Lecture notes on Monte Carlo methods

François Portier

March 30, 2020

Contents

,

Introduction

Let $S \subset \mathbb{R}^d$ be a set and $\mathcal{S}$ be a σ-algebra over S. Let μ denote a non-negative measure on $\mathcal{S}$. Given an integrable function $g : S \to \mathbb{R}$, i.e., g is measurable and $\int |g| \,\mathrm{d}\mu < \infty$, the aim is to study some integration algorithm that returns an approximated value of

$$I_\mu(g) = \int g(x) \,\mathrm{d}\mu(x).$$

Of course when the value of $I_\mu(g)$ might be analytically computed, any integration algorithm is useless. The point is that most of the time, we are not able to compute $I_\mu(g)$ analytically. A classical example is the standard Gaussian density, given by, for every $x \in \mathbb{R}$,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Since we don't know any primitive of ϕ, the computation of $\Phi(y) = \int_{-\infty}^{y} \phi(x) \,\mathrm{d}x$ must be done with the help of approximation methods. The previous example corresponds to the case when μ is the standard Gaussian distribution and $g(x) = \mathbb{I}_{\{x \leq y\}}$. Many examples arise from multidimensional integration problem especially when the integration domain S is not a product of segments. For instance, if $S = \{x \in \mathbb{R}^d \,:\, h(x) \leq 1\}$, $\lambda(S) = \int_S \mathrm{d}x$ might be unknown. In some applications, we do not even have any analytic expression for the integrand φ. In such case, φ is called a "black-box" function, i.e., one can only evaluate the function φ at some points of the domain. For instance, if $t(x)$ denotes the temperature at the location x of the ambient space, then we usually don't have an exact formula. It is then tempting to evaluate the temperature at some points $x_1, \ldots, x_n$, of the domain S, which gives $t(x_1), \ldots, t(x_n)$ and use those evaluation to approximate the average temperature $\int_S t(x) \,\mathrm{d}x$.

When a closed-formula is available for g it might be preferable to rely on some approximations of g based for instance on analytic function theory. For instance, if we know the integral value of $\underline{g}$ and $\overline{g}$, and if we have $\underline{g} \leq g \leq \overline{g}$, where $\int |\underline{g} - \overline{g}| \,\mathrm{d}\mu \leq \epsilon$, then the value of the integral is known with precision $\epsilon/2$ (consider $\tilde{g} = (\overline{g} + \underline{g})/2$).

Consequently, when no analytic expression for the function function g is known or no approximation is sufficiently accurate, we are bounded to the following *Monte Carlo* types of procedures:

1. Choose *randomly* some points, called *nodes* or *particles*, $X_1, \ldots, X_n$ in S, $n \in \mathbb{N}^*$.

2. Evaluate $g(X_1), \ldots, g(X_n)$.

3. Compute an approximation of $I_\mu(g)$ based on $((X_1, g(X_1)), \ldots, (X_n, g(X_n)))$.

In this course, we only focus on the case when g is evaluated exactly, i.e., without any noise. Hence we focus on the first step and the last step. Moreover, we shall only be concerned by stochastic integration methods, also called Monte Carlo methods, where the points $X_1, \ldots, X_n$ are random variables. In contrast, deterministic methods consider point grids that are fixed by the user and the main differences between both approaches should be given in the first chapter.

Monte Carlo integration methods are widely used in many domains of sciences including Biology, Physics, Economics and Finance. For instance, in Physics sensors are often used to measure some physical attribute (e.g., the temperature) at different locations (hence g is unknown). In Finance, option prices are expectations under the risk-neutral measure (Hull and White, 1988). Other important fields of application are Statistics and Machine-Learning where many algorithms are calibrated using Monte Carlo methods. It includes for instance stochastic gradient descent (Wang et al., 2013), *anomaly detection*, *bootstrap resampling methods* and the branch of *Bayesian statistics* (Robert and Casella, 2004; Oates et al., 2017) where complex models often provides expectations analytically intractable.

Running any Monte Carlo algorithm is associated to some computational time. The computation time can be expressed in terms of elementary operations needed to produce the approximated value of the integral. In some cases, for each $x \in S$, the evaluation of $g(x)$ can be given by a single elementary operation. In some other cases, the evaluation of g is heavy. The same can be stated concerning the generation of random variables according to μ. The previous rules depend of course on the context. For instance, generating random variables from a particular distribution can involve some difficulty, e.g., rejection methods, Metropolis-Hastings algorithm or computing g might take more than a single elementary operation, e.g., crash test simulation.

We shall have a particular interest in the following aspects:

- The analysis of the integration error regarding the number of nodes. The estimation of the integration error.
- The adaptation to the function to integrate.
- The computation time of the algorithms.

Chapter 1

The Monte Carlo method

In this chapter, we introduce and study the classical Monte Carlo method. It is the occasion to recall some basic probability concepts on asymptotic convergence. These concepts shall be useful in the other chapters.

1.1 Definition of Monte Carlo and basic properties

Let $(S, \mathcal{S}, \mu)$ be a probability space. The Monte Carlo method is dedicated to approximate quantities of the form $I_\mu(g) = \int g \,\mathrm{d}\mu$ where g is an integrable function with respect to μ. The Monte Carlo method follows from the law of large numbers, i.e., whenever $X_1, \ldots, X_n$ is an independent and identically distributed (i.i.d.) sequence of random variables with common distribution μ, we have that, almost surely,

$$n^{-1} \sum_{i=1}^{n} g(X_i) \to \int g \,\mathrm{d}\mu, \quad \text{as } n \to \infty.$$

The Monte Carlo algorithm is as follows.

Algorithm 1 (Monte Carlo)**.**
Input*: the sample number $n \in \mathbb{N}^*$.*

(i) Let $n \in \mathbb{N}^*$. Generate $X_1, \ldots, X_n$ independently with common distribution μ.

(ii) Compute

$$\hat{I}_n^{(\mathrm{mc})}(g) = n^{-1} \sum_{i=1}^{n} g(X_i).$$

The first properties associated to the Monte Carlo method are listed in the following proposition and results from basic calculations and the strong law of large numbers.

Proposition 1.1.1. *Suppose that $\int |g| \,\mathrm{d}\mu < \infty$, then*

- *$\hat{I}_n^{(\mathrm{mc})}(g)$ is an unbiased estimator of $I_\mu(g)$,*
- *$\hat{I}_n^{(\mathrm{mc})}(g)$ is strongly consistent estimating $I_\mu(g)$.*

Based-on the central limit theorem, we can obtain the rate of convergence of $\hat{I}_n^{(\mathrm{mc})}$ as claimed in the following proposition.

Proposition 1.1.2. *Suppose that $\int |g|^2 \,\mathrm{d}\mu < \infty$, then*

- $\mathrm{var}(\hat{I}_n^{(\mathrm{mc})}(g)) = n^{-1}\sigma_\mu^2(g)$ *where* $\sigma_\mu^2(g) = \mathrm{var}(g(X_1))$,
- *the random sequence* $n^{1/2}(\hat{I}_n^{(\mathrm{mc})}(g) - I_\mu(g))$ *converges in distribution to* $\mathcal{N}(0, \sigma_\mu^2(g))$.

1.2 Estimation error

1.2.1 Asymptotic confidence intervals

We have just seen that the variance associated to Monte Carlo is $n^{-1}\sigma^2$. In order to have an idea (through confidence interval) of the accuracy of Monte Carlo, it is tempting to estimate the variance. The most classical estimator of $\sigma^2(g)$ is given by

$$\hat{\sigma}_n^2(g) = (n-1)^{-1}\sum_{i=1}^n (g(X_i) - \hat{I}_n(g))^2.$$

Using some algebra and Slutsky's Lemma, one can extend Proposition 1.1.1 to the analysis of $\hat{\sigma}_n^2(g)$.

Proposition 1.2.1. *Suppose that* $\int |g|^2 \,\mathrm{d}\mu < \infty$, *then*

- $\hat{\sigma}_n^2(g)$ *is an unbiased and strongly consistent estimator of* $\sigma^2(g)$,
- *the random sequence* $\left(n^{1/2}/\hat{\sigma}_n(g)\right)(\hat{I}_n^{(\mathrm{mc})}(g) - I_\mu(g))$ *converges in distribution to* $\mathcal{N}(0,1)$.

A consequence of the last point is that we are able to build asymptotically consistent confidence intervals. Indeed defining, for every $\alpha \in (0,1)$,

$$\hat{C}(\alpha) = \left[\hat{I}_n^{(\mathrm{mc})}(g) - \left(\frac{\hat{\sigma}_n(g)}{n^{1/2}}\right)\Phi^-(1-\alpha/2), \hat{I}_n^{(\mathrm{mc})}(g) - \left(\frac{\hat{\sigma}_n(g)}{n^{1/2}}\right)\Phi^-(\alpha/2)\right],$$

some algebra gives that

$$\mathbb{P}\left(I_\mu(g) \in \hat{C}(\alpha)\right) \to 1-\alpha, \quad \text{as } n \to \infty.$$

1.2.2 Concentration inequalities

The limitation of the previous approach is that the confidence intervals are based on the asymptotic distribution. If for a given $n \in \mathbb{N}^*$, one prefers to obtain confidence intervals for which the probability that the true value of the integral lies in the interval truly larger than $1-\alpha$, one should rely on inequalities valid at finite sample size. The most basic of such inequalities, follows from Markov's inequality, i.e., for any random vector X, any number $k \geq 1$ and $\epsilon > 0$, $\mathbb{P}(|X| > \epsilon) \leq \mathbb{E}[|X|^k]/\epsilon^k$. The previous applied with $k = 2$ and $X = \hat{I}_n^{(\mathrm{mc})}(g) - I_\mu(g)$, gives that for any $\epsilon > 0$ and any $n \geq 1$,

$$\mathbb{P}(|\hat{I}_n^{(\mathrm{mc})}(g) - I_\mu(g)| > \epsilon) \leq \frac{\sigma_\mu^2(g)}{n\epsilon^2}.$$

It follows that

$$\hat{C}_2(\alpha) = \left[\hat{I}_n^{(\mathrm{mc})}(g) - \left(\frac{\sigma_\mu(g)}{n^{1/2}}\right)\frac{1}{\sqrt{\alpha}}, \hat{I}_n^{(\mathrm{mc})}(g) + \left(\frac{\sigma_\mu(g)}{n^{1/2}}\right)\frac{1}{\sqrt{\alpha}}\right].$$

Then we have

$$\mathbb{P}(I_\mu(g) \in \hat{C}_2(\alpha)) = 1 - \mathbb{P}\left(|\hat{I}_n^{(\mathrm{mc})}(g) - I_\mu(g)| > \left(\frac{\sigma_\mu(g)}{n^{1/2}}\right)\frac{1}{\sqrt{\alpha}}\right) \geq 1-\alpha.$$

Following the previous approach the resulting interval is generally larger than the asymptotic confidence interval ($1/\sqrt{\alpha} > \Phi^{-}(1-\alpha/2)$ for all $\alpha \in (0,1)$) and requires to know the variance or a bound for the variance. Applications include the case where $g(x) = 1_{\{\mathcal{S}\}}(x)$ for which $\sigma_\mu^2(g) = I_\mu(g)(1 - I_\mu(g)) \leq 1/4$. In such a case, we obtain the interval

$$\left[\hat{I}_n^{(\mathrm{mc})}(g) - \left(\frac{1}{2\sqrt{n\alpha}}\right), \hat{I}_n^{(\mathrm{mc})}(g) + \left(\frac{1}{2\sqrt{n\alpha}}\right)\right].$$

A more accurate approach for bounded random variables relies on Hoeffding's inequality, stated in the following theorem.

Theorem 1.2.2. *Let $X_1, \ldots, X_n$ be independent real-valued random variables such that for all $1 \leq i \leq n$, $a \leq X_i \leq b$ almost surely, then*

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| > \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right).$$

Proof. We admit that for any random variable $Y \in \mathbb{R}$, such that $c \leq Y \leq d$ almost surely and $\mathbb{E}Y = 0$, we have for all $s > 0$,

$$\mathbb{E}[e^{sY}] \leq e^{s^2(d-c)^2/8}$$

Then using that for any $\epsilon > 0$, $x \in \mathbb{R}$, $s > 0$, it holds that $1_{\{x > \epsilon\}} \leq e^{(sx - s\epsilon)}$ we find

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \epsilon\right) &\leq \mathbb{E}\left[e^{(s(\sum_{i=1}^n (X_i - \mathbb{E}[X_i])) - s\epsilon)}\right] \\ &= e^{-s\epsilon}\prod_{i=1}^n \mathbb{E}\left[e^{s((X_i - \mathbb{E}[X_i]))}\right] \\ &\leq e^{ns^2(b-a)^2/8}e^{-s\epsilon},\end{aligned}$$

where the last line is because $d - c = b - a$ for $Y = X_i - \mathbb{E}[X_i]$. The minimum in s is achieved when $s^* = 4\epsilon/(n(b-a)^2)$, leading to the bound $e^{-\frac{2\epsilon^2}{n(b-a)^2}}$. Conclude by considering the event $\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) < -\epsilon$. □

Following the previous, we have that

$$\hat{C}_2(\alpha) = \left[\hat{I}_n^{(\mathrm{mc})}(g) - \sqrt{\frac{(b-a)^2\log(2/\alpha)}{2n}}, \hat{I}_n^{(\mathrm{mc})}(g) + \sqrt{\frac{(b-a)^2\log(2/\alpha)}{2n}}\right].$$

In the case where $g(x) = 1_{\{S\}}(x)$, we obtain

$$\left[\hat{I}_n^{(\mathrm{mc})}(g) - \sqrt{\frac{\log(2/\alpha)}{2n}}, \hat{I}_n^{(\mathrm{mc})}(g) + \sqrt{\frac{\log(2/\alpha)}{2n}}\right].$$

1.2.3 Illustration

We consider the estimation of $\lambda(S)$ where S is the set of points in $(x, y) \in \mathbb{R}^2$ such that

$$h(x,y) = \left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2 \leq \frac{1}{4}.$$

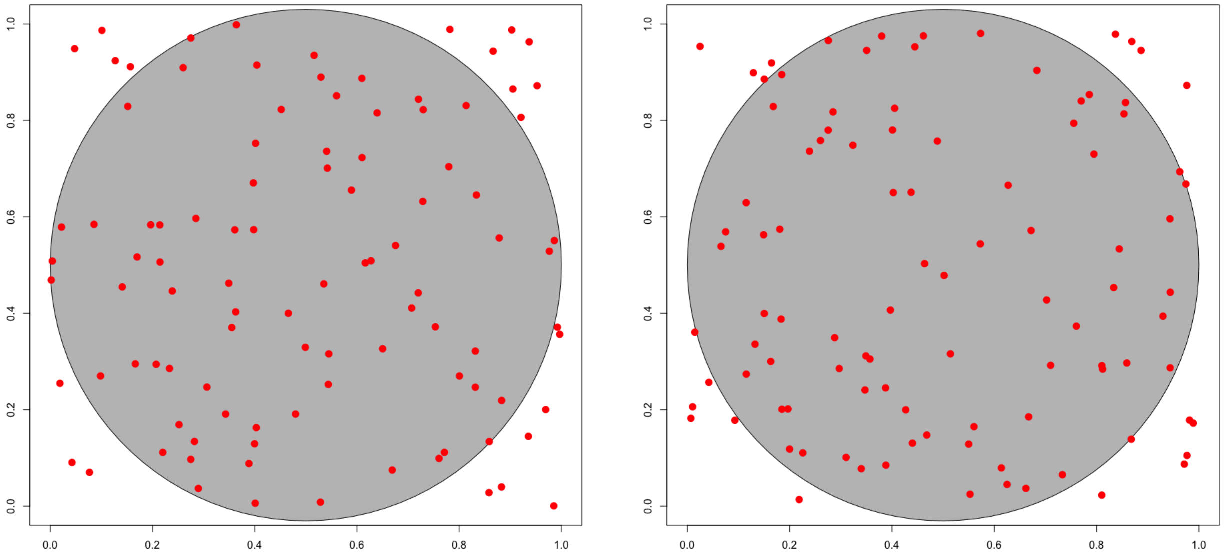

Figure 1.1: In red are the points of two i.i.d. samples of size 100 with comon distribution $\mathcal{U}[0,1]^2$. In grey is the unit sphere with center $(.5,.5)$.

