If X_t satisfies the SDE

$$dX_t = b(X_t, t)dt + \sigma(X_t, t) \circ dW_t \tag{5.1}$$

in the Stratonovich sense, then X_t satisfies the modified Ito SDE

$$dX_t = \Big(b(X_t, t) + \frac{1}{2}\partial_x\sigma\sigma(X_t, t)\Big)dt + \sigma(X_t, t)dW_t. \tag{5.2}$$

To understand this, we assume the solution X_t of the Stratonovich SDE satisfies

$$dX_t = \alpha(X_t, t)dt + \beta(X_t, t)dW_t. \tag{5.3}$$

Then by the definition of the Stratonovich integral

$$\int_0^t \sigma(X_s, s) \circ dW_s \approx \sum_j \frac{1}{2}(\sigma(X_{t_j}, t_j) + \sigma(X_{t_{j+1}}, t_{j+1}))(W_{t_{j+1}} - W_{t_j}).$$

From (5.3) we have

$$X_{t_{j+1}} = X_{t_j} + \alpha(X_{t_j}, t_j)\Delta t_j + \beta(X_{t_j}, t_j)\Delta W_{t_j} + h.o.t.,$$

and thus

$$\begin{aligned}\sum_j \sigma(X_{t_{j+1}}, t_{j+1})\Delta W_{t_j} = \sum_j \Big(&\sigma(X_{t_j}, t_j)\Delta W_{t_j} + \partial_t\sigma(X_{t_j}, t_j)\Delta t_j\Delta W_{t_j} \\ &+ \partial_x\sigma\alpha(X_{t_j}, t_j)\Delta t_j\Delta W_{t_j} + \partial_x\sigma\beta(X_{t_j}, t_j)\Delta W_{t_j}^2 + h.o.t.\Big) \\ &\to \int_0^t \sigma(X_s, s)dW_s + \int_0^t \partial_x\sigma\beta(X_s, s)ds\end{aligned}$$

from the fact $dW_t^2 = dt$. Summarizing the above results we obtain that X_t satisfies

$$dX_t = \Big(b(X_t, t) + \frac{1}{2}\partial_x\sigma\beta(X_t, t)\Big)dt + \sigma(X_t, t)dW_t. \tag{5.4}$$

To make (5.3) and (5.3) consistent, we take

$$\beta(x, t) = \sigma(x, t), \quad \alpha(x, t) = b(x, t) + \frac{1}{2}\partial_x\sigma\sigma(x, t).$$

In the high dimensions, one can derive similarly

$$d\boldsymbol{X}_t = \Big(\boldsymbol{b}(\boldsymbol{X}_t, t) + \frac{1}{2}\nabla_x\boldsymbol{\sigma} : \boldsymbol{\sigma}(\boldsymbol{X}_t, t)\Big)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t) \cdot d\boldsymbol{W}_t \tag{5.5}$$

where $(\nabla_x\boldsymbol{\sigma} : \boldsymbol{\sigma})_i := \sum_{jk} \partial_k\sigma_{ij}\sigma_{kj}$ in the index notation if $\boldsymbol{X}$ satisfies

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t) \circ d\boldsymbol{W}_t. \tag{5.6}$$

With this connection, we can check that the Stratonovich integral satisfies the Newton-Leibnitz chain rule

$$df(X_t) = f'(X_t) \circ dX_t = f'(X_t)b(X_t, t)dt + f'(X_t)\sigma(X_t, t) \circ dW_t$$

and its corresponding multi-dimensional form is

$$df(\boldsymbol{X}_t) = \nabla f(\boldsymbol{X}_t) \circ d\boldsymbol{X}_t = \nabla f(\boldsymbol{X}_t) \cdot \boldsymbol{b}(\boldsymbol{X}_t, t)dt + \nabla f(\boldsymbol{X}_t) \cdot \boldsymbol{\sigma}(\boldsymbol{X}_t, t) \circ d\boldsymbol{W}_t.$$

We finally remark here that the Ito isometry and mean zero property no longer hold for the Stratonovich integral, which can be easily observed from (5.2).

One reason that the Stratonovich interpretation is important is due to the following Wong-Zakai type theorem. The motivation is to intuitively understand the SDE (1.2) in the pathwise sense, i.e. for each fixed realization ω of W_t, we want to solve X_t by treating $W_\cdot(\omega)$ like a deterministic forcing term. But the issue is that the ordinary differential equation can not be solved in the classical case because of the rough property of the path of the Brownian motion. Since the C^1 functions on $[0, T]$ are dense in $C[0, T]$, so if we regularize the Brownian motion path from the following way

$$W^m \to W \text{ in } L^\infty[0, T] \text{ norm } \text{ as } m \to \infty,$$

where $W^m \in C^1[0, T]$, the differential equation

$$dX_t^m = b(X_t^m, t)dt + \sigma(X_t^m, t)dW_t^m$$

can be solved in the classical sense. We denote the solution as X_t^m. Then it can be proved that

$$X^m \to X \text{ in } L^\infty[0, T] \text{ norm}, \quad m \to \infty, \ a.s.$$

and the limit X_t is precisely the Stratonovich solution of the SDE (see [11, 12] for more details).

Now a rationale for why Stratonovich interpretation is useful in physics may be as follows. In realistic situations, the noise term $\dot{W}$ in (1.1) is usually not "white" but a smoothed colored noise since the idealistic white noise must be supplied with infinite energy from external environment. This smoothed colored noise exactly corresponds to some regularization of the white noise, which falls into the regime in the Wong-Zakai type smoothing argument.

Homeworks

1. Prove that with midpoint approximation

$$\int_0^t W_s dW_s \approx \sum_j W_{t_{j+\frac{1}{2}}}(W_{t_{j+1}} - W_{t_j}) \to \frac{W_t^2}{2}$$

and the rightmost approximation

$$\int_0^t W_s dW_s \approx \sum_j W_{t_{j+1}}(W_{t_{j+1}} - W_{t_j}) \to \frac{W_t^2}{2} + \frac{t}{2}$$

in $L_P^2(\Omega)$ as $|\Delta| \to 0$.

2. Prove the relation (3.3) through the following steps:

   (a) Prove that the Hermite polynomials satisfy

   $$\sum_{n=0}^{\infty} \frac{u^n}{n!} h_n(x) = \exp\left(ux - \frac{u^2}{2}\right)$$

   and

   $$\sum_{n=0}^{\infty} \frac{u^n}{n!} H_n(x,a) = \exp\left(ux - \frac{au^2}{2}\right),$$

   where $H_n(x,a) = a^{n/2} h_n(x/\sqrt{a})$ $(a > 0)$ and $H_n(x,0) = x^n$.

   (b) Prove that

   $$\left(\frac{1}{2}\frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial a}\right) H_n(x,a) = 0 \quad \text{and} \quad \frac{\partial}{\partial x} H_n(x,a) = n H_{n-1}(x,a).$$

   (c) Prove the relation (3.3) through Itô's formula.

3. Solving the SDE

   (a) $dX_t = -X_t/(1+t)dt + 1/(1+t)dW_t$ with initial $X_0 = 0$.

   (b) $dX_t = -X_t dt + e^{-t} dW_t$ with initial X_0.

4. For the multidimensional OU process

   $$d\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_t dt + \boldsymbol{\sigma} \cdot d\boldsymbol{W}_t,$$

   derive the relations that the stationary mean and covariance matrix should satisfy.

5. Prove that if one takes the right-most endpoint integral (backward stochastic integral) like

   $$\int_0^T f(t,\omega) * dW_t \approx \sum_j f(t_{j+1})(W_{t_{j+1}} - W_{t_j}).$$

   Then the SDE defined as

   $$dX_t = b(X_t,t)dt + \sigma(X_t,t) * dW_t \tag{5.7}$$

   can be related to the Ito SDE as

   $$dX_t = \Big(b(X_t,t) + \partial_x \sigma \sigma(X_t,t)\Big)dt + \sigma(X_t,t)dW_t.$$

References

[1] I.M. Gelfand and G.E. Shilov. *Generalized functions*, volume Vol. 1-5. Academic Press, New York and London, 1964-1966.

[2] T. Hida and Si Si. *Lectures on White noise functionals*. World Scientific Publishing Co., Singapore, 2008.

[3] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. North-Holland Publishing Company, Amsterdam, Oxford and New York, 1981.

[4] K. Ito. Stochastic integral. *Proc. Imp. Acad. Tokyo*, 20:519–524, 1944.

[5] I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, Berlin, Heidelberg and New York, 1991.

[6] P.R. Kramer and A.J. Majda. Stochastic mode reduction for particle-based simulation methods for complex microfluid systems. *SIAM J. Appl. Math.*, 64:401–422, 2003.

[7] R Kubo. The fluctuation-dissipation theorem. *Rep. Prog. Phys.*, 29:255, 1966.

[8] B. Oksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, New York, Berlin, Heidelberg and Tokyo, 4th edition edition, 1998.

[9] P.E. Protter. *Stochastic integral and differential equations*. Springer-Verlag, Berlin, Heidelberg and New York, Second edition edition, 2004.

[10] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Springer-Verlag, Berlin and Heidelberg, 3rd edition edition, 2005.

[11] H.J. Sussman. On the gap between deterministic and stochastic ordinary differential equations. *Ann. Prob.*, 60:19–41, 1978.

[12] E. Wong and M. Zakai. On the convergence of ordinary integrals to stochastic integrasl. *Ann. Math. Stat.*, 36:1560–1564, 1965.

Lecture 14 Connections with PDE *

Tiejun Li

1 Liouville equation

Consider N non-interacting particles moving according to the following deterministic ODEs

$$\frac{d\boldsymbol{X}_t^i}{dt} = \boldsymbol{b}(\boldsymbol{X}_t^i), \quad \boldsymbol{X}_t^i\big|_{t=0} = \boldsymbol{X}_0^i, \quad i = 1, 2, \ldots, N. \tag{1.1}$$

An interesting question is to ask what the transition rule for the distribution of these particles is in macroscopic viewpoint, that is, to describe its distributive law when the number of particles N goes to infinity. To investigate this, it is natural to consider its empirical distribution at time t at first

$$\mu^N(\boldsymbol{x}, t) = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{x} - \boldsymbol{X}_t^i),$$

where $\delta(\cdot)$ is the Dirac's δ-function. We have for any compactly supported smooth function $\phi(\boldsymbol{x}) \in C_c^\infty(\mathbb{R}^d)$

$$\begin{aligned}
\frac{d}{dt}(\mu^N, \phi) &= \frac{1}{N}\sum_{i=1}^N \frac{d}{dt}\int_{\mathbb{R}^d} \delta(\boldsymbol{x} - \boldsymbol{X}_t^i)\phi(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{1}{N}\sum_{i=1}^N \frac{d}{dt}\phi(\boldsymbol{X}_t^i) = \frac{1}{N}\sum_{i=1}^N \nabla_{\boldsymbol{x}}\phi(\boldsymbol{X}_t^i)\cdot\boldsymbol{b}(\boldsymbol{X}_t^i) \\
&= \Big(\mu^N, \boldsymbol{b}\cdot\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x})\Big),
\end{aligned}$$

where the notation $(\boldsymbol{f}, \boldsymbol{g}) := \int_{\mathbb{R}^d} \boldsymbol{f}(\boldsymbol{x})\cdot\boldsymbol{g}(\boldsymbol{x})d\boldsymbol{x}$ is the inner product of functions. Denote the space of probability measures on $\mathbb{R}^d$ as $\mathcal{M}(\mathbb{R}^d)$. Now let us suppose the initial distribution

$$\mu^N(\boldsymbol{x}, 0) := \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{x} - \boldsymbol{X}_0^i) \xrightarrow{*} \mu_0(\boldsymbol{x}) \in \mathcal{M}(\mathbb{R}^d) \text{ as } N \to \infty$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

in the sense that $(\mu^N, \phi) \to (\mu, \phi)$ for any $\phi \in C_c^\infty(\mathbb{R}^d)$. One can establish the limit $\mu^N(\boldsymbol{x}, t) \overset{*}{\to} \mu(\boldsymbol{x}, t)$ and indeed μ satisfies

$$\frac{d}{dt}(\mu, \phi) = (\mu, \boldsymbol{b} \cdot \nabla_{\boldsymbol{x}} \phi(\boldsymbol{x})), \quad \mu(\boldsymbol{x}, 0) = \mu_0(\boldsymbol{x}).$$

If we assume the probability measure μ has density $\psi(\boldsymbol{x}, t) \in C^1(\mathbb{R}^d \times [0, T])$, then we obtain the following hyperbolic equation after integration by parts

$$\partial_t \psi + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{b}\psi) = 0.$$

If the drift vector $\boldsymbol{b}$ satisfies $\nabla_{\boldsymbol{x}} \cdot \boldsymbol{b} = 0$, we get

$$\partial_t \psi + \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \psi = 0.$$

This is called the Liouville equation which is well-known in classical mechanics. The orbit of the equation

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{b}(\boldsymbol{x})$$

is called the characteristics of the above hyperbolic PDE.

