Monte-Carlo
MMD-MA, Université Paris-Dauphine

Xiaolu Tan
tan@ceremade.dauphine.fr

Septembre 2015

Contents

Chapter 1

Introduction

1.1 The principle

The principle of Monte-Carlo method is to use simulations of the random variables to estimate a quantity such as

$$\mathbb{E}\Big[f(X)\Big] \;=\; \int_{\mathbb{R}^d} f(x)\ \mu(dx), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a function, X is some random vector taking value in $\mathbb{R}^d$ with distribution μ. It may also be used to solve an optimization problem of the kind

$$\min_{\theta\in\Theta}\ \mathbb{E}\Big[f_\theta(X)\Big], \tag{1.2}$$

where $(f_\theta : \mathbb{R}^d \to \mathbb{R})_{\theta\in\Theta}$ is a family of functions.

To solve the basic problem (1.1), the method consists in simulating a sequence of i.i.d. random vectors $(X_k)_{k\geq 1}$ with the same distribution of X, and then estimate $\mathbb{E}[f(X)]$ by the empirical mean value

$$\overline{Y}_n \;:=\; \frac{1}{n}\sum_{k=1}^{n} Y_k \;:=\; \frac{1}{n}\sum_{k=1}^{n} f(X_k). \tag{1.3}$$

The advantages of the Monte-Carlo are usually its simplicity, flexibility and efficiency for high dimensional problems. It can also be served as an alternative method (or benchmark) for other numerical methods.

1.2 The error analysis of Monte-Carlo method

Theorem 1.1 (Law of large number) *Let $(Y_k)_{k\geq 1}$ be a sequence of i.i.d. random variables such that $\mathbb{E}[|Y|] < \infty$. Then with $\overline{Y}_n$ defined in* (1.3)*, one has*

$$\mathbb{P}\Big(\lim_{n\to\infty} \overline{Y}_n = \mathbb{E}[Y]\Big) \;=\; 1.$$

Theorem 1.2 (Central limit theorem) *Let $(Y_k)_{k\geq 1}$ be a sequence of i.i.d. random variables such that $\mathbb{E}[|Y|^2] < \infty$. Then*

$$\sqrt{n}\big(\overline{Y}_n - \mathbb{E}[Y]\big) \quad \Rightarrow \quad N\big(0, \mathrm{Var}(Y)\big).$$

And consequently,

$$\frac{\sqrt{n}}{\hat{\sigma}_n}\big(\overline{Y}_n - \mathbb{E}[Y]\big) \quad \Rightarrow \quad N\big(0,1\big), \qquad \text{where } \hat{\sigma}_n^2 \ := \ = \frac{1}{n}\sum_{k=1}^{n}\big(Y_k - \overline{Y}_n\big)^2. \tag{1.4}$$

Notice that $\hat{\sigma}_n^2$ defined in (1.4) is an estimator of the variance $\mathrm{Var}[Y]$ from the sequence $(Y_k)_{k\geq 1}$, and it admits the representation

$$\hat{\sigma}_n^2 \ = \ \frac{1}{n}\sum_{k=1}^{n} Y_k^2 \ - \ (\overline{Y}_n)^2.$$

The central limit theorem induces that the asymptotic confidence interval of level $p(R) := \int_{|x|\leq R}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$ of the estimator $\overline{Y}_n$ is given by

$$\left[\overline{Y}_n - \frac{R}{\sqrt{n}}\hat{\sigma}_n,\ \overline{Y}_n + \frac{R}{\sqrt{n}}\hat{\sigma}_n\right]. \tag{1.5}$$

More precisely, it means that

$$\mathbb{P}\left[\overline{Y}_n - \frac{R}{\sqrt{n}}\hat{\sigma}_n \ \leq \ \mathbb{E}[Y] \ \leq \ \overline{Y}_n + \frac{R}{\sqrt{n}}\hat{\sigma}_n\right] \ \to \ p(R), \quad \text{as } n \to \infty.$$

Remark 1.1 *In practice, for $p(R) = 95\%$, we know $R \approx 1.96$.*

Conclusions To utilize the Monte-Carlo method, the first issue is then how to simulate a sequence of random vector $(X_k)_{k\geq 1}$ given its law μ or given its definition based on other random elements, the second issue is how to improve the estimator by reducing its error. The error of Monte-Carlo method is measured by its confidence interval (1.5), whose length is given by $2R\hat{\sigma}_n/\sqrt{n}$. When the confidence level is fixed, R is fixed. One can then use a larger n, where the cost is the computation time which is proportional to n in general. Otherwise, one can reduce $\hat{\sigma}_n$ by find some other random variable $\tilde{Y}$ such that

$$\mathbb{E}\big[\tilde{Y}\big] \ = \ \mathbb{E}[Y] \quad \text{and} \quad \mathrm{Var}\big[\tilde{Y}\big] \ < \ \mathrm{Var}[Y].$$

We shall address these issues in the following of the course.

Chapter 2

Simulation of random vectors or stochastic processes

2.1 Random variable of uniform distribution on $[0, 1]$

Here we admits that we know how to simulate a random variable of uniform distribution $\mathcal{U}[0, 1]$. In particular, most of the programming environment are equipped with the computer with a generator of uniform distribution.

However, it worths noticing that a generator of random variables in a computer is a deterministic program, and hence it generates always a sequence of deterministic variables, in place of a sequence of independent random variables. In practice, we search for a generator such that the sequence of generated variables has "similar" performance statistically.

2.2 Inverse method

Let X be random variable, its distribution function, defined by

$$F(x) \quad := \quad \mathbb{P}(X \leq x),$$

is a right-continuous non-decreasing function from $\mathbb{R}$ to $[0, 1]$. We then define its right-continuous generalized inverse function by

$$F^{-1}(u) \quad := \quad \inf\big\{x \in \mathbb{R} \ : F(x) \geq u\big\}.$$

Theorem 2.1 *Let X be a random variable with distribution function F and U be a random variable of uniform distribution $\mathcal{U}[0, 1]$ on interval $[0, 1]$. Then*

$$X \quad \sim \quad F^{-1}(U) \qquad \textit{in law.}$$

Proof. Notice that for any $y \in \mathbb{R}$, we have

$$\text{“}F^{-1}(u) \le y\text{”} \quad \Longleftrightarrow \quad \text{“}u \le F(y)\text{”}.$$

It follows that for any $y \in \mathbb{R}$,

$$\mathbb{P}\big[F^{-1}(U) \le y\big] \;=\; \mathbb{P}\big[U \le F(y)\big] \;=\; F(y).$$

□

Example 2.1 (i) *Let X be a random variable of discrete distribution, $\mathbb{P}(X = x_k) = p_k$ where $x_k \in \mathbb{R}$, $p_k \ge 0$ for $k \in \mathbb{N}$ and $\sum_{k\in\mathbb{N}} p_k = 1$. Then let $U \sim \mathcal{U}[0,1]$, and Z be the random variable defined by*

$$Z \;:=\; x_n, \quad \text{if } U \in \left[\sum_{k=0}^{n-1} p_k,\ \sum_{k=0}^{n} p_k\right).$$

Then Z has the same distribution of X. The definition of Z can be interpreted as $F^{-1}(U)$ with the distribution function F of X.

