CONVEX OPTIMIZATION FOR MONTE CARLO: STOCHASTIC OPTIMIZATION FOR IMPORTANCE SAMPLING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF INSTITUTE FOR
COMPUTATIONAL AND MATHEMATICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ernest K. Ryu
May 2016

This dissertation is online at: http://purl.stanford.edu/kb323fs8835

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Stephen Boyd, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

John Duchi

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Art Owen

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Acknowledgments

First, I would like to thank my friends and family. It is no hyperbole for me to say I would not be where I am without them.

I would also like to thank the Cortlandt and Jean E. Van Rensselaer Engineering Fellowship Fund, Department of Energy and, the Simons Foundation for supporting my graduate studies.

Finally, I would like to thank my advisor Stephen Boyd. He is a great scholar, a great teacher, and a great advisor. I have been extremely fortunate to work with him, and I have tried to learn every bit I could from him during my time at Stanford.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In Monte Carlo simulations it is often essential that a method is accompanied by an appropriate variance reduction method. Reducing the variance of a Monte Carlo method is, at least conceptually, an optimization problem, and mathematical optimization has indeed been used as a theoretical and conceptual tool in this pursuit.

However, traditional Monte Carlo methods have only used numerical optimization sparingly, and convex optimization even less. Numerical optimization is study of algorithms for finding a solution to an optimization problem, as opposed to the study of analytical solutions of an optimization problem. Convex optimization is the study of convex optimization problems, a subclass of optimization problems for which efficient algorithms for finding the global optimum exists.

In this work we present a framework for using convex optimization for Monte Carlo. More specifically, we present a framework for using stochastic convex optimization for adaptive importance sampling, self-normalized importance sampling, and what-if simulations.

The main idea is to perform importance sampling and numerical optimization simultaneously. In particular, the numerical optimization does not rely on black-box optimization solvers, and this allows the computational cost of each iteration to remain cheap. Because the optimization is performed on a convex problem, we can establish convergence and optimality.

Previous work. The earliest uses of Monte Carlo simulations are attributed to Enrico Fermi, Stanislaw Ulam, and John von Neumann in the 1930s and 1940s [51] and the earliest formal publication to Metropolis and Ulam in 1949 [52]. In the early 1950s, importance sampling was discovered and studied by Herman Kahn. [40, 41, 42]. In 1956, Hale Trotter and John Tukey generalized the idea to self-normalized importance sampling and what-if simulations [72]. Since then there has been a large body of work on Monte Carlo methods and importance sampling. Adaptive importance sampling was first studied in the 1970s [71, 66, 32, 57].

Some previous work on adaptive importance sampling have used stochastic optimization methods such as stochastic subgradient descent without much regard to convexity [1, 2, 3, 28]. While these methods are applicable to a more general class of candiate sampling distributions, they have little theoretical guarantees on the variances of the estimators; this is not surprising since in nonconvex optimization it is difficult to prove anything beyond mere convergence to a stationary point, such as a rate of convergence or convergence to the global optimum.

Other previous work on adaptive importance sampling solves an optimization subproblem to update the sampling parameter each time, either with an off-the-shelf deterministic optimization algorithm or, especially in the case of the cross-entropy method, by focusing on special cases with analytic solutions [57, 25, 26, 49, 64, 65, 24, 23, 59, 27, 19, 21, 36]. While some these methods do exploit convexity to establish that the subproblems can be solved efficiently, these subproblems and the storage requirement to represent these subproblems grow in size with the number of iterations. One could loosely argue that the inefficiency is a consequence of separating the optimization and the importance sampling.

That convex optimization problems are the subset of optimization problems for which we can find efficient and reliable solution methods is well-known and is the basis of the field of convex optimization [54, 60, 8, 15, 56, 11, 18].

The optimization methods presented in this work are relatively standard within the field of optimization. However, we occasionally show convergence results as the standard results are not exactly in the form we need. (Sub)gradient descent was first studies by Augustin-Louis Cauchy in the 1840s and and stochastic (sub)gradient

descent by Hertbert Robbins and Sutton Monro in the 1950s. [20, 62]. Since then there has been a large body of work on this method [68, 60, 44, 14]. Mirror descent was introduced by Arkadi Nemirovski and David Yudin [54]. Since then there has been a large body of work on this method [7, 53]. The origins of sample average approximation can be traced back to maximum likelihood estimation, but was first studied as a stochastic optimization algorithm by Alexander Shapiro [67]. Stochastic saddle point subgradient descent was first studied by Kenneth Arrow, Leonid Hurwicz, and Arkadi Nemirovski [4, 55], and the generalization to stocahstic saddle point mirror descent was done by Nemirovski et al. [53]. Stochastic optimization with biased gradients was first studied by Boris Polyak [60].

Contributions. In this work we present a framework for using stochastic convex optimization for adaptive importance sampling, self-normalized importance sampling, and what-if simulations. In sections 2.3 and 2.4, we present the insight that log-concave families paired with a Rényi divergence results in a convex setup. How exactly these notions relate to importance sampling is illustrated in later sections. The connection between the Rényi divergence and importance sampling was first was discussed in [50]. The Rényi divergence was first introduced by Alfréd Rényi [61] as an information theoretic quantity, and a comprehensive analysis of its properties can be found in [73].

In Sections 3 and 4 (more specifically Sections 3.4 and 4.4) we present adaptive importance sampling and adaptive self-normalized importance sampling algorithms, accompanied by a theoretical analysis of their performance. Because of the algorithms use stochastic optimization, their iterations are computationally simple. Because of convexity we can theoretically establish convergence rates and make claims on optimality.

In Section 5 we extend this approach to what-if simulations. We present adaptive algorithms that take advantage of convexity and stochastic optimization, which estimates asymptotically have minimum maximum variance.

Outline. In Section 2, we introduce preliminary concepts not immediately related to Monte Carlo simulation. In Section 3, 4, 5 we respectively introduce importance sampling, self-normalized importance sampling, and what if simulations and show how the material of Section 2 and convex optimization can be applied to obtain adaptive methods with theoretical guarantees. In Section 6, we discuss topics omitted during Sections 3, 4, and 5.

Chapter 2

Preliminaries

2.1 Convexity, subgradients, and stochastic subgradients

A set $\Theta \subseteq \mathbf{R}^p$ is convex if $\eta\theta_1 + (1-\eta)\theta_2 \in \Theta$ for all $\theta_1, \theta_2 \in \Theta$ and $\eta \in [0,1]$. A function $U : \Theta \to \mathbf{R} \cup \{\infty\}$ is convex on Θ if Θ is a convex set and

$$U(\eta\theta_1 + (1-\eta)\theta_2) \leq \eta U(\theta_1) + (1-\eta)U(\theta_2)$$

for all $\theta_1, \theta_2 \in \Theta$ and $\eta \in [0,1]$. We use the convention that $\infty \leq \infty$. We say a function U is concave if $-U$ is convex and log-concave if $\log U$ is concave.

An optimization problem

$$\begin{array}{ll} \text{minimize} & U(\theta) \\ \text{subject to} & \theta \in \Theta, \end{array}$$

where $\theta \in \mathbf{R}^p$ is the optimization variable, is convex if the *constraint set* Θ is a convex set and the *objective function* U is a convex function. Loosely speaking, convex optimization problems can be solved efficiently while most non-convex optimization problems cannot.

Let $U : \Theta \to \mathbf{R} \cup \{\infty\}$ be a convex function on Θ. If U is differentiable at θ_0, then

$$U(\theta) \geq U(\theta_0) + \nabla U(\theta_0)^T(\theta - \theta_0)$$

for all $\theta \in \Theta$ [63, Theorem 25.1]. If U is not differentiable at θ_0, we use subgradients, a generalization of gradients to non-differentiable functions.

We call g a *subgradient* of U at θ_0 if it satisfies

$$U(\theta) \geq U(\theta_0) + g^T(\theta - \theta_0)$$

for all $\theta \in \Theta$. If (and only if) U is differentiable at θ_0, there is exactly one subgradient of U at θ_0, namely $\nabla U(\theta_0)$. When U is not differentiable at θ_0, there can be more than one subgradient at a given point, and we write $\partial U(\theta_0)$ for the *set* of subgradients of U at θ_0. Roughly speaking, $\partial U(\theta_0)$ is usually nonempty, i.e., a subgradient of U at θ_0 usually exists, provided $U(\theta_0) < \infty$ [63, Theorem 23.4].

Let g be a random variable on $\mathbf{R}^p$. If U is differentiable at θ_0 and

$$\mathbb{E}g = \nabla U(\theta_0)$$

we say g is a *stochastic gradient* of U at θ_0. If

$$\mathbb{E}g \in \partial U(\theta_0)$$

then we say g is a *stochastic subgradient* of U at θ_0.

Remark. All of the actual algorithms presented in this work do not explicitly use subgradients. In fact, simply assuming differentiability throughout and mentally replacing the notation ∂ with ∇ should cause little if any problems.

However, the use of subgradients are actually necesary for a rigorous discussion for the following reasons. First, derivatives are not defined on the boundary of a function's domain, and not all convex functions are differentiable. However, these are not serious issues. One could generalize the standard definition of derivatives to

address this technical detail on the boundary, and the non-differentiable functions we deal with turns out to be almost surely differentiable (c.f. Lemma 12).

The most important reason is that the expectation of a subgradient is a subgradient, but the expectation of a gradient need not be a gradient. Specifically, if $u(\theta; x)$ is a convex function of θ for all x, then the assertion

$$\int \nabla_\theta u(\theta; x)\, d\mu(x) = \nabla_\theta \int u(\theta; x)\, d\mu(x)$$

is usually true but requires justification. On the other hand, if $g(x) \in \partial_\theta u(\theta; x)$, then

$$\int g(x)\, d\mu(x) \in \partial_\theta \int u(\theta; x)\, d\mu(x)$$

is always true so long as the integrals are well-defined. See Section 8.1 for further discussion.

So by using subgradients, we can avoid the burden of establishing differentiability altogether. In any case, mulling over differentiability is entirely unnecessary as none of our results depend on differentiability.

2.2 Stochastic optimization

Consider the convex optimization problem

$$\begin{array}{ll} \text{minimize} & U(\theta) \\ \text{subject to} & \theta \in \Theta. \end{array} \tag{2.1}$$

Roughly speaking, one can efficiently find a global minimum of problem (2.1) through standard methods if one can compute $U(\theta)$, a subgradient in $\partial U(\theta)$, and the projection onto the set Θ. In later sections, however, we will encounter optimization problems where evaluating $U(\theta)$ or a subgradient for any given θ is not feasible.

For such problems, we use *stochastic optimization* methods, which only require stochastic information such as stochastic subgradients. Because U is convex, these methods find the global minimum with reasonable rates of convergence.