2 Fokker-Planck equation

If the deterministic equation (1.1) is replaced with the following SDEs

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t) \cdot d\boldsymbol{W}_t, \tag{2.1}$$

the same question on the probability distribution of $\boldsymbol{X}$ may be asked. To simplify the discussion, we assume the transition probability density function exists and is defined as $(t \geq s)$

$$p(\boldsymbol{x}, t|\boldsymbol{y}, s)d\boldsymbol{x} = \mathbb{P}\{\boldsymbol{X}_t \in [\boldsymbol{x}, \boldsymbol{x} + d\boldsymbol{x})|\boldsymbol{X}_s = \boldsymbol{y}\}.$$

For any function $f \in C_c^\infty(\mathbb{R}^d)$, the Ito formula gives

$$\begin{aligned} df(\boldsymbol{X}_t) &= \nabla f(\boldsymbol{X}_t) \cdot d\boldsymbol{X}_t + \frac{1}{2}(d\boldsymbol{X}_t)^T \cdot \nabla^2 f(\boldsymbol{X}_t) \cdot (d\boldsymbol{X}_t) \\ &= (\boldsymbol{b} \cdot \nabla f + \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f)dt + \nabla f \cdot \boldsymbol{\sigma} \cdot d\boldsymbol{W}_t. \end{aligned}$$

Integrating both sides from s to t we get

$$\begin{aligned} f(\boldsymbol{X}_t) - f(\boldsymbol{X}_s) &= \int_s^t \nabla f(\boldsymbol{X}_\tau) \cdot \{\boldsymbol{b}(\boldsymbol{X}_\tau, \tau)d\tau + \boldsymbol{\sigma}(\boldsymbol{X}_\tau, \tau)d\boldsymbol{W}_\tau\} \\ &\quad + \frac{1}{2}\int_s^t \sum_{i,j} \partial_{ij}^2 f(\boldsymbol{X}_\tau) a_{ij}(\boldsymbol{X}_\tau, \tau)d\tau, \end{aligned}$$

where the diffusion matrix $\boldsymbol{a}(\boldsymbol{x},t)=\boldsymbol{\sigma}(\boldsymbol{x},t)\boldsymbol{\sigma}^T(\boldsymbol{x},t)$. Now taking expectation on both sides and utilizing the initial condition $\boldsymbol{X}_s=\boldsymbol{y}$, we have

$$\mathbb{E}f(\boldsymbol{X}_t)-f(\boldsymbol{y})=\mathbb{E}\int_s^t \mathcal{L}f(\boldsymbol{X}_\tau,\tau)d\tau, \tag{2.2}$$

where the operator $\mathcal{L}$ is defined as

$$\mathcal{L}f(\boldsymbol{x},t)=\boldsymbol{b}(\boldsymbol{x},t)\cdot\nabla f(\boldsymbol{x})+\frac{1}{2}\sum_{i,j}a_{ij}(\boldsymbol{x},t)\partial^2_{ij}f(\boldsymbol{x}). \tag{2.3}$$

In the language of transition pdf $p(\boldsymbol{x},t|\boldsymbol{y},s)$, we have

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})p(\boldsymbol{x},t|\boldsymbol{y},s)d\boldsymbol{x}-f(\boldsymbol{y}) \;=\; \int_s^t\int_{\mathbb{R}^d}\mathcal{L}f(\boldsymbol{x},\tau)p(\boldsymbol{x},\tau|\boldsymbol{y},s)d\boldsymbol{x}d\tau.$$

This is exactly the definition of the weak solution of the PDE with respect to t and $\boldsymbol{x}$

$$\partial_t p=\mathcal{L}^*_{\boldsymbol{x}}p(\boldsymbol{x},t|\boldsymbol{y},s),\quad p(\boldsymbol{x},t|\boldsymbol{y},s)|_{t=s}=\delta(\boldsymbol{x}-\boldsymbol{y}),\quad t\geq s, \tag{2.4}$$

in the sense of distribution, where the operator $\mathcal{L}^*$ is the formal adjoint of $\mathcal{L}$ defined through

$$(\mathcal{L}f,g)_{L^2}=(f,\mathcal{L}^*g)_{L^2}.$$

The concrete form of $\mathcal{L}^*$ reads

$$\mathcal{L}^*f(\boldsymbol{x},t)=-\nabla_{\boldsymbol{x}}\cdot(\boldsymbol{b}(\boldsymbol{x},t)f(\boldsymbol{x}))+\frac{1}{2}\nabla^2_{\boldsymbol{x}}:(\boldsymbol{a}(\boldsymbol{x},t)f(\boldsymbol{x})), \tag{2.5}$$

where $\nabla^2_{\boldsymbol{x}}:(\boldsymbol{a}f)=\sum_{ij}\partial_{ij}(a_{ij}f)$. Indeed by assuming the solution $p(\boldsymbol{x},t|\boldsymbol{y},s)\in C^{2,1}(\mathbb{R}^d,[0,T])$, which means p is C^2 in $\boldsymbol{x}$-variable and C^1 in t-variable, we can directly obtain the PDE (2.4) through integration by parts. For the rigorous proof about the connection between the SDEs and above PDE, the readers may be referred to [**?**].

The Equation (2.4) is well-known as the *Kolmogorov's forward equation*, or the *Fokker-Planck equation* in physics. The "forward" means it is for the forward time variable $t>s$ and its corresponding space variable $\boldsymbol{x}$. When we consider the equation for the backward time variable $s<t$ and $\boldsymbol{y}$, we will call it *backward equation*, which will be considered in Section 4. The transition pdf $p(\boldsymbol{x},t|\boldsymbol{y},s)$ is simply the fundamental solution of this operator. By analogy with the deterministic case, the SDE (2.1) may be regarded as the "stochastic characteristics" of the parabolic equation (2.4). This viewpoint will be found to be very useful in many situations.

We finally remark that the joint distribution $p(\boldsymbol{x},t;\boldsymbol{y},s)$ and the distribution density $p(\boldsymbol{x},t)$ starting from some initial distribution both satisfy the forward Kolmogorov type equation with respect to $\boldsymbol{x}$ and t. The reason is straightforward since the derivation from $p(\boldsymbol{x},t|\boldsymbol{y},s)$ to $p(\boldsymbol{x},t;\boldsymbol{y},s)$ or $p(\boldsymbol{x},t)$ is simply by timing $p(\boldsymbol{y},s)$ and integrating with respect to $\boldsymbol{y}$.

Example 2.1 (Brownian motion)**.** *The SDE reads*

$$d\boldsymbol{X}_t = d\boldsymbol{W}_t, \quad \boldsymbol{X}_0 = 0.$$

So the Fokker-Planck equation is

$$\partial_t p = \frac{1}{2}\Delta p, \quad p(\boldsymbol{x}, 0) = \delta(\boldsymbol{x}). \tag{2.6}$$

It is well-known from PDE that its unique solution is the heat kernal

$$p(\boldsymbol{x}, t) = \frac{1}{\sqrt{2\pi t}} \exp\Big(-\frac{\boldsymbol{x}^2}{2t}\Big),$$

which is exactly the pdf of $N(0, t\boldsymbol{I})$. The PDE (2.6) *gives another characterization of the Brownian motion.*

Example 2.2 (Brownian dynamics)**.** *The SDE reads*

$$d\boldsymbol{X}_t = -\frac{1}{\gamma}\nabla V(\boldsymbol{X}_t)dt + \sqrt{\frac{2k_BT}{\gamma}}d\boldsymbol{W}_t. \tag{2.7}$$

So the Fokker-Planck equation is

$$\partial_t p - \nabla \cdot \Big(\frac{1}{\gamma}\nabla V(\boldsymbol{x})p\Big) = \frac{k_BT}{\gamma}\Delta p = D\Delta p, \tag{2.8}$$

where $D = k_BT/\gamma$ is the diffusion coefficient. Note that this also gives another understanding about the Einstein's relation in (**??**).

Alternatively (2.8) can be derived from the following recipe. Define the free energy associated with the pdf p as

$$\mathcal{F}(p) = \int_{\mathbb{R}^d} \Big(k_BTp(\boldsymbol{x})\ln p(\boldsymbol{x}) + V(\boldsymbol{x})p(\boldsymbol{x})\Big)d\boldsymbol{x}, \tag{2.9}$$

where the first term $k_B\int_{\mathbb{R}^d} p(\boldsymbol{x})\ln p(\boldsymbol{x})d\boldsymbol{x}$ corresponds to the negative entropy $-S$ in thermodynamics, and the second term $\int_{\mathbb{R}^d} V(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ is the internal energy U. The chemical potential μ is then given by

$$\mu = \frac{\delta\mathcal{F}}{\delta p} = k_BT(1 + \ln p(\boldsymbol{x})) + V(\boldsymbol{x}).$$

The current density is defined as

$$\boldsymbol{j}(\boldsymbol{x}) := p(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}) \tag{2.10}$$

with the velocity field $\boldsymbol{u}(\boldsymbol{x})$ given by the Fick's Law

$$\boldsymbol{u}(\boldsymbol{x}) = \frac{1}{\gamma}\boldsymbol{f} = -\frac{1}{\gamma}\nabla\mu,$$

where $\boldsymbol{f} = -\nabla\mu$ is the force field. Then the Smoluchowski's equation (2.8) is a consequence of the continuity equation

$$\partial_t p + \nabla \cdot \boldsymbol{j} = 0.$$

This approach via deterministic PDE to describe the Brownian dynamics is more common in physics.

Finally we want to mention that if the underlying stochastic dynamics is a Stratonovich SDE, we will have its transition pdf satisfies the following type of PDE

$$\partial_t p + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{b}p) = \frac{1}{2}\nabla_{\boldsymbol{x}} \cdot (\boldsymbol{\sigma} \cdot \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{\sigma}p)), \tag{2.11}$$

where $\nabla_{\boldsymbol{x}} \cdot (\boldsymbol{\sigma} \cdot \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{\sigma}p)) = \partial_i(\sigma_{ik}\partial_j(\sigma_{jk}p))$. If the underlying stochastic dynamics is defined through the backward stochastic integral,

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{x}, t)dt + \boldsymbol{\sigma}(\boldsymbol{x}, t) * d\boldsymbol{W}_t,$$

then $p(\boldsymbol{x}, t)$ satisfies

$$\partial_t p + \partial_i\Big[(b_i + \partial_k\sigma_{ij}\sigma_{kj})p\Big] = \frac{1}{2}\partial_{ij} : (\sigma_{ik}\sigma_{jk}p), \tag{2.12}$$

where the Einstein summation convention is assumed. In the one-dimensional case, it can be simplified to

$$\partial_t p + \partial_x(bp) = \frac{1}{2}\partial_x(\sigma^2\partial_x p). \tag{2.13}$$

The proof is straightforward and left as an exercise.

3 Boundary Condition

Many stochastic problems occur in a bounded domain, in which case the boundary conditions are needed. To pose suitable boundary conditions in different situations, we need to understand the probability current $\boldsymbol{j}(\boldsymbol{x}, t) = \boldsymbol{b}(\boldsymbol{x}, t)p(\boldsymbol{x}, t) - 1/2\nabla_{\boldsymbol{x}} \cdot (\boldsymbol{a}(\boldsymbol{x}, t)p(\boldsymbol{x}, t))$ in the Fokker-Planck equation

$$\partial_t p(\boldsymbol{x}, t) + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{j}(\boldsymbol{x}, t) = 0 \tag{3.1}$$

more intuitively at first. To do this, let us investigate the role of probability flux between regions R_1 and R_2 separated by a boundary S_{12} (see Fig. 1).

Consider the probability transfer from region R_1 to R_2 during the time t to $t + \delta t$, we have

$$P_{1\to 2} = \int_{R_2} d\boldsymbol{x} \int_{R_1} d\boldsymbol{y} p(\boldsymbol{x}, t + \delta t; \boldsymbol{y}, t),$$

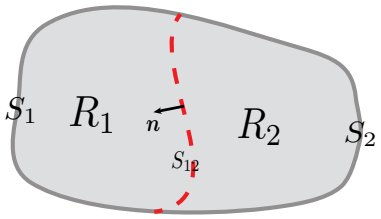


Figure 1: Probability flux across a boundary

and with the similar reason the probability transfer from region R_2 to R_1 has the form

$$P_{2\to 1} = \int_{R_1} d\boldsymbol{x} \int_{R_2} d\boldsymbol{y} p(\boldsymbol{x}, t+\delta t; \boldsymbol{y}, t).$$

Thus the net probability flow rate from R_2 to R_1 is

$$J_{2\to 1} = \lim_{\delta t\to 0}(P_{2\to 1} - P_{1\to 2})/\delta t.$$

With the equality

$$\int_{R_2} d\boldsymbol{x} \int_{R_1} d\boldsymbol{y} p(\boldsymbol{x}, t; \boldsymbol{y}, t) = 0,$$

we obtain

$$\begin{aligned} J_{2\to 1} &= \int_{R_1} d\boldsymbol{x} \int_{R_2} d\boldsymbol{y} \partial_t p(\boldsymbol{x}, t; \boldsymbol{y}, s=t) - \int_{R_2} d\boldsymbol{x} \int_{R_1} d\boldsymbol{y} \partial_t p(\boldsymbol{x}, t; \boldsymbol{y}, s=t) \\ &= \int_{R_2} d\boldsymbol{x} \nabla_{\boldsymbol{x}} \cdot \boldsymbol{j}(\boldsymbol{x}, t; R_1, t) - \int_{R_1} d\boldsymbol{x} \nabla_{\boldsymbol{x}} \cdot \boldsymbol{j}(\boldsymbol{x}, t; R_2, t) \\ &= \int_{S_{12}} dS \boldsymbol{n} \cdot (\boldsymbol{j}(\boldsymbol{x}, t; R_1, t) + \boldsymbol{j}(\boldsymbol{x}, t; R_2, t)), \end{aligned}$$

where $\boldsymbol{j}(\boldsymbol{x}, t; R_1, t) := \int_{R_1} d\boldsymbol{y} \boldsymbol{j}(\boldsymbol{x}, t; \boldsymbol{y}, t)$, $\boldsymbol{n}$ is the normal pointing from R_2 to R_1. The last equality is obtained by divergence theorem and the fact that $\boldsymbol{j}(\boldsymbol{x}, t; R_2, t) = 0$ when $\boldsymbol{x} \in S_1$ and $\boldsymbol{j}(\boldsymbol{x}, t; R_1, t) = 0$ when $\boldsymbol{x} \in S_2$. From the fact that $\boldsymbol{x} \in R_1 \cup R_2$ we have $\boldsymbol{j}(\boldsymbol{x}, t) = \int_{\mathbb{R}^d} d\boldsymbol{y} \boldsymbol{j}(\boldsymbol{x}, t; \boldsymbol{y}, t) = \boldsymbol{j}(\boldsymbol{x}, t; R_1, t) + \boldsymbol{j}(\boldsymbol{x}, t; R_2, t)$ and thus

$$J_{2\to 1} = \int_{S_{12}} dS \boldsymbol{n} \cdot \boldsymbol{j}(\boldsymbol{x}, t).$$

Recalling the probability flux defined as

$$J_{ij}^n = \mu_{n,i} p_{ij} - \mu_{n,j} p_{ji}$$

from state i to state j at time n in a discrete time Markov chain and

$$J_{ij}(t) = \mu_i(t) p_{ij} - \mu_j(t) p_{ji}$$

for a continuous time Markov chain, we have that $\boldsymbol{n}\cdot\boldsymbol{j}(\boldsymbol{x},t)$ is exactly the continuous space version of $J_{ij}(t)$ along a specific direction $\boldsymbol{n}$.

Three commonly used boundary conditions are as follows. It will be instructive for the readers to compare them with the boundary conditions for the Wiener process.

Reflecting barrier. In the microscopic sense, the reflecting barrier means that the particles will be reflected once it hits the boundary ∂D. Thus there will be no probability flux across ∂D and the reflecting boundary condition has the form

$$\boldsymbol{n}\cdot\boldsymbol{j}(\boldsymbol{x},t)=0 \quad \boldsymbol{x}\in\partial D. \tag{3.2}$$

Note that in this case the total probability is conserved since

$$\begin{aligned}\frac{d}{dt}\int_D p(\boldsymbol{x},t)d\boldsymbol{x} &= -\int_D \nabla_{\boldsymbol{x}}\cdot\boldsymbol{j}(\boldsymbol{x},t)d\boldsymbol{x}\\ &= -\int_{\partial D}\boldsymbol{n}\cdot\boldsymbol{j}(\boldsymbol{x},t)dS=0.\end{aligned}$$

Absorbing barrier. In the microscopic sense, the absorbing barrier means that the particles will be absorbed (or removed) once it hits the boundary ∂D. Thus the probability on the boundary ∂D will be zero. The absorbing boundary condition is

$$p(\boldsymbol{x},t)=0 \quad \boldsymbol{x}\in\partial D. \tag{3.3}$$

The total probability is no longer conserved in this case.