This could be generalized to any function h. We generate independent random variables $X_1,\ldots,X_n$ with common distribution $\mathcal{U}[0,1]^2$. Then we compute the Monte Carlo estimator of

$$\lambda(S) = \int 1_{\{h(x,y)\le 1/4\}} \frac{\mathrm{d}x\mathrm{d}y}{\lambda([0,1]^2)},$$

given by

$$\hat{I}_n^{(\mathrm{mc})}(S) = n^{-1}\sum_{i=1}^{n} 1_{\{h(X_i)\le 1/4\}}.$$

For the two samples of size 100 presented in Figure 1.1 and for the three methods discussed previously, asymptotic, Markov and Hoeffding (for Markov we used the bound $\sigma^2 = 1/4$), we give the confidence intervals with level $\alpha = 0.05$,

Asymptotic $\hat{C}(0.05)$	$[0.68, 0.84]$	$[0.74, 0.90]$
Markov $\hat{C}_1(0.05)$	$[0.54, 0.98]$	$[0.60, 1.04]$
Hoeffding $\hat{C}_2(0.05)$	$[0.62, 0.90]$	$[0.68, 0.96]$.

The true value $\lambda(S) = \pi/4 \simeq 0.79$.

1.3 The effect of the dimension

Monte Carlo procedures are often promoted over deterministic methods when the dimension exceeds 3. In high dimension, deterministic methods become hard to compute and their accuracies deteriorate quickly with the dimension. In the literature, this is often referred to as the *curse of dimensionality*. To illustrate our point, we consider the Riemann's sums method, which estimate $I_\mu(g) = \int_{[0,1]^d} g(x)\,\mathrm{d}x$ by

$$I_n^{(\mathrm{rs})}(g) = n^{-d} \sum_{x \in G} g(x),$$

where the summation is over x in the grid $G = \{(i_1, \ldots, i_d) \,:\, 1 \leq i_k \leq n, \forall k = 1, \ldots, d\}$. We have the following proposition.

Proposition 1.3.1. *Suppose that $g : [0,1]^d \to \mathbb{R}$ is a Lipschitz function, i.e.,*

$$|g(x) - g(y)| \leq L\|x - y\|, \quad \forall (x, y) \in [0,1]^d \times [0,1]^d,$$

then

$$|I_n^{(rs)}(g) - I_\mu(g)| \leq \frac{L\sqrt{d}}{n}.$$

Proof. Define for any $(i_1, \ldots, i_d) \in G$, the rectangle $R_{(i_1,\ldots,i_d)} = [(i_1 - 1)/n, i_1/n) \times \ldots \times [(i_d - 1)/n, i_d/n)$ and write

$$I_\mu(g) = \sum_{x \in G} \int_{R_x} g(x)\,\mathrm{d}x.$$

Then use the Lipschitz property to conclude. □

The natural competitor of the previous method is the following Monte-Carlo procedure with n^d generated random variables (assuming that generating according to $\mathcal{U}[0,1]^d$ is of the same order as evaluating g and represents one single operations). Such an estimator satisfies

$$\mathrm{var}(\hat{I}_{n^d}^{(\mathrm{mc})}(g)) = \frac{1}{n^d}\sigma_\mu^2(g).$$

Consequently,

$$\mathbb{E}[|\hat{I}_{n^d}^{(\mathrm{mc})}(g) - I_\mu(g)|] \leq \frac{\sigma_\mu(g)}{n^{d/2}}.$$

More generally, using a grid of n^d nodes, deterministic methods such as Riemann sums or Gaussian quadrature reach an accuracy of order n^{-s} (Novak, 2016, Theorem 1), where s stands for the regularity of the integrand. Monte Carlo methods are, in contrast, subjected to an optimal error bound of order $n^{-s}n^{-d/2}$ (Novak, 2016, Theorem 3). For instance, as demonstrated before, the naive Monte Carlo method, which does not use any regularity of the integrand, has an expected error bound of order $n^{-d/2}$.

Exercises

Exercise 1.3.1 (generation by cdf inversion).

1. *Let F be a cumulative distribution function on $\mathbb{R}$. We define the generalized inverse of F, for any $u \in (0,1)$,*

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

Show that for any $x \in \mathbb{R}$, $u \in (0,1)$, $F^{-}(u) \leq x$ iff $u \leq F(x)$. Deduce that if $U \sim \mathcal{U}[0,1]$, $F^{-}(U)$ has the same distribution as F.

2. *Propose a method to generate some random numbers with exponential distribution $\mathcal{E}(\lambda)$.*

3. *Let X be a real random variable and F its cumulative distribution function. Show that for any $a < b$ such that $F(b) - F(a) > 0$,*

$$F^{-1}\left(F(a) + U(F(b) - F(a))\right), \qquad \text{where} \qquad U \sim \mathcal{U}([0,1]),$$

is a random variable with distribution $\mathbb{P}(X \in \cdot | X \in]a, b])$.

4. *Write an algorithm based on the inversion method to sample from the truncated Gaussian distribution $\propto \exp(-0.5x^2)1_{[a,\infty)}(x)$*

Exercise 1.3.2 (generation by rejection sampling)**.** *a) Let f and g be two densities on the real line such that, for all $x \in \mathbb{R}$, $f(x) \leq cg(x)$ for some $c > 0$. Let Y be distributed according to g and U be uniformly distributed on $[0,1]$. Show that the conditional distribution of Y given $U \leq f(Y)/cg(Y)$ has density f.*

b) For each sample generated, what is the probability to reject? Discuss the limit of such an approach.

c) Let $Z \sim \mathcal{E}(1/2)$ (exponential distribution with parameter $1/2$) and $a > 0$. Show that the density of $\sqrt{a^2 + Z}$ is $x \exp(-0.5(x^2 - a^2))1_{[a,\infty)}(x)$.

d) Propose an (envelope) rejection algorithm to sample from the truncated Gaussian distribution $\propto \exp(-0.5x^2)1_{[a,\infty)}(x)$?

Exercise 1.3.3. *Let $f : [0,1] \to [0,1]$ a mesurable function such that $\int_{[0,1]} f \mathrm{d}\lambda < \infty$. Our goal is to compute $\theta = \int_0^1 f(x)\,\mathrm{d}x$. Let $U_1, \dots, U_n$ be a sequence of i.i.d. random variables distributed according to $\mathcal{U}[0,1]$.*

(a) Express $\hat{\theta}_n^{(1)}$, the Monte Carlo estimator based on (U_i), and compute its variance.

(b) Let $g(x,y) = 1\{y \leq f(x)\}$. Express $\int_0^1 g(x,y)\,\mathrm{d}y$ with the help of $f(x)$.

(c) We call U-statistique with kernel h any quantity of the type

$$\frac{1}{n(n-1)} \sum_{i \neq j} h(U_i, U_j).$$

Using the sequence (U_i), define a U-statistic $\hat{\theta}_n^{(2)}$ which is unbiased for $\int_0^1 f(x)\,\mathrm{d}x$. Verify it is unbiased.

Chapter 2

Importance sampling

Importance sampling relies on a simple change of measure. Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is a density function whose support is $S_f \subset \mathbb{R}^d$. Let X be a random variable whose distribution admits a density q (with respect to the Lebesgue measure) which support contains S_f. Then for any measurable function g, we have

$$\int g(x) f(x) \, \mathrm{d}x = \mathbb{E}_q \left[\frac{f}{q} g \right].$$

A Monte-Carlo estimator is then introduced to estimate the expectation that appears in the right-hand side. In this section, a particular interest should be dedicated to the optimal choice of q.

2.1 Change of measure

Let μ_f and μ_q be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Denote by $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $q : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ their respective density with respect to the Lebesgue measure. Hence, for any $A \in \mathcal{B}(\mathbb{R}^d)$, we have

$$\mu_f(A) = \int_A f d\lambda \qquad \text{and} \qquad \mu_q(A) = \int_A q d\lambda.$$

The following concept of dominated measures will be useful.

Definition 2.1.1. *The measure μ dominates ν, $\mu \gg \nu$, whenever for all $A \in \mathcal{B}(\mathbb{R}^d)$, $\mu(A) = 0$ implies that $\nu(A) = 0$.*

As soon as μ_q dominates μ_f, a change of measure will be valid as explained in the following proposition. For that, we introduce the importance function $w : \mathbb{R}^d \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ defined as

$$w(x) = \frac{f(x)}{q(x)}, \qquad x \in \mathbb{R}^d.$$

Proposition 2.1.1. *The following points are equivalent:*

(i) $\mathbb{E}_q[w] = 1$

(ii) $\mu_q \gg \mu_f$

(iii) for all measurable positive function g, $\mathbb{E}_q[wg] = \int g f d\lambda$

Proof. We have $\mathbb{E}[w] = \int_{q>0} f d\lambda = 1 - \int_{q=0} f d\lambda$. Hence (i) is equivalent to

$$\int_{q=0} f d\lambda = \mu_f(q = 0) = 0.$$

Since any A such that $\mu_q(A) = 0$ is contained in the set $q = 0$ we have that (i) implies (ii). Suppose now that (ii) holds and take $A = \{q = 0\}$. we have that $\mu_q(A) = 0$ and hence $0 = \mu_f(A) = \int_{q=0} f d\lambda$. The previous equation has been shown to be equivalent to (i). We finish the proof showing that (i) holds if and only if (iii) holds. The if part is obvious. For the only if part, we have that $\int_{q=0} f = 0$. Then for any measurable set A, we have $\int_{A,q=0} f d\lambda = 0$ which implies that $\mathbb{E}[w\mathbb{I}_A] = \int_{q>0} f\mathbb{I}_A d\lambda = \mu_f(A)$. As a consequence for any *simple function* f, (iii) holds. This can be extended to any positive function as follows. First, approximate f by f_n, a sequence of nondecreasing simple functions and then invoke the monotone convergence theorem. □

2.2 Importance sampling

Suppose that $g : \mathbb{R}^d \to \mathbb{R}$ is such that $\int |g| f < +\infty$. We are interested in estimating

$$I_f(g) = \int gf \,\mathrm{d}\lambda.$$

From the previous proposition, if q is a density with respect to the Lebesgue measure such that $q \gg f$. Then it holds that

$$I_f(g) = \mathbb{E}_q\left[gw\right].$$

Importance sampling follows from applying the Monte Carlo principle the the previous expectation. Let $X_1, ..., X_n$ be an i.i.d. sequence with common density q,

$$\hat{I}_n^{(\mathrm{is})} = n^{-1} \sum_{i=1}^{n} w(X_i) g(X_i)$$

The distribution associated to q is usually called the *sampling distribution*, the *sampler* or the *proposal.* The following result is simple consequence of *strong law of large numbers.*

Proposition 2.2.1. *Suppose that* $\int |g| f \,\mathrm{d}\lambda < \infty$ *and that* $q \gg f$*, then*

$$\lim_{n\to\infty} \hat{I}_n^{(\mathrm{is})} = I_f(g).$$

Define the variance

$$r_q^2(g, f) = \int \left(\frac{g(x)f(x)}{q(x)} - I_f(g)\right)^2 q(x)\,\mathrm{d}x$$

and its empirical estimate

$$\hat{r}_n^2 = n^{-1} \sum_{i=1}^{n} \left(w(X_i)g(X_i) - \hat{I}_n^{(\mathrm{is})}\right)^2.$$

The result that follows is a simple consequence of the *central limit theorem* and Slutsky's lemma.

Proposition 2.2.2. *Suppose that* $\int |g| f \,\mathrm{d}\lambda < \infty$ *and that* $q \gg f$*, if moreover,* $\int g(x)^2 f(x)^2 / q(x)\,\mathrm{d}x < \infty$*, then*

$$n^{1/2}(\hat{I}_n^{(\mathrm{is})} - I_f(g)) \rightsquigarrow \mathcal{N}(0, r_q^2(g, f))$$

and

$$(n/\hat{r}_n)^{1/2}\,(\hat{I}_n^{(\mathrm{is})} - I_f(g)) \rightsquigarrow \mathcal{N}(0, 1).$$

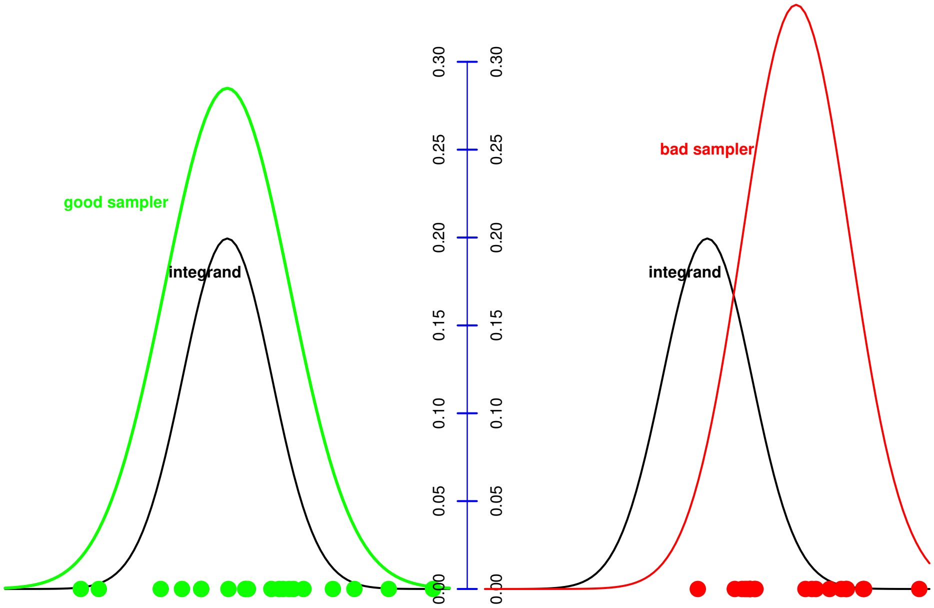

Figure 2.1: The two above samplers are likely to provide different results in estimating $I_f(g)$.

Example 2.2.1. *Suppose that $g = 1$ and f is the standard Gaussian density and that the sampler q is the density of the distribution $\mathcal{N}(\theta, 1)$. Then $I_f(g) = 1$ and*

$$
\begin{aligned}
r_q^2(g, f) + 1 &= \int \frac{f^2}{q} \\
&= \frac{1}{\sqrt{2\pi}} \int \frac{e^{-x^2}}{e^{-\frac{(x-\theta)^2}{2}}} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{-x^2-2x\theta-\theta^2+2\theta^2}{2}} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{-(x+\theta)^2}{2}+\theta^2} \mathrm{d}x \\
&= e^{\theta^2}
\end{aligned}
$$

The previous example is rather important as it permits to understand the two extreme cases of importance sampling:

1. if $\theta = 0$, i.e. $q = f$. Then $r_q^2(1, f) = 0$. More generally, when gf is not a density but a non-negative function, then taking $q \propto gf$ shall imply a variance equal to 0. Such a choice is not possible as it requires that $q = \frac{gf}{\int gf}$, i.e., the knowledge of $\int gf$, the quantity we are looking for.