2.2.1 Stochastic subgradient descent

The stochastic optimization method *stochastic subgradient descent* solves problem (2.1) with the algorithm

$$\theta_{n+1} = \Pi_\Theta(\theta_n - \alpha_n g_n), \tag{2.2}$$

where $\theta_1 \in \Theta$ is some starting point, g_n is a stochastic subgradient of U at θ_n, $\alpha_n > 0$ is a sequence of *step sizes*, and Π_Θ is the projection onto Θ. (Since θ_n is random, we mean $\mathbb{E}[g_n|\theta_n] \in \partial U(\theta_n)$ when we say g_n is a subgradient.) The intuition is that $-g_n$, although noisy, generally points towards a descent direction of U at θ_n, and therefore each step reduces the function value of U in expectation.

The method requires us to choose the step size α_n. There is a large body of research investigating the choices of step sizes that ensure convergence and their rates of convergence. For the sake of simplicity, we will only consider the choice $\alpha_n = C/\sqrt{n}$.

This still leaves us with us to choose the parameter C and the starting point θ_1. In general, there is no good way to choose these parameters. They should be chosen through several informal iterations of trying what works well.

As we will see later, the convergence result we need for the stochastic optimization method is

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}U(\theta_i) \to U(\theta_\star). \tag{2.3}$$

One may be tempted to first establish one of the usual notions of convergence such as

$$\theta_n \to \theta_\star \tag{2.4}$$

or

$$\mathbb{E}U(\theta_n) \to U(\theta_\star), \tag{2.5}$$

and then use these to establish (2.4). However, it is often better to show (2.3) directly. In fact, (2.4) and (2.3) are different notions of convergence that do not necessarily imply each other. Also, the rate we obtain by working with (2.3) directly is better than the rate one would obtain by showing rate on (2.5) first. See [44] for a discussion

on the different notions of convergence of stochastic gradient descent.

Convergence proof. Let us establish when stochastic subgradient descent converges.

Lemma 1. *Assume Θ is a nonempty convex compact set and U has a subgradient for all $\theta \in \Theta$. Also assume $\mathbb{E}[\|g_n\|_2^2|\theta_n] \leq G^2 < \infty$ for $n = 1, 2, \ldots$. (Also assume we use stepsize $\alpha_n = C/\sqrt{n}$.) Then algorithm (2.2) converges with rate*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}U(\theta_i) \leq U(\theta_\star) + \mathcal{O}(1/\sqrt{n}).$$

Proof. The assumption that U has a subgradient on all of Θ implies U is finite on all of Θ by definition of subgradients. The assumption also implies that U is lower-semicontinuous [63, Corollary 23.5.2], and a lower-semicontinuous function with compact domain always has a minimizer [10, Proposition 3.2.1]. We write $\theta_\star$ for a minimizer, and D for the diameter of Θ.

Then we have

$$\begin{aligned}
\|\theta_{i+1} - \theta_\star\|_2^2 &= \|\Pi(\theta_i - C/\sqrt{i}g_i) - \Pi(\theta_\star)\|_2^2 \\
&\leq \|\theta_i - C/\sqrt{i}g_i - \theta_\star\|_2^2 \\
&= \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}\|g_i\|_2^2 - 2\frac{C}{\sqrt{i}}g_i^T(\theta_i - \theta_\star),
\end{aligned}$$

where the first inequality follows from nonexpansivity of Π (c.f. Lemma 13 of the appendix). We take expectation conditioned on θ_i on both sides to get

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{i+1} - \theta_\star\|_2^2 \mid \theta_i\right] &\leq \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}\mathbb{E}\left[\|g_i\|_2^2 \mid \theta_i\right] - 2\frac{C}{\sqrt{i}}\mathbb{E}[g_i|\theta_i]^T(\theta_i - \theta_\star) \\
&\leq \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}G^2 - 2\frac{C}{\sqrt{i}}\mathbb{E}[g_i|\theta_i](\theta_i - \theta_\star) \\
&\leq \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}G^2 - 2\frac{C}{\sqrt{i}}(U(\theta_i) - U(\theta_\star)),
\end{aligned}$$

where the second inequality follows from the definition of G and the third inequality

follows from re-arranging the following consequence of g_i being a stochastic subgradient

$$U(\theta_\star) \geq U(\theta_i) + \mathbb{E}[g_i|\theta_i]^T(\theta_\star - \theta_i).$$

We take the full expectation on both sides and re-arrange to get

$$\mathbb{E}U(\theta_i) - U(\theta_\star) \leq \frac{\sqrt{i}}{2C}(\mathbb{E}\|\theta_i - \theta_\star\|_2^2 - \mathbb{E}\|\theta_{i+1} - \theta_\star\|_2^2) + \frac{C}{2\sqrt{i}}G^2.$$

We take a summation to get an "almost telescoping" series:

$$\begin{aligned} 2\sum_{i=1}^{n}(\mathbb{E}U(\theta_i) - U(\theta_\star)) &\leq \frac{1}{C}\sum_{i=1}^{n}(\sqrt{i} - \sqrt{i-1})\mathbb{E}\|\theta_i - \theta_\star\|_2^2 + CG^2\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \\ &\leq \frac{D^2}{C}\sum_{i=1}^{n}(\sqrt{i} - \sqrt{i-1}) + CG^2\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \\ &\leq \frac{D^2}{C}\sqrt{n} + 2CG^2\sqrt{n}, \end{aligned}$$

where the second inequality follows from the definition of D and the third inequality follows from

$$\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \leq \int_0^n \frac{1}{\sqrt{i}}\, di.$$

Finally, we divide both sides by $2n$ to get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U(\theta_i) \leq U(\theta_\star) + \left(\frac{D^2}{2C} + CG^2\right)\frac{1}{\sqrt{n}}.$$

□

Discussion of assumptions. That U has a subgradient on all of Θ is not a strong assumption so long as $U < \infty$ on Θ [63, Theorem 23.4] and is made to rule out pathologies. That Θ is closed is necessary for the projection to be well-defined but that Θ is bounded is usually unnecessary in practice. That the subgradients are bounded $\mathbb{E}\|g_i\|^2 \leq G < \infty$ is necessary, but this assumption will almost always hold in practice.

Loosely speaking, the only real assumption is that $U < \infty$ on Θ. We will see later that this corresponds to the assumption that the importance sampling estimator has finite variance. There is no simple way to prevent this problem, but all importance sampling methods face this same issue.

Batching. One common variation of stochastic subgradient descent is *batching*, which key insight is that an average of stochastic subgradients is a stochastic subgradient itself. *Stochastic subgradient descent with batch size m* solves Problem 2.1 with

$$g_n = \frac{1}{m} \sum_{i=1}^{m} g_{ni}$$
$$\theta_{n+1} = \Pi_\Theta(\theta_n - \alpha_n g_n),$$

where $g_{n1}, \ldots, g_{nm}$ are stochastic subgradients of U at θ_n. One could say that plain stochastic gradient descent has batch size 1.

Batching is essential in practice; not only does it make parallelization easier, but it also makes the stochastic optimization method more stable, loosely speaking. In fact, all numerical examples in this work use batching. There is a trade-off in choosing batch sizes. Given the same computational resources, a small batch size results in many inaccurate updates while a large batch size results in few accurate updates. In this work, we do not address where the best point within this trade-off is and refer interested readers to [48].

2.2.2 Stochastic mirror descent

A scrutiny of stochastic gradient descent and its convergence analysis reveals that the Euclidean norm plays a special role in it. Because of this, the performance of stochastic gradient descent can be poor when the geometry of the problem is not very Euclidean, loosely speaking.

Stochastic mirror descent is a generalization of stochastic gradient descent that can adapt to the geometry of the problem better than stochastic subgradient descent,

roughly speaking. Here we avoid an in depth discussion of stochastic mirror descent and merely present two versions of mirror descent that we later use.

An instance of stochastic mirror descent solves the problem

$$\begin{array}{ll} \text{minimize} & U(\theta) \\ \text{subject to} & \mathbf{1}^T\theta = 1, \quad \theta \succeq 0, \end{array} \tag{2.6}$$

where $\theta \in \mathbf{R}^p$ is the optimization variable and $\mathbf{1} \in \mathbf{R}^p$ is the vector containing all ones, with the algorithm

$$\begin{aligned} \theta^*_{n+1} &= \theta^*_n - \alpha_n g_n \\ \theta_{n+1} &\propto \exp(\theta^*_{n+1}) \end{aligned}$$

where $\theta^*_1 \in \mathbf{R}^p$ is some starting point, g_n is a stochastic subgradient of U at θ_n, and $\alpha_n > 0$ is a sequence of step sizes. The operation $\exp(\theta^*_{n+1})$ is evaluated element-wise, and the $\propto$ notation means

$$\theta_{n+1} = \frac{1}{\mathbf{1}^T \exp(\theta^*_{n+1})} \exp(\theta^*_{n+1}).$$

The variables $\theta^*_1, \theta^*_2, \ldots$ are called the *mirrored* variables of $\theta_1, \theta_2, \ldots$.

Also consider the optimization problem

$$\begin{array}{ll} \text{minimize} & U(\theta) \\ \text{subject to} & \theta \succ 0, \end{array}$$

where $\theta \in \mathbf{S}^p_{++}$ is the optimization variable and $\mathbf{S}^p_{++}$ denotes the set of $p \times p$ symmetric positive definite matrices. (So the optimization variable θ is a matrix.) An instance of stochastic mirror descent solves this problem with

$$\begin{aligned} \theta^*_{n+1} &= \theta^*_n - \alpha_n g_n \\ \theta_{n+1} &= \exp(\theta^*_{n+1}), \end{aligned}$$

where exp denotes the matrix exponential [34, §9.3].

Convergence proof. Let us prove that the the first algorithm solves problem (2.6).

Lemma 2. *Assume U has a subgradient at θ for all θ that satisfies $\mathbf{1}^T\theta = 1$ and $\theta \succeq 0$. Also assume $\mathbb{E}[\|g_n\|_\infty^2|\theta_n] \leq G^2 < \infty$ for $n = 1, 2, \ldots$, where $\|\cdot\|_\infty$ denotes the ℓ_∞ norm [39, §5.2]. Then* (2.6) *converges with rate*

$$\frac{1}{n}\sum_{i=1}^{n} U(\theta_i) = U(\theta_\star) + \mathcal{O}(\log n/\sqrt{n}).$$

Proof. As discussed in Section 2.2.1, that U has subgradients on all of Θ implies that U is finite within the constraint set and that a solution $\theta_\star$ exists.