(ii) *Let $X \sim \mathcal{E}(\lambda)$ be a random variable of exponential distribution of parameter $\lambda > 0$, i.e. the density function is given by $f(x) := \lambda e^{-\lambda x}\mathbf{1}_{x\ge 0}$, and the distribution is given by $F(x) := 1 - e^{-\lambda x}$. By direct computation, $F^{-1}(u) = -\lambda^{-1}\log(1-u)$ for every $u \in (0,1)$. Then for $U \sim \mathcal{U}(0,1)$,*

$$F^{-1}(U) \;=\; -\lambda^{-1}\log(1-U) \;\sim\; -\lambda^{-1}\log(U) \;\sim\; \mathcal{E}(\lambda),$$

since $1-U$ and U have the same distribution when $U \sim \mathcal{U}(0,1)$.

2.3 Transformation method

Proposition 2.1 (Box-Muller) *Suppose that U and V are independent random variables of uniform distribution on the interval $(0,1]$. Let*

$$X \;:=\; \sqrt{-2\log(U)}\cos(2\pi V) \quad \text{and} \quad Y \;:=\; \sqrt{-2\log(U)}\sin(2\pi V).$$

Then X and Y are two independent random variables of Gassian distribution $N(0,1)$.

Proof.

□

Exercise 2.1 *Let (U,V) be a random vector which is uniformly distributed on the disk $\{(u,v)\ : u^2+v^2 \le 1\}$. Let*

$$X \;:=\; U\sqrt{\frac{-2\log(U^2+V^2)}{U^2+V^2}} \quad \text{and } Y \;:=\; V\sqrt{\frac{-2\log(U^2+V^2)}{U^2+V^2}}.$$

Prove that $(X, Y) \sim N(0, I_2)$.

2.4 Reject method

Let $f : \mathbb{R}^d \to \mathbb{R}^+$ and $g : \mathbb{R}^d \to \mathbb{R}^+$ be two density functions such that, for some constant $\gamma > 0$, one has

$$f(x) \ \leq \ \gamma g(x), \quad \text{for all } x \in \mathbb{R}^d.$$

In practice, g is the density function of some distribution with well-known simulation method (such as Gaussian distribution, uniform distribution, exponential distribution, etc.), but f is the density function of some distribution without an easy simulation method. The objective is to use the simulations of random variables of distribution g, together with a rejection procedure, to simulate the random variable of distribution f.

Proposition 2.2 *Let* $(Y_k)_{k\geq 1}$ *be an i.i.d. sequence of random variables of density* g, *and* $(U_k)_{k\geq 1}$ *be an i.i.d. sequence of random variable of distribution* $\mathcal{U}[0,1]$. *Moreover,* $(Y_k)_{k\geq 1}$ *and* $(U_k)_{k\geq 1}$ *are also independent. Define a sequence* $(X_n)_{n\geq 1}$ *of random variables*

$$X_n \ := \ Y_{N_n}, \quad \textit{with } N_0 := 0, \ \textit{and} N_{n+1} \ := \ \min\Big\{k > N_n \ : U_k \leq \frac{f(X_k)}{\gamma g(X_k)}\Big\}.$$

Then $(X_n)_{n\geq 1}$ *is a sequence of i.i.d. random variable of density* f.

Proof.

□

Exercise 2.2 *Let* $f : \mathbb{R} \to \mathbb{R}^+$ *be defined by*

$$f(x) := (1 - |x|)^+.$$

Give a numerical algorithm (based on the above reject method) to simulate an i.i.d. sequence of random variables of density f.

2.5 Simulation of Gaussian vector

The case of dimension 2 Let $(Z_1, Z_2) \sim N(0, I_2)$ be two independent random variable of Gaussian distribution, $\rho \in [-1, 1]$ a constant. Define

$$X_1 \ := \ Z_1 \quad \text{and} \quad X_2 \ := \ \rho Z_1 + \sqrt{1-\rho^2} Z_2.$$

It is clear that $X_1 \sim X_2 \sim N(0,1)$ and $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2) = \rho$, which means that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

More generally, for $(Z_1, Z_2) \sim N(0, I_2)$, let

$$X_1 := \mu_1 + \sigma_1 Z_1 \quad \text{and} \quad X_2 := \mu_2 + \sigma_2 \Big(\rho Z_1 + \sqrt{1-\rho^2} Z_2 \Big),$$

then

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right). \tag{2.1}$$

Notice also that any Gaussian vector of dimension 2 can be written in the form (2.1).

General case: Cholesky's method Let $Z \sim N(0, I_d)$ be standard Gaussian random vector of dimension d, and A be a lower triangular matrix of dimension $d \times d$, i.e.

$$Z = \begin{pmatrix} Z_1 \\ \cdots \\ Z_d \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ A_{d1} & A_{d2} & \cdots & A_{dd.} \end{pmatrix}$$

Then the vector $X := AZ \sim N(\mathbf{0}, \Sigma)$ with variance-covariance matrix $\Sigma := AA^T$.

Cholesky's method consists in finding a lower triangular matrix A such that $AA^T = \Sigma$, where Σ is a given variance-covariance matrix. Let us write the equation $AA^T = \Sigma$ as

$$\begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ A_{d1} & A_{d2} & \cdots & A_{dd.} \end{pmatrix} \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{d1} \\ 0 & A_{22} & \cdots & A_{d2} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A_{dd.} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1d} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma_{d1} & \Sigma_{d2} & \cdots & \Sigma_{dd} \end{pmatrix}.$$

The solution is given by

$$\begin{cases} A_{11}^2 = \Sigma_{11} \\ A_{21} A_{11} = \Sigma_{21} \\ \cdots \\ A_{d1} A_{11} = \Sigma_{d1} \end{cases} \cdots \iff \begin{cases} A_{ii} = \sqrt{\Sigma_{ii} - \sum_{k=1}^{i-1} A_{ik}^2} \\ A_{ij} = \big(\Sigma_{ii} - \sum_{k=1}^{j-1} A_{ik} A_{jk} \big) / A_{jj}, \quad \forall j < i. \end{cases}$$

Exercise 2.3 *Provide a pseudo code for the algorithm of Cholesky's method.*

2.6 Simulation of Brownian motion

Definition 2.1 *A standard Brownian motion W is a stochastic process starting from 0, and having*
(i) *continuous paths (i.e. $t \mapsto W_t$ is almost surely continuous),*
(ii) *independent increments (i.e. $W_t - W_s \perp W_s - W_r$, $\forall 0 \leq r \leq s \leq t$),*
(iii) *stationary and Gaussian increments (i.e. $W_t - W_s \sim N(0, t-s)$).*

Forward simulation Using the the independent and stationary Gaussian increments property, one can simulate a path of a Brownian motion in a forward way. Let $0 = t_0 < t_1 < \cdots$ be a discrete grid of $\mathbb{R}^+$, $(Z_k)_{k\geq 1}$ be a sequence of i.i.d. random variable of Gaussian distribution $N(0,1)$, we define W by

$$W_0 \;:=\; 0 \quad \text{and} \quad W_{t_{k+1}} \;:=\; W_{t_k} \;+\; \sqrt{t_{k+1} - t_k} Z_{k+1}.$$

Then W is a sample of paths of the Brownian motion on the discrete grid $(t_k)_{k\geq 0}$.