2. If $\theta \gg 1$, then $r_q^2(1, f) \gg 1$. The observations $X_1, ..., X_n$ often falls in non interesting part of S_f, where f is small. This could lead to a very poor estimation of $I_f(1)$.

2.3 Minimum variance

The question raised in this section is the one of variance optimality: does it exist an optimal sampler q that would minimize the variance $r_q^2(g, f)$? We start by writing

$$r_q^2(g) = \int (gf)^2/q d\lambda - I_f(g)^2.$$

As the quantity in the right does not depend on q, what really matter when minimizing the variance is the value of $\int (gf)^2/q d\lambda$. Consequently, a key quantity thereafter is

$$C_\varphi(q) = \int \varphi^2/q d\lambda.$$

Lemma 2.3.1 (variance optimality)**.** *Let φ be a measurable function such that $0 < \int |\varphi| < \infty$. The minimum of C_φ over the set of densities q is achieved if and only if $q = |\varphi| / \int |\varphi| d\lambda$ a.e. and*

$$C_\varphi(q^*) = \left(\int |\varphi| d\lambda \right)^2.$$

Proof. Let $q^* = |\varphi| / \int |\varphi| d\lambda$. If q does not dominate q^* then $\int q^{*2}/q = +\infty$. If it does, using the Cauchy-Schwarz inequality, we obtain $1 = (\int q^*)^2 = (\int (q^*/\sqrt{q})\sqrt{q})^2 \leq \int q^{*2}/q$. As a consequence

$$\left(\int |\varphi| d\lambda \right)^2 \leq C_\varphi(q).$$

From this we deduce that $q = q^*$ is an argmin. If now q is such that $\int q^{*2}/q = 1$, then equality holds in the Cauchy-Schwarz inequality meaning that $q^* = \kappa q$ a.e. with $\kappa > 0$. But κ needs to be 1 because q and q^* are densities. □

Define the density

$$q^* = |g|f / \int |g| f d\lambda.$$

Proposition 2.3.2 (variance optimality)**.** *Let g be a measurable function such that $0 < \int |g| f < \infty$. The minimum of $r_q(g, f)$ over the set of densities q is achieved if and only if $q = q^*$ a.e. and*

$$r_{q^*}(g, f) = \left(\int |g| f d\lambda \right)^2 - \left(\int g f d\lambda \right)^2 .$$

Hence we have the following conclusion depending on the sign of g:

- If g changes its sign on a non-zero measure set, then it is not possible to reduce the variance to 0.
- Else choosing $q^* \propto |g|f$ gives a 0 variance.

2.4 Two-stage importance sampling

2.4.1 A parametric family of sampler

According to the previous section, a first way to proceed is to represent the function q^* on a graph and then to select the sampler according to this representation. Such an approach remains very limited as it will certainly fail in high dimension where the patterns of q^* are difficultly observable on a graph. Another strategy is to obtain an approximation of the best sampler among a parametric family. Hence we shall use some simulations to approximate the variance and then optimize it to get a parametric estimate of q^*. This approach is presented in Section 2.4.2. Another way to proceed, presented in Section 2.4.3 is to rely on the Kullback-Leibler (KL) divergence. We stress that the KL approach is more general than the variance approach in the sense that it works for any target density (not necessarily q^*) and especially densities that does not depend on g. This is of particular interest when several functions g need to be integrated. This suits well the Bayesian context where one is interested in computing $\mathbb{E}_P[X]$ where P stands for the *distribution a posteriori* and whose density is usually known up to a scale factor. Finally, in the same spirit as the KL approach, another approach follows from the *method of moments*. This is presented in Section 2.4.4.

In the following, we let

$$\mathcal{Q} = \{q_\theta \, : \, \theta \in \Theta\},$$

where $\Theta \subset \mathbb{R}^q$ and for each $\theta \in \Theta$, q_θ is a density with respect to the Lebesgue measure. We assume further on that for every $\theta \in \Theta$, $q_\theta \gg f$. Define the importance weights associated to θ as

$$w_\theta(x) = \frac{f(x)}{q_\theta(x)}, \qquad x \in \mathbb{R}^d.$$

For g integrable with respect to f, the fact that q_θ dominates f ensures the unbiasedness and the consistency of $n^{-1} \sum_{i=1}^n g(X_i) w_\theta(X_i)$ for each $\theta \in \Theta$.

2.4.2 The variance criterion

Let

$$\theta^* \in \operatorname{argmin}_{\theta \in \Theta} r_{q_\theta}^2(g, f) = \arg\min_{\theta \in \Theta} \int \frac{g^2 f^2}{q_\theta} \mathrm{d}\lambda.$$

A first problem is that, generally, we cannot evaluate the function $\theta \mapsto \int (g^2 f^2 / q_\theta) \, \mathrm{d}\lambda$ as it is an integral involving g and f. A variance estimator of the same type as $\hat{r}_n$ is not suitable as it would require to draw some points with respect to q_θ for each θ. We define in the following a simulation based estimator of θ^*. This shall give us a near-optimal $q_{\hat{\theta}_n}$ from which we are going to apply an importance sampling procedure as detailed in the introduction.

Algorithm 2 (Importance sampling *via* variance minimization).
Input: *the sample number $n \in \mathbb{N}^*$, the parametric family $\mathcal{Q}$, the initial density q_0.*

(i) Let $1 < n_1 < n$. Generate $X_1, \ldots, X_{n_1}$ independently with common distribution q_0. Compute $\hat{\theta}_1$ as the minimizer over $\theta \in \Theta$ of

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{(gf)(X_i)^2}{q_\theta(X_i) q_0(X_i)}.$$

(ii) Let $n_2 = n - n_1$. Generate $Z_1, \ldots, Z_{n_2}$ according to $q_{\hat{\theta}_1}$, and compute

$$n_2^{-1} \sum_{i=1}^{n_2} g(Z_i) w_{\hat{\theta}_1}(Z_i).$$

2.4.3 The Kullback-Leibler approach

In what follows, the target density is f. Define the likelihood function

$$L(\theta) = \int \log(q_\theta / f) f \, \mathrm{d}\lambda.$$

If the model $\mathcal{Q}$ is identifiable, i.e., $q_\theta = q_{\theta^*}$ almost everywhere implies that $\theta = \theta^*$, and if f belongs to $\mathcal{Q}$, it is well known that L is uniquely maximized at θ^* and that $q_{\theta^*} = f$ (van der Vaart, 1998). The proof is based on the use of the inequality $\log(x) \leq 2(\sqrt{x} - 1)$, $\forall x > 0$, to obtain the following inequality, which involves the Hellinger distance,

$$L(\theta) \leq - \int (\sqrt{q}_\theta - \sqrt{q})^2 \, \mathrm{d}\lambda.$$

Moreover if $f = \tilde{f} c$, we find

$$L(\theta) = c \left\{ \int (\log(q_\theta / \tilde{f}) - \log(c)) |\tilde{f}| \, \mathrm{d}\lambda \right\} = c \int \log(q_\theta / \tilde{f}) \tilde{f} \, \mathrm{d}\lambda - \log(c).$$

Consequently, to maximize L is equivalent to maximize $\int \log(q_\theta / \tilde{f}) \tilde{f} \, \mathrm{d}\lambda$. This can be estimated empirically in the same way as the variance in the previous section. We obtain the following procedure.

Algorithm 3 (Importance sampling *via* KL).
Input: *the sample number $n \in \mathbb{N}^*$, the parametric family $\mathcal{Q}$, the initial density q_0.*

(i) Let $1 < n_1 < n$. Generate $X_1, \ldots, X_{n_1}$ independently with common distribution q_0. Compute $\hat{\theta}_1$ as the maximizer over $\theta \in \Theta$ of

$$n_1^{-1} \sum_{i=1}^{n_1} \log \left(\frac{q_\theta(X_i)}{f(X_i)} \right) \frac{f(X_i)}{q_0(X_i)}.$$

(ii) Let $n_2 = n - n_1$. Generate $Z_1, \ldots, Z_{n_2}$ according to $q_{\hat{\theta}_1}$, and compute

$$n_2^{-1} \sum_{i=1}^{n_2} g(Z_i) w_{\hat{\theta}_1}(Z_i).$$

2.4.4 Generalized method of moments

The generalized method of moments (GMM) consists in minimizing a certain Euclidean distance, say $\| \cdot \|$, between the empirical moments associated to f, the targeted density, and the theoretical ones computed according to q_θ. The moments are computed according to the so called moment function $h : \mathbb{R}^d \to \mathbb{R}^q$. Applied to our problem and using the same notation as before, it consists in minimizing

$$\theta \mapsto \| \sum_{i=1}^{n_1} w_{i,n} h(X_i) - \int h q_\theta \|^2$$

where

$$w_{i,n} \propto \frac{f(X_i)}{q_0(X_i)} \qquad \text{s.t.} \qquad \sum_{i=1}^{n_1} w_{i,n} = 1.$$

Note that $\sum_{i=1}^{n_1} w_{i,n} h(X_i)$ is an estimate of $\int hf$. We have that

$$\hat\theta_{n_1} \in \operatorname{argmin}_{\theta \in \Theta} \sum_{i=1}^{n_1} \frac{f(X_i)}{q_0(X_i)} \left\| h(X_i) - \int h q_\theta \right\|^2 .$$

This can be shown by starting from the latter equation and writing $h(X_i) - \int hq_\theta = h(X_i) - \overline{w_n h} + \overline{w_n h} - \int hq_\theta$.

2.4.5 A unified view

Each of the approaches (variance, KL and GMM) is interested in estimating

$$\psi(\theta) = \int m_\theta \, \mathrm{d}\lambda,$$

with m_θ corresponding to, respectively,

$$(g^2 f^2 / q_\theta), \quad \log(q_\theta) f, \quad \|g - E_\theta(h)\|^2 f.$$

In each case, a natural estimator of ψ is given by

$$\hat\psi_{n_1}(\theta) = n_1^{-1} \sum_{i=1}^{n_1} \frac{m_\theta(X_i)}{q_0(X_i)},$$

where (X_i) are i.i.d. radnom variables generated according to q_0. The proof's heuristic that lies behind each approach, variance, KL and GMM, is as follows. From the law of large number,

$$|\hat\psi_{n_1}(\theta) - \psi(\theta)| \to 0 \qquad \text{as } n_1 \to \infty.$$

Therefore, we can expect that the minimizer of $\hat\psi_{n_1}$ will converge, as $n_1 \to \infty$ to the minimizer of ψ (see Lemma 2.5.1 below for more details). As a result, the first step consists in searching for a good sampler, For the sake of generality and because the technical details for the variance, the KL approach and the GMM are similar, we study all these approaches in the mean time by considering a function $\theta \mapsto \hat\psi_{n_1}(\theta)$ based on n_1 sample points, which converges pointwise to ψ a function that is minimized at θ^*.

Algorithm 4 (Unified view)**.**
***Input**: the sample number $n \in \mathbb{N}^*$, the parametric family $\mathcal{Q}$, the initial density q_0, the function $\hat\psi_{n_1}$.*

(i) Let $n_1 < n$. Generate $X_1, \ldots, X_{n_1}$ independently with common distribution q_0. Compute $\hat\theta_1$ as the minimizer over $\theta \in \Theta$ of $\hat\psi_{n_1}$.

(ii) Let $n_2 = n - n_1$. Generate $Z_1, \ldots, Z_{n_2}$ according to $q_{\hat\theta_1}$, and compute

$$\hat I_{n,1}^{(\mathrm{is})} = n_2^{-1} \sum_{i=1}^{n_2} g(Z_i) w_{\hat\theta_1}(Z_i).$$

2.5 Asymptotic optimality of the two-stage importance sampling

The aim of the section is to show that under mild condition, the asymptotic variance of two-stage importance sampling achieves the same variance as the target density.

The following result explains formally why the first step of Algorithm 4 is working.

Lemma 2.5.1. *Suppose that $\Theta \subset \mathbb{R}^d$ is a compact set and that the function ψ is continuous. Assume there exists a unique minimizer $\theta^* \in \Theta$ of ψ and that $\sup_{\theta\in\Theta} |\hat{\psi}_{n_1}(\theta) - \psi(\theta)| \to 0$ in probability (resp. a.s.) and that $\hat{\theta}_{n_1}$ minimizes $\hat{\psi}$. Then $|\hat{\theta}_{n_1} - \theta^*| \to 0$ in probability (resp. a.s.).*

Proof. We focus on the proof for the convergence in probability. Let $\epsilon > 0$, since the set $\Theta \cap \{|\theta - \theta^*| \geq \epsilon\}$ is a compact and the function ψ is continuous we have that

$$\min_{\theta\in\Theta\cap\{|\theta-\theta^*|\geq\epsilon\}} \psi(\theta) = \psi(\tilde{\theta}^*),$$

for some $\tilde{\theta}^* \in \Theta$. As θ^* is the unique minimizer $\psi(\tilde{\theta}^*) - \psi(\theta^*) = \alpha > 0$. As a result, for any $\theta \in \Theta$ such that $|\theta - \theta^*| \geq \epsilon$, we have

$$\psi(\theta) - \psi(\theta^*) \geq \alpha \;.$$

It follows that

$$\mathbb{P}(|\hat{\theta}_{n_1} - \theta^*| \geq \epsilon) \leq \mathbb{P}(\psi(\hat{\theta}_{n_1}) - \psi(\theta^*) \geq \alpha) \;.$$

Using that $\hat{\theta}_{n_1}$ minimizes $\hat{\psi}$, we have

$$\begin{aligned} 0 < \psi(\hat{\theta}_{n_1}) - \psi(\theta^*) &= (\psi(\hat{\theta}_{n_1}) - \hat{\psi}(\hat{\theta}_{n_1})) + (\hat{\psi}(\hat{\theta}_{n_1}) - \hat{\psi}(\theta^*)) + (\hat{\psi}(\theta^*) - \psi(\theta^*)) \\ &\leq (\psi(\hat{\theta}_{n_1}) - \hat{\psi}(\hat{\theta}_{n_1})) + (\hat{\psi}(\theta^*) - \psi(\theta^*)) \\ &\leq 2\sup_{\theta\in\Theta} |\hat{\psi}(\theta) - \psi(\theta)| \;. \end{aligned}$$

Hence we find that

$$\mathbb{P}(|\hat{\theta}_{n_1} - \theta^*| \geq \epsilon) \leq \mathbb{P}(\sup_{\theta\in\Theta} |\hat{\psi}(\theta) - \psi(\theta)| \geq \alpha/2) \;,$$

which goes to 0 by assumption. □

We continue with the following nice property that the function $\theta \mapsto C_\varphi(q_\theta)$ is continuous under mild conditions on the parametric family $\mathcal{Q}$.