We can rewrite the algorithm as

$$\theta_{n+1}(j) = \theta_n(j)\exp(-\alpha_n g_n(j)),$$

where $j = 1, \ldots, p$ corresponds to the p entries of the vectors. Define

$$B(a\|b) = \sum_{j=1}^{n} a(j)\log(a(j)/b(j)),$$

which is an instance of the *Bregman divergence* [16]. (We use the convention $0\log 0 = 0$.)

The following facts are easy to verify: $B(a\|b) \geq 0$, $B(a\|b) = 0$ if and only if $a = b$, and $B(\theta_\star\|\theta_1) < \infty$ since all entries of θ_1 are positive. It is not easy to verify

$$B(a\|b) \geq \|a - b\|_1^2,$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm, but it is true [7, Proposition 5.1].

After some algebra we get

$$\begin{aligned}
B(\theta_\star\|\theta_{i+1}) &= B(\theta_\star\|\theta_i) - \alpha_i(\theta_i - \theta_\star)^T g_i - \alpha_i(\theta_{i+1} - \theta_i)^T g_i - B(\theta_{i+1}\|\theta_i) \\
&\leq B(\theta_\star\|\theta_i) - \alpha_i(\theta_i - \theta_\star)^T g_i + \frac{\alpha_i^2}{4}\|g_i\|_\infty^2 + \|\theta_{i+1} - \theta_i\|_1^2 - B(\theta_{i+1}\|\theta_i) \\
&\leq B(\theta_\star\|\theta_i) - \alpha_i(\theta_i - \theta_\star)^T g_i + \frac{\alpha_i^2}{4}\|g_i\|_\infty^2
\end{aligned}$$

The first line can be verified through straightforward algebra, the second line follows from an instance of Young's inequality [74, 30]

$$a^T b \leq \|a\|_1^2 + \frac{1}{4}\|b\|_\infty^2,$$

and the third line follows from the inequality previously discussed.

We take conditional expectations on both sides to get

$$\mathbb{E}[B(\theta_\star\|\theta_{i+1}) \mid \theta_i] \leq B(\theta_\star\|\theta_i) - \alpha_i(U(\theta_i) - U(\theta_\star)) + \frac{\alpha_i^2}{4}G^2.$$

We take the full expectations on both sides to get

$$\mathbb{E}B(\theta_\star\|\theta_{i+1}) \leq \mathbb{E}B(\theta_\star\|\theta_i) - \alpha_i(\mathbb{E}U(\theta_i) - U(\theta_\star)) + \frac{\alpha_i^2}{4}G^2.$$

We sum both sides to get

$$\begin{aligned}\mathbb{E}B(\theta_\star\|\theta_{n+1}) &\leq B(\theta_\star\|\theta_1) - C\sum_{i=1}^{n-1}\frac{1}{\sqrt{i}}(\mathbb{E}U(\theta_i) - U(\theta_\star)) + C^2G^2\sum_{i=1}^{n}\frac{1}{i} \\ &\leq B(\theta_\star\|\theta_1) + C^2G^2(1+\log n)\end{aligned}$$

We take "almost telescoping" series in a similar way as before to get

$$\begin{aligned}\sum_{i=1}^{n}(\mathbb{E}U(\theta_i) - U(\theta_\star)) &\leq \frac{1}{C}\sum_{i=1}^{n}(\sqrt{i} - \sqrt{i-1})B(\theta_\star\|\theta_i) + \frac{CG^2}{4}\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \\ &\leq \frac{1}{C}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}B(\theta_\star\|\theta_i) + \frac{CG^2}{4}\sqrt{n} \\ &\leq \sum_{i=1}^{n}\frac{1}{\sqrt{i}}(B(\theta_\star\|\theta_1)/C + CG^2 + CG^2\log i) + \frac{CG^2}{4}\sqrt{n} \\ &\leq \frac{2B(\theta_\star\|\theta_1)}{C}\sqrt{n} + 2CG^2\sqrt{n}\log n + \frac{CG^2}{4}\sqrt{n}.\end{aligned}$$

The second inequality follows from the

$$(\sqrt{x} - \sqrt{x-1})(\sqrt{x} + \sqrt{x-1}) = 1 \leq 1 + \sqrt{1 - 1/x} = \frac{1}{\sqrt{x}}(\sqrt{x} + \sqrt{x-1})$$
$$\Rightarrow \sqrt{x} - \sqrt{x-1} \leq \frac{1}{\sqrt{x}},$$

for $x \geq 1$. The third is line follows from the inequality we just established. The fourth line follows from the inequalities

$$\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n}$$

and

$$\sum_{i=1}^{n} \frac{\log i}{\sqrt{i}} \leq \int_0^n \frac{\log x}{\sqrt{x}}\, dx = 2\sqrt{n}(\log n - 2).$$

Finally, we conclude

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}U(\theta_i) - U(\theta_\star)) \leq \left(\frac{2B(\theta_\star \| \theta_1)}{C} + \frac{CG^2}{4}\right)\frac{1}{\sqrt{n}} + 2CG^2\frac{\log n}{\sqrt{n}}.$$

□

2.2.3 Sample average approximation

The stochastic optimization method *sample average approximation* solves problem (2.1) with the algorithm

$$\theta_{n+1} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n} u_i(\theta),$$

where $u_1, u_2, \ldots$ are *random functions* that satisfy

$$\mathbb{E}u_i(\theta) = U(\theta)$$

for all $\theta \in \Theta$ and $i = 1, 2, \ldots$. In other words, u are unbiased estimates of U. If the minimizer is not unique, then θ_{n+1} is chosen to be any minimizer. Convergence

and its rate of convergence of sample average approximation can be established with techniques similar to those used to analyze the behavior of the maximum likelihood estimator.

The computational cost of the minimization is a significant disadvantage of sample average approximation. Sometimes, however, the minimization has a closed form solution and the computational cost per iteration of sample average approximation becomes comparable to that of stochastic subgradient descent or stochastic mirror descent.

The theoretical and empirical performance of sample average approximation is often much better than that of stochastic subgradient descent and stochastic mirror descent. In fact, under mild assumptions one can show

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U(\theta_i) = U(\theta_\star) + \mathcal{O}(1/n)$$

when using sample average approximation [67, 31]. (This is related to the fact that the maximum likelihood estimator attains the Cramér-Rao lower bound.)

So when the specific problem structure at hand allows the minimization step to be performed efficiently, sample average approximation should be chosen over stochastic subgradient descent or stochastic mirror descent.

2.3 Families with log-concave parameterization

Let $\mathcal{F}$ be a collection of probability measures on $\mathcal{X}$. We say $\mathcal{F}$ has a *log-concave parameterization* if $\mathcal{F}$ can be written as

$$\mathcal{F} = \{f_\theta d\mu \mid \theta \in \Theta\},$$

where $d\mu$ is some common base measure, Θ is convex, and $f_\theta(x)$ is log-concave in θ for μ-almost all $x \in \mathcal{X}$. This should not be confused with the convexity properties of f_θ as a function x. We write F_θ for the distribution that satisfies $dF_\theta = f_\theta d\mu$.

Mixtures. Let $p_1, p_2, \ldots, p_p$ be probability densities with respect to $d\mu$. The family

of their mixtures can be written as $\mathcal{F} = \{f_\theta d\mu \,|\, \theta \in \Theta\}$, where

$$f_\theta = \theta_1 p_1 + \theta_2 p_2 u + \cdots + \theta_p p_p, \quad \Theta = \{\theta \in \mathbf{R}^p \,|\, \theta_1, \ldots, \theta_p \geq 0,\, \theta_1 + \cdots + \theta_p = 1\}.$$

The mixture f_θ is linear and therefore log-concave in θ on Θ.

Exponential families. Define the function $T : \mathcal{X} \to \mathbf{R}^p$ and a nonempty convex set $\Theta \subseteq \mathbf{R}^p$. Then we have the p-dimensional exponential family $\mathcal{F} = \{f_\theta d\mu \,|\, \theta \in \Theta\}$, where

$$f_\theta(x) = \exp(\theta^T T(x) - A(\theta))$$

and

$$A(\theta) = \log \int \exp(\theta^T T(x))\, d\mu(x)$$

normalizes f_θ into a probability distribution.

Exponential families are extensively studied throughout the statistics literature, and it is well known that A is a convex function of θ on Θ [17, 46, 43]. Since

$$\log f_\theta(x) = \theta^T T(x) - A(\theta)$$

the parameterization is is log-concave.

Affine transformations of log-concave continuous distributions. Let $p(x)$ be a log-concave density of a continuous random variable in $\mathbf{R}^k$. (So $\log p(x)$ is a concave function of x.) If $X \sim p(x)dx$ then

$$A^{-1}(X - b) \sim p(Ax + b) \det(A) dx.$$

Since concavity is preserved under affine transformations, $\log p(Ax+b)$ is concave in $(A, b) \in \mathbf{S}^k \times \mathbf{R}^k$, where $\mathbf{S}^k$ denotes the set of $k \times k$ symmetric matrices. Restricted to $\mathbf{S}^k_{++}$, the set of $k \times k$ positive definite matrices, $\log \det A$ is concave.

We write $\theta = (A, b) \in \mathbf{S}^k \times \mathbf{R}^k$. The family of scalings with respect to a positive definite matrix and translations of $X \sim p(x)dx$ is given by $\mathcal{F} = \{f_\theta \,|\, \theta \in \Theta\}$, where

$$f_\theta = p(Ax + b) \det A$$

and $\Theta = \mathbf{S}^k_{++} \times \mathbf{R}^k$. As discussed, this parameterization is log-concave.

2.4 Rényi generalized divergence

Let P and Q be probability measures and α a parameter in $[1, \infty)$. When $P \ll Q$, i.e., P is absolutely continuous with respect to Q, we define the Rényi divergence of order α of Q from P as

$$D_\alpha(P\|Q) = \begin{cases} \displaystyle\int \log\left(\frac{dP}{dQ}\right)\, dP & \alpha = 1 \\ \displaystyle\frac{1}{\alpha - 1}\log \int \left(\frac{dP}{dQ}\right)^{\alpha-1} dP & 1 < \alpha < \infty. \end{cases}$$

If $P \not\ll Q$, then $D_\alpha(P\|Q) = \infty$ for all $\alpha \in [1, \infty)$. For our purposes, we shall view $D_\alpha(P\|Q)$ as a measure of distance between the probability distributions P and Q, just like the KL-divergence. In fact, D_α is the KL-divergence when $\alpha = 1$. We list a few properties of D_α:

- $0 \leq D_\alpha(P\|Q) \leq \infty$.
- $D_\alpha(P\|Q) = 0$ if and only if $P = Q$.
- $D_\alpha(P\|Q)$ is nondecreasing in α for fixed P and Q.
- $D_\alpha(P\|Q)$ is continuous on $\{\alpha \in [1, \infty) \mid D_\alpha(P\|Q) < \infty\}$ for fixed P and Q, i.e., it is continuous where finite.
- $D_\alpha(P\|Q)$ is not symmetric in P and Q.