Periodic boundary condition. In the periodic case with period L_j in the x_j-direction for $j=1,\ldots,d$, the boundary condition is

$$p(x_j+L_j,t)=p(x_j,t), \quad j=1,2,\ldots,d.$$

4 Backward equation

Now let us consider the equation for the transition pdf $p(\boldsymbol{x},t|\boldsymbol{y},s)$ with respect to variable $\boldsymbol{y}$ and s. Suppose $\boldsymbol{X}_t$ satisfies (2.1). For any given $f(\boldsymbol{x})\in C_c^\infty(\mathbb{R}^d)$, we define

$$u(\boldsymbol{y},s)=\mathbb{E}^{\boldsymbol{y},s}f(\boldsymbol{X}_t)=\int_{\mathbb{R}^d} f(\boldsymbol{x})p(\boldsymbol{x},t|\boldsymbol{y},s)d\boldsymbol{x}, \quad s\le t.$$

Assume that $p(\boldsymbol{x},t|\boldsymbol{y},s)$ is C^1 in s and C^2 in $\boldsymbol{y}$, then we have

$$du(\boldsymbol{X}_\tau,\tau)=(\partial_\tau u+\mathcal{L}u)(\boldsymbol{X}_\tau,\tau)d\tau+\nabla u\cdot\boldsymbol{\sigma}\cdot d\boldsymbol{W}_\tau$$

by Ito formula. Taking expectation we obtain

$$\begin{aligned}\lim_{t\to s}\frac{1}{t-s}(\mathbb{E}^{\boldsymbol{y},s}u(\boldsymbol{X}_t,t)-u(\boldsymbol{y},s)) &= \lim_{t\to s}\frac{1}{t-s}\int_s^t \mathbb{E}^{\boldsymbol{y},s}(\partial_\tau u+\mathcal{L}u)(\boldsymbol{X}_\tau,\tau)d\tau\\ &= \partial_s u(\boldsymbol{y},s)+\mathcal{L}u(\boldsymbol{y},s).\end{aligned}$$

On the other hand it is obvious that

$$\mathbb{E}^{\boldsymbol{y},s}u(\boldsymbol{X}_t,t)=\mathbb{E}^{\boldsymbol{y},s}f(\boldsymbol{X}_t)=u(\boldsymbol{y},s)$$

and thus

$$\partial_s u(\boldsymbol{y},s)+\mathcal{L}u(\boldsymbol{y},s)=0.$$

From the arbitrariness of f, we obtain

$$\partial_s p(\boldsymbol{x},t|\boldsymbol{y},s)+\mathcal{L}_{\boldsymbol{y}}p(\boldsymbol{x},t|\boldsymbol{y},s)=0,\quad p(\boldsymbol{x},t|\boldsymbol{y},t)=\delta(\boldsymbol{x}-\boldsymbol{y}),\quad s<t. \tag{4.1}$$

This is the well-know *Kolmogorov backward equation* for the transition density since the time variable s goes backward.

5 Invariant distribution and detailed balance

Consider the Fokker-Planck equation (3.1) for describing the evolution of the probability density. It is interesting to study the case when the system achieves a steady state: that is, the pdf is independent of the time, if the system admits such a solution. This situation is only meaningful when the drift $\boldsymbol{b}$ and diffusion coefficient $\boldsymbol{\sigma}$ does not depend on t. In this case, the process $\{\boldsymbol{X}_t\}$ is a time-homogeneous Markov process since the transition rule only depends on the states other than the time. The steady state pdf satisfies the following PDE

$$\nabla_{\boldsymbol{x}}\cdot(\boldsymbol{b}(\boldsymbol{x})p_s(\boldsymbol{x}))=\frac{1}{2}\nabla_{\boldsymbol{x}}^2:(\boldsymbol{a}(\boldsymbol{x})\ p_s(\boldsymbol{x})) \tag{5.1}$$

with suitable boundary conditions. This $p_s(\boldsymbol{x})$ is called the *stationary distribution* or *invariant distribution* of the considered system.

Specially for the Langevin equation (2.7), the invariant distribution satisfies

$$\nabla\cdot\boldsymbol{j}_s(\boldsymbol{x})=0,$$

where $\boldsymbol{j}_s$ is defined in (2.10). In particular, we are interested in the *equilibrium solution* with a stronger condition $\boldsymbol{j}_s=0$, i.e. the *detailed balance condition* in the continuous case, which implies the chemical potential

$$\mu=\text{constant}.$$

It is not difficult to deduce the following well-known Gibbs distribution for the equilibrium

$$p_s(\boldsymbol{x})=\frac{1}{Z}\exp\Big(-\frac{V(\boldsymbol{x})}{k_BT}\Big) \tag{5.2}$$

as long as the normalization constant

$$Z=\int_{\mathbb{R}^d}e^{-\frac{V(\boldsymbol{x})}{k_BT}}d\boldsymbol{x} \tag{5.3}$$

is finite.

6 Further topics on Diffusion Processes

All of the discussions in this section are considered for the time-homogeneous SDEs

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t) \cdot d\boldsymbol{W}_t. \tag{6.1}$$

where $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ are independent of time t. This time-homogeneity implies that the translational invariance of time for its transition kernel $p(\cdot, t|\boldsymbol{y}, s)$ (see pp. 110 in [2])

$$p(A, t+s|\boldsymbol{y}, s) = p(A, t|\boldsymbol{y}, 0), \quad s, t \geq 0$$

for any $\boldsymbol{y} \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$, where

$$p(A, t|\boldsymbol{y}, s) := \mathbb{E}^{\boldsymbol{y},s} 1_A(\boldsymbol{X}_t) = \int_A p(d\boldsymbol{x}, t|\boldsymbol{y}, s).$$

6.1 Semigroup and backward Equation

Define the operator T_t on any function $f \in C_0(\mathbb{R}^d)$ as

$$T_t f(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_t) = \int_{\mathbb{R}^d} f(\boldsymbol{z}) p(d\boldsymbol{z}, t|\boldsymbol{x}, 0).$$

Then we have $T_0 f(\boldsymbol{x}) = f(\boldsymbol{x})$ and the following semigroup property for any $t, s \geq 0$

$$\begin{aligned} T_t \circ T_s f(\boldsymbol{x}) &= \mathbb{E}^{\boldsymbol{x}}(\mathbb{E}^{\boldsymbol{X}_t} f(\boldsymbol{X}_s)) \\ &= \int p(d\boldsymbol{y}, t|\boldsymbol{x}, 0) \int f(\boldsymbol{z}) p(d\boldsymbol{z}, s|\boldsymbol{y}, 0) \\ &= \int f(\boldsymbol{z}) \int p(d\boldsymbol{z}, s+t|\boldsymbol{y}, t) p(d\boldsymbol{y}, t|\boldsymbol{x}, 0) \\ &= \mathbb{E}^{\boldsymbol{x}}(f(\boldsymbol{X}_{t+s})) = T_{t+s} f(\boldsymbol{x}). \end{aligned}$$

Under the condition that $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ are bounded and Lipschitz, one can further show $T_t : C_0(\mathbb{R}^d) \to C_0(\mathbb{R}^d)$ and it is strongly continuous (Theorem 18.11 in [1]) in the sense that

$$\lim_{t \to 0+} \|T_t f - f\|_\infty = 0, \quad \text{for any } f \in C_0(\mathbb{R}^d).$$

T_t is called *Feller semigroup* in the literature. With this setup, we can utilize the tools from semigroup theory to study T_t [3].

Definition 6.1. *The infinitesimal generator $\mathcal{A}$ of T_t is defined as*

$$\mathcal{A} f(\boldsymbol{x}) = \lim_{t \to 0+} \frac{\mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_t) - f(\boldsymbol{x})}{t},$$

where $f \in D(\mathcal{A}) := \{f \in C_0(\mathbb{R}^d)$ such that the limit exists$\}$.

For $f \in C_c^2(\mathbb{R}^d) \subset D(\mathcal{A})$ we have

$$\mathcal{A}f(\boldsymbol{x}) = \mathcal{L}f(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla f(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{\sigma}\boldsymbol{\sigma}^T) : \nabla^2 f(\boldsymbol{x}).$$

from Ito formula (2.2). We will show that $u(\boldsymbol{x}, t) = \mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_t)$ satisfies the *backward equation* for $f \in C_c^2(\mathbb{R}^d)$

$$\partial_t u = \mathcal{A}u(\boldsymbol{x}), \quad u|_{t=0} = f(\boldsymbol{x}). \tag{6.2}$$

Proof. At first it is not difficult to observe that $u(\boldsymbol{x}, t)$ is differentiable with respect to t from Ito's formula and the condition $f \in C_c^2(\mathbb{R}^d)$. For any fixed $t > 0$, define $g(\boldsymbol{x}) = u(\boldsymbol{x}, t)$. Then we have

$$\begin{aligned}
\mathcal{A}g(\boldsymbol{x}) &= \lim_{s\to 0+} \frac{1}{s}\Big(\mathbb{E}^{\boldsymbol{x}} g(\boldsymbol{X}_s) - g(\boldsymbol{x})\Big) \\
&= \lim_{s\to 0+} \frac{1}{s}\Big(\mathbb{E}^{\boldsymbol{x}}\mathbb{E}^{\boldsymbol{X}_s} f(\boldsymbol{X}_t) - \mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_t)\Big) \\
&= \lim_{s\to 0+} \frac{1}{s}\Big(\mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_{t+s}) - \mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_t)\Big) \\
&= \lim_{s\to 0+} \frac{1}{s}(u(\boldsymbol{x}, t+s) - u(\boldsymbol{x}, t)) = \partial_t u(\boldsymbol{x}, t).
\end{aligned}$$

This means $u(\cdot, t) \in D(\mathcal{A})$ and the proof is complete.

The readers can also derive the equation (6.2) from (4.1) if the transition pdf exists.

6.2 Feynman-Kac Formula

Theorem 6.2. *(Feynman-Kac Formula) Let $f \in C_0^2(\mathbb{R}^d)$ and $q \in C(\mathbb{R}^d)$. Assume that q is lower bounded, then*

$$v(\boldsymbol{x}, t) = \mathbb{E}^x\Big(\exp(\int_0^t q(\boldsymbol{X}_s)ds) f(\boldsymbol{X}_t)\Big)$$

satisfies the PDE

$$\partial_t v = \mathcal{A}v + qv, \quad v|_{t=0} = f(\boldsymbol{x}). \tag{6.3}$$

Intuitive explanation: In the absence of Brownian motion, the SDE becomes

$$\frac{d\boldsymbol{X}_t}{dt} = \boldsymbol{b}(\boldsymbol{X}_t), \quad \boldsymbol{X}_0 = \boldsymbol{x}$$

and the PDE becomes

$$\partial_t v = \boldsymbol{b} \cdot \nabla v + qv, \quad v|_{t=0} = f(\boldsymbol{x}).$$

The method of characteristics gives us

$$v(\boldsymbol{x}, t) = \exp(\int_0^t q(\boldsymbol{X}_s)ds) f(\boldsymbol{X}_t).$$

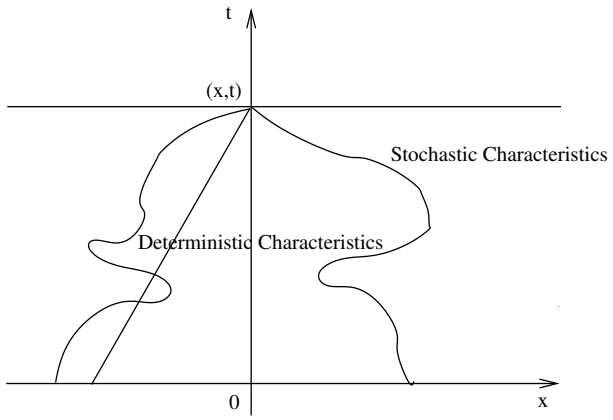


Figure 2: Schematics of Feynmann-Kac formula.

The Feynmann-Kac formula tells us the solution of that parabolic PDE (6.3) can be represented by the ensemble of solution for the ODEs with stochastic characteristics originated from $\boldsymbol{x}$.

Proof. Let $Y_t = f(\boldsymbol{X}_t), Z_t = \exp(\int_0^t q(\boldsymbol{X}_s)ds)$, define $v(\boldsymbol{x},t) = \mathbb{E}^{\boldsymbol{x}}(Y_t Z_t)$. With the similar reason as the previous section, we have $v(\boldsymbol{x},t)$ is differentiable with respect to t and

$$\begin{aligned}
\frac{1}{s}\Big(\mathbb{E}^{\boldsymbol{x}} v(\boldsymbol{X}_s,t) - v(\boldsymbol{x},t)\Big) &= \frac{1}{s}\Big(\mathbb{E}^{\boldsymbol{x}}\mathbb{E}^{\boldsymbol{X}_s} Z_t f(\boldsymbol{X}_t) - \mathbb{E}^{\boldsymbol{x}} Z_t f(\boldsymbol{X}_t)\Big) \\
&= \frac{1}{s}\Big(\mathbb{E}^{\boldsymbol{x}} \exp(\int_0^t q(\boldsymbol{X}_{r+s})dr) f(\boldsymbol{X}_{t+s}) - \mathbb{E}^{\boldsymbol{x}} Z_t f(\boldsymbol{X}_t)\Big) \\
&= \frac{1}{s}\mathbb{E}^{\boldsymbol{x}}\Big(\exp(-\int_0^s q(\boldsymbol{X}_r)dr) Z_{t+s} f(\boldsymbol{X}_{t+s}) - Z_t f(\boldsymbol{X}_t)\Big) \\
&= \frac{1}{s}\mathbb{E}^{\boldsymbol{x}}\Big(Z_{t+s} f(\boldsymbol{X}_{t+s}) - Z_t f(\boldsymbol{X}_t)\Big) \\
&+ \frac{1}{s}\mathbb{E}^{\boldsymbol{x}}\Big(Z_{t+s} f(\boldsymbol{X}_{t+s})(\exp(-\int_0^s q(\boldsymbol{X}_r)dr) - 1)\Big) \\
&\to \partial_t v - q(\boldsymbol{x}) v(\boldsymbol{x},t) \quad \text{ as } s \to 0.
\end{aligned}$$

The left hand side is $\mathcal{A}v(\boldsymbol{x},t)$ by definition. The proof is complete.

6.3 First exit time

Theorem 6.3. *Suppose $D \subset \mathbb{R}^d$ is a bounded open set and the boundary ∂D is of C^2 type. The coefficients $\boldsymbol{b}, \boldsymbol{\sigma}$ of the SDEs satisfy the Lipschitz condition on $\bar{D}$ and the diffusion matrix $\boldsymbol{a}$ is coercive which is defined as*

$$\sum_{i,j} a_{ij}(\boldsymbol{x})\xi_i\xi_j \geq K|\xi|^2 \quad \textit{for } \boldsymbol{x} \in D,\ \xi \in \mathbb{R}^d,\ K > 0.$$

Then for $f \in C(\partial D)$, the solution of PDE

$$\mathcal{A}u = 0 \text{ in } D, \quad u = f(\boldsymbol{x}) \text{ on } \partial D$$

can be represented as

$$u(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{x}}\big(f(\boldsymbol{X}_{\tau_D})\big),$$

where τ_D is the first exit time from domain D defined as

$$\tau_D := \inf_t\{t \geq 0, \boldsymbol{X}_t \notin D\}$$

and thus $\boldsymbol{X}_{\tau_D}$ is the first exit point. Specially, if $\mathcal{A}u = \Delta u$, then $u(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{x}}\big(f(\boldsymbol{W}_{\tau_D})\big)$.