Lemma 2.5.2 (The variance is a continuous map)**.** *Let φ be a measurable function such that $0 < \int |\varphi| < \infty$. Suppose that for $|\varphi|$-almost every $x \in \mathbb{R}^d$, $\theta \mapsto q_\theta(x)$ is continuous on Θ, for any $\theta \in \Theta$, $q_\theta \gg |\varphi|$ and there exists $\eta > 0$ such that $\sup_{\theta\in\Theta} \int |\varphi|^{2+\eta}/q_\theta^{1+\eta} < +\infty$ then $C_\varphi : \theta \mapsto C_\varphi(q_\theta) = \int \varphi^2/q_\theta$ is continuous on Θ.*

Proof. Let $(\theta_n)_{n\geq1} \subset \Theta$ such that $\theta_n \to \theta \in \Theta$. We have

$$C_\varphi(q_{\theta_n}) = \left(\int |\varphi| \,\mathrm{d}\lambda\right) \mathbb{E}_\varphi[f_n]$$

$$C_\varphi(q_\theta) = \left(\int |\varphi| \,\mathrm{d}\lambda\right) \mathbb{E}_\varphi[f]$$

where $f_n = (|\varphi|/q_{\theta_n})$, $f = (|\varphi|/q_\theta)$ and $\mathbb{E}_\varphi$ denote the expectation with respect to $q^* = |\varphi|/\int|\varphi|$. The proof uses the concept of uniform integrability applied to the sequence $(f_n(X))_{n\geq1}$, with $X \sim q^*$. By Markov inequality, we have that

$$\mathbb{E}_\varphi[f_n 1_{\{f_n > x\}}] \leq x^{-\eta}\mathbb{E}_\varphi[f_n^{1+\eta}].$$

It follows that $\lim_{x\to\infty}\sup_{n\geq 1}\mathbb{E}_\varphi[f_n 1_{\{f_n>x\}}]=0$, which means that the sequence $(f_n(X))_{n\geq 1}$ is uniformly integrable. In addition, $f_n(X)\to f(X)$ almost surely by asumption. Then Proposition A.0.1 gives that $\mathbb{E}_\varphi[f_n]\to\mathbb{E}_\varphi[f]$ implying that $C_\varphi(q_{\theta_n})\to C_\varphi(q_\theta)$.

□

Theorem 2.5.3. *Suppose that for q^*-almost every $x\in\mathbb{R}^d$, $\theta\mapsto q_\theta(x)$ is continuous, that $0<\int|g|f<+\infty$ and there exists $\eta>0$ such that $\sup_{\theta\in\Theta}\int|gf|^{2+\eta}/q_\theta^{1+\eta}<+\infty$. Suppose that $\hat\theta_1\to\theta^*$ almost surely. Then, as $(n_1,n_2)\to\infty$,*

$$\sqrt{n_2}(\hat I_{n,1}^{(is)}-I)\rightsquigarrow\mathcal{N}(0,v^*),$$

with $v^*=\int g^2f^2/q_{\theta^*}-I_f(g)^2$.

Proof. The proof follows from the application of Theorem A.0.3 with $Y_{n,i}=([gf](Z_i)/q_{\hat\theta_1}(Z_i)-I)/\sqrt{n_2}$ according to the probability measure conditional on $(X_1,\ldots,X_{n_1})$. The mean zero property is obtained using the support condition that $q_\theta\gg f\gg f|g|$ for any $\theta\in\Theta$ which is a consequence of the assumptions. The Lindeberg condition is obtained using the Markov inequality and that $\sup_{\theta\in\Theta}\int[gf]^{2+\eta}/q_\theta^{1+\eta}<\infty$. The convergence of the variance is obtained by the almost sure convergence of $\hat\theta_1$ and the continuity of the map $\theta\mapsto\int[gf]^2/q_\theta$, provided by Lemma 2.5.2. □

Remark 2.5.1. *As the previous result indicates, the error committed on θ^* is not of first importance. We shall see that it only appears as a second-order error associated to the estimator.*

2.6 On the choice of the allocation

The computation of the expectation and the variance is helpful to provide answers to the following questions:

- How should we split the computational resources between first and second step of the Algorithm 4?
- Shall we use the points generated in the first step?

The last question is related to this estimate

$$\hat I_{n,2}^{(\mathrm{is})}=n^{-1}\left(\sum_{i=1}^{n_1}g(X_i)w_0(X_i)+\sum_{i=1}^{n_2}g(Z_i)w_{\hat\theta_1}(Z_i)\right).$$

The comparison is made with respect to the oracle estimator $\hat I_n^{(\mathrm{is}^*)}$ which requires the knowledge of q_{θ^*}. Let $(Z_1^*,\ldots,Z_n^*)$ be independently distributed with distribution q_{θ^*} and define

$$\hat I_n^{(\mathrm{is}^*)}=\frac{1}{n}\sum_{i=1}^n g(Z_i^*)w_{\theta^*}(Z_i^*).$$

An analysis of the bias and the variance of $\hat I_{n,k}^{(\mathrm{is})}$, $k=1,2$, under reasonable conditions gives some answers. We rely principally on the property that given $X_1,\ldots,X_{n_1}$, the distribution of Z_1 is fixed and equal $q_{\hat\theta_1}$ and that given $X_1,\ldots,X_{n_1}$, the sequence $Z_1,\ldots,Z_{n_2}$ is an independent sequence of random variables. Both estimators $\hat I_{n,1}^{(\mathrm{is})}$ and $\hat I_{n,2}^{(\mathrm{is})}$ have the same expectation.

Proposition 2.6.1. *Suppose that $0<\int|g|f\,\mathrm{d}\lambda<\infty$ and that $q_\theta\gg f$ for all $\theta\in\Theta$. For each $k\in\{1,2\}$, the estimator $\hat I_{n,k}^{(\mathrm{is})}$ is an unbiased estimator of $\int gf\,\mathrm{d}\lambda$.*

Proof. We have

$$\begin{aligned}
\mathbb{E}[\hat{I}_{n,2}^{(\mathrm{is})}] &= \frac{1}{n}\left(\sum_{i=1}^{n_1}\mathbb{E}\left[\frac{[gf](X_i)}{q_0(X_i)}\right]+\sum_{i=1}^{n_2}\mathbb{E}\left[\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)}\right]\right)\\
&= \frac{1}{n}\left(n_1\mathbb{E}\left[\frac{[gf](X_1)}{q_0(X_1)}\right]+\sum_{i=1}^{n_2}\mathbb{E}\left[\mathbb{E}\left[\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)}\mid (X_1,\ldots,X_{n_1})\right]\right]\right)\\
&= \frac{1}{n}\left(n_1\int [gf]\,\mathrm{d}\lambda+n_2\int [gf]\,\mathrm{d}\lambda\right)\\
&= \int gf\,\mathrm{d}\lambda\ .
\end{aligned}$$

From similar calculations we obtain that $\mathbb{E}[\hat{I}_{n,1}^{(\mathrm{is})}]=\int gf\,\mathrm{d}\lambda$. □

As given by the following proposition the variance of $\hat{I}_{n,2}^{(\mathrm{is})}$ and $\hat{I}_{n,1}^{(\mathrm{is})}$ are different. Define

$$u(\theta)=\int [gf]^2/q_\theta .$$

Proposition 2.6.2. *Suppose that $q_\theta \gg f$ for all $\theta\in\Theta$. Let $v_0=\mathrm{var}([gf](X_1)/q_0(X_1))$ and suppose that $v_0<+\infty$ and $\sup_{\theta\in\Theta}\int\frac{[gf](x)^2}{q_\theta(x)}\,\mathrm{d}x<+\infty$. We have*

$$\begin{aligned}
\mathrm{var}(\hat{I}_{n,1}^{(\mathrm{is})}) &= \frac{(\mathbb{E}[u(\hat{\theta}_1)]-I_f(g)^2)}{n_2},\\
\mathrm{var}(\hat{I}_{n,2}^{(\mathrm{is})}) &= \frac{n_1v_0+n_2(\mathbb{E}[u(\hat{\theta}_1)]-I_f(g)^2)}{n^2}.
\end{aligned}$$

Proof. Write

$$\hat{I}_{n,2}^{(\mathrm{is})}-\int [gf]\,\mathrm{d}\lambda=\frac{1}{n}\left(\sum_{i=1}^{n_1}\left(\frac{[gf](X_i)}{q_0(X_i)}-I_f(g)\right)+\sum_{i=1}^{n_2}\left(\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)}-I_f(g)\right)\right).$$

As we have seen before, the bias is 0. The variance is therefore equal to the expectation of the square. As it holds that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n_1}\left(\frac{[gf](X_i)}{q_0(X_i)}-I_f(g)\right)\sum_{i=1}^{n_2}\left(\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)}-I_f(g)\right)\right]\\
&=\mathbb{E}\left[\sum_{i=1}^{n_1}\left(\frac{[gf](X_i)}{q_0(X_i)}-I_f(g)\right)\mathbb{E}\left[\sum_{i=1}^{n_2}\left(\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)}-I_f(g)\right)\mid (X_1,\ldots,X_{n_1})\right]\right]=0,
\end{aligned}$$

and using conditional independence, we find that

$$\begin{aligned}
n^2\mathrm{var}\left(\hat{I}_{n,2}^{(\mathrm{is})}-I_f(g)\right) &= \mathbb{E}\left[\sum_{i=1}^{n_1}\left(\frac{[gf](X_i)}{q_0(X_i)}-I_f(g)\right)^2\right]+\mathbb{E}\left[\sum_{i=1}^{n_2}\left(\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)}-I_f(g)\right)^2\right]\\
&= n_1\mathbb{E}\left[\left(\frac{[gf](X_1)}{q_0(X_1)}-I_f(g)\right)^2\right]+n_2\mathbb{E}\left[\left(\frac{[gf](Z_1)}{q_{\hat{\theta}_1}(Z_1)}-I_f(g)\right)^2\right]\\
&= n_1v_0+n_2\mathbb{E}\left[\mathbb{E}\left[\left(\frac{[gf](Z_1)}{q_{\hat{\theta}_1}(Z_1)}-I_f(g)\right)^2\mid (X_1,\ldots,X_{n_1})\right]\right]\\
&= n_1v_0+n_2(\mathbb{E}[u(\hat{\theta}_1)]-I_f(g)^2)\ .
\end{aligned}$$

For the second statement, we use the previous calculus to obtain

$$
\begin{aligned}
\operatorname{var}(\hat{I}^{(\mathrm{is})}_{n,1}) &= \operatorname{var}\left(\hat{I}^{(\mathrm{is})}_{n,1} - I_f(g)\right) \\
&= \frac{1}{n_2^2}\mathbb{E}\left[\sum_{i=1}^{n_2}\left(\frac{[gf](Z_i)}{q_{\hat{\theta}_1}(Z_i)} - I_f(g)\right)^2\right] \\
&= \frac{1}{n_2}(\mathbb{E}[u(\hat{\theta}_1)] - I_f(g)^2)\ .
\end{aligned}
$$

□

If the function u satisfies some regularity conditions, it holds that

$$
|u(\hat{\theta}_{n_1}) - u(\theta^*)| = O_p(n_1^{-1})\ .
$$

Indeed, the previous results from M and Z-estimation theory in which conditions are given to ensure that $\hat{\theta}_n - \theta^* = O_p(n_1^{-1/2})$. Examples includes maximum likelihood estimators, least squares estimators, GLM (van der Vaart, 1998). Suppose that θ^* is an interior point and that u is differentiable, and for some neighbourhood $v(\theta^*)$, there exists $\kappa > 0$ such that for all $\theta \in v(\theta^*)$,

$$
|u(\theta) - u(\theta^*) - \partial_\theta u(\theta^*)(\theta - \theta^*)| \le \kappa|\theta - \theta^*|^2\ .
$$

Since u is minimized at θ^*, $\partial_\theta u(\theta^*) = 0$ and as $\mathbb{P}(\hat{\theta}_{n_1} \in v(\theta^*))$ goes to 1 from Lemma 2.5.1, we have

$$
\mathbb{P}(|u(\hat{\theta}_{n_1}) - u(\theta^*)| \le \kappa|\hat{\theta}_{n_1} - \theta^*|^2) \to 1\ .
$$

In the following, we make the slightly stronger condition that $\mathbb{E}[u(\hat{\theta}_{n_1}) - u(\theta^*)] = \sigma^2/n_1 + o(n_1^{-1})$.

Proposition 2.6.3. *Assume that $\mathbb{E}[u(\hat{\theta}_1) - u(\theta^*)] = \sigma^2/n_1 + o(n_1^{-1})$ and that $v_0 > v(\theta^*) = u(\theta^*) - I_f(g)^2$. We have that, whenever $v(\theta^*) > 0$,*

$$
\lim_{n\to+\infty} n^{1/2} \inf_{1\le n_1\le n}\left\{\mathbb{E}[(\sqrt{n}(\hat{I}^{(\mathrm{is})}_{n,1} - I_f(g))^2] - \mathbb{E}[(\sqrt{n}(\hat{I}^{(\mathrm{is}^*)}_{n} - I_f(g))^2]\right\} = 2\sigma\sqrt{v(\theta^*)},
$$

whenever $v(\theta^*) = 0$,

$$
\lim_{n\to+\infty} n \inf_{1\le n_1\le n}\left\{\mathbb{E}[(\sqrt{n}(\hat{I}^{(\mathrm{is})}_{n,1} - I_f(g))^2]\right\} = 4\sigma^2.
$$

Moreover,

$$
\lim_{n\to+\infty} n^{1/2} \inf_{1\le n_1\le n}\left\{\mathbb{E}[(\sqrt{n}(\hat{I}^{(\mathrm{is})}_{n,2} - I_f(g))^2] - \mathbb{E}[(\sqrt{n}(\hat{I}^{(\mathrm{is}^*)}_{n} - I_f(g))^2]\right\} = 2\sigma\sqrt{(v_0 - v(\theta^*))}.
$$

Proof. The variance of $\sqrt{n}(\hat{I}^{(\mathrm{is}^*)}_{n} - I_f(g))$ is indeed $v(\theta^*)$. By way of comparison, the variance of $\sqrt{n}(\hat{I}^{(\mathrm{is})}_{n,2} - I_f(g))$ is given by $\{n_1 v_0 + n_2(\mathbb{E}[u(\hat{\theta}_1)] - I_f(g)^2)\}/n$. The difference between both variances expresses as

$$
\begin{aligned}
\frac{n_1(v_0 - v(\theta^*)) + n_2(\mathbb{E}[u(\hat{\theta}_1) - u(\theta^*)])}{n} &= \frac{n_1(v_0 - v(\theta^*)) + n_2\sigma^2/n_1}{n} + o(n_1^{-1}) \\
&= \frac{n_1(v_0 - v(\theta^*)) + (n - n_1)\sigma^2/n_1}{n} + o(n_1^{-1})\ .
\end{aligned}
$$

We suppose the remainder to have no effect. Fixing n, optimizing over n_1 gives the value $n_1 = \sqrt{\frac{\sigma^2 n}{(v_0 - v(\theta^*))}}$. Plugging-in this value gives the first statement. For $\hat{I}_{n,1}^{(\mathrm{is})}$, the difference between variances is

$$\begin{aligned}
\frac{n\mathbb{E}[u(\hat{\theta}_1) - I_f(g)^2]}{n_2} - v(\theta^*) &= \frac{n_1}{n_2}\mathbb{E}[u(\hat{\theta}_1) - I_f(g)^2] + \mathbb{E}[u(\hat{\theta}_1) - u(\theta^*)] \\
&= \frac{n_1}{n - n_1}\mathbb{E}[u(\hat{\theta}_1) - I_f(g)^2] + \frac{\sigma^2}{n_1} + o(n_1^{-1}) \\
&= \frac{n_1}{n - n_1}(v(\theta^*) + \sigma^2/n_1) + \frac{\sigma^2}{n_1} + o(n_1^{-1})
\end{aligned}$$