2.4.1 Convexity of Rényi divergence

In later sections, we encounter the optimization problem

$$\begin{array}{ll} \text{minimize} & D_\alpha(P\|F_\theta) \\ \text{subject to} & \theta \in \Theta \end{array}$$

for some P and α. This can be interpreted as finding the distribution F_θ closest (measured by D_α) to a target distribution P. Here we look at a sufficient condition that allows us to solve this problem via convex optimization.

Lemma 3. *Assume $\mathcal{F}$ has a log-concave parameterization. Then $D_1(P\|F_\theta)$ and $\exp((\alpha-1)D_\alpha(P\|F_\theta))$ for $\alpha \in (1,\infty)$ are convex in θ on Θ.*

We note that $D_\alpha(P\|Q)$ is convex in Q, but this is not why this lemma is true. Rather, we establish convexity by building up the function of interest from known convex functions with operations that perserve convexity [15, §3.2].

Proof. With the parameterization, we can write

$$D_1(P\|F_\theta) = \int \left(\log \frac{dP}{d\mu} - \log f_\theta\right)\, dP.$$

Since $-\log f_\theta$ is convex in θ and since a nonnegative weighted integral of convex functions is convex, $D_1(P\|F_\theta)$ is convex in θ.

Now let's look at the $\alpha > 1$ case. Since composition of a convex increasing function with a convex function preserves convexity, $1/f_\theta^{\alpha-1} = \exp(-(\alpha-1)\log f_\theta)$ is convex. Therefore

$$\exp((\alpha-1)D_\alpha(p\|f_\theta)) = \int \frac{1}{f_\theta^{\alpha-1}} \left(\frac{dP}{d\mu}\right)^{\alpha-1}\, dP$$

is convex. □

Since $\exp((\alpha-1)D_\alpha(p\|f_\theta))$ is a monotone transformation of $D_\alpha(p\|f_\theta)$ when $\alpha > 1$, minimizing one is equivalent to minimizing the other. Therefore we work with the convex one.

Stochastic subgradients. The convex functions $D_1(P\|F_\theta)$ and $\exp((\alpha-1)D_\alpha(P\|F_\theta))$ are useful not only becuase they are convex, but also because we have access to their stochastic subgradients. Let us see how to get a stochastic subgradient of $\exp D_2(P\|F_\theta)$. Stochastic subgradients of $D_1(P\|F_\theta)$ and $\exp((\alpha-1)D_\alpha(P\|F_\theta))$ for $\alpha \neq 2$ can be obtained in a similar fashion.

For simplicity, assume differentiability throughout. Let $X \sim F_{\theta_0}$ and

$$g = -\frac{1}{f_{\theta_0}(X)} \left(\frac{dP}{dF_{\theta_0}}(X) \right)^2 \nabla_\theta f_{\theta_0}(X).$$

Then g is a subgradient of $\exp D_2(P\|F_\theta)$, i.e.,

$$\mathbb{E}g \in \nabla_\theta \exp D_2(P\|F_\theta).$$

To see why, assume we can evaluate the gradient under the integral. Then we have

$$\begin{aligned}
\nabla_\theta \exp D_2(P\|F_\theta) &= \int \nabla_\theta \frac{dP}{dF_\theta}(x)\ dP(x) \\
&= -\int \frac{1}{f_\theta^2(x)} \frac{dP}{d\mu}(x) \nabla_\theta f_\theta(x)\ dP(x) \\
&= -\int \frac{1}{f_\theta(x)} \left(\frac{dP}{dF_\theta}(x) \right)^2 \nabla_\theta f_\theta(x)\ dF_\theta(x) \\
&= \mathbb{E}_{F_\theta} \left[-\frac{1}{f_\theta(X)} \left(\frac{dP}{dF_\theta}(X) \right)^2 \nabla_\theta f_\theta(X) \right].
\end{aligned}$$

While this is not a rigorous argument, it does illustrate the main idea. We show a more general and rigorous version of this as Lemma 11 of the appendix.

2.4.2 Rényi divergence and moments

We note that $D_\alpha(P\|Q) < \infty$ if and only if the α-th moment of $(dP/dQ)(X)$, where $X \sim Q$, is finite. In fact, one could say $D_\alpha(P\|Q)$ is a measure of how large the α-th moment of $(dP/dQ)(X)$ is. So we will write $D_\alpha(P\|Q) < \infty$ as a shorthand for saying that the α-th moment of $(dP/dQ)(X)$ is finite.

Chapter 3

Adaptive importance sampling

Consider the problem of approximating the expected value (or integral)

$$I = \mathbf{E}\phi(X) = \int \phi(x)\, dF(x),$$

where $X \sim F$ is a random variable on $\mathcal{X}$ and $\phi : \mathcal{X} \to \mathbf{R}$. To avoid pathologies, we assume $0 < \mathbb{E}|\phi(X)| < \infty$; if $\mathbb{E}|\phi(X)| = 0$ the problem is trivial and if $\mathbb{E}|\phi(X)| = \infty$, the problem is not well defined. We call F the *nominal distribution*.

The *plain Monte Carlo* algorithm computes I with the algorithm

$$\begin{aligned} X_n &\sim F \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i). \end{aligned}$$

The analysis of this method is straightforward:

$$\mathbb{E}\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\phi(X_i) = I$$

and

$$\begin{aligned}\mathbf{Var}\hat{I}_n &= \mathbb{E}(\hat{I}_n - I)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(\phi(X_i) - I)^2 + \frac{2}{n}\sum_{1\le i<j\le n}\mathbb{E}(\phi(X_i) - I)(\phi(X_j) - I) \\ &= \frac{1}{n}\mathbf{Var}_F\phi(X).\end{aligned}$$

So the estimator is *unbiased*, i.e., $\mathbb{E}\hat{I}_n = I$, and has variance

$$\mathbf{Var}\hat{I}_n = \frac{1}{n}\mathbf{Var}_F\phi(X) = \frac{1}{n}\left(\int \phi^2(x)\, dF(x) - I^2\right).$$

$\hat{I}_n$ is guaranteed to converge to I as $n \to \infty$ (in L^2 as long as $\mathbf{Var}_F\phi(X) < \infty$). Sometimes this approach is good enough to compute I to necessary precision for the problem at hand. If so there is no need to seek improved algorithms.

Often, however, this method is not fast enough, i.e., $\mathbf{Var}\hat{I}_n$ is not small enough, and one must employ a *variance reduction* technique to compute I to necessary precision in a practical amount of computation time. The variance $\mathbf{Var}\hat{I}_n$ contains the factors $1/n$ and $\mathbf{Var}_F\phi(X)$. Loosely speaking, the factor $1/n$ is considered hard to improve upon when using a Monte Carlo method. So we explore approaches to reduce $\mathbf{Var}_F\phi(X)$.

3.1 Importance sampling

One variance reduction technique is called *importance sampling*, which is based on the following key insight:

$$I = \int \phi(x)\, dF(x) = \int \phi(x)\frac{dF}{d\tilde{F}}(x)\, d\tilde{F}(x) = \mathbb{E}_{\tilde{F}}\left[\phi(X)\frac{dF}{d\tilde{F}}(X)\right],$$

where $\phi dF \ll d\tilde{F}$ and $dF/d\tilde{F}$ denotes the Radon-Nokodym derivative. This recasting of I as an expectation under $\tilde{F}$ motivates importance sampling:

$$\begin{aligned} X_n &\sim \tilde{F} \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{dF}{d\tilde{F}}(X_i). \end{aligned}$$

The distribution $\tilde{F}$ is called a *sampling* or *importance distribution* and has to satisfy $\phi dF \ll d\tilde{F}$.

The analysis of this algorithm is similar to what we had before:

$$\mathbb{E}\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{F}}\phi(X_i)\frac{dF}{d\tilde{F}}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_F\phi(X_i) = I$$

and

$$\begin{aligned} \mathbf{Var}\hat{I}_n &= \mathbb{E}(\hat{I}_n - I)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{\tilde{F}}\left(\phi(X_i)\frac{dF}{d\tilde{F}}(X_i) - I\right)^2 \\ &\quad + \frac{2}{n}\sum_{1\le i<j\le n} \mathbb{E}\left(\phi(X_i)\frac{dF}{d\tilde{F}}(X_i) - I\right)\left(\phi(X_j)\frac{dF}{d\tilde{F}}(X_j) - I\right) \\ &= \frac{1}{n}\mathbf{Var}_{\tilde{F}}\left(\phi(X_i)\frac{dF}{d\tilde{F}}(X)\right). \end{aligned}$$

So again the estimator is unbiased, i.e., $\mathbb{E}\hat{I}_n = I$, and has variance

$$\mathbf{Var}\hat{I}_n = \frac{1}{n}\mathbf{Var}_{\tilde{F}}\left(\frac{\phi(X)f(X)}{\tilde{f}(X)}\right) = \frac{1}{n}\left(\int \phi^2(x)\left(\frac{dF}{d\tilde{F}}(x)\right)^2 d\tilde{F}(x) - I^2\right).$$

Note that importance sampling reduces to plain Monte Carlo when $\tilde{F} = F$.

Per-sample variance. We call

$$\mathbf{Var}_{\tilde{F}}\left(\phi(X)\frac{dF}{d\tilde{F}}(X)\right)$$

the *per-sample variance* of the sampling distribution $\tilde{F}$, as it is the variance contributed by each sample from $\tilde{F}$. When the per-sample variance of $\tilde{F}$ is smaller than the per-sample variance of F, importance sampling indeed provides variance reduction. To extract the most variance reduction, we should choose a sampling distribution $\tilde{F}$ with the smallest per-sample variance.

Define the distribution $F_\star$ such that

$$dF_\star = \frac{1}{J}|\phi|dF,$$

where

$$J = \int |\phi(x)|dF(x)$$

is a normalizing factor. Simple algebra gives us

$$\mathbf{Var}_{\tilde{F}}\left(\phi(X)\frac{dF}{d\tilde{F}}(X)\right) = J^2 \exp D_2(F_\star\|\tilde{F}) - I^2.$$

Since $J^2 > 0$ and exp is a monotone transformation, reducing the per-sample variance is equivalent to reducing $D_2(F_\star\|\tilde{F})$.

So given a family of probability distributions $\mathcal{F}$, a solution to the optimization problem

$$\begin{array}{ll}\text{minimize} & D_2(F_\star\|\tilde{F}) \\ \text{subject to} & \tilde{F} \in \mathcal{F}\end{array}$$

has the smallest per-sample variance among distributions in $\mathcal{F}$. (The constraint $\phi dF \ll d\tilde{F}$ is implicitly enforced since $D_2(F_\star\|\tilde{F}) = \infty$ if $F_\star \not\ll \tilde{F}$.) This also tells us that $F_\star$ has the smallest per-sample variance among all distributions, although there is no good way to generate samples from $F_\star$ in practice. This optimization problem formalizes the well-known notion that sampling distributions "close to $F_\star$" have small variance.