Heuristic proof. From PDE theory, one has the solution $u \in C^2(D) \cap C(\bar{D})$ (c.f. Chapter 6 in [**?**]). So we can apply the Ito's formula to $u(\boldsymbol{X}_t)$ and take expectation

$$\mathbb{E}^{\boldsymbol{x}} u(\boldsymbol{X}_{\tau_D}) - u(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{x}} \int_0^{\tau_D} \mathcal{A}u(\boldsymbol{X}_t)dt = 0. \tag{6.4}$$

Thus

$$u(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{x}} u(\boldsymbol{X}_{\tau_D}) = \mathbb{E}^{\boldsymbol{x}}\big(f(\boldsymbol{X}_{\tau_D})\big).$$

Note that in the above derivations we naively take the expectation of the stochastic integral term to be zero. But this is not true in general because τ_D is a random time. In fact, it is the result of the following useful Dynkin's formula.

Lemma 6.4 (Dynkin's formula)**.** *Let $f \in C_0^2(\mathbb{R}^d)$. Suppose τ is a stopping time with $\mathbb{E}^{\boldsymbol{x}}\tau < \infty$, then*

$$\mathbb{E}^{\boldsymbol{x}} f(\boldsymbol{X}_\tau) = f(\boldsymbol{x}) + \mathbb{E}^{\boldsymbol{x}} \int_0^{\tau} \mathcal{A}u(\boldsymbol{X}_t)dt.$$

To prove $\mathbb{E}^{\boldsymbol{x}}\tau_D < \infty$, we define an auxiliary function $h(\boldsymbol{x}) = -A\exp(\lambda x_1)$. Then for sufficiently large $A, \lambda > 0$ we have

$$\mathcal{A}h(\boldsymbol{x}) = \sum_{ij} a_{ij}(\boldsymbol{x})\partial_{ij}h(\boldsymbol{x}) + \sum_i b_i(\boldsymbol{x})\partial_i h(\boldsymbol{x}) \leq -1, \quad \boldsymbol{x} \in D.$$

By Itô's formula

$$\mathbb{E}^{\boldsymbol{x}} h(\boldsymbol{X}_{\tau_D \wedge T}) - h(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{x}} \int_0^{\tau_D \wedge T} \mathcal{A}h(\boldsymbol{X}_s)ds \leq -\mathbb{E}^{\boldsymbol{x}}(\tau_D \wedge T)$$

for any fixed $T > 0$. Since $|h(\boldsymbol{x})| \leq C$ for $\boldsymbol{x} \in D$, we have

$$\mathbb{E}^{\boldsymbol{x}}(\tau_D \wedge T) \leq 2C.$$

Taking $T \to \infty$ and using the monotone convergence theorem we obtain $\mathbb{E}^{\boldsymbol{x}}(\tau_D) \leq 2C$.

Homeworks

1. Derive the equations (2.11) and (2.12).

2. Derive the detailed balance condition for the multidimensional OU process:

$$dX_t = BX_t dt + \sigma dW_t$$

   if the invariant distribution has mean 0 and covariance matrix Σ.

References

[1] O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, 1997.

[2] B. Oksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, New York, Berlin, Heidelberg and Tokyo, 4th edition edition, 1998.

[3] K. Yosida. *Functional analysis*. Springer-Verlag, Berlin and Heidelberg, sixth edition edition, 1995.

Lecture 16 Numerical SDEs: Basics *

Tiejun Li

1 Schemes

As most of the SDEs can not be solved in analytical form, we should appeal to numerical computations for practical purpose. Below we illustrate the basic idea of constructing the numerical schemes for solving the SDEs

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t. \tag{1.1}$$

Most of the ideas can be extended to the multidimensional SDEs with coefficients involving t explicitly.

With Ito's formula, we define

$$df(X_t) = f'(X_t)dX_t + \frac{1}{2}f''(X_t)(dX_t)^2 = (L_1 f)(X_t)dt + (L_2 f)(X_t)dW_t, \tag{1.2}$$

where

$$(L_1 f)(x) = b(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x), \quad (L_2 f)(x) = \sigma(x)f'(x).$$

Taking integral from t_n to t_{n+1} to both sides of (1.1), and taking $f(x) = b(x)$ and $\sigma(x)$, we have

$$\begin{aligned}
X_{t_{n+1}} =& X_{t_n} + \int_{t_n}^{t_{n+1}} b(X_s)ds + \int_{t_n}^{t_{n+1}} \sigma(X_s)dW_s \\
=& X_{t_n} + b(X_{t_n})\delta t_n + \sigma(X_{t_n})(W_{t_{n+1}} - W_{t_n}) && (1.3)\\
&+ \int_{t_n}^{t_{n+1}} dW_s \int_{t_n}^{s} (L_2\sigma)(X_\tau)dW_\tau && (1.4)\\
&+ \int_{t_n}^{t_{n+1}} dW_s \int_{t_n}^{s} (L_1\sigma)(X_\tau)d\tau + \int_{t_n}^{t_{n+1}} ds \int_{t_n}^{s} (L_2 b)(X_\tau)dW_\tau && (1.5)\\
&+ \int_{t_n}^{t_{n+1}} ds \int_{t_n}^{s} (L_1 b)(X_\tau)d\tau, && (1.6)
\end{aligned}$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

where $\delta t_n = t_{n+1} - t_n$. The above procedure can be further carried on by replacing $L_i b(X_\tau)$, $L_i\sigma(X_\tau)$ as $L_i b(X_t)$, $L_i\sigma(X_t)$ and getting higher order iterative integrals correspondingly. The obtained series is usually called Itô-Taylor expansion for SDEs. It is not difficult to find that each term in the Ito-Taylor expansion has the form

$$I_{\boldsymbol{i}}(g) = \int_{t_n}^{t_{n+1}} dW_{s_1}^{i_1} \int_{t_n}^{s_1} dW_{s_2}^{i_2} \cdots \int_{t_n}^{s_{k-1}} dW_{s_k}^{i_k} g(X_{s_k})$$

with some $k \in \{1, 2, \ldots\}$. Here the characteristic index of the integral $\boldsymbol{i} = (i_1, i_2, \ldots, i_k)$ and $i_j \in \{0, 1\}$ for $j = 1, 2, \ldots, k$. The integrand g is the action of some compositions of operators L_1 and L_2 on function b or σ. We take the convention $W_t^0 := t$ and $W_t^1 := W_t$ for the ease of notation. This set-up can be extended to the system driven by multidimensional Brownian easily.

Now similar with solving deterministic ODEs, we truncate the Ito-Taylor series to different orders to obtain different schemes. For example, if we only keep terms until (1.3), then we have

(1) Euler-Maruyama scheme

$$X_{n+1} = X_n + b(X_n)\delta t_n + \sigma(X_n)\delta W_n, \tag{1.7}$$

where $\delta W_n \sim N(0, \delta t_n)$. The Euler-Maruyama scheme is the most commonly used numerical scheme for its simplicity.

From the basic intuition $dW_t \sim \sqrt{dt}$, we have that roughly (1.8)$\sim O(\delta t)$, (1.5)$\sim O(\delta t^{3/2})$ and (1.6)$\sim O(\delta t^2)$. By extracting the leading order term (1.8), we obtain

$$\begin{aligned}\int_{t_n}^{t_{n+1}} dW_s \int_{t_n}^{s} (L_2\sigma)(X_\tau) dW_\tau &\approx (L_2\sigma)(X_{t_n}) \int_{t_n}^{t_{n+1}} dW_s \int_{t_n}^{s} dW_\tau \\ &= \frac{1}{2}(L_2\sigma)(X_{t_n})[(\delta W_n)^2 - \delta t_n].\end{aligned}$$

Substitute this into the Ito-Taylor expansion we obtain the well-known Milstein scheme.

(2) Milstein scheme

$$X_{n+1} = X_n + b(X_n)\delta t_n + \sigma(X_n)\delta W_n + \frac{1}{2}(\sigma\sigma')(X_n)[(\delta W_n)^2 - \delta t_n]. \tag{1.8}$$

We should remark that although Milstein scheme is more accurate than the Euler-Maruyama scheme in some sense, it is only practical for the SDEs driven by single Wiener process. That is because the explicit characterization

$$\int_{t_n}^{t_{n+1}} \int_{t_n}^{s} dW_s dW_\tau = \frac{1}{2}[(\delta W_n)^2 - \delta t_n]$$

is only valid in one dimensional case. In multi-dimensions, when $i \neq j$ it is impossible to get an explicit sampling form of

$$\int_{t_n}^{t_{n+1}} \int_{t_n}^{s} dW_s^i dW_\tau^j,$$

where W_t^i, W_t^j are independent Wiener processes. Though some strategies are proposed to approximate the above random variables, they are not so common in practical applications. The readers may be refereed to [1] for more details.

Although Milstein scheme is more accurate, it is not so popular in practice for the term $\sigma'(X_t)$ may be too complicate to compute even in 1D case. To overcome this issue, one can take the following type of schemes by borrowing the idea from Runge-Kutta method for solving ODEs.

(3) Runge-Kutta scheme

$$\begin{aligned} \hat{X}_n =& X_n + \sigma(X_n)\sqrt{\delta t_n}, \\ X_{n+1} =& X_n + b(X_n)\delta t_n + \sigma(X_n)\delta W_n \\ &+ \frac{1}{2}\frac{1}{\sqrt{\delta t_n}}[\sigma(\hat{X}_n) - \sigma(X_n)][(\delta W_n)^2 - \delta t_n]. \end{aligned} \tag{1.9}$$

If we formally take higher order Itô-Taylor expansion, specifically applying formula (1.2) to

$$(L_1 b)(X_\tau),\ (L_2 b)(X_\tau),\ (L_1\sigma)(X_\tau),\ (L_2\sigma)(X_\tau)$$

and dropping higher order terms, we have the following higher order scheme.

(4) Higher order scheme

$$\begin{aligned} X_{n+1} =& X_n + b\delta t_n + \sigma\delta W_{t_n} + \frac{1}{2}\sigma\sigma'\{(\delta W_n)^2 - \delta t_n\} \\ &+ \sigma b'\Delta Z_n + \frac{1}{2}(bb' + \frac{1}{2}\sigma^2 b'')\delta t_n^2 \\ &+ (b\sigma' + \frac{1}{2}\sigma^2\sigma'')(\delta W_n \delta t_n - \delta Z_n) \\ &+ \frac{1}{2}\sigma\Big[\sigma\sigma'' + (\sigma')^2\Big]\Big[\frac{1}{3}(\delta W_n)^2 - \delta t_n\Big]\delta W_n, \end{aligned} \tag{1.10}$$

where

$$\Delta Z_n := \int_{t_n}^{t_{n+1}} \int_{t_n}^{s} dW_\tau ds$$

is a Gaussian R.V. satisfying $\mathbb{E}(\Delta Z_n) = 0,\ \mathbb{E}((\Delta Z_n)^2) = \delta t_n^3/3, \mathbb{E}(\delta Z_n \delta W_n) = \delta t_n^2/2$.

The convergence of the discretized solution of SDEs has two senses according to the needs of realistic applications. They are called *strong convergence* and *weak convergence*, respectively. Define $\{X_t^{\delta t}\}$ a numerical solution of SDEs with maximal stepsize δt, X_t is the exact solution, then we have the following definition.

Definition 1.1 (Convergence of numerical solutions)**.** *We have the following two typical concepts of convergence for the numerical solution of SDEs.*

(1) Strong convergence (mean-square convergence) *If*

$$\max_{0\le t\le T}\mathbb{E}|X_t^{\delta t}-X_t|^2\le C(\Delta t)^{2\alpha},$$

where C is a constant independent of δt, then we call $\{X_t^{\Delta t}\}$ strongly converges, or converges in the mean-square sense, to X_t with order α.

(2) Weak convergence (convergence w.r.t. expectation) *If*

$$\max_{0\le t\le T}|\mathbb{E}f(X_t^{\delta t})-\mathbb{E}f(X_t)|\le C_f(\delta t)^{\beta},$$

for any $f\in C_b^\infty(\mathbb{R}^n)$, where C_f is a constant independent of δt but may depend on f, then we call $\{X_t^{\delta t}\}$ weakly converges to X_t with order β.

A straightforward result about the convergence order is below.

Proposition 1.2. *When the considered function f in the weak convergence has the property $\|f'\|_\infty\le K$, we have $\beta\ge\alpha$.*

Proof. By the mean value theorem and the Hölder's inequality, we obtain

$$|\mathbb{E}f(X_t^{\delta t})-\mathbb{E}f(X_t)|\le\mathbb{E}|f(X_t^{\delta t})-f(X_t)|\le K\mathbb{E}|X_t^{\delta t}-X_t|\le K(\mathbb{E}|X_t^{\delta t}-X_t|^2)^{\frac{1}{2}}.$$

The above proposition gives a rationale why the former is called *strong convergence* compared with the other one in some sense. Before introducing the convergence analysis, let us state the main theorem about the convergence of numerical schemes

Theorem 1.3 (Convergence order)**.** *Define the length of the multi-index $\boldsymbol{i}=(i_1,i_2,\dots,i_k)$ as*

$$l(\boldsymbol{i}):=k,\quad n(\boldsymbol{i}):=\{\textit{the number of zeros in }\boldsymbol{i}\}.$$

and the set of indices

$$\mathcal{S}_\alpha=\left\{\boldsymbol{i}|\ l(\boldsymbol{i})+n(\boldsymbol{i})\le 2\alpha\quad or\quad l(\boldsymbol{i})=n(\boldsymbol{i})=\alpha+\frac{1}{2}\right\}\quad for\quad \alpha\in\left\{\frac{1}{2},1,\frac{3}{2},\cdots\right\},$$

$$\mathcal{W}_\beta=\{\boldsymbol{i}|l(\boldsymbol{i})\le\beta\}\qquad for\quad \beta\in\{1,2,3,\cdots\}.$$

Then with mild smoothness conditions on b, σ and the function f in weak approximation, the scheme derived by truncating the Ito-Taylor expansion up to all indices with $\boldsymbol{i}\in\mathcal{S}_\alpha$ has strong order α; the scheme derived by truncating the Ito-Taylor expansion up to terms with $\boldsymbol{i}\in\mathcal{W}_\beta$ has weak order β.

The proof and detailed requirements about the smoothness conditions on b, σ and f may be found in [1] (Theorems 10.6.3 and 14.5.1). Applying this theorem to the constructed schemes in this section, we have Table 1.

	Strong order	Weak order
Euler-Maruyama	1/2	1
Milstein	1	1
Scheme (1.9)	1	1
Scheme (1.10)	2	2

Table 1: The convergence order of some numerical schemes for SDE.