We distinguish between the two cases $v(\theta^*) = 0$ and $v(\theta^*) > 0$. In the first case, the previous equals

$$\frac{\sigma^2 n}{(n - n_1)n_1} \ .$$

It is minimized for $n_1 = n/2$. In the other case, denote by q the function to minimize, we have

$$f'(x) = \frac{x^2(nv(\theta^*) + \sigma^2) - \sigma^2(n - x)^2}{x^2(n - x)^2} \ .$$

As n is large enough, $nv(\theta^*) - \sigma^2 > 0$ implying that a "zero" of q' verifies

$$x\sqrt{(nv(\theta^*) + \sigma^2)} - \sigma(n - x) = 0,$$

equivalently,

$$x(\sqrt{(nv(\theta^*) + \sigma^2)} + \sigma) = \sigma n \ .$$

Hence, the value of n_1 which minimizes the leading term is

$$\sqrt{\frac{\sigma^2 n}{v(\theta^*)}}(\sqrt{1 - \sigma^2/v(\theta^*)n} + \sqrt{\sigma^2/v(\theta^*)n})^{-1} = \sqrt{\frac{\sigma^2 n}{v(\theta^*)}} + O(1) \ .$$

The minimum is then given by

$$\sqrt{\sigma^2 v(\theta^*)}/\sqrt{n} + O(1/n) + \sqrt{\sigma^2 v(\theta^*)}/\sqrt{n} + O(1/n) \ .$$

□

Remark 2.6.1. *According to the previous proposition, both estimators $\hat{I}_{n,1}^{(\mathrm{is})}$ and $\hat{I}_{n,2}^{(\mathrm{is})}$ behave similarly when $v(\theta^*) > 0$. The rates of convergence are both in $n^{-1/2}$. Their variance are very close, i.e., $O(n^{-1/2})$, from the variance of the oracle estimator. The difference relies in the constants $\sqrt{v_0 - v(\theta^*)}$ and $\sqrt{v(\theta^*)}$.*

Remark 2.6.2. *When g is positive and the density $gf/\int gf \,\mathrm{d}\lambda$ belongs to the class $\{q_\theta : \theta \in \Theta\}$, we have that $v(\theta^*) = 0$. When $v(\theta^*) = 0$,*

$$\lim_{n\to+\infty} n^{3/2} \inf_{1 \le n_1 \le n} \mathbb{E}[(\hat{I}_{n,2}^{(\mathrm{is})} - I_f(g))^2] = 2\sigma\sqrt{v_0} \ .$$

For $\hat{I}_{n,1}^{(\mathrm{is})}$, we have already showed that the rate of convergence of the MSE is in n^{-2}. Consequently, when $v(\theta^*) = 0$ we shall prefer $\hat{I}_{n,1}^{(\mathrm{is})}$ before $\hat{I}_{n,2}^{(\mathrm{is})}$. In both cases, the convergence rate is no longer in $n^{-1/2}$ (as Monte Carlo) but in $n^{-3/4}$ and n^{-1}, respectively.

Remark 2.6.3. *From the previous proposition, for $\hat{I}_{n,2}^{(\mathrm{is})}$ the value that we should take is $n_1 = \lfloor \sqrt{cn} \rfloor$, for some $c = \frac{\sigma^2}{(v_0 - v(\theta^*))}$. Unfortunately, the number c is unknown. A first possibility is to estimate c but this should involve additional calculation. A maybe more reasonable way is to choose arbitrary, e.g., $c = 2$. As the appropriate rate of convergence is already provided by $n_1 = O(\sqrt{n})$, the influence of c shall not be of fundamental importance. For $\hat{I}_{n,1}^{(\mathrm{is})}$ when $v(\theta^*) > 0$ one should take $n_1 = \lfloor \sqrt{cn} \rfloor$, for some $c = \frac{\sigma^2}{v(\theta^*)}$. But when $v(\theta^*) = 0$, the value n_1 to take is simply $n/2$.*

Remark 2.6.4. *Another way around this problem is to apply the Lindeberg central limit theorem in order to provide a central limit theorem with the specified variance. This approach is more difficult.*

2.7 Computational consideration

Running any of the previous algorithms is associated to some computation time. The computing time is given by the number of elementary operations needed to produce the approximated value of the integral. To compute this we rely on some calculation rules. We follow these rules to evaluate the computational complexity:

- Generate X_1 is 1 elementary operation.
- Generate $Z_{1,k}$ is 1 elementary operation for each k.
- Evaluate $[gf](X_1)$ is 1 elementary operation.

The previous rules depend of course on the context. For instance, generating random variables from a particular distribution can involve some difficulty (e.g. rejection methods, Metropolis-Hastings algorithm) or computing gf might take more than a single elementary operation (e.g. crash test simulation). When this arises, the amount of observations $([gf](X_1), \ldots, [gf](X_n))$ is fixed and all the methods are computed with the same amount of observation.

We additionally assume that the optimization step requires k evaluations of $\hat{\psi}'$ to produce a precision $|\theta_k - \theta^\star| \leq 1/\sqrt{k}$. Such an assumption is somehow optimistic as it essentially covers cases when the function is differentiable and convex. This should not be necessarily the case in practice.

Computation time for $\hat{I}_{n,2}^{(\mathrm{is})}$. Based on the previous rules, we evaluate the number of elementary operations required to compute the method when $n_1 = \lfloor (cn)^{1/2} \rfloor$, for some $c > 0$. Suppose that $\hat{\theta}_1$ is known. We generate $\lfloor (cn)^{1/2} \rfloor$ random variable $Z_1, Z_2, \ldots$ and $n - \lfloor (cn)^{1/2} \rfloor$ random variables $X_1, X_2, \ldots$, for some $c > 0$. This is $O(n)$ operations. We also compute $[gf](X_1)/q_0(X_1), \ldots, [gf](X_{n_1})/q_0(X_{n_1})$ and $[gf](Z_1)/q_{\hat{\theta}_1}(Z_1), \ldots, [gf](Z_{n_2})/q_{\hat{\theta}_1}(Z_{n_2})$. This is $O(n)$ more operations. Moreover, we have to conduct an optimization step to compute $\hat{\theta}_1$. This is the computation of $\hat{\theta}_1$ define as the minimizer of

$$\hat{\psi}(\theta) := n_1^{-1} \sum_{i=1}^{n_1} \frac{[gf](X_i)^2}{q_\theta(X_i)^2} .$$

With k evaluations of ψ' (each evaluations is $O(n_1)$ operations; to evaluate and to compute the sum), we obtain $\hat{\theta}_1^{opt}$ such that $|\hat{\theta}_1^{opt} - \hat{\theta}_1| = 1/\sqrt{k}$. Consequently, with $k = n_1$, we have a precision in $1/\sqrt{n_1}$ and this requires approximately $n_1 \times n_1$ operations. This choice of k is such that the stochastic error and the optimisation error have the same order. From question 17, this translates in a precision of

$$|\psi(\hat{\theta}_1^{opt}) - \psi(\theta^*)| = O_p(n_1^{-1}) .$$

In summary, we obtain a total number of operations in $O(n + n_1^2) = O(n)$.

Exercises

Exercise 2.7.1.

(a) Let $\Sigma \in \mathbb{R}^{d\times d}$ be a given matrix. Derive the first step importance sampling estimate (corresponding to $\hat{\theta}_1$) in the case of the KL approach for the parametric model $\{\mathcal{N}(\mu, \Sigma) \,:\, \mu \in \mathbb{R}^d\}$.

(b) Derive the first step importance sampling estimate (corresponding to $\hat{\theta}_1$) in the case of the KL approach for the parametric model $\{\mathcal{N}(\mu, \Sigma) \,:\, \mu \in \mathbb{R}^d,\, \Sigma \in \mathbb{R}^{d\times d}\}$ (hint: $(\partial/\partial A)\log(|A|) = A^{-T}$ and $(\partial/\partial A)\operatorname{tr}(AB) = (\partial/\partial A)\operatorname{tr}(BA) = B^T$).

Exercise 2.7.2.

Let $(f_k)_{k=1,\dots,K}$ be a sequence of densities supported on $\mathbb{R}$, i.e., strictly positive functions such that $\int_{\mathbb{R}} f_k \mathrm{d}\lambda = 1$. We suppose that we can generate according to each f_k. Our goal is to compute $\int_{\mathbb{R}} g \mathrm{d}\lambda$, for $g : \mathbb{R} \to \mathbb{R}$.

(a) Let $\alpha = (\alpha_1, \dots, \alpha_K) \in [0, 1]^K$ be such that $\sum_{k=1}^K \alpha_k = 1$. Based on a uniform random variable generation : $U \sim \mathcal{U}[0, 1]$, show how to generate $B = (B_1, \dots, B_K)$ such that each $B_k \sim \mathcal{B}(\alpha_k)$ and $\sum_{k=1}^K B_k = 1$.

(b) Using the previous question, write an algorithm to generate X according to $f_\alpha = \sum_{k=1}^K \alpha_k f_k$. Justify that X is distributed according to f_α.

(c) Let $(\alpha_k^{(\ell)})$ be such that for each $\ell = 1, \dots L$, $\alpha^{(\ell)} = (\alpha_1^{(\ell)}, \dots, \alpha_K^{(\ell)}) \in [0, 1]^K$ and $\sum_{k=1}^K \alpha_k^{(\ell)} = 1$. Let $(n_1, \dots, n_L) \in \mathbb{N}^L$ and $N = \sum_{\ell=1}^L n_\ell$. Let $(X_i^{(\ell)})_{i=1,\dots,n_\ell} \overset{i.i.d.}{\sim} f_{\alpha^{(\ell)}}$. Based on $\hat{S}_\ell = \sum_{i=1}^{n_\ell} g(X_i^{(\ell)})/f_{\alpha^{(\ell)}}(X_i^{(\ell)})$, $\ell = 1, \dots, L$, give an unbiased estimator of $\int_{\mathbb{R}} g \mathrm{d}\lambda$ (hint : one may look for estimate of the type $N^{-1}\sum_{\ell=1}^L \gamma_\ell \hat{S}_\ell$ for some (γ_ℓ) known).

(d) Express the variance of the estimator you defined in the previous question. Give a condition that guarantee the variance to be smaller than C/N, for some constant $C > 0$.

(e) In order to select optimally the weights $(\alpha_k^{(\ell)})$, provide an algorithm with $\ell = 1, \dots, L$ steps that updates at each step the weights $\alpha_k^{(\ell)}$ (hint : you can propose to optimize some criterion at each step justifying your choice. The algorithm should be clearly described).

Chapter 3

Control variates

Control variates is based on the following one-sentence principle: if you wish to evaluate the (unknown) integral of a certain function you better use functions of which you know the integral. The control variates method consists in incorporating this new piece of information, the known integral value of some "'control functions", in the basic Monte Carlo framework. The aim is to reduce the variance of the traditional Monte Carlo estimate. In this chapter we first consider antithetic variate methods which is a particular method among the family of control variates methods.

3.1 The antithetic variate method

The method of antithetic variate attempts to reduce the variance of Monte Carlo by introducing negative dependence between pairs of replications. If we generate a independent sample of uniform random variables $U_1, \cdots, U_n$ we also use the so called antithetic sample $(1-U_1), \cdots, (1-U_n)$ (which has the same distribution) so that large values of one ample is balanced by small values of the other. The asymptotic study of the method and the construction of confidence intervals are the same as for Monte Carlo. Consequently, we rather focus in the following on comparing the variance between the antithetic technique and Monte Carlo.

3.1.1 Variance analysis

Let Z be a random variable with distribution μ. Assume that there exists a measurable map $L : S \to S$ such that $L(Z)$ and Z have the same distribution. Let $g : S \to \mathbb{R}$ be such that $\mathbb{E}(g^2(Z)) < +\infty$. Given $\{Z_k, k \geq 1\}$ independent and identically distributed random variables with common distribution μ. The approximation of $I_\mu(g) = \mathbb{E}[g(Z)]$ by the antithetic variate method is provided by the estimate

$$\hat{I}_n^{(\text{av})}(g) = \frac{1}{2n} \sum_{i=1}^{n} \left\{ g(Z_i) + g(L(Z_i)) \right\}.$$

An example of such a map L is $L(U) = a+b-U$ when U is a uniform random variable on $[a, b]$. If $Z \sim \mathcal{N}(\mu, 1)$, then $2\mu - Z \sim \mathcal{N}(\mu, 1)$. More generally, if $Z \sim \mathcal{N}_d(\mu, \Sigma)$ where $\Sigma > 0$, then $\mu - \Sigma^{1/2} H \Sigma^{-1/2}(Z - \mu) \sim \mathcal{N}_d(\mu, \Sigma)$, where $HH^T = I$. The variance of $\hat{I}_n^{(\text{va})}$ can be computed for each values of $n \in \mathbb{N}^*$, leading to

$$\text{var}(\hat{I}_n^{(\text{av})}(g)) = \frac{1}{n} \text{var}\left(\frac{g(Z) + g(L(Z))}{2} \right) = \frac{1}{2n} \text{var}(g(Z)) \left\{ 1 + \text{corr}(g(Z), g(L(Z))) \right\},$$

where $\text{corr}(X, Y) = \text{cov}(X, Y)/\sqrt{\text{var}(X)\,\text{var}(Y)}$.

Considering that evaluating $g \circ L$ represents the same budget as evaluating g and that generating Z is negligible. The computational effort to furnish to compute $\hat{I}_n^{(\text{av})}$ is approximately twice the effort to compute

Monte Carlo with n samples, as g is evaluated $2n$-times. Hence a natural competitor of the method is $\overline{\mu}_{2n}$ given by

$$\hat{I}_{2n}^{(\mathrm{mc})}(g) = \frac{1}{2n}\sum_{i=1}^{2n} g(Z_i).$$

A condition on the variance characterizes whether the antithetic variate method over-perform Monte Carlo.

Proposition 3.1.1. *If* $\mathbb{E}|g(Z)|^2 < \infty$*, then*

- $\mathrm{var}(\hat{I}_{2n}^{(\mathrm{mc})}(g)) \geq \mathrm{var}(\hat{I}_{2n}^{(\mathrm{av})}(g)) \quad \Leftrightarrow \quad \mathrm{cov}(g(Z), g(L(Z))) \leq 0,$
- *if moreover, g is an increasing real function and $g \circ L$ is decreasing (or conversely), then* $\mathrm{cov}(g(Z), g(L(Z))) \leq 0$.

Proof. The first assertion is obtained by direct computation. For the second assertion, we apply Lemma 3.1.2 with $h = g$ and $\tilde{h} = g \circ L$. □

Lemma 3.1.2. *Let Z be a real random variable and suppose that $g : \mathbb{R} \to \mathbb{R}$ is increasing with* $\mathbb{E}[h(Z)^2] < \infty$ *and $\tilde{h} : \mathbb{R} \to \mathbb{R}$ is decreasing with* $\mathbb{E}[\tilde{h}(Z)^2] < \infty$*, then* $\mathrm{cov}(h(Z), \tilde{h}(Z)) \leq 0$.