If $\mathcal{F}$ has a log-concave parameterization, then the distribution with smallest per-sample variance among $\mathcal{F}$ can be found by the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \exp D_2(F_\star \| F_\theta) \\ \text{subject to} & \theta \in \Theta. \end{array} \tag{3.1}$$

A suboptimal approach. At this point, one might consider the following approach.

1. Choose $\mathcal{F}$ with a log-concave parameterization.
2. Find a solution $\theta_\star$ of the convex optimization problem (3.1).
3. Perform importance sampling with sampling distribution $F_{\theta_\star}$.

Indeed, convexity makes solving optimization problem (3.1) feasible. Furthermore, we have an optimality statement: the importance sampling with $F_{\theta_\star}$ achieves the smallest variance among all choices within $\mathcal{F}$.

Unfortunately, performing step 2, solving problem (3.1), is usually no more easier than computing I, the unknown quantity of iterest. In fact, evaluating the objective $\exp D_2(F_\star \| F_\theta)$ at a given point θ requires a Monte Carlo simulation of its own. One could could try solving problem (3.1) approximately before moving on to the importance sampling. However, this creates a trade-off of how much time one should spend on step 2 before moving on to step 3. Furthermore, the optimality statement is lost, although one could argue that this is only of theoretical interest.

In the next section, we will present a method that performs step 2 and step 3 simultaneously. By performing optimization and importance sampling simultaneously, we eliminate the trade-off and retain the optimality statement.

3.2 Adaptive importance sampling

Adaptive importance sampling is based on the key insight that the sampling distribution need not be the same for all iterations. The adaptive importance sampling

algorithm computes I with

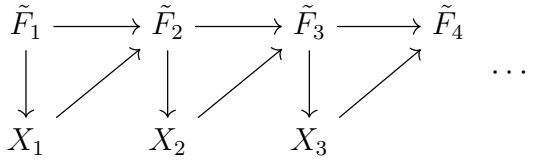
$$\begin{aligned} X_n &\sim \tilde{F}_n \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i) \\ \tilde{F}_{n+1} &= \text{update with } X_1, \ldots, X_n, \tilde{F}_1, \ldots, \tilde{F}_n. \end{aligned}$$

As we show soon, the estimator is *unbiased*, i.e., $\mathbb{E}\hat{I}_n = I$, and has variance

$$\mathbf{Var}\hat{I}_n = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbf{Var}_{\tilde{F}_i}\left(\phi(X)\frac{dF}{d\tilde{F}_i}(X)\right)\right).$$

Since $\tilde{F}_n$ is determined based on the past random information, $\tilde{F}_n$, a probability distribution, is itself random. The variance of $\hat{I}_n$ may look different from that of importance sampling, but they are similar in spirit: $\mathbf{Var}\hat{I}_n$ is $1/n$ times the average expected per-sample variance of $\tilde{F}_1, \ldots, \tilde{F}_n$. If $\tilde{F}_1 = \cdots = \tilde{F}_n = \tilde{F}$, this algorithm reduces to importance sampling.

Analysis of adaptive importance sampling. The analysis is essentially the same as before when we must diligently handle the conditionally dependencies. The conditional dependencies of our sequences $\tilde{F}_1, \tilde{F}_2, \ldots$ and $X_1, X_2, \ldots$ are



So X_n is independent of the entire past conditioned on $\tilde{F}_n$ for all i. With this we have

$$\mathbb{E}\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i)\middle|\tilde{F}_i\right]\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[I\right] = I$$

and

$$\begin{aligned}
\mathbf{Var}\hat{I}_n &= \mathbb{E}(\hat{I}_n - I)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i) - I\right)^2 \\
&\quad + \frac{2}{n^2}\sum_{1\le i<j\le n}\mathbb{E}\left(\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i) - I\right)\left(\phi(X_j)\frac{dF}{d\tilde{F}_j}(X_j) - I\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left(\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i) - I\right)^2\middle|\tilde{F}_i\right]\right] \\
&\quad + \frac{2}{n^2}\sum_{1\le i<j\le n}\mathbb{E}\left[\mathbb{E}\left[\left(\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i) - I\right)\left(\phi(X_j)\frac{dF}{d\tilde{F}_j}(X_j) - I\right)\middle|X_i,\tilde{F}_i,\tilde{F}_j\right]\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\mathbf{Var}_{\tilde{F}_i}\left(\phi(X)\frac{dF}{d\tilde{F}_i}(X)\right).
\end{aligned}$$

This gives the stated result.

3.2.1 Main framework

In a sense, adaptive importance sampling is a meta algorithm; determining how to update $\tilde{F}_n$ fully specifies the method and its performance. The specific adaptive importance sampling methods we consider in this work are determined by the following 3 choices:

- Choose a family of distributions.
- Choose an objective to minimize.
- Choose a stochastic optimization method to perform the minimization.

So using the parameterization, the algorithm will be of the form:

$$\begin{aligned} X_n &\sim F_{\theta_n} \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{dF}{dF_{\theta_i}}(X_i) \\ \theta_{n+1} &= \text{Stochastic optimization with } X_1, \ldots, X_n, \theta_1, \ldots, \theta_n. \end{aligned} \tag{3.2}$$

Write $V_\star$ for the minimum per-sample variance for the family $\mathcal{F}$, i.e., $V_\star$ is the optimal value for

$$\begin{array}{ll} \text{minimize} & \mathbf{Var}_{F_\theta}\left(\phi(X)\frac{dF}{dF_\theta}(X)\right) \\ \text{subject to} & \theta \in \Theta. \end{array}$$

Then

$$\frac{1}{n}V_\star \leq \mathbf{Var}\hat{I}_n.$$

If algorithm (3.2) is run with a stochastic optimization algorithm such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbf{Var}_{F_{\theta_i}}\left(\phi(X)\frac{dF}{dF_{\theta_i}}(X)\right) \leq V_\star + o(1),$$

then

$$\mathbf{Var}\hat{I}_n \leq \frac{1}{n}V_\star + o\left(\frac{1}{n}\right).$$

Putting the two together, we can say

$$\mathbf{Var}\hat{I}_n \approx \frac{1}{n}V_\star,$$

i.e., the variance of $\hat{I}_n$ is asymptotically optimal with respect to the family $\mathcal{F}$.

3.3 Central limit theorem

In the previous section, we established a condition under which the asymptotic variance of $\hat{I}_n$ becomes optimal. In this section, we establish a sufficient condition under which $\hat{I}_n$ is asymptotically normally distributed.

Lemma 4. *Assume algorithm* (3.2) *uses a stochastic optimization method that yields performance*

$$\mathbf{Var}\hat{I}_n = \frac{1}{n}V_\star + o\left(\frac{1}{n}\right).$$

Furthermore, assume there is a $\varepsilon > 0$ *such that* $D_{2+\varepsilon}(F_\star \| F_\theta)$ *is bounded for all* $\theta \in \Theta$. *(So the* $(2+\varepsilon)$*-th moments of the estimators are bounded.) Then we have*

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

Proof. First define

$$Y_{ni} = \begin{cases} \frac{1}{\sqrt{n}}\left(\frac{\phi(X_i)f(X_i)}{f_{\theta_i}(X_i)} - I\right) & \text{for } i \le n \\ 0 & \text{otherwise} \end{cases}$$

and

$$J_{nm} = \sum_{i=1}^{m} Y_{ni}.$$

Also define the σ-algebras

$$\mathcal{G}_m = \sigma(\theta_1, \theta_2, \ldots, \theta_{m+1}, X_1, X_2, \ldots, X_m)$$

for all m. Then for any given n, the process $J_{n1}, J_{n2}, \ldots$ is a martingale with respect to $\mathcal{G}_1, \mathcal{G}_2, \ldots$ and to we have to prove

$$J_{nn} = \sum_{i=1}^{n} Y_{ni} = \sum_{i=1}^{\infty} Y_{ni} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

Define

$$\sigma_{ni}^2 = \mathbb{E}\left[Y_{ni}^2 \mid \mathcal{G}_{i-1}\right] = \begin{cases} \frac{1}{n}\mathbf{Var}_{F_{\theta_i}}\left(\phi(X)\frac{dF}{dF_{\theta_i}}(X)\right) & \text{for } i \le n \\ 0 & \text{otherwise.} \end{cases}$$

Then a form of the Martingale CLT, *c.f.*, Theorem 35.12 of [12], states that if

$$\sum_{i=1}^{n} \sigma_{ni}^2 \xrightarrow{\mathcal{P}} V_\star$$

and

$$\sum_{i=1}^{n} \mathbb{E} Y_{ni}^2 I_{\{|Y_{ni}| \geq \delta\}} \to 0$$

for each $\delta > 0$, then $J_{nn} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star)$.

Since

$$\sum_{i=1}^{n} \sigma_{ni}^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Var}_{F_{\theta_i}} \left(\phi(X) \frac{dF}{dF_{\theta_i}}(X) \right),$$

and since, by assumption, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E} \mathbf{Var}_{F_{\theta_i}} \left(\phi(X) \frac{dF}{dF_{\theta_i}}(X) \right) - V_\star \right) &= \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{Var}_{F_{\theta_i}} \left(\phi(X) \frac{dF}{dF_{\theta_i}}(X) \right) - V_\star \right| \\ &= o(1) \to 0, \end{aligned}$$

i.e., $\sum_{i=1}^{n} \sigma_{ni}^2$ converges to $V_\star$ in L^1, we have

$$\sum_{i=1}^{n} \sigma_{ni}^2 \xrightarrow{\mathcal{P}} V^*.$$

The second condition follows from the fact that the $(2+\varepsilon)$-th moent is bounded:

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{E} Y_{ni}^2 I_{\{|Y_{ni}| \geq \delta\}} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left(\phi(X_i) \frac{dF}{dF_{\theta_i}}(X_i) - I \right)^2 I_{\left\{ |\phi(X_i) dF / dF_{\theta_i}(X_i)|^\varepsilon \geq \delta^\varepsilon n^{\varepsilon/2} \right\}} \\ &\leq \frac{1}{n^{1+\varepsilon/2} \delta^\varepsilon} \sum_{i=1}^{n} \mathbb{E} \left(\phi(X_i) \frac{dF}{dF_{\theta_i}}(X_i) - I \right)^{2+\varepsilon} \\ &\leq \frac{B}{n^{\varepsilon/2} \delta^\varepsilon} \to 0. \end{aligned}$$

So the 2 conditions are met and we have the desired CLT. □

3.4 Examples

3.4.1 Stochastic subgradient descent with exponential family

In this section, we consider the adaptive importance sampling algorithm we get when we choose an exponential family for the family $\mathcal{F}$ (as defined in Section 2.3), the per-sample variance, $D_2(F_\star \| F_\theta)$, for the objective to minimize, and stochastic subgradient descent for the stochastic optimization algorithm.