Brownian bridge The forward simulation method consists in simulating $W_{t_{k+1}}$ knowing the value of W_{t_k}. There is backward simulation method, i.e. one simulates first the variable W_{t_n}, and then simulates the variables $W_{t_{n-1}}$, $W_{t_{n-2}}$, $\cdots$, W_{t_2}, W_{t_1} recursively.

Proposition 2.3 *Let $0 = t_0 < t_1 < \cdots$ be a discrete grid, then the conditional distribution of W_k knowing $\big(W_{t_i},\ i \neq k\big)$ is a Gaussian distribution $N(\mu, \sigma^2)$ with*

$$\mu = \frac{t_{k+1} - t_t}{t_{k+1} - t_{k-1}} W_{t_{k-1}} + \frac{t_k - t_{k-1}}{t_{k+1} - t_{k-1}} W_{t_{k+1}} \quad \textit{and} \quad \sigma^2 = \frac{(t_{k+1} - t_k)(t_k - t_{k-1})}{t_{k+1} - t_{k-1}},$$

in particular,

$$\mathcal{L}\big(W_{t_k} \;\big|\; W_{t_{k-1}} = x,\ W_{t_{k+1}} = y\big) \;=\; N\Big(x + \frac{t_k - t_{k-1}}{t_{k+1} - t_{k-1}} y,\ \frac{(t_{k+1} - t_k)(t_k - t_{k-1})}{t_{k+1} - t_{k-1}}\Big).$$

Proof.

□

Exercise 2.4 *Give the backward simulation algorithm for a Brownian motion on $[0,1]$, using the above results.*

Chapter 3

Variance reduction techniques

Recall that the principle of the Monte-Carlo method is to estimate

$$\mathbb{E}[Y], \quad \big(\text{where } Y \ := \ f(X)\big)$$

by

$$\overline{Y}_n \ := \ \frac{1}{n}\sum_{k=1}^{n} Y_k \ := \ \frac{1}{n}\sum_{k=1}^{n} f(X_k), \tag{3.1}$$

with simulations $(Y_k)_{k\geq 1}$ (or more precisely $(X_k)_{k\geq 1}$) of random variables Y. In view of the confidence interval (1.5), it is clear that to reduce the error, one should either augment the simulation number n (in cost of computation time), or reduce the variance $\hat{\sigma}_n^2$. More precisely, since the variance $\mathrm{Var}[Y]$ of Y is fixed, the real issue is to find some other random variable $\tilde{Y}$ satisfying

$$\mathbb{E}\big[\tilde{Y}\big] \ = \ \mathbb{E}[Y] \quad \text{and} \quad \mathrm{Var}\big[\tilde{Y}\big] \ < \ \mathrm{Var}[Y]. \tag{3.2}$$

In most of cases, $\tilde{Y}$ admits the representation $\tilde{Y} := g(X)$ with some function $g : \mathbb{R}^d \to \mathbb{R}$. Then using the simulations of $\tilde{Y}$, one could expect an estimator of $\mathbb{E}[Y](= \mathbb{E}\big[\tilde{Y}\big])$ with smaller error.

3.1 The antithetic variable

For many random variables (vectors) X, their distributions have some symmetric property and admits a simple transformation $A : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$A(X) \quad \text{and} \quad X \ \text{ have the same distribution.}$$

We call $A(X)$ the antithetic variable of X. For example, let $X \sim \mathcal{U}[0,1]$, then $A(X) := 1 - X \sim \mathcal{U}[0,1]$; let $X \sim N(0,\sigma^2)$, then $A(X) := -X \sim N(0,\sigma^2)$. It follows that

$\mathbb{E}\big[f(A(X))\big] = \mathbb{E}[f(X)]$, and hence

$$\mathbb{E}\big[\tilde{Y}\big] \;=\; \mathbb{E}[Y] \quad \text{with} \quad \tilde{Y} \;:=\; \frac{f(X)+f\big(A(X)\big)}{2}.$$

Then a new Monte-Carlo estimator can be given by

$$\tilde{Y}_n \;:=\; \frac{1}{n}\sum_{k=1}^{n}\frac{f(X_k)+f(A(X_k))}{2} \;=\; \frac{1}{2n}\sum_{k=1}^{n}\Big(f(X_k)+f(A(X_k))\Big). \tag{3.3}$$

In some context, we can expect that $\mathrm{Var}[\tilde{Y}]$ much smaller than $\mathrm{Var}[Y]$ (see the criteria (3.2)).

Example 3.1 (Naive Examples) (i) *Let $f(x) := x$ and $X \sim N(0,\sigma^2)$ be a Gaussian r.v., then $Y := f(X) = X$. The random variable X admits an antithetic variable $-X$. Then $\tilde{Y} := \frac{f(X)+f(-X)}{2} \equiv 0$ and it is clear that*

$$\mathbb{E}\big[\tilde{Y}\big] \;=\; \mathbb{E}[Y] \quad \textit{and} \quad \mathrm{Var}[Y] \;>\; \mathrm{Var}\big[\tilde{Y}\big] \;=\; 0.$$

i.e. (3.2) *is true for this example.*

(ii) *Let $f(x) := x$, $Y := f(U)$ and $U \sim \mathcal{U}[0,1]$ which admits an antithetic variable $1-U$. Then $\tilde{Y} := \frac{f(U)+f(1-U)}{2} \equiv \frac{1}{2}$ and it is clear that* (3.2) *holds true in this context.*

Exercise 3.1 *Let $U \sim \mathcal{U}[0,1]$, then $X := -\frac{\log(U)}{\lambda} \sim \mathcal{E}(\lambda)$ and $\tilde{X} := -\frac{\log(1-U)}{\lambda} \sim \mathcal{E}(\lambda)$. Then*

$$\mathbb{E}\big[X\big] \;=\; \mathbb{E}\Big[\frac{X+\tilde{X}}{2}\Big].$$

Comparer the variance of X and that of $\frac{X+\tilde{X}}{2}$.