2 Strong convergence

We will analyze the mean-square convergence of the Euler-Maruyama scheme under the assumption that $b(x)$ satisfies global Lipschitz and linear grow condition with constant L and $\sigma = 1$, i.e. the additive noise case.

Now suppose the SDE takes the form

$$dX_t = b(X_t)dt + dW_t \tag{2.1}$$

with the Euler-Maruyama discretization

$$X_{n+1} = X_n + b(X_n)\delta t_n + \delta W_n. \tag{2.2}$$

Introduce the "linear stochastic" interpolation of X_n as

$$d\bar{X}_t = b(X_n)dt + dW_t, \quad t \in [t_n, t_{n+1}).$$

where the driving term W_t is assumed to be the same as that in continuous form. Then $\bar{X}_{t_n} = X_n$ and we have the so called "discrete Ito formula" for $f \in C^2(\mathbb{R})$

$$df(\bar{X}_t) = f'(\bar{X}_t)d\bar{X}_t + \frac{1}{2}f''(\bar{X}_t)(d\bar{X}_t)^2,$$

i.e.

$$f(\bar{X}_t) = f(X_n) + \int_{t_n}^{t} \Big[f'(\bar{X}_s)b(X_n) + \frac{1}{2}f''(\bar{X}_s)\Big]ds + \int_{t_n}^{t} f'(\bar{X}_s)dW_s, \quad t \in [t_n, t_{n+1}).$$

Lemma 2.1. *Let $\delta t = \max_n \delta t_n$. We have the following bounds for X_t*

$$\sup_{t\le T} \mathbb{E}|X_t|^2 \le K_1(T), \qquad \sup_{t\in[t_n,t_{n+1})} \mathbb{E}|X_t - X_{t_n}|^2 \le K_2(T)\delta t,$$

where the constant $K_1(T)$ depends on T, L and $\mathbb{E}|X_0|^2$, and $K_2(T)$ depends on $L, \delta t$ and $K_1(T)$.

Proof. Applying Ito formula to $|X_t|^2$, we have

$$d|X_t|^2 = 2X_t \cdot (b(X_t) + dW_t) + dt.$$

Integrating from 0 to t and taking expectation we have

$$\mathbb{E}|X_t|^2 = \mathbb{E}|X_0|^2 + 2\mathbb{E}\int_0^t X_s \cdot b(X_s)ds + 2\mathbb{E}\int_0^t X_s dW_s + t.$$

Taking advantage of (**??**) for the Ito integral and the inequality $2ab \leq a^2 + b^2$, we obtain

$$\mathbb{E}|X_t|^2 \leq \mathbb{E}|X_0|^2 + T + \int_0^t \mathbb{E}|X_s|^2 ds + L\int_0^t (1 + \mathbb{E}|X_s|^2)ds.$$

The Gronwall inequality gives

$$\sup_{t\leq T} \mathbb{E}|X_t|^2 \leq (\mathbb{E}|X_0|^2 + T + LT)\exp((L+1)T).$$

For the second inequality, we have from SDE

$$X_t - X_{t_n} = \int_{t_n}^t b(X_s)ds + (W_t - W_{t_n}).$$

Squaring both sides and taking expectation we get

$$\mathbb{E}|X_t - X_{t_n}|^2 \leq 2\mathbb{E}\left(\int_{t_n}^t b(X_s)ds\right)^2 + 2\delta t.$$

From Hölder's inequality we obtain

$$\mathbb{E}|X_t - X_{t_n}|^2 \leq 2L\delta t\int_{t_n}^t (1 + \mathbb{E}|X_s|^2)ds + 2\delta t \leq 2L\delta t^2(1 + K_1(T)) + 2\delta t, \quad t \in [t_n, t_{n+1}).$$

Proposition 2.2 (Half order mean-square convergence)**.** *The Euler-Maruyama scheme is of strong order 1/2.*

Proof. From (2.1) we have

$$X_{t_{n+1}} = X_{t_n} + \int_{t_n}^{t_{n+1}} b(X_t)dt + \delta W_n,$$

and the Equation (2.2) can be rewritten as

$$X_{n+1} = X_n + \int_{t_n}^{t_{n+1}} b(X_n)dt + \delta W_n.$$

Define the error $e_{n+1} = X_{t_{n+1}} - X_{n+1}$, then

$$e_{n+1} = e_n + \int_{t_n}^{t_{n+1}} (b(X_t) - b(X_n))dt.$$

Squaring both sides and from the inequality $2ab \le a^2\delta t + b^2/\delta t$, we obtain

$$\begin{aligned}
|e_{n+1}|^2 &= |e_n|^2 + \Big[\int_{t_n}^{t_{n+1}} (b(X_t) - b(X_n))dt\Big]^2 + 2e_n \cdot \Big[\int_{t_n}^{t_{n+1}} (b(X_t) - b(X_n))dt\Big] \\
&\le |e_n|^2(1+\delta t) + \left(1 + \frac{1}{\delta t}\right)\left[\int_{t_n}^{t_{n+1}} (b(X_t) - b(X_n))dt\right]^2 \\
&\le |e_n|^2(1+\delta t) + L^2(1+\delta t)\int_{t_n}^{t_{n+1}} |X_t - X_n|^2 dt, \qquad (2.3)
\end{aligned}$$

where the last inequality is from Hölder's inequality and Lipschitz condition.

From the inequality $|X_t - X_n|^2 \le 2|X_t - X_{t_n}|^2 + 2|X_{t_n} - X_n|^2$ we have

$$\mathbb{E}|e_{n+1}|^2 \le \mathbb{E}|e_n|^2(1 + L_1\delta t) + L_2\delta t^2,$$

where $L_1 = 1+2L^2(1+\delta t)$ and $L_2 = 2L^2(1+\delta t)K_2(T)$ can be bounded by positive constants independent of δt if δt is small.

The discrete Gronwall's inequality then guarantees

$$\mathbb{E}|e_n|^2 \le \mathbb{E}|e_0|^2(1 + L_1\delta t)^n + L_2\delta t^2 \frac{(1 + L_1\delta t)^n - 1}{L_1\delta t} \le \frac{L_2}{L_1}(e^{L_1 T} - 1)\delta t$$

if we assume $e_0 = 0$. The proof is complete.

We want to remark here that in the considered additive noise case, the Euler-Maruyama scheme is exactly the Milstein scheme since $\sigma'\sigma = 0$ and thus the last term in (1.8) diminishes! From Theorem 1.3, we can prove it is of strong order 1 in principle. It is indeed true but the proof will be more tedious with higher smoothness condition on b. We leave the proof as an exercise to the reader.

3 Weak Convergence

Now let us consider the weak convergence of the Euler-Maruyama scheme with the tools from PDEs. We will only consider the 1D case with $\sigma = 1$ for simplicity. But the essential part of the proof is the same for high dimensional case. The weak convergence is to analyze the error

$$e_n = \mathbb{E}f(X_n) - \mathbb{E}f(X_{t_n})$$

for smooth function f. From the stated result in Theorem 1.3, we know that the Euler-Maruyama scheme is of weak order 1. Before we go to the rigorous proof, let us give a more transparent observation on this point by elementary deductions.

Suppose $X_0 = x$, f is a smooth enough function. Formally in order to consider the weak convergence of a numerical scheme to approximate Markov process X_t, we start from the

weak Ito-Taylor expansion

$$\mathbb{E}^x f(X_h) = f(x) + \int_0^h \mathcal{A} f(X_t) dt \sim \sum_{n=0} \frac{\mathcal{A}^n f(x)}{n!} h^n, \tag{3.1}$$

where $\mathcal{A}$ is the infinitesimal generator of X_t and h is the time stepsize. Correspondingly for the numerical solution X_t^N, we have

$$\mathbb{E}^x f(X_h^N) = f(x) + \int_0^h \mathcal{A} f(X_t^N) dt \sim \sum_{n=0} \frac{\mathcal{A}_N^n f(x)}{n!} h^n, \tag{3.2}$$

where $\mathcal{A}_N$ is the infinitesimal generator of X_t^N. To have an idea about the global weak convergence order, we need to figure out the local weak truncation order at first.

Take the diffusion process as a specific example. Now

$$dX_t = b(X_t)dt + dW_t$$

and the Euler-Maruyama scheme reads

$$X_{n+1} = X_n + b(X_n)\Delta t + \Delta W_n.$$

Define the continuous extension of the numerical solution as

$$dX_t^N = b(X_n)dt + dW_t, \quad t \in [t_n.t_{n+1}).$$

We have the infinitesimal generator

$$\mathcal{A} f(y) = b(y) f'(y) + \frac{1}{2} f''(y)$$

and

$$\mathcal{A}^2 f(y) = b(y)\Big[b(y)f'(y) + \frac{1}{2}f''(y)\Big] + \frac{1}{2}\Big[b(y)f'(y) + \frac{1}{2}f''(y)\Big]''.$$

Correspondingly for the numerical solution X_t^N we have

$$\mathcal{A}_N f(y) = b(x) f'(y) + \frac{1}{2} f''(y)$$

and

$$\mathcal{A}_N^2 f(y) = b(x)\Big[b(x)f'(y) + \frac{1}{2}f''(y)\Big] + \frac{1}{2}\Big[b(x)f'(y) + \frac{1}{2}f''(y)\Big]'',$$

where x is the initial condition. Now it is obvious that

$$\mathbb{E}^x f(X_h^N) - \mathbb{E}^x f(X_h) = O(h^2) \tag{3.3}$$

and thus the weak local truncation error is of second order and we can expect that the Euler-Maruyama scheme is of weak order 1. In fact if one can bound the expansion terms in

(3.1) and (3.2) up to corresponding order, the above formal derivations based on the local error analysis can be made rigorous.

It will become clear soon that the weak convergence analysis essentially relies on some estimates about the solution of the backward equation. Let us first consider the partial differential equation

$$\partial_t u(x,t) = \mathcal{L}u(x,t) = b(x)\partial_x u + \frac{1}{2}\partial_{xx} u, \quad u(x,0) = f(x). \tag{3.4}$$

Define the notation $C_P^m(\mathbb{R}^d, \mathbb{R})$ the space of functions $w \in C^m(\mathbb{R}^d, \mathbb{R})$ for which all partial derivatives up to order m have polynomial growth. More concretely, there exist a constant $K > 0$, and $m, p \in \mathbb{N}$ such that

$$|\partial_{\boldsymbol{x}}^{\boldsymbol{j}} w(\boldsymbol{x})| \leq K(1 + |\boldsymbol{x}|^{2p}), \quad \forall\ |\boldsymbol{j}| < m$$

for any $\boldsymbol{x} \in \mathbb{R}^d$, where $\boldsymbol{j}$ is a d-multi-index. Here we have $d = 1$ and we will simply denote $C_P^m(\mathbb{R}^d, \mathbb{R})$ as C_P^m in later texts.

The following important lemma can be found in [1] (Theorem 4.8.6, pp. 153).

Lemma 3.1. *Suppose that $f \in C_P^{2\beta}$ for some $\beta \in \{2, 3, \ldots\}$, $\boldsymbol{X}_t$ is time-homogeneous and $b \in C_P^{2\beta}$ with uniformly bounded derivatives. Then $\partial u/\partial t$ is continuous and*

$$u(\cdot, t) \in C_P^{2\beta}, \quad t \leq T$$

for any fixed $T < \infty$.

Theorem 3.2 (Weak convergence)**.** *Assume that b is Lipschitz and the conditions in Lemma 3.1 also hold for b and f, then the Euler-Maruyama scheme is of weak order 1 .*

Proof. Define the backward operator

$$\tilde{\mathcal{L}} = \partial_t + b(x)\partial_x + \frac{1}{2}\partial_{xx},$$

and denote by v the solution of

$$\tilde{\mathcal{L}} v = 0, \quad t \in (0, t_n) \tag{3.5}$$

with the final condition $v(x, t_n) = f(x)$. It is straightforward that $v(x,t) = u(x, t_n - t)$ for the solution u of (3.4).

By Itô's formula we have

$$\mathbb{E}v(X_0, 0) = \mathbb{E}v(X_{t_n}, t_n) = \mathbb{E}f(X_{t_n}).$$

Hence

$$\begin{aligned}|e_n| &= |\mathbb{E}f(X_n) - \mathbb{E}f(X_{t_n})| \\ &= |\mathbb{E}v(X_n, t_n) - \mathbb{E}v(X_0, 0)| \\ &= \Big|\mathbb{E}\Big(\int_0^{t_n} \big(\partial_t v(\bar{X}_s, s) + b(X_{n_s})\partial_x v(\bar{X}_s, s) + \frac{1}{2}\partial_{xx} v(\bar{X}_s, s) - \tilde{\mathcal{L}} v(\bar{X}_s, s)\big) ds\Big)\Big|,\end{aligned}$$

where $n_s := \{m | t_m \le s < t_{m+1}\}$, and $\bar{X}_s$ is the continuous extension of X_n defined as

$$d\bar{X}_s = b(X_{n_s})ds + dW_s, \quad x \in [t_m, t_{m+1}).$$

With this definition, we obtain

$$\begin{aligned}|e_n| &= \Big|\mathbb{E}\Big(\int_0^{t_n} \big(b(X_{n_s})\partial_x v(\bar{X}_s, s) - b(\bar{X}_s)\partial_x v(s, \bar{X}_s)\big) ds\Big)\Big| \\ &\le \Big|\mathbb{E}\Big(\int_0^{t_n} \big(b(X_{n_s})\partial_x v(X_{n_s}, t_{n_s}) - b(\bar{X}_s)\partial_x v(\bar{X}_s, s)\big) ds\Big)\Big| \\ &+ \Big|\mathbb{E}\Big(\int_0^{t_n} b(X_{n_s})\big(\partial_x v(\bar{X}_s, s) - \partial_x v(X_{n_s}, t_{n_s})\big) ds\Big)\Big| \\ &= \Big|\mathbb{E}\sum_m \int_{t_m}^{t_{m+1}} \Big(b(X_m)\partial_x v(X_m, t_m) - b(\bar{X}_s)\partial_x v(\bar{X}_s, s)\Big) ds\Big| \\ &+ \Big|\mathbb{E}\sum_m \int_{t_m}^{t_{m+1}} b(X_m)\big(\partial_x v(\bar{X}_s, s) - \partial_x v(X_m, t_m)\big) ds\Big|. \qquad (3.6)\end{aligned}$$

Using Itô's formula again, we have for any function $g(x, t)$

$$\begin{aligned}g(\bar{X}_t, t) - g(X_m, t_m) &= \int_{t_m}^t \Big[\partial_t g(\bar{X}_s, s) + b(X_m)\partial_x g(\bar{X}_s, s) + \frac{1}{2}\partial_{xx} g(\bar{X}_s, s)\Big] ds \\ &+ \int_{t_m}^t \partial_x g(\bar{X}_s, s) dW_s, \quad t \in [t_m, t_{m+1}).\end{aligned}$$

Using this with $g = b\partial_x v$ and $g = \partial_x v$ in (3.6), we have the highest derivatives $\partial_{xxx} v, \partial_{xx} b \in C_P^{2\beta}$ as long as $\beta \ge 2$. Notice that $b(X_m)$ is independent of $\int_{t_m}^t \partial_x g(\bar{X}_s, s) dW_s$ conditional on X_m, together with the fact that $\mathbb{E}|X_m|^{2r}$ and $\mathbb{E}|\bar{X}_t|^{2r} \le C$ for any $r \in \mathbb{N}$ (Exercise 3), we get

$$|e_n| \le C\sum_m \Delta t^2 \le C\Delta t,$$

which is the desired estimate. □

Example 3.3 (Weak approximation)**.** *For the SDE*

$$dX_t = -\frac{1}{2}X_t dt + dW_t, \quad X_0 = 0,$$

compute $u = \mathbb{E}X_t^2|_{t=1}$ *with the Euler-Maruyama scheme.*

Solution. The exact solution of u is

$$u = \mathbb{E}X_t^2|_{t=1} = 1 - e^{-1} \approx 0.632.$$

In order to compute the expectation numerically, we take the Euler-Maruyama scheme

$$X_{n+1,k} = (1 - \frac{\Delta t}{2})X_{n,k} + \sqrt{\Delta t} \cdot R_{n,k}, \quad k = 1, 2, \ldots, N,$$

where $\Delta t = 1/M$, $n = 0, 1, \ldots, M-1$ and $R_{n,k}$ are i.i.d. $N(0,1)$ random variables. So the approximate solution

$$u_{N,\Delta t} = \frac{1}{N}\sum_{k=1}^{N}(X_{M,k})^2.$$

We take $M = 2000$, and compute u with different sample size N as follows.