Assume Θ is compact and $\Theta \subset \mathbf{int}\{\theta \mid D_4(F_\star \| F_\theta) < \infty\}$. (So Θ is in the interior of the set for which the estimators have finite 4th moments.)

The adaptive importance sampling method with these choices is

$$\begin{aligned}
&X_n \sim F_{\theta_n} \\
&\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i) \frac{dF}{dF_{\theta_i}}(X_i) \\
&g_n = \phi^2(X_n) \left(\frac{dF}{dF_{\theta_0}} \right)^2 (X_n)(\nabla A(\theta) - T(X_n)) \\
&\theta_{n+1} = \Pi_\Theta(\theta_n - (C/\sqrt{n}) g_n).
\end{aligned}$$

The estimator $\hat{I}_n$ is unbiased, and has performance

$$\frac{1}{n} V_\star \leq \mathbf{Var} \hat{I}_n \leq \frac{1}{n} V_\star + \mathcal{O}\left(\frac{1}{n^{3/2}} \right)$$

and

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

The adaptive importance sampling method with batch size m is

$$\begin{aligned}
&X_{n1}, X_{n2}, \ldots, X_{nm} \sim F_{\theta_n} \\
&\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\phi(X_{ij})\frac{dF}{dF_{\theta_i}}(X_{ij}) \\
&g_n = \frac{1}{m}\sum_{j=1}^{m}\phi^2(X_{nj})\left(\frac{dF}{dF_\theta}\right)^2(X_{nj})(\nabla A(\theta) - T(X_{nj})) \\
&\theta_{n+1} = \Pi_\Theta(\theta_n - (C/\sqrt{n})g_n).
\end{aligned}$$

The estimator $\hat{I}_n$ is unbiased, and has performance

$$\frac{1}{nm}V_\star \le \mathbf{Var}\hat{I}_n \le \frac{1}{nm}V_\star + \mathcal{O}\left(\frac{1}{n^{3/2}m}\right)$$

and

$$\sqrt{nm}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

Discussion of assumptions. While one can find realistic setups for $D_4(F_\star\|F_\theta)$ is not bounded, this assumption is not unreasonable; any adaptive importance sampling method would require that the second moments their estimators are finite, i.e., $D_2(F_\star\|F_\theta) < \infty$, and that their fourth moments are finite, i.e., $D_4(F_\star\|F_\theta) < \infty$, is not asking for too much more.

The assumption implies $D_{2+\varepsilon}(F_\star\|F_\theta) \le B_1 < \infty$, which is necessary for the central limit theorem. It also implies $\exp D_2(F_\star\|F_\theta)$, the function we wish to minimize, is finite on all of Θ. Finally, it implies that $\mathbb{E}\|g_i\|_2^2 \le G^2 < \infty$ by Lemma 6, which we show below.

Lemma 5 (Theorem 2.7.1 of [47])**.** *Let $\psi(x)$ be any function on $\mathcal{X}$. Then the function*

$$f(\theta) = \int \psi(x)\exp(\theta^T T(x))\, d\mu(x)$$

is smooth (infinitely differentiable) on $\mathbf{int}\{\theta \mid f(\theta) < \infty\}$.

A direct consequence of this lemma is that $A(\theta)$, $\exp(D_2(F_\star\|F_\theta))$, and $\exp(3D_4(F_\star\|F_\theta))$

are smooth and all derivatives can be evaluated under their integrals on the interiors of the sets on which they are finite.

Lemma 6. *That* $\Theta \subseteq \mathbf{int}\{\theta \mid D_4(F_\star \| F_\theta) < \infty\}$ *implies* $\mathbb{E}\|g_i\|_2^2$ *is bounded for* $i = 1, 2, \ldots$.

Proof. For all $i \in \{1, 2, \ldots, p\}$,

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \exp(3D_4(F_\star \| F_\theta))/J^4 &= 3 \int \left(\frac{\partial}{\partial \theta_i} A(\theta) - T_i(x) \right) \phi^4(x) \left(\frac{dF}{dF_\theta} \right)^3 \, dF(x) \\ &= 3 \exp(3D_4(F_\star \| F_\theta))/J^4 \frac{\partial}{\partial \theta_i} A(\theta) - 3 \int T_i(x) \left(\frac{dF}{dF_\theta} \right)^3 \, dF(x)\end{aligned}$$

exists and is and smooth on $\mathbf{int}\{\theta \mid D_4(F_\star \| F_\theta) < \infty\}$. by Lemma 5. As we already know the first term is smooth by Lemma 5, this tells us the second term is smooth.

Repeating this, we have

$$\begin{aligned}\frac{\partial^2}{\partial \theta_i^2} \frac{\exp(3D_4(F_\star \| F_\theta))}{J^4} = \int \left(3 \frac{\partial^2}{\partial \theta_i^2} A(\theta) + 9 \left(\frac{\partial}{\partial \theta_i} A(\theta) \right)^2 - 18 T_i(x) \frac{\partial}{\partial \theta_i} A(\theta) + 9 T_i^2(x) \right) \\ \phi^4(x) \left(\frac{dF}{dF_\theta} \right)^3 \, dF(x)\end{aligned}$$

We know the first 3 terms are smooth from what we just proved and Lemma 5. So we conclude that

$$\int T_i^2(x) \phi^4(x) \left(\frac{dF}{dF_\theta} \right)^3 \, dF(x)$$

is a smooth function of θ on $\mathbf{int}\{\theta \mid D_4(F_\star \| F_\theta) < \infty\}$.

Finally, we conclude that

$$\begin{aligned}&\mathbb{E}_{F_\theta} \left\| (\nabla A(\theta) - T(X)) \phi^2(X) \left(\frac{dF(X)}{dF_\theta(X)} \right)^2 \right\|_2^2 \\ &\quad = \|\nabla A(\theta)\|_2^2 \exp(3D_4(F_\star \| F_\theta))/J^4 - 2 \nabla A(\theta)^T \int T(X) \phi^4(x) \left(\frac{dF}{dF_\theta} \right)^3 \, dF(x) \\ &\qquad + \int \|T(X)\|_2^2 \phi^4(x) \left(\frac{dF}{dF_\theta} \right)^3 \, dF(x)\end{aligned}$$

is a continuous function on the compact set Θ. So the supremum over the compact set Θ is finite. □

3.4.2 Cross-entropy method

In this section, we show that the *cross-entropy method*, a widely used adaptive importance sampling method, is a special case of our adaptive importance sampling framework. Cross-entropy method is the adaptive importance sampling algorithm we get when we choose exponential family for the family $\mathcal{F}$ (as defined in Section 2.3), $D_1(F_\star \| F_\theta)$, not $D_2(F_\star \| F_\theta)$, for the objective to minimize, and sample average approximation for the stochastic optimization algorithm.

We first note that

$$\begin{aligned} D_1(F_\star \| F_\theta) &= \int \log\left(\frac{dF_\star}{d\mu}\right)\, dF_\star + \frac{1}{J}\int |\phi| \frac{dF}{dF_{\theta_0}}(-\log f_\theta)\, dF_{\theta_0} \\ &= \int \log\left(\frac{dF_\star}{d\mu}\right)\, dF_\star + \frac{1}{J}\mathbb{E}_{F_{\theta_0}}\left[-|\phi(X)| \frac{dF}{dF_{\theta_0}}(X) \log f_\theta(X)\right]. \end{aligned}$$

Since constants can be ignored, we can view

$$-|\phi(X)| \frac{dF}{dF_{\theta_0}}(X) \log f_\theta(X)$$

with $X \sim F_{\theta_0}$ as an unbiased estimate of the function $D_1(F_\star \| F_\theta)$.

The adaptive importance sampling method with these choices is

$$\begin{aligned} X_n &\sim F_{\theta_n} \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) \frac{dF}{dF_{\theta_i}}(X_i) \\ \theta_{n+1} &= \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} |\phi(X_i)| \frac{dF}{dF_{\theta_i}}(X_i) \log f_\theta(X_i). \end{aligned}$$

Since $\mathcal{F}$ is an exponential family, the maximization is

$$\operatorname*{argmax}_{\theta\in\Theta} \sum_{i=1}^{n} |\phi(X_i)| \frac{dF}{dF_{\theta_i}}(X_i)(\theta^T T(X_i) - A(\theta)).$$

Since A is differentiable (Lemma 5), the maximizer satisfies

$$\nabla A(\theta_{n+1}) = \left(\sum_{i=1}^{n} |\phi(X_i)| \frac{dF}{dF_{\theta_i}}(X_i) T(X_i)\right) / \left(\sum_{i=1}^{n} |\phi(X_i)| \frac{dF}{dF_{\theta_i}}(X_i)\right).$$

Quite often, the function ∇A is easily invertible and θ_{n+1} can computed via an analytical solution [43].

Putting these together the method simplifies to the cross-entropy method

$$\begin{aligned} X_n &\sim F_{\theta_n} \\ w_n &= |\phi(X_n)| \frac{dF}{dF_{\theta_n}}(X_n) \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} w_i \\ \theta_{n+1} &= (\nabla A)^{-1}\left(\left(\sum_{i=1}^{n} w_i T(X_i)\right) / \left(\sum_{i=1}^{n} w_i\right)\right). \end{aligned}$$

That the cross-entropy method uses sample average approximation is both an advantage and disadvantage. When the minimization has a closed-form solution (i.e., when ∇A is invertible) sample average approximation is a superior method compared to stochastic subgradient descent or stochastic mirror descent as discussed in Section 2.2.3. However, the minimization often does not have a closed-form solution, and sample average approximation becomes inefficient.

The cross-entropy method inherits these strength and limitation. When using certain exponential families for the sampling distributions, the cross-entropy method will work well. In the sense of having minimum asymptotic variance, the cross-entropy method is suboptimal. The $\theta_\star$ that minimizes $D_1(F_\star \| F_\theta)$ does not necessarily minimize $D_2(F_\star \| F_\theta)$, the per-sample variance. In practice, however, the difference

between $D_1(F_\star \| F_\theta)$ and $D_2(F_\star \| F_\theta)$ should be small and the faster convergence rate of sample average approximation should matter more when the iteration count n is not too large.

Finally, we point out that batching can also be used with the cross-entropy method. In practice, it is necessary to withold updating θ_n for the first few iterations. Batching is a way to accomplish this. Also, if one is using a numerical optimization solver to compute θ_n in the absense of an analytical solution, batching is a good way to reduce the computational cost of the optimization.