Variance analysis By direct computation, we have

$$\begin{aligned}\mathrm{Var}\big[\tilde{Y}\big] \;&=\; \frac{1}{4}\Big(\mathrm{Var}[f(X)] \;+\; 2\mathrm{Cov}\big[f(X),f(A(X))\big] \;+\; \mathrm{Var}[f(A(X))]\Big)\\ &=\; \frac{1}{2}\mathrm{Var}[Y] \;+\; \frac{1}{2}\mathrm{Cov}\big[f(X),f(A(X))\big].\end{aligned}$$

Then one has

$$\mathrm{Var}\big[\tilde{Y}\big] \;\le\; \frac{1}{2}\mathrm{Var}[Y] \tag{3.4}$$

whenever

$$\mathrm{Cov}\big[f(X),f(A(X))\big] \;\le\; 0.$$

Intuitively, since $A(X)$ is the "antithetic" variable, we can expect that $A(X)$ has a negative correlation with X. In practice, the computation error of estimators $\tilde{Y}_n$ (in (3.3))

and $\overline{Y}_{2n}$ (in (3.1)) should be the same, and under the condition (3.4), one has

$$\mathrm{Var}\big[\tilde{Y}_n\big] \quad \leq \quad \mathrm{Var}\big[\overline{Y}_{2n}\big].$$

Remark 3.1 *It is very important to use the same simulation X_k for estimator $\tilde{Y}_n$ in* (3.3)*. Otherwise, image that $(\tilde{X}_k)_{k\geq 1}$ is i.i.d with $(X_k)_{k\geq 1}$, and consider*

$$\hat{Y}_n \quad := \quad \frac{1}{n}\sum_{k=1}^{n}\frac{f(X_k)+f(A(\tilde{X}_k))}{2}.$$

Then one has

$$\mathrm{Var}\big[\hat{Y}_n\big] \quad = \quad \mathrm{Var}\big[\overline{Y}_{2n}\big],$$

which means that the estimator $\hat{Y}_n$ is not better than the classical estimator.

Case of Gaussian distribution When X is of Gaussian distribution, we can provide more precise criteria for condition (3.4).

Proposition 3.1 *Let $X \sim N(\mu,\sigma^2)$, which admits an antithetic variable $A(X) := 2\mu - X$. Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function, then*

$$\mathrm{Cov}\big[f(X), f(A(X))\big] \quad \leq \quad 0.$$

Proof. Without loss of generality, we can suppose that $X \sim N(0,1)$. Let X_1, X_2 be two independent r.v. of distribution $N(0,1)$, then for a monotone function, one has

$$\big(f(X_1)-f(X_2)\big)\big(f(-X_1)-f(-X_2)\big) \quad \leq \quad 0.$$

And hence

$$\mathbb{E}\Big[\big(f(X_1)-f(X_2)\big)\big(f(-X_1)-f(-X_2)\big)\Big] \quad \leq \quad 0.$$

By direct computation, it follows that

$$\begin{aligned} 0 \quad &\geq \quad \mathbb{E}\Big[\big(f(X_1)-f(X_2)\big)\big(f(-X_1)-f(-X_2)\big)\Big] \\ &= \quad 2\ \mathrm{Cov}\big[f(X_1), f(-X_1)\big] \quad = \quad 2\ \mathrm{Cov}\big[f(X), f(-X)\big]. \end{aligned}$$

□

Example 3.2 *In application of finance, a problem may be*

$$\mathbb{E}\big[e^{-rT}(S_T-K)^+\big] \quad \textit{with} \quad S_T := S_0 e^{(r-\sigma^2/2)T+\sigma W_T},$$

where W is a Brownian motion, i.e. $W_T \sim N(0,T)$. In this case, it is clear that $Y := e^{-rT}(S_T-K)^+$ can be expressed as an increasing function of W_T, and one can then use the antithetic variable technique in the Monte-Carlo method.

3.2 Variate control method

We recall that the random variable takes the form $f(X)$ with some random vector X and $f : \mathbb{R}^d \to \mathbb{R}$. Suppose that there is some other function $g : \mathbb{R}^d \to \mathbb{R}$ (close to f) and such that the constant

$$m \;:=\; \mathbb{E}[g(X)]$$

can be computed explicitly. Then for every constant $b \in \mathbb{R}$, one has

$$\mathbb{E}[Y] \;=\; \mathbb{E}[\tilde{Y}(b)] \quad \text{with} \quad \tilde{Y}(b) \;:=\; f(X) - b\big(g(X) - m\big).$$

It follows another Monte-Carlo estimator of $\mathbb{E}[Y]$, with simulations $(X_k)_{k\geq 1}$,

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{Y}_k(b) \quad \text{where} \quad \tilde{Y}_k(b) \;:=\; f(X_k) - b\big(g(X_k) - m\big). \tag{3.5}$$

Example 3.3 *Let $X \sim \mathcal{U}[0,1]$, $f : [0,1] \to \mathbb{R}$ be some function, and $Y := f(X)$. By approximation, one may find some polynomial function $g : [0,1] \to \mathbb{R}$ such that $f \approx g$. Besides, the constant $m := E[g(X)]$ is known explicitly whenever g is a polynomial. Take $b = 1$, it follows that*

$$\mathbb{E}[Y] \;=\; \mathbb{E}\big[f(X) - g(X) + m\big]$$

and we can expect that

$$\mathrm{Var}\big[f(X) - g(X) + m\big] \;=\; \mathrm{Var}\big[f(X) - g(X)\big] \;<\; \mathrm{Var}\big[f(X)\big],$$

since g is an approximation of f.

Variance analysis By direct computation, it follows that

$$\begin{aligned}\mathrm{Var}\big[\tilde{Y}(b)\big] &= \mathrm{Var}\big[f(X) - b\big(g(X) - m\big)\big] \\ &= \mathrm{Var}\big[f(X)\big] - 2b\ \mathrm{Cov}\big[f(X), g(X)\big] + b^2\mathrm{Var}\big[g(X)\big].\end{aligned}$$

We then minimize the variance on the control variable b:

$$\min_{b\in\mathbb{R}} \mathrm{Var}\big[\tilde{Y}(b)\big] \;=\; \mathrm{Var}[Y] - \frac{\big(\mathrm{Cov}\big[f(X), g(X)\big]\big)^2}{\mathrm{Var}\big[g(X)\big]} \;=\; \mathrm{Var}[Y]\big(1 - \rho^2\big[f(X), g(X)\big]\big),$$

with the optimal control variable

$$b^* \;:=\; \frac{\mathrm{Cov}\big[f(X), g(X)\big]}{\mathrm{Var}\big[g(X)\big]}. \tag{3.6}$$

Remark 3.2 (i) *The above computation shows that to use the variate control method, one should search for a function $g : \mathbb{R}^d \to \mathbb{R}$ such that*

$$m := \mathbb{E}\big[g(X)\big] \quad \textit{is known explicitly, and} \quad |\rho\big(f(X), g(X)\big)| \ \textit{is big.}$$

(ii) *As in Remark 3.1, it is very important to use the same simulation X_k for estimator $\tilde{Y}_n$ in* (3.5)*. Otherwise, image that $(\tilde{X}_k)_{k\geq 1}$ is i.i.d with $(X_k)_{k\geq 1}$, and consider*

$$\hat{Y}_n \quad := \quad \frac{1}{n}\sum_{k=1}^{n}\big(f(X_k) - b(g(\tilde{X}_k) - m)\big),$$

Then one has

$$\rho\big(f(X), g(\tilde{X})\big) \quad = \quad 0,$$

which means that the estimator $\hat{Y}_n$ is not better than the classical estimator.