Proof. Note that, by assumption, $(h(x) - h(y))(\tilde{h}(x) - \tilde{h}(y)) \leq 0$ for all $(x, y) \in \mathbb{R}^2$ and that, whenever X and Y are independent copies of Z,

$$\begin{aligned}\mathbb{E}[(h(X) - h(Y))(\tilde{h}(X) - \tilde{h}(Y))] &= 2\mathbb{E}[h(Z)\tilde{h}(Z)] - 2\mathbb{E}[h(Z)]\mathbb{E}[\tilde{h}(Z)] \\ &= 2\,\mathrm{cov}(h(Z), \tilde{h}(Z)).\end{aligned}$$

□

3.1.2 Examples

The previous proposition yields a general method for the construction of antithetic variate to evaluate $\mathbb{E}[Z]$ when F_Z, the cumulative distribution function of Z, is known. Define the generalized inverse of F_Z by

$$F_Z(u) = \inf\{x \in \mathbb{R} \,:\, F(x) \geq u\}.$$

We are going to use the following result that if $U \sim \mathcal{U}[0,1]$, then $F_Z^{-}(U) \sim Z$, which proof can be done as an exercise (see Exercise 1.3.1). This leads to the following algorithm.

Algorithm 5 (Antithetic variate)**.**
Input: *the sample number $n \in \mathbb{N}^*$.*

(i) Let $n \in \mathbb{N}^*$. Generate $U_1, \ldots, U_n$ independently with common uniform distribution on $[0,1]$.

(ii) Compute

$$\hat{I}_n = \frac{1}{2n}\sum_{i=1}^{n}\left\{F_Z^{-1}(U_i) + F_Z^{-1}(1 - U_i)\right\}.$$

Another example is for computing integrals of an increasing function g over the segment line $[a, b]$. The antithetic variate estimate is given by

$$\frac{1}{2n}\sum_{i=1}^{n}\left\{g(U_i) + g((a + b) - U_i)\right\},$$

where $U_1, \ldots, U_n$ are independent random variables with common distribution $\mathcal{U}[a, b]$.

In the special case when generating the random variables Z is expensive with respect to evaluating g, the two methods, antithetic variate and Monte Carlo, might be compared with the same sample number. Define the symmetric and antisymmetric part of g, respectively given by

$$g_0(z) = \frac{g(z) + g(L(z))}{2}\ , \qquad g_1(z) = \frac{g(z) - g(L(z))}{2}.$$

We have that $\operatorname{cov}(g_0(Z), g_1(Z)) = 0$ implying that

$$\operatorname{var}(g(Z)) = \operatorname{var}(g_0(Z)) + \operatorname{var}(g_1(Z)).$$

Hence

$$\operatorname{var}(\hat{I}_n^{(\mathrm{mc})}(g)) = \operatorname{var}(\hat{I}_n^{(\mathrm{av})}(g)) + n^{-1} \operatorname{var}(g_1(Z))\ .$$

Consequently, in such a case, the antithetic variate method always reduces the variance of the naïve Monte Carlo.

3.2 The control variates method

Let $((X_1, Z_1), \ldots, (X_n, Z_n))$ be an independent and identically distributed sequence of random variables valued in $S \times \mathbb{R}$ and suppose that $g : S \to \mathbb{R}$ is such that $\mathbb{E}[|g(X_1)|] < \infty$ and that $\mathbb{E}[Z_1]$ is known. The distribution of X_1 is denoted by μ. The aim of the control variate method is to estimate $I_\mu = \mathbb{E}[g(X_1)]$ (the dependence in g is removed in this section for ease of reading) using the knowledge of $\mathbb{E}[Z_1]$. Since $\mathbb{E}[Z_1]$ is known, we can suppose without any loss of generality that $\mathbb{E}[Z_1] = 0$. The control variates class of estimator is given by

$$\hat{I}_n^{(\mathrm{cv})} = n^{-1} \sum_{i=1}^n (g(X_i) - Z_i).$$

Some basic properties are easily derived.

Proposition 3.2.1. *Suppose that* $\mathbb{E}[|g(X_1)|] < \infty$, $\mathbb{E}[|Z_1|] < \infty$,

- $\hat{I}_n^{(\mathrm{cv})}$ *is unbiased and strongly consistent.*

Suppose moreover that $\mathbb{E}[|g(X_1)|^2] < \infty$, $\mathbb{E}[|Z_1|^2] < \infty$,

- $\operatorname{var}(\hat{I}_n^{(\mathrm{cv})}) = \operatorname{var}(g(X_1) - Z_1)/n$ *and* $\hat{I}_n^{(\mathrm{cv})}$ *is asymptotically normal with variance* $s^2 = \operatorname{var}(g(X_1) - Z_1)$, *i.e.,* $\sqrt{n}(\hat{I}_n^{(\mathrm{cv})} - I_\mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, s^2)$.
- *The estimation of the variance can be done consistently by*

$$\hat{s}^2 = (n-1)^{-1} \sum_{i=1}^n \{(g(X_i) - Z_i) - \hat{I}_n^{(\mathrm{cv})}\}^2.$$

Remark 3.2.1. *The control variates is an extension of Monte Carlo, as when* $Z_1 = 0$ *we recover the Monte Carlo estimate. It also includes antithetic variates methods, when taking* $Z_1 = (g(X_1) - (g \circ L)(X_1))/2$.

Remark 3.2.2. *So far we can not be sure that the introduction of the control variates* (Z_i) *reduces the variance over Monte Carlo as it is not guaranteed that* $\operatorname{var}(g(X_1) - Z_1) \leq \operatorname{var}(g(X_1))$. *Hence it makes sense to parametrize the control variate estimate in order to play on the influence of the control variates on the estimation. This leads to the following control variates estimate*

$$n^{-1} \sum_{i=1}^n (g(X_i) - \beta Z_i),$$

Control variates e.g., antithetic variate	Importance sampling e.g., stratified sampling
Idea : approximate g by control variates Post processing schemes as $X_1, ..., X_n$ are fixed	Idea : change the underlying measure new sampling $X_1, ..., X_n$

Figure 3.1: Control variates and importance sampling

where $\beta \in \mathbb{R}$. For this estimate, we have the same properties as the one stated in Proposition 3.2.1. According to the variance, which is a quadratic function of β, the best possible choice of β is $\mathbb{E}[g(X_1)Z_1]/\mathbb{E}[Z_1^2]$. The questions that naturally follows are:

- *How to estimate β ?*
- *Does the choice of β influences the estimation?*
- *Does the choice of β influences the computation time?*

Remark 3.2.3. *In many examples, one should deal with the observation of several control variates $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ where for each $i = 1, \ldots, n$, $\boldsymbol{Z}_i \in \mathbb{R}^m$. This leads to the following estimator*

$$\hat{I}_n^{(\mathrm{cv})}(\beta) = n^{-1} \sum_{i=1}^{n} (g(X_i) - \beta^T \boldsymbol{Z}_i),$$

where $\beta \in \mathbb{R}^m$. According to the variance, the best possible choice of β is the solution of $\mathbb{E}[\boldsymbol{Z}_1 \boldsymbol{Z}_1^T]\beta = \mathbb{E}[\boldsymbol{Z}_1 g(X_1)]$. The same questions as before are still interesting. We shall see in the following that all that matter is the accuracy of the approximation in L_2 of $g(X_1)$ by elements of the form $\boldsymbol{\beta}^T Z_1$.

Remark 3.2.4. *Importance sampling and control variates actually form 2 distinct groups of methods based on different ideas as detailed in Figure 3.1.*

Examples. We conclude this section by providing some examples.

1. One can use an antithetic variates to construct Z, i.e., by taking $\frac{1}{2}(g - g \circ L)$. The problem is that functions like L are not always available. The canonical example is when computing $\mathbb{E}[W]$ for which F_W is known (see previous section). Then one can take $g(X_i)$ equal to $F_W^{-1}(U_i)$ and Z_i equal to $(F_W^{-1}(U_i) - F_W^{-1}(1 - U_i))/2$.
2. In Finance, calculation of the price of an Asian option (see exercise class, TD2).
3. Numerical integration. We wish to evaluate $\int_{[0,1]^d} g(x)\mathrm{d}x$. Let $f_1, ..., f_m$ be a family of functions such that, for all $k \in \{1, \ldots, m\}$,
$$\int_{[0,1]^d} f_k \,\mathrm{d}\lambda = 0,$$
e.g., polynomials, Fourier, Splines.

3.3 Asymptotics

In this section we are interested in the choice of β in the multivariate case, i.e., we have at our disposal $m \in \mathbb{N}^*$ control variates. We observe $((X_1, g(X_1), \boldsymbol{Z}_1), (X_n, g(X_n), \boldsymbol{Z}_n))$. For each $\beta \in \mathbb{R}^m$, the control variates estimate

$$\hat{I}_n^{(\mathrm{cv})}(\beta) = n^{-1} \sum_{i=1}^{n} (g(X_i) - \beta^T \boldsymbol{Z}_i),$$

is an unbiased estimator. Hence among this class of estimator, it is tempting to characterize the one with the smallest variance. It is the one associated with β^* defined as the minimizer of

$$\mathbb{E}\left[\left(g(X_1) - \sum_{k=1}^{m} \beta_k Z_{1,k}\right)^2\right].$$

This can be estimated without bias (for each β) by

$$\frac{1}{(n-1)} \sum_{i=1}^{n} \left(g(X_i) - \sum_{k=1}^{m} \beta_k Z_{k,i} - \hat{I}_n^{(\mathrm{cv})}(\beta)\right)^2$$

Hence, we define $\hat{\beta}_n$ as a minimizer of the previous quantity. By the Hilbert projection theorem, we obtain that $\hat{\beta}_n$ verifies the equation

$$(\boldsymbol{Z}_{n,m}^T \boldsymbol{Z}_{n,m})\hat{\beta}_n = \boldsymbol{Z}_{n,m}^T \boldsymbol{g}_n,$$

where $\boldsymbol{Z}_{n,m} = (\boldsymbol{Z}_1 - \overline{\boldsymbol{Z}}^n, \ldots, \boldsymbol{Z}_n - \overline{\boldsymbol{Z}}^n)^T$, $\boldsymbol{g}_n = (g(X_1), \ldots, g(X_n))^T$, $\overline{\boldsymbol{Z}}^n = n^{-1} \sum_{i=1}^n \boldsymbol{Z}_i$. Among the solutions of the previous equation, we define $\hat{\beta}_n$ as

$$\hat{\beta}_n = (\boldsymbol{Z}_{n,m}^T \boldsymbol{Z}_{n,m})^+ \boldsymbol{Z}_{n,m}^T \boldsymbol{g}_n,$$

where A^+ denotes the generalized inverse. The resulting control variate estimator is given by

$$\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) = n^{-1} \sum_{i=1}^{n} \left(g(X_i) - \hat{\beta}_n^T \boldsymbol{Z}_i\right).$$

This estimator is biased. We are going to study the convergence properties of $\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n)$. The convergence in probability and the asymptotic normality are obtained in the following proposition.

Proposition 3.3.1. *Suppose that $\mathbb{E}[|g(X_1)|] < \infty$, $\mathbb{E}[|g(X_1)Z_{k,1}|] < \infty$, $\forall k = 1, \ldots, m$, and that $\mathbb{E}[\boldsymbol{Z}_1 \boldsymbol{Z}_1^T]$ is invertible, then*

- *the estimator $\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n)$ is strongly consistent,*

if moreover $\mathbb{E}[|g(X_1)|^2] < \infty$, then

- *$\sqrt{n}(\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) - \mathbb{E}[g(X_1)]) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_m^2)$, where $\sigma_m^2 = \operatorname{argmin}_{\beta \in \mathbb{R}^m} \operatorname{var}(g(X_1) - \beta^T \boldsymbol{Z}_1)$*

Proof. First we show that $\hat{\beta}_n \to \beta^*$, almost surely. We have, by the law of large number that, almost surely,

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T \to \mathbb{E}[\boldsymbol{Z}_1 \boldsymbol{Z}_1^T],$$
$$n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i g(X_i) \to \mathbb{E}[\boldsymbol{Z}_1 g(X_i)].$$

Hence there exists $N = N(\omega)$ such that $\forall n \geq N$, the previous estimated matrix $\boldsymbol{Z}_{n,m}^T \boldsymbol{Z}_{n,m}/n$ is invertible. Using the co-factor formula for the inverse, we get that $(\boldsymbol{Z}_{n,m}^T \boldsymbol{Z}_{n,m}/n)^+ \to \mathbb{E}[\boldsymbol{Z}_1 \boldsymbol{Z}_1^T]^{-1}$ almost surely. Hence it follows that with probability 1,

$$\left(n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T\right)^{-1} \left(n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i g(X_i)\right) \to \beta^* = \mathbb{E}[\boldsymbol{Z}_1 \boldsymbol{Z}_1^T]^{-1} \mathbb{E}[\boldsymbol{Z}_1 g(X_i)].$$

An appropriate expression of the estimator is

$$\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) = \begin{pmatrix} 1 & -\hat{\beta}_n^T \end{pmatrix} n^{-1} \sum_{i=1}^{n} \begin{pmatrix} g(X_i) \\ \boldsymbol{Z}_i \end{pmatrix}$$

From the law of large number again, we deduce that, almost surely,

$$\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) \longrightarrow \begin{pmatrix} 1 & -\beta^{*T} \end{pmatrix} \begin{pmatrix} \mathbb{E}[g(X_1] \\ 0 \end{pmatrix} = \mathbb{E}[g(X_1].$$

For the asymptotic normality, we rely on a similar decomposition as before,

$$\sqrt{n}(\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) - \mathbb{E}[g(X_1]) = \begin{pmatrix} 1 & -\hat{\beta}_n^T \end{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} g(X_i) - \mathbb{E}[g(X)] \\ \boldsymbol{Z}_i \end{pmatrix}.$$

It follows that

$$\sqrt{n}(\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) - \hat{I}_n^{(\mathrm{cv})}(\beta^*)) = (\beta^* - \hat{\beta}_n)^T n^{-1/2} \sum_{i=1}^{n} \boldsymbol{Z}_i.$$

Conclude using Slutsky's Lemma and the central limit theorem $\sqrt{n}(\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) - \mathbb{E}[g(X_1])$ has the same asymptotic law as $\sqrt{n}(\hat{I}_n^{(\mathrm{cv})}(\beta^*) - \mathbb{E}[g(X_1])$ The asymptotic variance is $\mathrm{var}(g(X_1) - \beta^{*T}\boldsymbol{Z}_1) = \sigma_m^2$. □

Remark 3.3.1. *The estimation of $\hat{\beta}_n$ has no effect on the asymptotics. Other estimators can be defined by estimating differently $\mathrm{var}(g(X_1) - \beta^T \boldsymbol{Z}_1)$. They share the same properties as the one stated in Proposition 3.3.1 as long as they estimate consistently $\mathbb{E}[g(X_1)\boldsymbol{Z}_1]$ and $\mathbb{E}[\boldsymbol{Z}_1\boldsymbol{Z}_1^T]$.*

Remark 3.3.2. *Suppose that you have a sequence of control variates $Z_{1,1}, Z_{1,2}, \ldots, Z_{1,m}, Z_{1,m+1}, \ldots$. Then $\sigma_{m+1} \leq \sigma_m$ for any $m \geq 0$. Note that the Monte Carlo variance correspond to σ_0.*

An estimator of the variance of $\hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n)$ is given by

$$\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \left(g(X_i) - \hat{\beta}_n^T \boldsymbol{Z}_i - \hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n) \right)^2.$$

Proposition 3.3.2. *Suppose that $\mathbb{E}[|g(X_1)|^2] < \infty$, $\mathbb{E}[|g(X_1)Z_{k,1}|] < \infty$, $\forall k = 1, \ldots, m$, and that $\mathbb{E}[\boldsymbol{Z}_1\boldsymbol{Z}_1^T]$ is invertible, then $\hat{\sigma}_n^2 \to \sigma_m^2$.*

Proof. Define

$$\tilde{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \left(g(X_i) - \beta^{*T} \boldsymbol{Z}_i - \hat{I}_n^{(\mathrm{cv})}(\beta^*) \right)^2.$$

Write

$$\hat{\sigma}_n^2 - \tilde{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \left(2g(X_i) - (\hat{\beta}_n + \beta^*)^T \boldsymbol{Z}_i - \hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n + \beta^*) \right) \left((\hat{\beta}_n - \beta^*)^T \boldsymbol{Z}_i - \hat{I}_n^{(\mathrm{cv})}(\hat{\beta}_n - \beta^*) \right).$$

Then show that it goes to 0 almost surely, using that $(\hat{\beta}_n - \beta^*) \to 0$ almost surely (as shown in the proof of Proposition 3.3.1). Then we conclude by showing that $\tilde{\sigma}_n^2 \to \sigma_m$ almost surely. □

3.4 Computational complexity

We recall the rules to evaluate the computational complexity.