3.4.3 Option pricing

Consider the pricing of an arithmetic Asian call option on an underlying asset under standard Black-Scholes assumptions [33]. We write S^0 for the initial price of the underlying asset, r and σ for the interest rate and volatility of the Black-Scholes model, and T for the maturity time. Under the Black-Scholes model, the price of the asset at time jT/k is

$$S^j(X) = S^0 \exp\left[(r - \frac{1}{2}\sigma^2)j\frac{T}{n} + \sigma\sqrt{\frac{T}{n}}\sum_{i=1}^{j} X(i)\right]$$

for $j = 1, \ldots, k$, where $X \in \mathbf{R}^k$ is random with independent standard normal entries $X(1), \ldots, X(k)$. The discounted payoff of the option with strike K is given by

$$\phi(X) = \exp^{-rT} \max\left\{\frac{1}{k}\sum_{j=1}^{k} S^j(X) - K, 0\right\},$$

and we wish to compute $I = \mathbb{E}\phi(X)$.

For this problem, we compare the performance of four different methods. The first method is plain Monte Carlo where the samples are taken from the nominal distribution. The second method performs adaptive importance sampling with multivariate Gaussians with their means as the parameters for the family $\mathcal{F}$, the per-sample variance, $D_2(F_\star \| F_\theta)$, for the objective to minimize, and stochastic mirror descent for the stochastic optimization algorithm. The third method performs adaptive importance

sampling with multivariate Gaussians with their means and covariance for the parameters for the family $\mathcal{F}$, the per-sample variance, $D_2(F_\star \| F_\theta)$, for the objective to minimize, and stochastic mirror descent for the stochastic optimization algorithm. The fourth method is the Cross-entropy method with multivariate Gaussians with their means for the parameters for the family $\mathcal{F}$.

The first method, plain Monte Carlo, has the form

$$\begin{aligned} X_n &\sim \mathcal{N}(0, I) \\ \hat{I}_n &= \frac{1}{n} \sum_{i=1}^{n} \phi(X_n). \end{aligned}$$

The third method, adaptive importance sampling with the mean and covariance as the parameters, has the form

$$\begin{aligned} &L_n L_n^T = \Sigma_n \quad \text{(Cholesky factorization)} \\ &Y_{n1}, Y_{n2}, \ldots, Y_{nm} \sim \mathcal{N}(0, I) \\ &X_{nj} = L_n^T Y_{nj} + \mu_n, \quad j = 1, \ldots, m \\ &w_{nj} = \frac{\phi(X_{nj})}{\sqrt{\det(S_n)}} \exp(-\|X_{nj}\|_2^2/2 + (X_{nj} - \mu_n)^T S_n (X_{nj} - \mu_n)/2) \quad j = 1, \ldots, m \\ &\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} w_{ij} \\ &g_n^S = \frac{1}{2m} \sum_{j=1}^{m} w_{ij}^2 (X_{nj} X_{nj}^T - \mu_n \mu_n^T - \Sigma_n) \\ &g_n^b = \frac{1}{m} \sum_{j=1}^{m} w_{ij}^2 (\mu_n - X_{nj}) \\ &S_{n+1}^* = S_n^* - (C_1/\sqrt{n}) g_n^S \\ &S_{n+1} = \exp S_{n+1}^* \\ &b_{n+1} = \Pi_{C_b}(b_n - (C_2/\sqrt{n}) g_n^b) \\ &\Sigma_{n+1} = S_{n+1}^{-1} \\ &\mu_{n+1} = S_{n+1}^{-1} b_{n+1}. \end{aligned}$$

with $C_b = [-2, 2]^k$. The second method, adaptive importance sampling with the mean but not the covariance as the paramter, is similar to the third method, but has S_n fixed as the identity matrix and does not spend time computing g_n^S and updating S_n. The fourth method, cross entropy with the mean as the paramter, has the form

$$
\begin{aligned}
&Y_{n1}, Y_{n2}, \ldots, Y_{nm} \sim \mathcal{N}(0, I) \\
&X_{nj} = Y_{nj} + \mu_n, \quad j = 1, \ldots, m \\
&w_{nj} = \phi(X_{nj}) \exp(-\|X_{nj}\|_2^2/2 + (X_{nj} - \mu_n)^T (X_{nj} - \mu_n)/2) \quad j = 1, \ldots, m \\
&\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} w_{ij} \\
&\mu_n = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} X_{ij} \right) / \left(\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} \right).
\end{aligned}
$$

We use the parameters $n = 10^5$, $m = 10^3$, $S^0 = 50$, $K = 70$, $r = 0.5$, $\sigma = 0.3$, $k = 64$, $T = 1$, and $C_1 = C_2 = 0.03$. The computations give us $I \approx 0.1807$.

Figure 3.1 shows the average per-sample variance for each method. As we can expect, the first method has a large constant per-sample variance. The third method has the smallest asymptotic per-sample variance, since it has an additional parameter Σ to optimize the per-sample variance over. The fourth method has fast convergence to its optimum, but the third method eventually catches up. Since third methods optimum actually minimizes the per-sample variance, it is better than the fourth method's optimum, but the difference is negligeable, which can be detected only when the plot is inspected carefully.

Figure 3.2 shows $\mathbf{Var}\hat{I}_n$ for the four methods wit the x-axis now in units of computation time. Really, Figures 3.1 and 3.2 are presenting the same data with different scalings. When computation time is factored in, the third method does worse than the second method; the extra computation cost the third method pays compared to the second method is more than the benefit the reduced per-sample variance brings. The computation cost of the second and fourth methods are comparable, and we can see that the second method eventually catches up as expected by the theory, although the fourth method is better non-asymptotically.

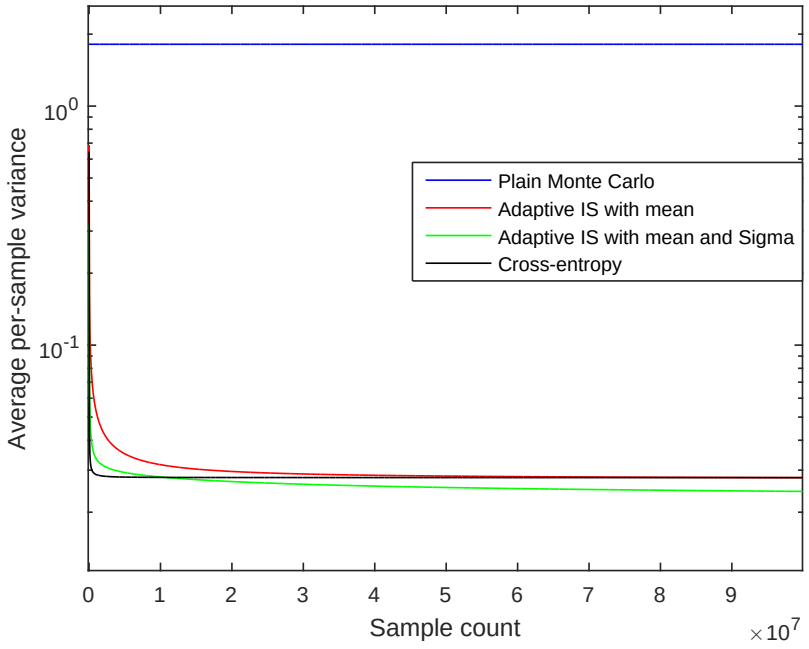


Figure 3.1: Average per-sample variances for the option pricing.

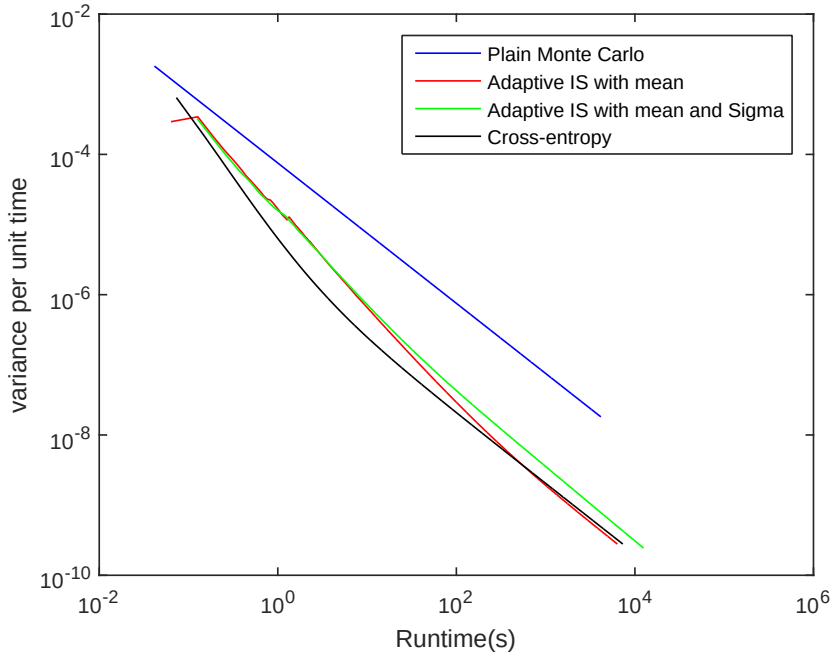


Figure 3.2: Estimator's variance as a function of runtime for the option pricing.

3.4.4 Stochastic mirror descent with mixtures

In this section, we consider the adaptive importance sampling algorithm we get when we choose a mixture of distributions for the family $\mathcal{F}$ (as defined in Section 2.3), the per-sample variance, $D_2(F_\star \| F_\theta)$, for the objective to minimize, and stochastic mirror descent for the stochastic optimization algorithm.

Assume

$$\int \frac{\phi^4(x)}{f_\theta^2(x)} \left(\frac{dF}{dF_\theta}\right)^3 \left(\sum_{i=1}^p p_i^2(x)\right) \, dF(x) < \infty \tag{3.3}$$

for all $\theta \in \Theta$.