Estimation of the optimal control variable b^* In practice, we use Monte-Carlo method to compute $\mathbb{E}[f(X)]$ since it can be computed explicitly. Then there is no reason we know how to compute $\mathrm{Cov}\big[f(X), g(X)\big]$, and hence we need to estimate it to obtain an estimation of b^* in (3.6). A natural estimator of b^* should then be

$$\hat{b}_n \quad := \quad \frac{\sum_{k=1}^{n}(Y_k - \overline{Y}_n)(g(X_k) - \overline{G}_n)}{\sum_{k=1}^{n}(g(X_k) - \overline{G}_n)^2}, \qquad \text{with } \overline{G}_n := \frac{1}{n}\sum_{k=1}^{n} g(X_k). \tag{3.7}$$

Further, to avoid the correlation between the estimator $\hat{b}_n$ and the simulations $\tilde{Y}_k(\hat{b}_n)$ in (3.5), we can estimate first $\hat{b}_n$ with a small number n of simulations of $(X_k)_{1\leq k\leq n}$, then use a large number m of simulations $(X_k)_{n+1\leq k\leq n+m}$ to estimate $\mathbb{E}[Y]$, i.e. to obtain the estimator

$$\frac{1}{m}\sum_{k=1}^{m}\tilde{Y}_{n+k}(\hat{b}_n).$$

Multi-variate controls On can also consider several functions $(g_k : \mathbb{R}^d \to \mathbb{R})_{k=1,\cdots,n}$. Denote $Z_k := g_k(X)$, $Z = (Z_1, \cdots, Z_n)$ and suppose that $\mathbb{E}[Z]$ is known explicitly, we can then have a new variate control candidate

$$\tilde{Y}(b) \ := \ Y - \langle b, Z - \mathbb{E}[Z]\rangle, \qquad \forall b = (b_1, \cdots, b_n) \in \mathbb{R}^n.$$

It is clear that $\mathbb{E}[Y] = \mathbb{E}\big[\tilde{Y}(b)\big]$, and by similar computation, one has

$$\min_{b\in\mathbb{R}^n}\mathrm{Var}\big[\tilde{Y}(b)\big] \quad = \quad \min_{b\in\mathbb{R}^d}\Big(\sigma_Y^2 - 2b\Sigma_{YZ} + b^T\Sigma_{ZZ}b\Big),$$

where Σ_Y, Σ_{YZ} and Σ_{ZZ} are given by

$$\mathrm{Var}\begin{bmatrix} Y \\ Z \end{bmatrix} \;=\; \begin{pmatrix} \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{YZ} & \Sigma_{ZZ} \end{pmatrix}.$$

The optimal control b^* is provided by

$$b^* \;:=\; \Sigma_{ZZ}^{-1}\Sigma_{ZY}.$$

3.3 Stratification

Recall that X is random vector and $Y := f(X)$ for some function $f : \mathbb{R}^d \to \mathbb{R}$. Let $g : \mathbb{R}^d \to \mathbb{R}^m$ be some function and denote $Z := g(X)$ another random vector. Let $(A_k)_{1\le k\le K}$ be partition of the support of Z in $\mathbb{R}^n$, i.e. $A_1, \cdots, A_K$ are disjoints such that

$$\sum_{k=1}^{K} p_k \;=\; 1 \quad \text{with} \quad p_k \;:=\; \mathbb{P}(Z \in A_k),\ \forall k = 1, \cdots K.$$

It follows by Bayes's formula that

$$\mathbb{E}\big[Y\big] \;=\; \sum_{k=1}^{K} p_k\, \mathbb{E}\big[Y\big|Z \in A_k\big].$$

Assumption 3.1 (i) *The values of probability $(p_k)_{1\le k\le K}$ are known explicitly.*
(ii) *One knows how to simulate a random variable following the conditional distribution $\mathcal{L}(Y|Z \in A_k)$.*

Under Assumption 3.1, we can propose another Monte-Carlo estimator of $\mathbb{E}\big[Y\big]$: for each $k = 1, \cdots, K$, let $(Y_i^{(k)})_{i\ge1}$ be a sequence of i.i.d random variable such that $\mathcal{L}(Y_1^{(k)}) = \mathcal{L}(Y|Z \in A_k)$, then for $n = (n_1, \cdots, n_K) \in \mathbb{N}^K$, denote

$$\hat{Y}_n \;:=\; \sum_{k=1}^{K} p_k \,\Big(\frac{1}{n_k}\sum_{i=1}^{n_k} Y_i^{(k)}\Big). \tag{3.8}$$

It is clear that

$$\mathbb{E}\big[\hat{Y}_n\big] \;=\; \mathbb{E}[Y] \quad \text{and} \quad \hat{Y}_n \to \mathbb{E}[Y] \text{ as } (n_1, \cdots, n_K) \to \infty.$$

Simulation of conditional distribution (i)Suppose that X is a random variable with distribution function F, $Z = X$ and $A_k = (a_k, a_{k+1}]$ for some constant $a_1, a_2, \cdots, a_{K+1}$. Let

$$X^{(k)} \;:=\; F^{-1}\Big(F(a_k) \;+\; U\big(F(a_{k+1}) - F(a_k)\big)\Big) \quad \text{where} \quad U \sim \mathcal{U}[0,1],$$

and $Y^{(k)} := f(X^{(k)})$. Then

$$\mathcal{L}\big(X^{(k)}\big) \;=\; \mathcal{L}(X|X\in A_k) \quad\text{and}\quad \mathcal{L}\big(Y^{(k)}\big)\big) \;=\; \mathcal{L}\big(f(X^{(k)})\big) \;=\; \mathcal{L}(Y|X\in A_k).$$

(ii) Suppose that X is a random vector of density $\rho : \mathbb{R}^d \to \mathbb{R}$ and $Z = X$. Define

$$\rho_k(x) \;:=\; \frac{1}{p_k}\rho(x)\mathbf{1}_{x\in A_k} \quad\text{with}\quad p_k \;:=\; \mathbb{P}(X\in A_k) \;=\; \int_{A_k}\rho(x)dx.$$

Then ρ_k is the density function of the conditional distribution of $\mathcal{L}(X|X\in A_k)$.