N	100	200	300	400	500	600
u	0.6586	0.6563	0.6785	0.6234	0.6407	0.6320
Error	0.0265	0.0242	0.0464	0.0087	0.0086	0.0001

Table 2: Weak approximation with Euler-Maruyama scheme

Homeworks

1. Give a sampling method for the random variables

$$\Delta Z_1 := \int_{t_n}^{t_{n+1}} \int_{t_n}^{s} dW_\tau ds, \quad \Delta Z_2 := \int_{t_n}^{t_{n+1}} \int_{t_n}^{s} d\tau dW_s.$$

   and ΔW_n.

2. Prove the Euler-Maruyama scheme is of strong order 1 for the SDE (2.1) with additive noise and higher smoothness condition on b.

3. Prove that for the Euler-Maruyama scheme

$$\mathbb{E}|X_n|^{2r},\ \mathbb{E}|X_t|^{2r},\ \mathbb{E}|\bar{X}_t|^{2r} \le C$$

   for $t \le T$, $n \le N$ and any $r \in \mathbb{N}$.

References

[1] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Springer-Verlag, 1992.

Lecture 17 Numerical SDEs: Advanced topics *

Tiejun Li

1 Implicit scheme

To overcome the stiffness issue, one can also apply implicit schemes, e.g. simplest implicit Euler:

$$X_{n+1} = X_n + b(X_{n+1})\delta t_n + \sigma(X_n)\delta W_n$$

or semi-implicit scheme

$$X_{n+1} = X_n + \Big[\alpha b(X_n) + (1-\alpha)b(X_{n+1})\Big]\delta t_n + \sigma(X_n)\delta W_n$$

for $\alpha \in (0,1)$.

The fully implicit scheme is also considered but not very successful although one can transform the Ito SDE form into right-most endpoint form at first. For example

$$X_{n+1} = X_n + \Big[b(X_{n+1}) - c(X_{n+1})\Big]\delta t_n + \sigma(X_{n+1})\delta W_n$$

where

$$c_i(x) = \sum_{jk} \frac{\partial \sigma_{ij}}{\partial x_k}\sigma_{kj}$$

is from the transformation. If $b = 0, \sigma(x) = x$, the above scheme implies

$$X_{n+1} = \frac{X_n}{1 - \delta W_n}$$

It is possible that $1 - \delta W_n = 0$ and indeed $\mathbb{E}|X_{n+1}| = \infty$!

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

2 Extrapolation method

Talay and Tubaro proposed the following extrapolation method based on the error expansion:

$$\begin{aligned} e(\delta) &= \mathbb{E}g(X_T^\delta) - \mathbb{E}g(X_T) = C_{g,\beta}\delta^\beta + C_{g,\beta+1}\delta^{\beta+1} \\ e\left(\frac{\delta}{2}\right) &= \mathbb{E}g(X_T^{\frac{\delta}{2}}) - \mathbb{E}g(X_T) = C_{g,\beta}(\frac{\delta}{2})^\beta + C_{g,\beta+1}(\frac{\delta}{2})^{\beta+1} \\ 2^{-\beta}e(\delta) - e(\frac{\delta}{2}) &= \mathbb{E}g(X_T^\delta) - \mathbb{E}g(X_T) = \tilde{C}_{g,\beta+1}\delta^{\beta+1} \end{aligned}$$

See details in Stoch. Anal. Appl. 8 (1990), 483-509.

3 Multilevel Monte Carlo method

So far we only considered the bias error of the approximation, i.e. the error brought by the time discretization. But a real approximation also involves Monte Carlo samplings. Since 2008, M. Giles proposed the general framework of multilevel Monte Carlo methods for SDEs [1], which approximates the expectation in an efficient way. This method stimulates a lot of follow-up works in different fields [2].

We have already known that the Euler-Maruyama scheme is of weak order 1 in computing $Y_E = \mathbb{E}f(X_T)$ for the SDE

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t$$

on $[0, T]$. In real computations, we take the weak approximator

$$Y_{h,N} = \frac{1}{N}\sum_{k=1}^{N} f(X_n^{(k)}), \qquad n = T/h \in \mathbb{N} \tag{3.1}$$

with stepsize h and N independent samples, where X_n is obtained by the Euler-Maruyama scheme. The mean square error has the bias-variance decomposition

$$\begin{aligned} \text{MSE} = \mathbb{E}(Y_E - Y_{h,N})^2 &\le 2|Y_E - \mathbb{E}f(X_n)|^2 + 2\mathbb{E}|\mathbb{E}f(X_n) - Y_{h,N}|^2 \\ &\le C_1h^2 + C_2N^{-1}. \end{aligned} \tag{3.2}$$

by the weak order 1 convergence and Monte Carlo estimate.

The above computation has the cost C_3Nh^{-1}. The cost-accuracy tradeoff

$$\min_{h,N} \text{MSE} \qquad \text{subject to a given cost } K = C_3Nh^{-1} \gg 1$$

gives the optimal choice

$$N \sim O(Kh), \quad h \sim O(K^{-\frac{1}{3}}) \quad \text{and} \quad \text{MSE} \sim O(K^{-\frac{2}{3}}). \tag{3.3}$$

This means that if we require the accuracy MSE $\sim O(\varepsilon^2)$, we must have $h \sim O(\varepsilon)$, $N \sim O(\varepsilon^{-2})$ and thus the cost $K \sim O(\varepsilon^{-3})$. The multilevel Monte Carlo method achieves the same accuracy with cost $K \sim O(\varepsilon^{-2}(\ln \varepsilon)^2)$, which is a typical fast algorithm.

The construction of multilevel Monte Carlo method is as follows. Define the L-level grids with time stepsize $h_l = M^{-l}T$ for $l = 0, 1, \ldots, L$. Denote by $F_l = f(X_{l,M^l})$ the approximation of $f(X_T)$ at the level l, where X_{l,M^l} is the approximation of X_T with stepsize h_l. We have

$$\mathbb{E}F_L = \sum_{l=0}^{L} \mathbb{E}(F_l - F_{l-1}) \qquad \text{where } F_{-1} := 0. \tag{3.4}$$

Take N_l realizations for each summand in (3.4), and define

$$Y_l = \frac{1}{N_l} \sum_{k=1}^{N_l} \big(F_l^{(k)} - F_{l-1}^{(k)}\big), \qquad l = 0, 1, \ldots, L.$$

Correspondingly define the final estimator

$$\hat{Y}_L = \sum_{l=0}^{L} Y_l. \tag{3.5}$$

From Monte Carlo estimate we have $\mathrm{var}(Y_l) = V_l/N_l$, where $V_l := \mathrm{var}(F_l - F_{l-1})$ for $l = 0, 1, \ldots, L$. With independent sampling in (3.5), we get

$$\mathrm{var}(\hat{Y}_L) = \sum_{l=0}^{L} \mathrm{var}(Y_l) = \sum_{l=0}^{L} \frac{V_l}{N_l} \tag{3.6}$$

with computational cost

$$K \sim O\left(\sum_{l=0}^{L} N_l h_l^{-1}\right).$$

The key point of multilevel Monte Carlo is that with the decomposition (3.4), the term $F_l - F_{l-1}$ has smaller fluctuations, i.e. smaller variance, at higher levels provided that the realizations of $F_l - F_{l-1}$ come from two discrete approximations with different time stepsizes but same Brownian paths. This property suggests that we can use less Monte Carlo simulations for higher levels, i.e. finer grids, but more simulations for lower levels, i.e. coarser grids. This cost-accuracy tradeoff is the origin of the efficiency of multilevel Monte Carlo method.

Now let us consider the minimization

$$\min_{N_l} \mathrm{var}(\hat{Y}_L) = \sum_{l=0}^{L} \frac{V_l}{N_l} \qquad \text{subject to the cost } K = \sum_{l=0}^{L} N_l h_l^{-1} \gg 1.$$

This is generally a very difficult problem so we relax N_l to be continuous. Upon introducing Lagrange multiplier we get the minimizer

$$N_l = \lambda\sqrt{V_l h_l}, \qquad \text{where} \quad \lambda = K\left(\sum_{l=0}^{L}\sqrt{V_l h_l^{-1}}\right)^{-1}. \tag{3.7}$$

From the strong and weak convergence result of Euler-Maruyama Scheme, we have

$$|\mathbb{E}(F_l) - Y_E| = O(h_l), \quad \mathbb{E}|X_T - X_{l,M^l}|^2 = O(h_l).$$

By assuming the Lipschitz continuity of f, we obtain

$$\mathrm{var}(F_l - f(X_T)) \le \mathbb{E}|f(X_{l,M^l}) - f(X_T)|^2 \le C\mathbb{E}|X_T - X_{l,M^l}|^2 = O(h_l)$$

and thus

$$V_l = \mathrm{var}(F_l - F_{l-1}) \le 2\mathrm{var}(F_l - f(X_T)) + 2\mathrm{var}(F_{l-1} - f(X_T)) = O(h_l)$$

since $h_{l-1} = Mh_l$ and $M \sim O(1)$.

For a given tolerance $\varepsilon \ll 1$, take

$$N_l = O(\varepsilon^{-2}Lh_l), \tag{3.8}$$

according to the optimal choice (3.7), we get the variance estimate

$$\mathrm{var}(\hat{Y}_L) = O(\varepsilon^2). \tag{3.9}$$

from (3.6). Further take $L = \ln\varepsilon^{-1}/\ln M$, we have

$$h_L = M^{-L} = O(\varepsilon).$$

So the bias error

$$|\mathbb{E}F_L - Y_E| = O(h_L) = O(\varepsilon). \tag{3.10}$$

Combing (3.9) and (3.10), we obtain the overall mean square error

$$\mathrm{MSE} = \mathbb{E}(Y_E - \hat{Y}_L)^2 = O(\varepsilon^2)$$

and the computational complexity

$$K = \sum_{l=0}^{L} N_l h_l^{-1} = O(\varepsilon^{-2}L^2) = O\left(\varepsilon^{-2}(\ln\varepsilon)^2)\right).$$

The optimal choice of M can be made by minimizing the prefactor in the estimate of the computational cost [1].

References

[1] M Giles. Multilevel monte carlo path simulation. *Operations Research*, 56:607–617, 2008.

[2] M Giles. Multilevel monte carlo methods. In J Dick, FY Kuo, JW Peters, and IH Sloan, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 79–98. Springer-Verlag, Heidelberg, 2014.

Lecture 18 Path integral *

Tiejun Li

1 Wiener Measure

The path integral, which can be dated to R. Feynman to construct a new formulation to understand quantum mechanics [2], gives very powerful formal approach to deal with the probability measures on path space and compute the expectation for some functionals of Wiener paths. Briefly speaking, path integral is a formal infinite dimensional limit of the considered stochastic process under finite dimensional approximations. Let us start with the formal representation of the Wiener measure P_* defined on the canonical space $(C[0,1], \mathcal{B}(C[0,1]))$ for the standard Wiener process.

From the definition of Wiener process, we have the joint pdf for $(W_{t_1}, W_{t_2}, \ldots, W_{t_n})$

$$p_n(w_1, w_2, \ldots, w_n) = \frac{1}{Z_n} \exp(-I_n(w)),$$

where $0 < t_1 < t_2 < \cdots < t_n \leq 1$ and

$$Z_n = (2\pi)^{\frac{n}{2}} \big[t_1(t_2 - t_1) \cdots (t_n - t_{n-1})\big]^{\frac{1}{2}},$$

$$I_n(w) = \frac{1}{2} \sum_{j=1}^{n} \Big(\frac{w_j - w_{j-1}}{t_j - t_{j-1}}\Big)^2 (t_j - t_{j-1}), \quad t_0 := 0, w_0 := 0.$$

Now we take the formal limit as $n \to \infty$, we obtain

$$p_n dw_1 dw_2 \cdots dw_n \to \frac{1}{Z} \exp(-I[w]) \delta(w_0) \mathcal{D}w, \tag{1.1}$$

where the δ-function $\delta(w_0)$ is to fix $w_0 = 0$, $I[w]$ is called the *action functional* of the Wiener process defined as

$$I[w] = \frac{1}{2} \int_0^1 \dot{w_t}^2 dt.$$

$\mathcal{D}w$ is a shortcut for $\prod_{0 \leq t \leq 1} dw_t$, which is the formal volume element in the path space $C[0,1]$. Z is the normalization factor. For notations, we use the lowercase w_t for dumb

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

variables, but the uppercase W_t for the stochastic process. This convention will be taken in this whole chapter.