- Generate X_1 is 1 elementary operation.
- Generate $Z_{1,k}$ is 1 elementary operation for each k.
- Evaluate $g(X_1)$ is 1 elementary operation.

Both complexities of Monte Carlo and Control variates are linear in n, but the constant for Control variates depends in m. Hence m can not be chosen as large as we want. Allowing m^2n point for Monte-Carlo we obtain a variance equal to $(nm^2)^{-1}\operatorname{var}(g(X_1)$ whereas, with n points for the control variates we have an asymptotic variance $n^{-1}\sigma_m^2$. Hence the cornerstone of the control variate method is the accuracy of the L_2-approximation of $g(X_1)$ by $\sum_{k=1}^m \alpha_k Z_{k,1}$ with respect to m. The condition for control variate to overperform Monte Carlo (without considering the constant in the computational time) is

$$\frac{\sigma_m}{\sigma_0} \leq m^{-1}.$$

Monte Carlo	number of operations
$X_1, ..., X_n$	n
$g(X_1), ..., g(X_n)$	n
$\sum g(X_i)$	n
total	n

Computation of $\hat{\beta}_n$	number of operations
$\boldsymbol{Z}_{n,m}^T g_n$	mn
$\boldsymbol{Z}_{n,m}^T \boldsymbol{Z}_{n,m}$	m^2n
Solving $Ax = b$	m^3
total	$m^2n + m^3$

Control variates (given $\hat{\beta}_n$)	number of operations
$X_1, ..., X_n$	n
$\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$	mn
$g(X_i) - \hat{\beta}^T \boldsymbol{Z}_i, i = 1 \dots n$	mn
$\sum(g(X_i) - \hat{\beta}_n^T \boldsymbol{Z}_i)$	n
total	mn

Figure 3.2: Computation time associated to control variates

Exercises

Exercise 3.4.1 (Stratified Monte Carlo)**.** *This exercise investigates a Monte Carlo method for the approximation of $\mathcal{I} = \mathbb{E}[\phi(Z)]$. Let X be an* auxiliary *random variable taking values in $\mathcal{S}$. Let $\{S_k, k \in \{1, \cdots, K\}\}$ be a finite partition of $\mathcal{S}$ such that $p_k = \mathbb{P}(X \in S_k) > 0$. Let $\{Z_i^{(k)}, i \geq 1\}$ be i.i.d. random variables with distribution $\mathbb{P}(Z \in \bullet | X \in S_k)$. We assume that the r.v. $\{Z_i^{(k)}, i \geq 1, k \in \{1, \cdots, K\}\}$ are independent.*

1. *a) Show that*

$$\mathbb{E}[\phi(Z)] = \sum_{k=1}^{K} p_k \mathbb{E}[\phi(Z) | X \in S_k].$$

b) Let $n > 0$ and an allocation *policy $\{q_k, k \in \{1, \cdots, K\}\}$ i.e. $\{q_k, k \in \{1, \cdots, K\}\}$ is a probability distribution on $\{1, \cdots, I\}$. Propose a Monte Carlo estimator based on n_k r.v. $Z_i^{(k)}$, $k \in \{1, \cdots, K\}$ where*

$$n_1 = \lfloor nq_1 \rfloor \qquad n_k = \lfloor n \sum_{j=1}^{k} q_j \rfloor - \lfloor n \sum_{j=1}^{k-1} q_j \rfloor.$$

Hereafter, we assume that for all $k \in \{1, \cdots, K\}$, $q_k > 0$ and n is large enough so that $n_k > 0$.

2. *What is the bias of this estimator ? Is it asymptotically consistant ?*

3. a) *Show that its variance is given by*

$$\frac{1}{n}\sum_{k=1}^{K}\frac{p_k^2}{q_k}\operatorname{var}(\phi(Z)|X \in S_k) + \sum_{k=1}^{K} p_k^2 \left(\frac{1}{n_k} - \frac{1}{nq_k}\right)\operatorname{var}(\phi(Z)|X \in S_k)$$

$$\sim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{I}\frac{p_k^2}{q_k}\operatorname{var}(\phi(Z)|X \in S_k)\ .$$

   b) *Show that the stratified estimator applied with* proportional allocation *($q_k = p_k$) reduces the variance of the naïve Monte Carlo estimator.*

   c) *What is the* optimal allocation *i.e. the allocation minimizing the variance ?*

4. a) *Show that the stratified sampler satisfies a Central Limit Theorem. Provide (asymptotic) confidence intervals.*

   b) *How to compute such confidence intervals when* $\operatorname{var}(\phi(Z)|X \in S_k)$ *is unknown ?*

Exercise 3.4.2.

We are in the same context as Exercise 2.7.2 and we suppose to have at our disposal a good sequence of weights $(\alpha_k^{(\ell)})$ which is now considered to be fixed (non random). Our goal now is to reduce the variance of the procedure using the f_k as control variables. For each $\ell = 1, \dots, L$, let $(X_i^{(\ell)})_{i=1,\dots,n_\ell} \overset{i.i.d.}{\sim} f_{\alpha^{(\ell)}}$, $i = 1, \dots, n_\ell$, and define

$$\hat{I}_\ell = \left(n_\ell^{-1}\sum_{i=1}^{n_\ell}\left\{\frac{g(X_i^{(\ell)}) - f_{\beta^{(\ell)}}(X_i^{(\ell)})}{f_{\alpha^{(\ell)}}(X_i^{(\ell)})}\right\}\right) + c(\beta^{(\ell)}),$$

where $\beta^{(\ell)} = (\beta_1^{(\ell)}, \dots, \beta_K^{(\ell)}) \in \mathbb{R}^K$.

(a) *Give the value of $c(\beta^{(\ell)})$ such that $\hat{I}_\ell$ is an unbiased estimate of $\int_{\mathbb{R}} g \mathrm{d}\lambda$.*

(b) *Let $w_i^\ell = 1/f_{\alpha^{(\ell)}}(X_i^{(\ell)})^2$. Show that*

$$\operatorname{var}(\hat{I}_\ell) = n_\ell^{-1}\left(\mathbb{E}\left[w_i^{(\ell)}\left(g(X_i^{(\ell)}) - \sum_{k=1}^{K}\beta_k^{(\ell)} Z_{k,i}^{(\ell)}\right)^2\right] - \left(\int g \mathrm{d}\lambda\right)^2\right),$$

where $Z_{k,i}^{(\ell)}$ is to be specified (hint : one might use the identity $\operatorname{var}(T) = \mathbb{E}[(T - \mathbb{E}T)^2] = \mathbb{E}[T^2] - \mathbb{E}[T]^2$*).*

(c) *Justify that $\hat{\beta}^{(\ell)}$ minimizing*

$$(\beta_1^{(\ell)}, \dots, \beta_K^{(\ell)}) \mapsto \sum_{i=1}^{n_\ell} w_i^{(\ell)}(g(X_i^{(\ell)}) - \sum_{k=1}^{K}\beta_k^{(\ell)} Z_{k,i}^{(\ell)})^2$$

is a reasonable choice.

(d) *Derive an expression for $\hat{\beta}^{(\ell)}$ using suitable matrix notations.*

(e) *Using the $\hat{\beta}^{(\ell)}$, derive an estimator of $\int g \mathrm{d}\lambda$.*

Chapter 4

Dependent sampling

4.1 Adaptive importance sampling

4.1.1 Presentation

Adaptive importance sampling (AIS) uses past samples to update the *sampling policy* q_t. Each stage t is formed with two steps : (i) to explore the space with n_t points according to q_t and (ii) to exploit the current amount of information to update the sampling policy. The very fundamental question raised in this section concerns the behavior of empirical sums based on AIS. Without making any assumption on the *allocation policy* n_t, the theory developed involves no restriction on the split of computational resources between the explore (i) and the exploit (ii) step. It is shown that AIS is asymptotically optimal : the asymptotic behavior of AIS is the same as some "oracle" strategy that knows the targeted sampling policy from the beginning. From a practical perspective, weighted AIS is introduced, a new method that allows to forget poor samples from early stages.

The adaptive choice of a sampling policy lies at the heart of many fields of *Machine Learning* where former Monte Carlo experiments guide the forthcoming ones. This includes for instance *reinforcment learning* Jie and Abbeel (2010); Peters et al. (2010); Schulman et al. (2015) where the optimal policy maximizes the reward; inference in *Bayesian* Del Moral et al. (2006) or *graphical models* Lou et al. (2017); *optimization* based on stochastic gradient descent Zhao and Zhang (2015) or without using the gradient Hashimoto et al. (2018); *rejection sampling* Erraqabi et al. (2016). *Adaptive importance sampling* (AIS) Oh and Berger (1992); Portier and Delyon (2018), which extends the basic Monte Carlo integration approach, offers a natural probabilistic framework to describe the evolution of sampling policies.

Suppose we are interested in computing some integral value $\int gf$, where $g : \mathbb{R}^d \to \mathbb{R}$ is called the integrand. The importance sampling estimate of $\int gf$ based on the sampling policy q, is given by

$$n^{-1} \sum_{i=1}^{n} \frac{[gf](x_i)}{q(x_i)}, \tag{4.1}$$

where $(x_1, \dots x_n) \overset{\text{i.i.d.}}{\sim} q$. The previous estimate is unbiased. It is well known (Evans and Swartz, 2000) (see also Chapter 2), that the optimal sampling policy, regarding the variance, is when q is proportional to $|g|f$. A slightly different context where importance sampling still applies is Bayesian estimation. Here the targeted quantity is $\int gf$ and we only have access to an unnormalized version f_u of the density $f = f_u / \int f_u$. Estimators usually employed are

$$\sum_{i=1}^{n} \frac{g(x_i) f_u(x_i)}{q(x_i)} \Bigg/ \sum_{i=1}^{n} \frac{f_u(x_i)}{q(x_i)} \ . \tag{4.2}$$

In this case, the optimal sampling policy q is proportional to $|g - \int gf|f$ (Douc et al., 2007).

Because appropriate policies naturally depend on g or f, we generally cannot simulate from them. They are then approximated adaptively, by densities from which we can simulate, using the information gathered from the past stages. This is the very spirit of AIS. At each stage t, the value I_t, standing for the current estimate, is updated using i.i.d. new samples $x_{t,1},\ldots x_{t,n_t}$ from q_t, where q_t is a probability density function that might depend on the past stages $1,\ldots t-1$. The distribution q_t, called the *sampling policy*, targets some optimal, at least suitable, sampling policy. The sequence $(n_t)\subset\mathbb{N}^*$, called the *allocation policy*, contains the number of particles generated at each stage.

The following algorithm describes the AIS schemes for the classical integration problem. For the Bayesian problem, it suffices to change the estimate according to (4.2). This is a generic representation of AIS as no explicit update rule is specified (this will be discussed just below).

Algorithm 6 (AIS).
***Inputs**: The number of stages $T\in\mathbb{N}^*$, the allocation policy $(n_t)_{t=1,\ldots T}\subset\mathbb{N}^*$, the sampler update procedure, the initial density q_0.*

```
Set $S_0 = 0$, $N_0 = 0$. For $t$ in $1,\ldots T$ :
  (i) (Explore) Generate $(x_{t,1},\ldots x_{t,n_t})$ from $q_{t-1}$
  (ii) (Exploit)
      (a) Update the estimate:
              $S_t = S_{t-1} + \sum_{i=1}^{n_t} \frac{[gf](x_{t,i})}{q_{t-1}(x_{t,i})}$
              $N_t = N_{t-1} + n_t$
              $I_t = N_t^{-1} S_t$
      (b) Update the sampler $q_t$
```

The theoretical properties of adaptive schemes are difficult to derive due to the recycling of the past samples at each stage and hence to the lack of independence between samples. Recently, a more realistic asymptotic regime was considered in Marin et al. (2012) in which the allocation policy (n_t) is a fixed growing sequence of integers. The authors establish the consistency of the estimate when the update is conducted with respect to a parametric family but depends *only* on the last stage. They focus on multiple adaptive importance sampling Cornuet et al. (2012) which is different than AIS (see Remark 4.1.2 below for more details).

4.1.2 Central limit theorems for AIS

The aim of the section is to provide conditions on the sampling policy (q_t) under which a central limit theorem holds for AIS and normalized AIS.

For the sake of generality and because it will be useful in the treatment of normalized estimators, we consider the multivariate case where $g=(g_1,\ldots g_p):\mathbb{R}^d\to\mathbb{R}^p$. In the whole paper, $\int gf$ is with respect to the Lebesgue measure, $\|\cdot\|$ is the Euclidean norm, $\mathcal{I}_p$ is the identity matrix of size (p,p).

To study the AIS algorithm, it is appropriate to work at the sample time scale as described below rather than at the sampling policy scale as described in the introduction. The sample $x_{t,i}$ (resp. the policy q_t) of the previous section (t is the block index and i the sample index within the block) is now simply denoted x_j (resp. q_j), where $j=n_1+\ldots n_t+i$ is the sample index in the whole sequence $1,\ldots n$, with $n=N_T$. The following algorithm is the same as Algorithm 6 (no explicit update rule is provided) but is expressed at the sample scale.

Algorithm 7 (AIS at sample scale).

***Inputs**: The number of stages $T \in \mathbb{N}^*$, the allocation policy $(n_t)_{t=1,\dots T} \subset \mathbb{N}^*$, the sampler update procedure, the initial density q_0.*

```
Set $S_0 = 0$. For $j$ in $1, \dots n$ :
  (i) (Explore) Generate $x_j$ from $q_{j-1}$
  (ii) (Exploit)
    (a) Update the estimate:   $S_j = S_{j-1} + \frac{[gf](x_j)}{q_{j-1}(x_j)}$
                               $I_j = j^{-1} S_j$
    (b) Update the sampler $q_j$ whenever $j \in \{N_t = \sum_{s=1}^t n_s : t \geq 1\}$
```

The martingale property

Define Δ_j as the j-th centered contribution to the sum S_j: $\Delta_j = \varphi(x_j)/q_{j-1}(x_j) - \int \varphi$. Define, for all $n \geq 1$,

$$M_n = \sum_{j=1}^n \Delta_j.$$

The filtration we consider is given by $\mathcal{F}_n = \sigma(x_1, \dots x_n)$. The quadratic variation of M is given by $\langle M \rangle_n = \sum_{j=1}^n \mathbb{E}\big[\Delta_j \Delta_j^T \mid \mathcal{F}_{j-1}\big]$. Set

$$V(q, \varphi) = \int \frac{\left(\varphi(x) - q(x) \int \varphi\right) \left(\varphi(x) - q(x) \int \varphi\right)^T}{q(x)} dx. \tag{4.3}$$

Lemma 4.1.1. *Assume that for all $1 \leq j \leq n$, the support of q_j contains the support of φ, then the sequence $(M_n, \mathcal{F}_n)$ is a martingale. In particular, I_n is an unbiased estimate of $\int \varphi$. In addition, the quadratic variation of M satisfies $\langle M \rangle_n = \sum_{j=1}^n V(q_{j-1}, \varphi)$.*

A central limit theorem for AIS

The following theorem describes the asymptotic behavior of AIS. The conditions will be verified for parametric updates (Portier and Delyon, 2018) in which case the asymptotic variance V_* is explicitly given.