Variance analysis Denote $\mu_k := \mathbb{E}[Y^{(k)}]$, $\sigma_k^2 := \mathrm{Var}[Y^{(k)}]$, $q_k := \frac{n_k}{n}$ with $n := \sum_{k=1}^K n_k$. Then

$$\begin{aligned}\mathrm{Var}\big[\hat{Y}_n\big] \;&=\; \sum_{k=1}^K p_k^2\,\frac{1}{n_k}\,\mathrm{Var}\big[Y^{(k)}\big] \;=\; \sum_{k=1}^K p_k^1\,\frac{1}{nq_k}\sigma_k^2\\ &=\; \frac{1}{n}\sum_{k=1}^K\frac{p_k^2}{q_k}\sigma_k^2 \;=\; \frac{1}{n}\,\sigma^2(q), \qquad \text{where } \sigma^2(q) \;:=\; \sum_{k=1}^K\frac{p_k^2}{q_k}\sigma_k^2.\end{aligned}$$

Recall also that $\mathrm{Var}\big[\overline{Y}_n\big] = \frac{1}{n}\mathrm{Var}\big[Y\big] = \frac{1}{n}\sigma^2$. Then one compare the value $\sigma^2(q)$ and σ^2.
(i) Proportional allocation: Let $\frac{n_k}{n} =: q_k^* = p_k$, then $\sigma^2(q^*) := \sum_{k=1}^K p_k\sigma_k^2$, and one has

$$\sigma^2 \;=\; \sigma^2(q^*) \;+\; \sum_{k=1}^K p_k\mu_k^2 - \Big(\sum_{k=1}^K p_k\mu_k\Big)^2 \;\ge\; \sigma^2(q^*), \tag{3.9}$$

where the last inequality follows by Jensen's inequality.

Remark 3.3 *Let us define a random variable η by*

$$\eta \;:=\; \sum_{k=1}^K k\mathbf{1}_{Z\in A_k}.$$

Then we have $\mu_k := \mathbb{E}[Y^{(k)}] = \mathbb{E}[Y|\eta = k]$ and $\sigma_k^2 = \mathrm{Var}[Y^{(k)}] = \mathrm{Var}[Y|\eta = k]$. Moreover,

$$\mathbb{E}\big[\mathrm{Var}[Y|\eta]\big] \;=\; \sum_{k=1}^K p_k\sigma_k^2 \quad\textit{and}\quad \mathrm{Var}\big[\mathbb{E}[Y|\eta]\big] \;=\; \sum_{k=1}^K p_k\mu_k^2 - \Big(\sum_{k=1}^K p_k\mu_k\Big)^2.$$

Then the equality (3.9) *can be interpreted as a variance decomposition:*

$$\sigma^2 \;=\; \mathrm{Var}[Y] \;=\; \mathbb{E}\big[\mathrm{Var}[Y|\eta]\big] \;+\; \mathrm{Var}\big[\mathbb{E}[Y|\eta]\big].$$

(ii) Optimal allocation: Let us consider the minimal variance problem

$$\min_q \sum_{k=1}^K \frac{p_k^2}{q_k}\sigma_k^2 \quad \text{subject to} \quad q \ : q_k \geq 0, \ \sum_{k=1}^K q_k = 1.$$

The Lagrange multiplier is given by

$$L(\lambda, q_1, \cdots, q_K) \ := \ \sum_{k=1}^K \frac{p_k^2}{q_k}\sigma_k^2 + \lambda\Big(\sum_{k=1}^K q_k - 1\Big).$$

Then the first order condition gives

$$\frac{\partial L}{\partial q_k} \ = \ -\frac{p_k^2\sigma_k^2}{q_k^2} + \lambda \ = \ 0,$$

which implies that

$$\frac{p_k\sigma_k}{q_k} \ = \ \sqrt{\lambda}, \qquad \forall k = 1, \cdots, K.$$

Thus $q_k = \sqrt{\lambda} p_k \sigma_k$ for all $k = 1, \cdots, K$, and it follows that

$$q_k \ = \ \frac{p_k\sigma_k}{\sum_{i=1}^K p_i\sigma_i}.$$

Application: Let S_t be defined by

$$S_t \ = \ S_0 e^{-\sigma^2/t + \sigma W_t},$$

where S_0 and σ are some positive constant. Denote $X := W_T/\sqrt{T} \sim N(0,1)$, motivated by its application in finance, we usually need to compute the value

$$\mathbb{E}\big[Y\big] \quad \text{with} \quad Y \ := \ f(S_T) \ = \ g(X),$$

for some function $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ (notice that S_T can be expressed as a function of X). Let $\Phi : \mathbb{R} \to (0,1)$ be the distribution function of the Gaussian distribution $N(0,1)$, and take $A_k = (a_k, a_{k+1}]$ for some constant $(a_k)_{1\leq k\leq K+1}$,

$$X^{(k)} \ := \ \Phi^{-1}\big(a_k + (a_{k+1} - a_k)U\big) \qquad Y^{(k)} \ := \ g(X^{(k)}).$$

We then obtain the following algorithm:

Algorithm 3.1 (i) *Choose the sequence of stratification* $(a_k)_{1\leq k\leq K+1}$.
(ii) *For each* $k = 1, \cdots, K$*, simulate a sequence of i.i.d. random variable* $(U_i^k)_{i\geq 1}$ *of uniform distribution* $\mathcal{U}[0,1]$*, and let* $X_i^{(k)} \ := \ \Phi^{-1}\big(a_k + (a_{k+1} - a_k)U_i^k\big)$.

(iii) *Estimate* $\mathbb{E}[Y]$ *by*

$$\sum_{k=1}^{K} \big(a_{k+1} - a_k\big) \Big(\frac{1}{n_k} \sum_{i=1}^{n_k} g(X_i^k)\Big).$$

3.4 Importance sampling method

For the importance sampling, let us begin with a simple example. Suppose that $X \sim N(0,1)$ and $h : \mathbb{R} \to \mathbb{R}$ is some function, then

$$\begin{aligned}
\mathbb{E}\big[h(X)\big] &= \int_{\mathbb{R}} h(x) \frac{1}{\sqrt{2\pi}} e^{x^2/2} dx \quad = \quad \int_{\mathbb{R}} h(x) e^{-x^2/2 + (x-\mu)^2/2} \frac{1}{\sqrt{2\pi}} e^{(x-\mu)^2/2} dx \\
&= \int_{\mathbb{R}} h(x) e^{-\mu x + \mu^2/2} \frac{1}{\sqrt{2\pi}} e^{(x-\mu)^2/2} dx \\
&= \mathbb{E}\Big[h(Y) e^{-\mu Y + \mu^2/2}\Big] \qquad (\text{where } Y \sim N(\mu, 1)) \\
&= \mathbb{E}\Big[h(X+\mu) e^{-\mu X - \mu^2/2}\Big]. \qquad (3.10)
\end{aligned}$$

In some context, we can expect that

$$\mathrm{Var}\Big[h(X)\Big] \quad > \quad \mathrm{Var}\Big[h(X+\mu) e^{-\mu X - \mu^2/2}\Big],$$

then we can use the latter expectation to propose a Monte-Carlo estimator. To deduce the equality (3.10), the main idea is to divide the function $h(x)$ by some density function and then multiple it. We can use this idea in a more general context.