To give a formal understanding on the Wiener measure (1.1), we note that

$$Z_n = \Big(\frac{2\pi}{n}\Big)^{\frac{n}{2}} \to 0$$

if $t_j - t_{j-1} = 1/n$. At the same time we have $\int_0^1 \dot{w}_t{}^2 dt \to +\infty$ because W_t is almost surely "half order" differentiable. This means $\exp(-\int_0^1 \dot{w}_t{}^2 dt) \to 0$ as the subdivision is infinitely refined. These two infinitesimals balance each other in the limit process and leads to a nontrivial limit which is the volume element in the path space $C[0,1]$. With this understanding,

$$\frac{1}{Z}\exp(-I[w])\delta(w_0) = \frac{\mathcal{D}P_*}{\mathcal{D}w}$$

may be thought of as the pdf of the Wiener process in the space $C[0,1]$. The probability of the event $\{W_\cdot \in A\}$, where $A \in \mathcal{B}(C[0,1])$, can be obtained as

$$\mathbb{P}(W_\cdot \in A) = \int_A \frac{1}{Z}\exp(-I[w])\delta(w_0)\mathcal{D}w.$$

We should emphasize that this interpretation is purely formal and all of the results induced by the path integral need to be reproved in rigorous mathematical language before we want to use it as an theorem. One reason to understand it is only formal is that we have no infinite dimensional Lebesgue measure [1]. To see this, let us consider an infinite dimensional Hilbert space H with orthonormal basis $\{e_1, e_2, \ldots\}$. Define the balls

$$B_n = B_{\frac{1}{2}}(e_n) = \{x | \|x - e_n\| \le 1/2\}, \quad B = B_2(0) = \{x | \|x\| \le 2\}.$$

As a Lebesgue measure, it should be translation invariant and finite for bounded sets. If the Lebesgue measure on H exists as $\mu(\cdot)$, then we have

$$0 < \mu(B_1) = \mu(B_2) = \cdots = \mu(B_n) = \cdots < \infty, \quad 0 < \mu(B) < \infty.$$

However from the disjointness of $\{B_n\}$ and $B_n \subset B$ for any n, we obtain

$$\mu(B) \ge \sum_n \mu(B_n) = \infty,$$

which is a contradiction! Thus the notation $\mathcal{D}w$ is totally meaningless! But the glamor of path integral is that it can give some extremely insightful results in a very efficient way. That is why it is also useful for applied mathematicians.

2 Expectation of a Wiener Functional

Example 2.1. *Compute the expectation*

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big).$$

Solution. Note that it is not straightforward to compute this expectation since the integrand involves the whole Wiener path, i.e. a Wiener functional. From the Karhunen-Loeve expansion,

$$\begin{aligned}\int_0^1 W_t^2 dt &= \int_0^1 \sum_{k,l} \sqrt{\lambda_k\lambda_l}\alpha_k\alpha_l\phi_k(t)\phi_l(t)dt \\ &= \sum_k \int_0^1 \lambda_k\alpha_k^2\phi_k^2(t)dt = \sum_k \lambda_k\alpha_k^2.\end{aligned}$$

Then

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) = \mathbb{E}\Big(\prod_k \exp(-\frac{1}{2}\lambda_k\alpha_k^2)\Big) = \prod_k \mathbb{E}\exp(-\frac{1}{2}\lambda_k\alpha_k^2).$$

From the identity

$$\mathbb{E}\exp(-\frac{1}{2}\lambda_k\alpha_k^2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot e^{-\frac{1}{2}\lambda_k x^2} dx = \sqrt{\frac{1}{1+\lambda_k}}$$

we obtain

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) = \prod_k \sqrt{\frac{1}{1+\lambda_k}} := M,$$

where

$$M^{-2} = \prod_{k=1}^{\infty}\Big(1+\frac{4}{(2k-1)^2\pi^2}\Big).$$

From the identities for infinite product series we have

$$\cosh(x) = \prod_{n=1}^{\infty}\Big(1+\frac{4x^2}{(2n-1)^2\pi^2}\Big),$$

where $\cosh(x) = (e^x + e^{-x})/2$. Thus

$$M = (\cosh(1))^{-\frac{1}{2}} = \sqrt{\frac{2e}{1+e^2}}.$$

Here we show how to apply the path integral approach to compute the expectation of this Wiener functional. The path integral approach to compute the expectation is composed of the following two steps.

Step 1. Discretize the problem into finite dimensions.

At first let us take finite dimensional approximation to the functional

$$\exp\left(-\frac{1}{2}\int_0^1 W_t^2 dt\right) \approx \exp\Big(-\frac{1}{2}\sum_{j=1}^n W_{t_j}^2 \Delta t\Big) = \exp\left(-\frac{1}{2}\Delta t X^T A X\right),$$

where $\Delta t = t_j - t_{j-1}$ for $j = 1, 2, \ldots, n$, $A = I$, and $X = (W_{t_1}, W_{t_2}, \cdots, W_{t_n})^T$. Thus

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) \approx \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\Delta t \boldsymbol{x}^T A \boldsymbol{x}\right) \cdot \frac{1}{Z_n} \exp\left(-\frac{1}{2}\Delta t \boldsymbol{x}^T B \boldsymbol{x}\right) d\boldsymbol{x}, \qquad (2.1)$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, $Z_n = (2\pi)^{\frac{n}{2}} (\det(\Delta t B)^{-1})^{\frac{1}{2}}$, and

$$B = \frac{1}{\Delta t^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

From equation (2.1), we have

$$\begin{aligned}\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) &\approx \frac{(2\pi)^{\frac{n}{2}} (\det(\Delta t(A+B))^{-1})^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}} (\det(\Delta t B)^{-1})^{\frac{1}{2}}} \\ &= \left(\frac{\det(B)}{\det(A+B)}\right)^{\frac{1}{2}} = \left(\frac{\prod_i \lambda_i^B}{\prod_i \lambda_i^{A+B}}\right)^{\frac{1}{2}},\end{aligned}$$

where $\lambda_i^B, \lambda_i^{A+B}$ are eigenvalues of B and $A+B$, respectively.

Step 2. Take the formal limit as $n \to \infty$.

If we take the formal limit as $n \to +\infty$, the matrix B will converge to the differential operator $\mathcal{B} = -d^2/dt^2$ with zero Dirichlet boundary condition at $t = 0$ and free Neumann boundary condition at $t = 1$. Thus the eigenvalues of $\mathcal{B}$ corresponds to the following Sturm-Liouville boundary value problem

$$-\frac{d^2 u}{dt^2} = \lambda u(t), \quad u(0) = 0, u'(1) = 0.$$

With the observation

$$\int_0^1 W_t^2 dt = (\mathcal{A} W_t, W_t),$$

where $\mathcal{A} = I$ and $(f, g) := \int_0^1 fg dt$, we have the formal path integral limit

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) = \int \exp\Big(-\frac{1}{2}(\mathcal{A} w_t, w_t)\Big) \cdot \frac{1}{Z} \exp\Big(-\frac{1}{2}(\mathcal{B} w_t, w_t)\Big) \delta(w_0) \mathcal{D} w$$

where the operator $\mathcal{B}u(t) := d^2u/dt^2$ and

$$Z = \int \exp\Big(-\frac{1}{2}(\mathcal{B}w_t, w_t)\Big)\delta(w_0)\mathcal{D}w.$$

Now we formally apply the Gaussian integrals in infinite dimensions to obtain

$$\mathbb{E}\exp\Big(-\frac{1}{2}\int_0^1 W_t^2 dt\Big) = \Big(\frac{\det \mathcal{B}}{\det(\mathcal{A}+\mathcal{B})}\Big)^{\frac{1}{2}},$$

where $\det \mathcal{B},\ \det(\mathcal{A}+\mathcal{B})$ mean the products of all eigenvalues for the following boundary value problems:

$$\begin{cases} \mathcal{B}u = \lambda u, & \text{or} \quad (\mathcal{A}+\mathcal{B})u = \lambda u, \\ u(0) = 0, & u'(1) = 0. \end{cases}$$

This yield the same result as before.

3 Girsanov Transformation

We have seen that the Wiener measure over $[0,1]$ can be formally expressed as

$$d\mu_W = Z^{-1}\exp\left(-\frac{1}{2}\int_0^1 \dot{w}_t^2 dt\right)\delta(w_0)\mathcal{D}w.$$

The solution of the SDE

$$dX_t = b(X_t,t) + \sigma(X_t,t)dW_t, \quad X_0 = 0.$$

can be viewed as a map between the Wiener path $\{W_t\}$ and $\{X_t\}$:

$$\{W_t\} \xrightarrow{\Phi} \{X_t\}.$$

Consequently, the mapping Φ induces another measures on $C[0,1]$, which is nothing but the distribution of $\{X_t\}$.

We now ask the question how the measure $d\mu_W$ changes under the mapping Φ? Let us first consider the case when $\sigma = 1$ in one dimension. The more general conditions can be derived in a similar way. We will perform the path integral through two steps as in the previous section: that is, making discretization first and then taking the formal continuum limit.

Step 1. Discretize the problem into finite dimensions.

With the Euler-Maruyama discretization, we obtain

$$X_{t_{j+1}} = X_{t_j} + b(X_{t_j}, t_j)(t_{j+1} - t_j) + (W_{t_{j+1}} - W_{t_j}). \tag{3.1}$$

In matrix form we have

$$B \cdot \begin{pmatrix} X_{t_1} \\ X_{t_2} \\ \vdots \\ X_{t_n} \end{pmatrix} - \begin{pmatrix} b(X_{t_0}, t_0)(t_1 - t_0) \\ b(X_{t_1}, t_1)(t_2 - t_1) \\ \vdots \\ b(X_{t_{n-1}}, t_{n-1})(t_n - t_{n-1}) \end{pmatrix} = B \cdot \begin{pmatrix} W_{t_1} \\ W_{t_2} \\ \vdots \\ W_{t_n} \end{pmatrix},$$

where $t_0 = 0$, $X_{t_0} = 0$, and the matrix B has the form

$$B = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}_{n \times n}.$$

The equation (3.1) indeed introduces a finite dimensional transformation Φ_n as

$$\{W_{t_1}, W_{t_2}, \cdots, W_{t_n}\} \xrightarrow{\Phi_n} \{X_{t_1}, X_{t_2}, \cdots, X_{t_n}\}.$$

With dumb variables representation for (3.1), we have

$$x_{j+1} = x_j + b(x_j, t_j)(t_{j+1} - t_j) + (w_{j+1} - w_j), \quad j = 0, \ldots, n-1 \tag{3.2}$$

where $w_0 = 0$ and x_0 is fixed. It is not difficult to find that the Jacobian of the transformation

$$\frac{\partial(w_1, \ldots, w_n)}{\partial(x_1, \ldots, x_n)} = 1. \tag{3.3}$$

Suppose we want to compute the average $\langle F[X_t] \rangle$, then

$$\langle F[X_t] \rangle \approx \langle F(X_{t_1}, X_{t_2}, \cdots, X_{t_n}) \rangle = \langle G(W_{t_1}, W_{t_2}, \cdots, W_{t_n}) \rangle,$$

where $G = F \circ \Phi_n$. Furthermore with transformation of variables

$$\begin{aligned} \langle F[X_t] \rangle &\approx \int G(w_1, w_2, \cdots, w_n) \frac{1}{Z_n} \exp(-I_n(\boldsymbol{w})) dw_1 dw_2 \cdots dw_n \\ &= \int F(x_1, x_2, \cdots, x_n) \frac{1}{Z_n} \exp(-\tilde{I}_n(\boldsymbol{x})) dx_1 dx_2 \cdots dx_n, \end{aligned} \tag{3.4}$$

where the transformation holds because of (3.3), and $\tilde{I}_n(\boldsymbol{x}) = I_n \circ \Phi_n^{-1}(\boldsymbol{x})$ by definition (3.2)

$$\begin{aligned} \tilde{I}_n(\boldsymbol{x}) =& \frac{1}{2} \sum_{j=1}^{n} \left(\frac{x_j - x_{j-1}}{t_j - t_{j-1}} \right)^2 (t_j - t_{j-1}) + \frac{1}{2} \sum_{j=1}^{n} b^2(x_{j-1}, t_{j-1})(t_j - t_{j-1}) \\ & - \sum_{j=1}^{n} (x_j - x_{j-1}) \cdot b(x_{j-1}, t_{j-1}). \end{aligned}$$

Changing the dumb variables x_i to w_i, we obtain

$$\begin{aligned}\langle F[X_t]\rangle &\approx \int F(w_1,w_2,\cdots,w_n)\frac{1}{Z_n}\exp(-I_n(\boldsymbol{w}))\exp\Big(-\frac{1}{2}\sum_{j=1}^n b^2(w_{j-1},t_{j-1})(t_j-t_{j-1})\Big)\\ &\quad\cdot\exp\Big(\sum_{j=1}^n b(w_{j-1},t_{j-1})\cdot(w_j-w_{j-1})\Big)dw_1dw_2\cdots dw_n\\ &= \Big\langle F(W_{t_1},W_{t_2},\cdots,W_{t_n})\exp\Big(-\frac{1}{2}\sum_{j=1}^n b^2(W_{t_{j-1}},t_{j-1})(t_j-t_{j-1})\Big)\\ &\quad\cdot\exp\Big(\sum_{j=1}^n b(W_{t_{j-1}},t_{j-1})\cdot(W_{t_j}-W_{t_{j-1}})\Big)\Big\rangle.\end{aligned}$$

Step 2. Take the formal limit as $n\to\infty$.

Now with the finite dimensional discretization, we can take formal continuum limit

$$\langle F[X_t]\rangle = \Big\langle F[W_t]\exp\Big(-\frac{1}{2}\int_0^1 b^2(W_t,t)dt+\int_0^1 b(W_t,t)dW_t\Big)\Big\rangle. \tag{3.5}$$

Since (3.5) is valid for arbitrary F, in mathematical language, this asserts that the distribution μ_X is absolutely continuous with respect to μ_W, and

$$\frac{d\mu_X}{d\mu_W} = \exp\Big(-\frac{1}{2}\int_0^1 b^2(W_t,t)dt+\int_0^1 b(W_t,t)dW_t\Big).$$

The above derivations can be done directly with continuum version if one gets familiar enough

$$\begin{aligned}\langle F[X_t]\rangle &= \langle G[W_t]\rangle \quad (\text{where } G=F\circ\Phi)\\ &= \int G[w_t]\cdot\frac{1}{Z}\exp\Big(-\frac{1}{2}\int_0^1\dot{w}_t^2dt\Big)\delta(w_0)\mathcal{D}w\\ &= \int F[x_t]\cdot\frac{1}{Z}\exp\Big(-\frac{1}{2}\int_0^1\dot{x}_t^2dt-\frac{1}{2}\int_0^1 b^2(x_t,t)dt+\int_0^1 b(x_t,t)\dot{x}_tdt\Big)\delta(x_0)\mathcal{D}x\\ &= \int F[w_t]\cdot\frac{1}{Z}\exp\Big(-\frac{1}{2}\int_0^1\dot{w}_t^2dt-\frac{1}{2}\int_0^1 b^2(w_t,t)dt+\int_0^1 b(x_t,t)\dot{w}_tdt\Big)\delta(w_0)\mathcal{D}w\\ &= \Big\langle F[W_t]\exp\Big(-\frac{1}{2}\int_0^1 b^2(W_t,t)dt+\int_0^1 b(W_t,t)dW_t\Big)\Big\rangle.\end{aligned}$$

A special case of this representation is the *Cameron-Martin formula*, for the transformation

$$\boldsymbol{X}_t = \boldsymbol{W}_t+\boldsymbol{\phi}(t) \tag{3.6}$$

where $\boldsymbol{\phi}$ is a smooth function. This can be obtained from SDE with $\boldsymbol{b}(X_t,t)=\dot{\boldsymbol{\phi}}(t)$. In this case, we get

$$\frac{d\mu_X}{d\mu_W} = \exp\Big(-\frac{1}{2}\int_0^1\dot{\boldsymbol{\phi}}^2(t)dt+\int_0^1\dot{\boldsymbol{\phi}}(t)d\boldsymbol{W}_t\Big). \tag{3.7}$$