Theorem 4.1.2 (central limit theorem for AIS)**.** *Assume that the sequence q_n satisfies*

$$V(q_n, gf) \to V_*, \qquad a.s. \tag{4.4}$$

for some $V_* \geq 0$ and that there exists $\eta > 0$ such that

$$\sup_{j \in \mathbb{N}} \int \frac{\|gf\|^{2+\eta}}{q_j^{1+\eta}} < \infty, \qquad a.s. \tag{4.5}$$

Then we have

$$\sqrt{n}\left(I_n - I_f(g)\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_*).$$

Remark 4.1.1 (zero-variance estimate)**.** *Suppose that $p = 1$ (recalling that $g : \mathbb{R}^d \to \mathbb{R}^p$). Theorem 4.1.2 includes the degenerate case $V_* = 0$. This happens when the integrand has constant sign and the sampling policy is well chosen, i.e. $q_n \to |g|f / \int |g|f$. In this case, we have that $\sqrt{n}(I_n - I_f(g)) = o_p(1)$, meaning that the standard Monte Carlo convergence rate $(1/\sqrt{n})$ has been improved. This is inline with the results presented in Zhang (1996) where fast rates of convergence (compared to standard Monte Carlo) are obtained under restrictive conditions on the allocation policy (n_t). Note that other techniques such as control variates, kernel smoothing or Gaussian quadrature can achieve fast convergence rates Oates et al. (2017); Portier and Segers (2018); Delyon et al. (2016); Bardenet and Hardy (2016).*

Remark 4.1.2 (adaptive multiple importance sampling)**.** *Another way to compute the importance weights, called multiple adaptive importance sampling, has been introduced in Veach and Guibas (1995) and has been successfully used in Owen and Zhou (2000); Cornuet et al. (2012). This consists in replacing q_{j-1} in the computation of S_j by $\bar{q}_{j-1} = \sum_{i=1}^{j} q_{i-1}/j$, x_j still being drawn under q_{j-1}. The intuition is that this averaging will reduce the effect of exceptional points x_j for which $f(x_j) \gg q_{j-1}(x_j)$ (but $|f(x_j)| \not\gg \bar{q}_{j-1}(x_j)$). Our approach is not able to study this variant, simply because the martingale property described previously is not anymore satisfied.*

Normalized AIS

The normalization technique described in (4.2) is designed to compute $\int gf$, where f is a density. It is useful in the Bayesian context where f is only known up to a constant. As this technique seems to provide substantial improvements compared to unnormalized estimates (as described in the previous section), we recommend to use it even when the normalized constant of f is known. Normalized estimators are given by

$$I_n^{(\text{norm})} = \frac{I_n(gf)}{I_n(f)}, \qquad \text{with} \quad I_n(\psi) = n^{-1} \sum_{j=1}^{n} \psi(x_j)/q_{j-1}(x_j).$$

Interestingly, normalized estimators are weighted least-squares estimates as they minimize the function $a \mapsto \sum_{j=1}^{n} (f(x_j)/q_{j-1}(x_j))(g(x_j) - a)^2$. In contrast with I_n, $I_n^{(\text{norm})}$ has the following shift-invariance property : whenever g is shifted by μ, $I_n^{(\text{norm})}$ simply becomes $I_n^{(\text{norm})} + \mu$. Because $I_n(gf)$ and $I_n(f)$ are of the same kind as I_n defined in the second AIS algorithm, a straightforward application of Theorem 4.1.2 (with $(g^T f, f)^T$ in place of gf).

Theorem 4.1.3 (central limit theorem for normalized AIS)**.** *Suppose that (4.4) and (4.5) hold with $(g^T f, f)^T$ (in place of gf). Then we have*

$$\sqrt{n}\Big(I_n^{(norm)} - I_f(g)\Big) \xrightarrow{\text{d}} \mathcal{N}(0, UV_*U^T),$$

with $U = (I_p, -I_f(g))$.

4.2 Metropolis-Hasting Algorithm

Bayesian estimation requires to compute moments of the so called *posterior distribution* whose probability density function f is given by

$$f(\theta) = \frac{\mathcal{L}(\theta)}{\int \mathcal{L}(\theta) d\theta} \qquad \theta \in \mathbb{R}^d,$$

where $\mathcal{L}$ is a positive function which stands for the likelihood of the observed data. The (unknown) quantities of interest writes as $\int gf$, for some given measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. A particular feature in this framework is that the integral at the denominator of f is unknown and difficult to compute making impossible to generate observations directly from f. Markov chains Monte Carlo (MCMC) methods aim to produce samples $X_1, \ldots, X_n$ in $\mathbb{R}^d$ that are approximately distributed according to f. Then $\int g df$ is classically approximated by the empirical average over the chain :

$$n^{-1} \sum_{i=1}^{n} g(X_i).$$

For inference, Bayesian credible intervals are usually computed using the quantiles the coordinate chains (see below). We refer to Robert and Casella (2004) for a complete description of MCMC methods. In what follows, we focus on the special MCMC method called Metropolis-Hasting (MH) algorithm which is one of the state of the art method in computational statistics and is frequently used to compute Bayesian estimators (Robert and Casella, 2004).

Let us introduce the MH algorithm with target density $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and proposal $Q(x, dy) = q(x, y)dy$, where q is a positive function defined on $\mathbb{R}^d \times \mathbb{R}^d$ satisfying $\int q(x, y)dy = 1$. Define for any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\rho(x, y) = \begin{cases} \min\left(1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right) & \text{if } f(x)q(x, y) > 0, \\ 1 & \text{if } f(x)q(x, y) = 0. \end{cases}$$

The MH chain starts at $X_0 \sim \nu$ and moves from X_n to X_{n+1} according to the following rule:

(i) Generate

$$Y \sim Q(X_n, dy) \qquad \text{and} \qquad W \sim \mathcal{B}(\rho(X_n, Y)).$$

(ii) Set

$$X_{n+1} = \begin{cases} Y & \text{if} \quad W = 1, \\ X_n & \text{if} \quad W = 0. \end{cases}$$

In the particular case that $q(x, y) = q_0(x - y)$, the previous algorithm is refereed to as the random walk MH.

Appendix A

Convergence results

In the context of importance sampling theory, the concept of uniform integrability is useful to obtain the continuity of some integral with respect to some parameter. Actually, it permits to alleviate some assumptions that would be needed if the Lebesgue dominated convergence theorem would be used. Let $(\omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence of random variable $(X_n)_{n\geq 1}$ is said to be uniformly integrable whenever

$$\lim_{x\to\infty} \sup \sup_{n\geq 1} \mathbb{E}[|X_n| 1_{\{|X_n|>x\}}] = 0.$$

The following property caracterizes the L_1 convergence with the help of uniform integrability.

Proposition A.0.1. *A sequence $(X_n)_{n\geq 1}$ is uniformly integrable and $X_n \to X$ in probability if an only if $\mathbb{E}[|X_n - X|] \to 0$.*

We present here a simple way to obtain the law of large numbers.

Theorem A.0.2. *Let $U_n, n \geq 1$ be a sequence of random variables and $S_n = U_1 + U_2 + ...U_n$ such that:*

$$\begin{aligned} &U_n \geq 0 \quad w.p.1 \\ &n^{-1}\mathbb{E}[S_n] \longrightarrow l \\ &\text{var}(S_n) \leq cn \end{aligned}$$

for some real numbers $c \geq 0$ and $l \geq 0$, then

$$\frac{S_n}{n} \longrightarrow l \quad w.p.1.$$

Proof. The trick in this proof is to first derive the result for S_{n^2}/n^2. Then a sandwich formula will permit to conclude for S_n/n. We have

$$\mathbb{E}\left[\sum_n \left(\frac{S_{n^2} - \mathbb{E}[S_{n^2}]}{n^2}\right)^2\right] \leq \sum_n \frac{c}{n^2} < \infty.$$

Thus

$$\sum_n \left(\frac{S_{n^2} - \mathbb{E}[S_{n^2}]}{n^2}\right)^2 \text{ is finite w.p.1,}$$

implying that $(S_{n^2} - \mathbb{E}[S_{n^2}])/n^2$ converges to zero, almost surely. Hence S_{n^2}/n^2 converges to l. Notice that if $n^2 \leq k \leq (n+1)^2$:

$$\frac{S_{n^2}}{n^2}\frac{n^2}{(n+1)^2} \leq \frac{S_k}{k} \leq \frac{S_{(n+1)^2}}{(n+1)^2}\frac{(n+1)^2}{n^2}$$

and since both side terms tend to l, the result is proved. □

The following theorem will be usefull to deal with 2-steps importance sampling estimate. This theorem is given in van der Vaart (1998).

Theorem A.0.3 (Lindeberg-Feller central limit theorem)**.** *For each $n \geq 1$, let $Y_{n,1}, \ldots, Y_{n,n}$ be independent random vectors with finite variance* $\operatorname{var}(Y_{i,n}) < \infty$ *such that*

$$\sum_{i=1}^{n} E[\|Y_{n,i}\|^2 1_{\{\|Y_{n,i}\| > \epsilon\}}] \to 0, \qquad \forall \epsilon > 0,$$
$$\sum_{i=1}^{n} \operatorname{var}(Y_{n,1}) \to \Sigma,$$

then, $\sum_{i=1}^{n}(Y_{n,i} - E[Y_{n,i}]) \rightsquigarrow \mathcal{N}(0, \Sigma)$.

The following result is a central limit theorem for martingale arrays. A reference textbook is (Hall and Heyde, 2014, Corollary 3.1). It will be useful to prove central limit theorems for adaptive importance sampling schemes.

Theorem A.0.4 (central limit theorem for martingales)**.** *Let $(W_{n,i})_{1 \leq i \leq n,\, n \geq 1}$ be a triangular array of random variables such that*

$$\mathbb{E}[W_{n,i} \mid \mathcal{F}_{i-1}] = 0, \quad \textit{for all } 1 \leq i \leq n,$$
$$\sum_{i=1}^{n} \mathbb{E}[W_{n,i}^2 \mid \mathcal{F}_{i-1}] \to v^* \geq 0, \quad \textit{in probability,}$$
$$\sum_{i=1}^{n} \mathbb{E}[W_{n,i}^2 \mathrm{I}_{\{|W_{n,i}| > \varepsilon\}} \mid \mathcal{F}_{i-1}] \to 0, \quad \textit{in probability,}$$

then, $\sum_{i=1}^{n} W_{n,i} \rightsquigarrow \mathcal{N}(0, v^*)$.

The following result is a concentration inequality for martingale arrays. It is a modification of (Freedman et al., 1975, Theorem 4.1), allowing the martingale increments to be unbounded. This result takes into account the rate of decrease of the quadratic variation (v appears in the bound), it will play a crucial role to control the behavior of kernel estimator (in proving Theorem **??**) for which the quadratic variation will depend on the bandwidth h_n.

Theorem A.0.5. *Let $(Y_i)_{1 \leq i \leq n}$ be random variables such that*

$$\mathbb{E}[Y_i \mid \mathcal{F}_{i-1}] = 0, \quad \textit{for all } 1 \leq i \leq n,$$

then, for all $t \geq 0$ and $v, m > 0$,

$$\mathbb{P}\left(\Big|\sum_{i=1}^{n} Y_i\Big| \geq t,\ \max_{i=1,\ldots,n} |Y_i| \leq m,\ \sum_{i=1}^{n} \mathbb{E}[Y_i^2 \mid \mathcal{F}_{i-1}] \leq v\right) \leq 2\exp\left(-\frac{t^2}{2(v + tm/3)}\right).$$

Bibliography

Bardenet, R. and A. Hardy (2016). Monte carlo with determinantal point processes. *arXiv preprint arXiv:1605.00361*.

Cornuet, J., J.-M. Marin, A. Mira, and C. P. Robert (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics 39*(4), 798–812.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(3), 411–436.

Delyon, B., F. Portier, et al. (2016). Integral approximation by kernel smoothing. *Bernoulli 22*(4), 2177–2208.

Douc, R., A. Guillin, J.-M. Marin, and C. P. Robert (2007). Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics 11*, 427–447.

Erraqabi, A., M. Valko, A. Carpentier, and O. Maillard (2016). Pliable rejection sampling. In *International Conference on Machine Learning*, pp. 2121–2129.

Evans, M. and T. Swartz (2000). *Approximating integrals via Monte Carlo and deterministic methods.* Oxford Statistical Science Series. Oxford University Press, Oxford.

Freedman, D. A. et al. (1975). On tail probabilities for martingales. *the Annals of Probability 3*(1), 100–118.

Hall, P. and C. C. Heyde (2014). *Martingale limit theory and its application.* Academic press.

Hashimoto, T. B., S. Yadlowsky, and J. C. Duchi (2018). Derivative free optimization via repeated classification. *arXiv preprint arXiv:1804.03761*.

Hull, J. and A. White (1988). The use of the control variate technique in option pricing. *Journal of Financial and Quantitative analysis 23*(03), 237–251.

Jie, T. and P. Abbeel (2010). On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*, pp. 1000–1008.

Lou, Q., R. Dechter, and A. T. Ihler (2017). Dynamic importance sampling for anytime bounds of the partition function. In *Advances in Neural Information Processing Systems*, pp. 3199–3207.

Marin, J.-M., P. Pudlo, and M. Sedki (2012). Consistency of the adaptive multiple importance sampling. *arXiv preprint arXiv:1211.2548*.

Novak, E. (2016). Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pp. 161–183. Springer.

Oates, C. J., M. Girolami, and N. Chopin (2017). Control functionals for Monte Carlo integration. *J. R. Statist. Soc. B 79*(3), 695–718.

Oh, M.-S. and J. O. Berger (1992). Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation 41*(3-4), 143–168.

Owen, A. and Y. Zhou (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc. 95*(449), 135–143.

Peters, J., K. Mülling, and Y. Altun (2010). Relative entropy policy search. In *AAAI*, pp. 1607–1612. Atlanta.

Portier, F. and B. Delyon (2018). Asymptotic optimality of adaptive importance sampling. In *Advances in Neural Information Processing Systems*, pp. 3138–3148.

Portier, F. and J. Segers (2018). Monte carlo integration with a growing number of control variates. *arXiv preprint arXiv:1801.01797*.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. Springer-Verlag, New York.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz (2015). Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Veach, E. and L. J. Guibas (1995). Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 419–428. ACM.

Wang, C., X. Chen, A. J. Smola, and E. P. Xing (2013). Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, pp. 181–189.

Zhang, P. (1996). Nonparametric importance sampling. *J. Amer. Statist. Assoc. 91*(435), 1245–1253.

Zhao, P. and T. Zhang (2015). Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pp. 1–9.