Importance sampling method Let X be a random vector of density function $f : \mathbb{R}^d \to \mathbb{R}^+$ and $h : \mathbb{R}^d \to \mathbb{R}$, the objective is to compute $\mathbb{E}\big[h(X)\big]$. Suppose that there is some other density function $g : \mathbb{R}^d \to \mathbb{R}^+$ such that $g(x) > 0$ for every $x \in \mathbb{R}^d$ such that $f(x) > 0$. Then by direct computation, we have

$$\mathbb{E}\Big[h(X)\Big] \quad = \quad \int_{\mathbb{R}^d} h(x) f(x) dx \quad = \quad \int_{\mathbb{R}^d} h(x) \frac{f(x)}{g(x)} g(x) dx \quad = \quad \mathbb{E}\Big[h(Z) \frac{f(Z)}{g(Z)}\Big],$$

where Z is a random vector of density function g. Then an importance sampling estimator for $\mathbb{E}[h(X)]$ is given, with a sequence $(Z_k)_{k \geq 1}$ of i.i.d. simulations of Y, by

$$\frac{1}{n} \sum_{k=1}^{n} h(Z_k) \frac{f(Z_k)}{g(Z_k)}. \qquad (3.11)$$

Variance analysis Let us compute the variance of the new estimator.

$$\begin{aligned}\mathrm{Var}\Big[h(Z)\frac{f(Z)}{g(Z)}\Big] &= \mathbb{E}\Big[h^2(Z)\frac{f^2(Z)}{g^2(Z)}\Big] - \Big(\mathbb{E}\Big[h(Z)\frac{f(Z)}{g(Z)}\Big]\Big)^2 \\ &= \int_{\mathbb{R}^d} h^2(z)\frac{f^2(z)}{g^2(z)}g(z)dz - \Big(\mathbb{E}\big[h(X)\big]\Big)^2 \\ &= \mathbb{E}\Big[h^2(X)\frac{f(X)}{g(X)}\Big] - \Big(\mathbb{E}\big[h(X)\big]\Big)^2.\end{aligned}$$

And hence the problem of minimizing the variance turns to be

$$\min_g \ \mathrm{Var}\Big[h(Z)\frac{f(Z)}{g(Z)}\Big] \iff \min_g \ \mathbb{E}\Big[h^2(X)\frac{f(X)}{g(X)}\Big]. \tag{3.12}$$

Example 3.4 (i) *Suppose that $h(x) = \mathbf{1}_A(x)$ for some subset $A \subset \mathbb{R}^d$. Then the minimization problem*

$$\min_g \ \mathrm{Var}\Big[h(Z)\frac{f(Z)}{g(Z)}\Big] = \min_g \ \mathrm{Var}\Big[\mathbf{1}_A(Z)\frac{f(Z)}{g(Z)}\Big],$$

is solved by $g(z) := \frac{f(z)\mathbf{1}_A(z)}{\alpha}$, where α is the constant making g a density function.
(ii) *Suppose that h is positive, then the minimization problem* (3.12) *is solved by $g(z) := \frac{1}{\alpha}f(z)h(z)$, where α is the constant making g a density function.*

The above two examples can not be implemented since to make g a density function, we need to choose

$$\alpha := \int_{\mathbb{R}^d} f(z)h(z)dz = \mathbb{E}\big[h(X)\big],$$

which is not known a priori. Therefore, the minimum variance problem (3.12) is not a well posed problem. In practice, we consider a family of density functions $\big(g_\theta(\cdot)\big)_{\theta\in\Theta}$, and then solve the minimum variance problem:

$$\min_{\theta\in\Theta} \mathrm{Var}\Big[h(Z)\frac{f(Z)}{g_\theta(Z)}\Big] \iff \min_{\theta\in\Theta} \ \mathbb{E}\Big[h^2(X)\frac{f(X)}{g_\theta(X)}\Big].$$

Gaussian vector case Let $X = (X_1, \cdots X_n) \sim N(0, \sigma^2 I_n)$, which admits density function

$$f(x_1, \cdots, x_n) := \Pi_{k=1}^n \rho(x_k), \ \text{ with } \rho(x) := \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}.$$

Let $\theta = (\theta_1, \cdots, \theta_n) \in \mathbb{R}^n$, and

$$g_\theta(x_1, \cdots, x_n) := \Pi_{k=1}^n \rho_{\theta_k}(x_k), \ \text{ with } \rho_{\theta_k}(x_k) := \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta_k)^2}{2\sigma^2}}.$$

Then the ratio of the density function turns to be

$$\frac{f(x)}{g_\theta(x)} \;=\; \exp\Big(\frac{-\sum_{k=1}^n \mu_k x_k + \frac{1}{2}\sum_{k=1}^n \mu_k^2}{\sigma^2}\Big).$$

Then

$$\mathbb{E}\big[h(X_1,\cdots,X_n)\big] \;=\; \mathbb{E}\Big[h\big(X_1+\theta_1,\cdots,X_n+\theta_n\big)e^{\left(-\sum_{k=1}^n \mu_k X_k - \frac{1}{2}\sum_{k=1}^n \mu_k^2\right)/\sigma^2}\Big].$$

Example 3.5 *In application in finance, one usually considers a Brownian motion W, and denote $\Delta W_k := W_{t_k} - W_{t_{k-1}}$ on the discrete time grid $0 = t_0 < t_1 < \cdots$ and one needs to compute $\mathbb{E}\big[h\big(\Delta W_1,\cdots,\Delta W_n\big)\big]$ for some function h. Let $X_k = \Delta W_k$, $\sigma^2 = \Delta t$ and $\mu_k = \theta_k/\Delta t$, then*

$$\begin{aligned} & \mathbb{E}\big[h\big(\Delta W_1,\cdots,\Delta W_n\big)\big] \\ = \;& \mathbb{E}\Big[h\big(\Delta W_1+\mu_1\Delta t,\cdots,\Delta W_n+\mu_n\Delta t\big)\exp\Big(-\sum_{k=1}^n \mu_k \Delta W_k - \frac{1}{2}\mu_k^2\Delta t\Big)\Big]. \end{aligned}$$

Chapter 4

Stochastic gradient algorithm

The objective is solve an optimization problem of the form

$$\min_{\theta\in\Theta} \mathbb{E}\Big[F(\theta, X)\Big], \tag{4.1}$$

where $(F(\theta,\cdot))_{\theta\in\Theta}$ is a family of functions.

An iterative algorithm to find the root

Proposition 4.1 *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ a bounded continuous function and $\theta^* \in \mathbb{R}^d$ such that $f(\theta^*) = 0$ and*

$$(\theta - \theta^*) \cdot f(\theta) > 0, \qquad \forall \theta \in \mathbb{R}^d \setminus \{\theta^*\}. \tag{4.2}$$

Let $(\gamma_n)_{n\geq 1}$ be a sequence of numbers satisfying

$$\gamma_n > 0, \ \forall n \geq 1, \quad \textit{and} \ \sum_{n\geq 1} \gamma_n = \infty, \quad \sum_{n\geq 1} \gamma_n^2 < \infty. \tag{4.3}$$

Further, with some $\theta_0 \in \mathbb{R}^d$, define a sequence $(\theta_n)_{n\geq 1}$ by

$$\theta_{n+1} = \theta_n - \gamma_{n+1} f(\theta_n), \quad \forall n \geq 0.$$

Then, $\theta_n \to \theta^*$ as $n \to \infty$.