A slight generalization is the *Girsanov formula*. Consider two SDE's:

$$\begin{cases} d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t)d\boldsymbol{W}_t, \\ d\boldsymbol{Y}_t = (\boldsymbol{b}(\boldsymbol{Y}_t, t) + \boldsymbol{\gamma}(t, \omega))dt + \boldsymbol{\sigma}(\boldsymbol{Y}_t, t)d\boldsymbol{W}_t, \end{cases}$$

where $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{b}, \boldsymbol{\gamma} \in \mathbb{R}^n$, $\boldsymbol{W} \in \mathbb{R}^m$ and $\boldsymbol{\sigma} \in \mathbb{R}^{n\times m}$. Assume that $\boldsymbol{X}_0 = \boldsymbol{Y}_0 = \boldsymbol{x}$. Then the distributions of $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{Y}_t\}$ over $[0, 1]$ are absolutely continuous with respect to each other. Moreover the Radon-Nikodym derivative is given by

$$\frac{d\mu_Y}{d\mu_X}[X.] = \exp\left(-\frac{1}{2}\int_0^1 |\boldsymbol{\phi}(t, \omega)|^2 dt + \int_0^1 \boldsymbol{\phi}(t, \omega)d\boldsymbol{W}_t\right), \tag{3.8}$$

where $\boldsymbol{\phi}$ is the solution of

$$\boldsymbol{\sigma}(\boldsymbol{X}_t, t)\boldsymbol{\phi}(t, \omega) = \boldsymbol{\gamma}(t, \omega).$$

Mathematically, the above two results have another formulation whose idea can be explained as follows. Suppose we have n independent standard Gaussian random variables $Z_1, Z_2, \ldots, Z_n \sim N(0, 1)$ on probability space $(\Omega, \mathcal{F}, P)$. Given a vector $(\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^n$, the new random variables with translation

$$\tilde{Z}_k = Z_k + \mu_k, \quad k = 1, 2 \ldots, n$$

are no longer $N(0, 1)$ distributed. But we can define another probability measure

$$\tilde{P}(d\omega) = \exp\Big(-\sum_{k=1}^n \mu_k Z_k(\omega) - \frac{1}{2}\sum_{k=1}^n \mu_k^2\Big)P(d\omega).$$

Then we have

$$\begin{aligned}
&\tilde{P}\left(\tilde{Z}_1 \in [\tilde{z}_1, \tilde{z}_1 + d\tilde{z}_1), \ldots, \tilde{Z}_n \in [\tilde{z}_n, \tilde{z}_n + d\tilde{z}_n)\right) \\
&= \exp\Big(-\sum_{k=1}^n \mu_k(\tilde{z}_k - \mu_k) - \frac{1}{2}\sum_{k=1}^n \mu_k^2\Big) P\left(\tilde{Z}_1 \in [\tilde{z}_1, \tilde{z}_1 + d\tilde{z}_1), \ldots, \tilde{Z}_n \in [\tilde{z}_n, \tilde{z}_n + d\tilde{z}_n)\right) \\
&= \exp\Big(-\sum_{k=1}^n \mu_k(\tilde{z}_k - \mu_k) - \frac{1}{2}\sum_{k=1}^n \mu_k^2\Big) \cdot (2\pi)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2}\sum_{k=1}^n (\tilde{z}_k - \mu_k)^2\Big) d\tilde{z}_1 \cdots d\tilde{z}_n \\
&= (2\pi)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2}\sum_{k=1}^n \tilde{z}_k^2\Big) d\tilde{z}_1 \cdots d\tilde{z}_n.
\end{aligned}$$

This reveals that the variables $\{\tilde{Z}_k\}_{k=1,\ldots,n}$ are again independent $N(0, 1)$ random variables on space $(\Omega, \mathcal{F}, \tilde{P})$. If we take

$$Z_k = \frac{\Delta W_k}{\sqrt{\Delta t_k}}, \quad \tilde{Z}_k = \frac{\Delta \tilde{W}_k}{\sqrt{\Delta t_k}}, \quad \mu_k = \phi_k\sqrt{\Delta t_k}$$

and take the formal limit as $n \to \infty$, where $\Delta W_k = W_{t_{k+1}} - W_{t_k}$ and W_t is the standard Wiener process on $(\Omega, \mathcal{F}, P)$, we may claim that

$$\tilde{W}_t = W_t + \int_0^t \phi(s) ds$$

is again a standard Wiener process on $(\Omega, \mathcal{F}, \tilde{P})$ with

$$\tilde{P}(d\omega) = \exp\Big(- \int_0^t \phi(s) dW_s - \frac{1}{2} \int_0^t \phi^2(s) ds \Big) P(d\omega). \tag{3.9}$$

This claim is indeed true even for multidimensional case and the translation $\phi(t)$ can be ω-dependent.

Theorem 3.1 (Girsanov theorem I)**.** *For Itô process*

$$d\tilde{\boldsymbol{W}}_t = \boldsymbol{\phi}(t, \omega) dt + d\boldsymbol{W}_t, \quad \tilde{\boldsymbol{W}}_0 = 0, \tag{3.10}$$

where $\boldsymbol{W} \in \mathbb{R}^d$ is a d-dimensional standard Wiener process on $(\Omega, \mathcal{F}, \mathbb{P})$. Define

$$Z_t(\omega) = \exp\Big(- \int_0^t \boldsymbol{\phi}(s, \omega) d\boldsymbol{W}_s - \frac{1}{2} \int_0^t \boldsymbol{\phi}^2(s, \omega) ds \Big). \tag{3.11}$$

Assume $\boldsymbol{\phi}(t, \omega)$ satisfies Novikov's condition

$$\mathbb{E} \exp\Big(\frac{1}{2} \int_0^T |\boldsymbol{\phi}|^2(s, \omega) ds \Big) < \infty, \tag{3.12}$$

where $T \leq \infty$ is a fixed constant. Define $\tilde{\mathbb{P}}$ as

$$\tilde{\mathbb{P}}(d\omega) = Z_T(\omega) \mathbb{P}(d\omega), \tag{3.13}$$

then we have $\tilde{\boldsymbol{W}}$ is a d-dimensional Wiener process with respect to $(\Omega, \mathcal{F}_T, \tilde{\mathbb{P}})$ for $t \leq T$.

The Novikov's condition is to ensure the process Z_t in (3.11) is an *exponential martingale*. The rigorous proof of Theorem 3.1 may be referred to [3, 4]. The definition (3.11) does not contradict (3.7). Indeed, they are consequences of each other. To see this, we note that for any functional F

$$\begin{aligned}
\Big\langle F[\tilde{\boldsymbol{W}}_t] \Big\rangle_{\tilde{\mathbb{P}}} &= \Big\langle F[\tilde{\boldsymbol{W}}_t] Z_T \Big\rangle_{\mathbb{P}} \\
&= \Big\langle F[\tilde{\boldsymbol{W}}_t] \exp\Big(- \int_0^T \boldsymbol{\phi}(s, \omega) d\tilde{\boldsymbol{W}}_s + \frac{1}{2} \int_0^T \boldsymbol{\phi}^2(s, \omega) ds \Big) \Big\rangle_{\mathbb{P}} \\
&= \Big\langle F[\boldsymbol{W}_t] \exp\Big(- \int_0^T \boldsymbol{\phi}(s, \omega) d\boldsymbol{W}_s + \frac{1}{2} \int_0^T \boldsymbol{\phi}^2(s, \omega) ds \Big) \frac{d\mu_{\tilde{\boldsymbol{W}}}}{d\mu_{\boldsymbol{W}}} \Big\rangle_{\mathbb{P}} \\
&= \Big\langle F[\boldsymbol{W}_t] \Big\rangle_{\mathbb{P}}.
\end{aligned}$$

It can also be verified by path integrals as follows

$$\begin{aligned}
\langle F[\boldsymbol{W}_t]\rangle_{\mathbb{P}} &= \int F[\boldsymbol{w}_t]\cdot\frac{1}{Z}\exp\Big(-\frac{1}{2}\int_0^T \dot{\boldsymbol{w}}_t^2 dt\Big)\delta(\boldsymbol{w}_0)D\boldsymbol{w}\\
&= \int F[\tilde{\boldsymbol{w}}_t]\cdot\frac{1}{Z}\exp\Big(-\frac{1}{2}\int_0^T \dot{\tilde{\boldsymbol{w}}}_t^2 dt\Big)\delta(\tilde{\boldsymbol{w}}_0)D\tilde{\boldsymbol{w}}\\
&= \int F\circ\Phi[\boldsymbol{w}_t]\cdot\frac{1}{Z}\exp\Big(-\frac{1}{2}\int_0^T \dot{\boldsymbol{w}}_t^2 dt-\frac{1}{2}\int_0^T \boldsymbol{\phi}^2 dt-\int_0^T \boldsymbol{\phi}(t)\dot{\boldsymbol{w}}_t dt\Big)\delta(\boldsymbol{w}_0)D\boldsymbol{w}\\
&= \Big\langle G[\boldsymbol{W}_t]\exp\Big(-\frac{1}{2}\int_0^T \boldsymbol{\phi}^2(t)dt-\int_0^T \boldsymbol{\phi}(t)d\boldsymbol{W}_t\Big)\Big\rangle_{\mathbb{P}}\\
&= \Big\langle F[\tilde{\boldsymbol{W}}_t]\exp\Big(-\frac{1}{2}\int_0^T \boldsymbol{\phi}^2(t)dt-\int_0^T \boldsymbol{\phi}(t)d\boldsymbol{W}_t\Big)\Big\rangle_{\mathbb{P}}\\
&= \Big\langle F[\tilde{\boldsymbol{W}}_t]Z_T\Big\rangle_{\mathbb{P}} = \Big\langle F[\tilde{\boldsymbol{W}}_t]\Big\rangle_{\tilde{\mathbb{P}}}.
\end{aligned}$$

Corresponding to (3.8), we have another form of Girsanov theorem.

Theorem 3.2 (Girsanov theorem II)**.** *For Itô processes $\boldsymbol{X},\boldsymbol{Y}$ satisfy*

$$\begin{cases} d\boldsymbol{X}_t=\boldsymbol{b}(\boldsymbol{X}_t,t)dt+\boldsymbol{\sigma}(\boldsymbol{X}_t,t)d\boldsymbol{W}_t, \quad \boldsymbol{X}_0=\boldsymbol{x},\\ d\boldsymbol{Y}_t=(\boldsymbol{b}(\boldsymbol{Y}_t,t)+\boldsymbol{\gamma}(t,\omega))dt+\boldsymbol{\sigma}(\boldsymbol{Y}_t,t)d\boldsymbol{W}_t, \quad \boldsymbol{Y}_0=\boldsymbol{x}, \end{cases}$$

where $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{b},\boldsymbol{\gamma}\in\mathbb{R}^n$, $\boldsymbol{W}\in\mathbb{R}^m$ and $\boldsymbol{\sigma}\in\mathbb{R}^{n\times m}$, and assume $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ satisfy the usual conditions in Theorem **??**. *Suppose there exists unique $\boldsymbol{\phi}(t,\omega)$ such that*

$$\boldsymbol{\sigma}(\boldsymbol{X}_t,t)\boldsymbol{\phi}(t,\omega)=\boldsymbol{\gamma}(t,\omega)$$

and the Novikov's condition holds

$$\mathbb{E}\exp\Big(\frac{1}{2}\int_0^T|\boldsymbol{\phi}|^2(s,\omega)ds\Big)<\infty. \tag{3.14}$$

Define $\tilde{\boldsymbol{W}}_t$, Z_t and $\tilde{\mathbb{P}}$ as in Theorem 3.1, then $\tilde{\boldsymbol{W}}$ is a standard Wiener process under $(\Omega,\mathcal{F}_T,\tilde{\mathbb{P}})$ and $\boldsymbol{Y}$ satisfies

$$d\boldsymbol{Y}_t=\boldsymbol{b}(\boldsymbol{Y}_t,t)dt+\boldsymbol{\sigma}(\boldsymbol{Y}_t,t)d\tilde{\boldsymbol{W}}_t, \quad \boldsymbol{Y}_0=\boldsymbol{x}, \quad t\le T.$$

Thus the law of $\boldsymbol{Y}_t$ under $\tilde{\mathbb{P}}$ is the same that of $\boldsymbol{X}_t$ under $\mathbb{P}$ for $t\le T$.

The readers may be referred to [3, 4] for proof details.

4 Feynman-Kac Formula: Revisited

Earlier we have known that the solution of PDE

$$\partial_t v=\frac{1}{2}\Delta v+q(x)v, \quad v|_{t=0}=f(x)$$

can be represented as

$$v(x,t) = \mathbb{E}^x\Big(\exp\Big(\int_0^t q(W_s)ds\Big) f(W_t)\Big).$$

In path integral form

$$v(x,t) = \int \delta(w_0 - x)\frac{1}{Z}\exp\Big(-\int_0^t \big(\frac{1}{2}\dot{w}_s^2 - q(w_s)\big)ds\Big) f(w_t)\mathcal{D}w,$$

where the delta-function $\delta(w_0 - x)$ is to shift the starting point of the Wiener process to x. Feynmann-Kac formula originates from Feynmann's interpretation of quantum mechanics, namely that solution of linear Schrödinger equation

$$i\hbar\partial_t\psi = -\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi, \quad \psi|_{t=0} = \psi_0(x) \tag{4.1}$$

can be expressed formally as

$$\psi(x,t) = \int \delta(w_0 - x)\frac{1}{Z}\exp\Big(\frac{i}{\hbar}I[w]\Big)\psi_0(w_t)\mathcal{D}w, \tag{4.2}$$

where $I[\cdot]$ is the Lagrangian defined as

$$I[w] = \int_0^t \Big(\frac{m}{2}\dot{w}_s^2 - V(w_s)\Big)ds.$$

Formally if we take

$$m = 1, \quad \hbar = -i$$

in (4.1) and (4.2), we exactly obtain the above formulation for Feynman-Kac problem! Indeed, that is the real story on how Feynman-Kac formula is created.

Feynman's formally expression is yet to be made rigorous. However, Kac's reinterpretation for the heat equation instead of Schrödinger's equation can be readily proved. The Feynman-Kac formula can also be generalized to the case when Δ is replaced by more general second order differential operator as we did in previous Chapter.

Homeworks

1. Derive the infinite dimensional characteristic function for Wiener process W_t

$$\Big\langle \exp\Big(i\int_0^1 \xi(t)dW_t\Big)\Big\rangle = \exp\Big(-\frac{1}{2}\int_0^1 |\xi|^2 dt\Big).$$

References

[1] T. Sauer B. Hunt and J.A. Yorke. Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces. *Bull. Amer. Math. Soc.*, 27:217–238, 1992.

[2] R.P. Feynman. Space-time approach to non-relativistic quantum mechanics. *Rev. Mod. Phys.*, 20:367–387, 1948.

[3] I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, Berlin, Heidelberg and New York, 1991.

[4] B. Oksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, New York, Berlin, Heidelberg and Tokyo, 4th edition edition, 1998.