Proof. (i) First, by its definition, we have

$$\begin{aligned} |\theta_{n+1} - \theta^*|^2 &= |\theta_n - \theta^*|^2 \ + \ 2(\theta_n - \theta^*) \cdot (\theta_{n+1} - \theta_n) \ + \ |\theta_{n+1} - \theta_n|^2 \\ &= |\theta_n - \theta^*|^2 \ - \ 2\gamma_{n+1} f(\theta_n) \cdot (\theta_n - \theta^*) \ + \ \gamma_n^2 |f(\theta_n)|^2 \\ &\leq |\theta_n - \theta^*|^2 \ + \ \gamma_n^2 |f(\theta_n)|^2, \end{aligned}$$

where the last inequality follows by (4.3). Define

$$x_n \;:=\; |\theta_n - \theta^*|^2 \;-\; \sum_{k=1}^{n} \gamma_k^2 |f((\theta_{k-1})|^2.$$

Then the sequence $(x_n)_{n\geq 1}$ is non-increasing. Moreover, it is bounded from below since $x_n \geq -|f|_\infty \sum_{k\geq 1} \gamma_k$. Therefore, there is some x_∞ such that $x_n \searrow x_\infty$ and hence

$$|\theta_n - \theta^*|^2 \;\to\; \ell := x_\infty + \sum_{k\geq 1} \gamma_k^2 |f(\theta_{k-1})|^2.$$

It is clear $\ell \geq 0$ since it is the limit of $|\theta_n - \theta^*|^2$. We claim that $\ell = 0$, which can conclude the proof.

(ii) We now prove $\ell = 0$ by contradiction. Assume that $\ell > 0$, then there is some $N > 0$ such that $\ell/2 \leq |\theta - \theta^*|^2 \leq 2\ell$ for every $n \geq N$. Besides, by the continuity of f and (4.2), we have

$$\eta \;:=\; \inf_{\ell/2 \leq |\theta-\theta^*|^2 \leq 2\ell} (\theta - \theta^*) \cdot f(\theta) \;>\; 0.$$

Therefore,

$$\sum_{n\geq 1} \gamma_n f(\theta_{n=1}) \cdot (\theta_{n-1} - \theta^*) \;\geq\; \sum_{n\geq N} \gamma_n f(\theta_{n=1}) \cdot (\theta_{n-1} - \theta^*) \;\geq\; \eta \sum_{n\geq N} \gamma_n \;=\; \infty.$$

However, we have also

$$\begin{aligned}
\sum_{n\geq 1} \gamma_n f(\theta_{n=1}) \cdot (\theta_{n-1} - \theta^*) \;&=\; -\sum_{n\geq 1} (\theta_n - \theta_{n-1}) \cdots (\theta_{n-1} - \theta^*) \\
&=\; -\frac{1}{2} \sum_{n\geq 1} \Big(|\theta_n - \theta^*|^2 - |\theta_n - \theta_{n-1}|^2 - |\theta_{n-1} - \theta^*|^2 \Big) \\
&=\; \frac{1}{2} \Big(\sum_{n\geq 1} \gamma_n^2 |f(\theta_{n-1})| - \ell + |\theta_0 - \theta^*|^2 \Big) \;<\; \infty.
\end{aligned}$$

This is a contradiction, and hence the claim $\ell = 0$ is true. □

Stochastic gradient algorithm

Theorem 4.1 *In the context of Proposition 4.1, suppose that* $f(\theta) = \mathbb{E}\big[F(\theta, X)\big]$ *for some function* $F : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ *and some random vector* X*. Suppose that* f *satisfies* (4.2) *and a sequence of numbers* $(\gamma)_{n\geq 1}$ *satisfies* (4.3)*. Then, with some* $\theta_0 \in \mathbb{R}^d$ *and a sequence* $(X_n)_{n\geq 1}$ *of i.i.d. simulations of* X*, we define a sequence* $(\theta_n)_{n\geq 1}$ *by*

$$\theta_{n+1} = \theta_n - \gamma_{n+1} F(\theta_n, X_{n+1}). \tag{4.4}$$

Then, $\theta_n \to \theta^*$ *almost surely as* $n \to \infty$*.*

Proof.

□

Application: optimal importance sampling In the context of Section 3.4, we solve the minimum variance problem

$$\begin{aligned} \min_{\mu\in\mathbb{R}} \mathrm{Var}\Big[h(X+\mu)e^{-\mu X-\mu^2/2}\Big] &\iff \min_{\mu\in\mathbb{R}} \mathbb{E}\Big[h^2(X+\mu)e^{-2\mu X-\mu^2}\Big] \\ &\iff \min_{\mu\in\mathbb{R}} \mathbb{E}\Big[h^2(X)e^{-\mu X+\mu^2/2}\Big] \end{aligned} \tag{4.5}$$

Let us denote

$$L(\mu, X) \;:=\; h^2(X)e^{-\mu X+\mu^2/2} \quad \text{and} \quad \ell(\mu) \;:=\; \mathbb{E}\big[L(\mu, X)\big],$$

and

$$F(\mu, X) \;:=\; \frac{\partial L}{\partial \mu}(\mu, X) := (\mu - X)h^2(X)e^{-\mu X+\mu^2/2} \quad \text{and} \quad f(\mu) \;:=\; \mathbb{E}\big[F(\mu, X)\big].$$

Then the minimum variance problem (4.5) is equivalent to find the μ^* such that $f(\mu^*) = \ell'(\mu^*) = 0$. Notice that

$$f'(\mu) \;=\; \ell''(\mu) \;=\; \mathbb{E}\Big[\big(1+(\mu-X)^2\big)h^2(X)e^{-\mu X+\mu^2/2}\Big] \;>\; 0,$$

and hence such a μ^* is separate for f. Therefore, we can use the stochastic gradient algorithm (4.4) to find the optimal μ^*.

Algorithm 4.1 (i) *Simulate a sequence $(X_n)_{n\geq 1}$ of i.i.d. simulations of X.*
(ii) *With $\mu_0 = 0$, use the iteration:*

$$\mu_{n+1} = \mu_n - \gamma_{n+1}F(\mu_n, X_{n+1})$$

(iii) *The estimator of $\mathbb{E}[h(X)]$ is given by*

$$\begin{aligned} \overline{Y}_{n+1} \;&:=\; \frac{1}{n+1}\sum_{k=1}^{n+1}\Big(h(X_k+\mu_{k-1})e^{-\mu_{k-1}X_k-\mu_{k-1}^2/2}\Big) \\ &=\; \frac{n}{n+1}\overline{Y}_n \;+\; \frac{1}{n+1}h(X_{n+1}+\mu_n)e^{-\mu_n X_{n+1}-\mu_n^2/2}. \end{aligned}$$

The advantage of the above algorithm is that one does not need to memorized the simulation $(X_n)_{n\geq 1}$ in the iteration.