The adaptive importance sampling method with these choices is

$$\begin{aligned}
&X_n \sim F_{\theta_n} \\
&\hat{I}_n = \frac{1}{n} \sum_{i=1}^n \phi(X_i) \frac{dF}{dF_{\theta_i}}(X_i) \\
&g_n = -\frac{\phi^2(X_n)}{f_{\theta_n}(X_n)} \left(\frac{dF}{dF_{\theta_n}}\right)^2 (X_n) \begin{pmatrix} p_1(X_n) \\ \vdots \\ p_p(X_n) \end{pmatrix} \\
&\theta_{n+1}^* = \theta_n^* - (C/\sqrt{n}) g_n \\
&\theta_{n+1} \propto \exp(\theta_{n+1}^*).
\end{aligned}$$

The estimator $\hat{I}_n$ is unbiased, and has performance

$$\frac{1}{n} V_\star \le \mathbf{Var} \hat{I}_n \le \frac{1}{n} V_\star + \mathcal{O}\left(\frac{\log n}{n^{3/2}}\right)$$

and

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

The adaptive importance sampling method with batch size m is

$$
\begin{aligned}
&X_{n1}, X_{n2}, \ldots, X_{nm} \sim F_{\theta_n} \\
&\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\phi(X_{ij})\frac{dF}{dF_{\theta_i}}(X_{ij}) \\
&g_n = -\frac{1}{m}\sum_{j=1}^{m}\frac{\phi^2(X_{nj})}{f_{\theta_n}(X_{nj})}\left(\frac{dF}{dF_{\theta_n}}\right)^2(X_{nj})\begin{pmatrix} p_1(X_{nj}) \\ \vdots \\ p_p(X_{nj}) \end{pmatrix} \\
&\theta^*_{n+1} = \theta^*_n - (C/\sqrt{n})g_n \\
&\theta_{n+1} \propto \exp(\theta^*_{n+1}).
\end{aligned}
$$

The estimator $\hat{I}_n$ is unbiased, and has performance

$$\frac{1}{nm}V_\star \leq \mathbf{Var}\hat{I}_n \leq \frac{1}{nm}V_\star + \mathcal{O}\left(\frac{\log n}{n^{3/2}m}\right)$$

and

$$\sqrt{nm}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

Discussion of assumptions. Assumption (3.3) implies all the conditions necessary for the stochastic optimization and central limit theorem to work as shown in Lemma 7.

However, one can reasonably argue that assumption (3.3) is too strong. In particular, it forbids setups that require defensive importance sampling; if $\phi dF \not\ll p_i d\mu$ for any $i \in \{1, 2, \ldots, p\}$ the assumption is violated.

Lemma 7. *Assumption* (3.3) *implies that* $\mathbb{E}\|g_i\|_2^2$ *is bounded, that* $D_{2+\varepsilon}(F_\star\|F_\theta)$ *is bounded for all* $\theta \in \Theta$*, and that* $D_2(F_\star\|F_\theta)$ *is finite for all* $\theta \in \Theta$*.*

Proof. Let us define

$$K(\theta) = \int \frac{\phi^4(x)}{f_\theta^2(x)}\left(\frac{dF}{dF_\theta}\right)^3\left(\sum_{i=1}^{p}p_i^2(x)\right)\,dF(x).$$

By the same argument as Lemma 3, K is a convex function on Θ. Since $\Theta = \mathbf{conv}\{e_1, \ldots, e_p\}$, where e_i denotes the ith unit vector for $i = 1, \ldots, p$ and **conv** denotes the convex hull,

$$\sup_{\theta \in \Theta} K(\theta) = \max_{i=1,\ldots,p} K(\theta_i) = B < \infty$$

by Theorem 32.2 of [63]. In other words, K is not only finite but also bounded on Θ. We immediately get $\mathbb{E}\|g_i\|_2^2$ is bounded since

$$K(\theta) = \mathbb{E}_{F_\theta}\|g\|^2.$$

By the Cauchy-Schwartz inequality and the equivalence of the ℓ_1 and ℓ_2 norms, we have

$$f_\theta^2(x) \leq \|\theta\|_2^2 \sum_{i=1}^p p_i(x) \leq p \sum_{i=1}^p p_i(x).$$

(Note that p is an integer and p_i is a probability density function.) So

$$\begin{aligned}\frac{1}{pJ^4} D_4(F_\star \| F_\theta) &= \frac{1}{p} \int \phi^4(x) \left(\frac{dF}{dF_\theta}\right)^3 dF(x) \\ &\leq \int \frac{\phi^4(x)}{f_\theta^2(x)} \left(\frac{dF}{dF_\theta}\right)^3 \left(\sum_{i=1}^p p_i^2(x)\right) dF(x) \leq B < \infty.\end{aligned}$$

This further implies that $D_{2+\varepsilon}(F_\star \| F_\theta)$ and $D_2(F_\star \| F_\theta)$ are bounded for all $\theta \in \Theta$. $\square$

Comparison with stochastic subgradient descent and defensive importance sampling. Had we chosen stochastic subgradient descent instead for our stochastic

optimization algorithm, the method would be

$$\begin{aligned}
&X_n \sim F_{\theta_n} \\
&\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{dF}{dF_{\theta_i}}(X_i) \\
&g_n = -\frac{\phi^2(X_n)}{f_{\theta_n}(X_n)}\left(\frac{dF}{dF_{\theta_n}}\right)^2(X_n)\begin{pmatrix} p_1(X_n) \\ \vdots \\ p_p(X_n)\end{pmatrix} \\
&\theta_{n+1} = \Pi_{\Delta_p}(\theta_n - (C/\sqrt{n})g_n).
\end{aligned}$$

The projection onto Δ_p is computationally simple c.f. Lemma 14 of the appendix. When assumption (3.3) is satisfied, this method converges as well.

Implementing the few lines of code for the projection onto Δ_p can be cumbersome. This is a trivial issue but it can be a reason to prefer stochastic mirror descent over stochastic subgradient descent.

There is, however, a good reason to use stochastic mirror descent in this setting. When using stochastic subgradient descent, it is likely that some entries of θ_n will be 0 for some n, by nature of the projection. When using stochastic mirror descent, however, all entries of θ_n remain positive for all finite n.

In the context of importance sampling with mixtures of $p_1, p_2, \ldots, p_p$, it may often be the case that

$$\phi dF \ll (p_1 + p_2 + \cdots + p_p)d\mu$$

but $\phi dF \not\ll p_i d\mu$ for some i. In this case, if θ does not have all positive entries, it is possible $\phi dF \not\ll f_\theta d\mu$, and the importance sampling fails.

If, say, $\phi dF \ll p_p d\mu$ and $p_p d\mu$ has finite per-sample variance, then one can use *defensive importance sampling* which requires that $\theta \in \Theta_p$ satisfies $\theta(p) \geq \varepsilon$ for some $\varepsilon > 0$. This ensures that $\phi dF \ll f_\theta d\mu$ and the per-sample variance of $f_\theta d\mu$ is finite [37, 38, 36]. One can perform defensive adaptive importance sampling with stochastic

subgradient descent since the projection onto

$$\{\theta \in \mathbf{R}^p \mid \mathbf{1}^T\theta = 1,\, \theta_1, \ldots, \theta_{p-1} \geq 0,\, \theta_p \geq \varepsilon\}$$

is also computationally simple. Modifying the proof of Lemma 14 of the appendix leads to an algorithm.

In most cases, however, it is probably simpler to use stochastic mirror descent. Since $\theta_n \succ 0$ for all n, all iterations are well-defined, i.e., $\phi dF \ll f_{\theta_n} d\mu$ and that per-sample variance of $f_\theta d\mu$ is finite, when using stochastic mirror descent, even though we do not have a rigorous proof of convergence.

Example. Consider the problem of computing

$$I = \mathbb{E}\left[\frac{\exp(2|X|)}{\sqrt{|X|}}\right] = \int_0^\infty \frac{\exp(2x)}{\sqrt{x}} \frac{\sqrt{2}\exp(-x^2/2)}{\sqrt{\pi}}\, dx,$$

where X is a standard normal. Plain Monte Carlo on this setup will not work well since the random variable $\exp(2|X|)/\sqrt{|X|}$ has infinite variance. (Even though the estimate will converge to I by strong law of large numbers.)

For this problem, we consider the adaptive importance sampling algorithm we get when we choose a mixture of distrubutions P_1, P_2, and P_3 for the family $\mathcal{F}$, the per-sample variance, $D_2(F_\star \| F_\theta)$, for the objective to minimize, and stochastic mirror descent for the stochastic optimization algorithms. The distributions to mix are

$$\begin{aligned} dP_1(x) &= \frac{1}{2\sqrt{x}} I_{[0,1]}(x)dx \\ dP_2(x) &= \exp(-x) I_{[0,\infty)}(x)dx \\ dP_3(x) &= \frac{1}{5}\exp(-x/5) I_{[0,\infty)}(x)dx. \end{aligned}$$

As a comparison, we also consider non-adaptive importance sampling with the parameter $\theta = (1/3, 1/3, 1/3)$. We run the simulation with paramters $n = 10^3$, $m = 10^3$, and $C = 10^{-3}$. We get $I \approx 11.7874$, and the estimators' variances are shown in Figure 3.3. The optimal mixture weights turn out to be $\theta_\star = (0.01, 0.53, 0.46)$.

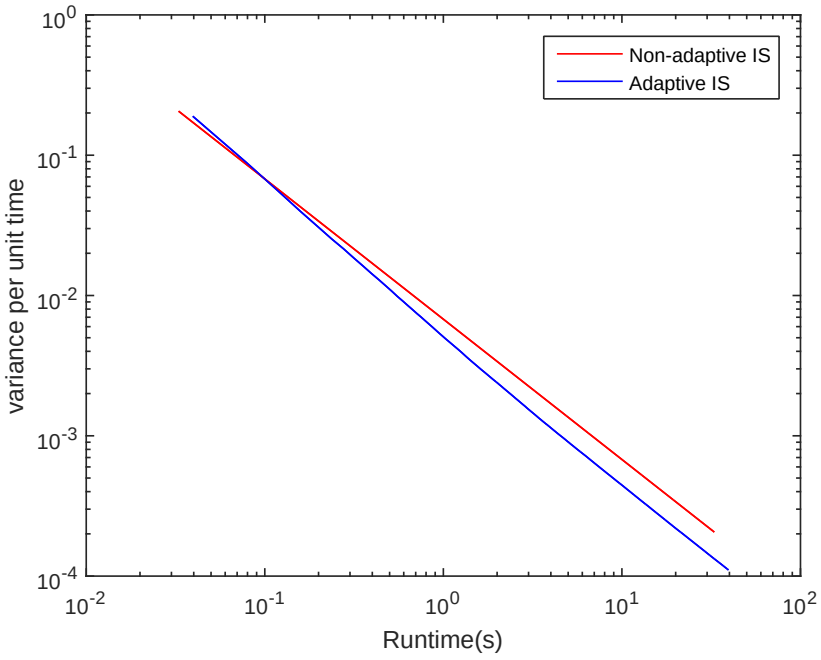


Figure 3.3: Estimator's variance as a function of runtime for the mixture example.

This is an example where we observe convergence without meeting assumption (3.3). In particular, $\phi dF \not\ll p_1 dx$ and if $\theta(1) = 0$ the per-sample variance of $f_\theta dx$ is infinite. Nevertheless, the method works well for reasons we have discussed.

3.4.5 Error rate of a communication channel

Consider the problem of finding the error rate of a communication channel using 8-PSK (phase-shift keying) subject to addative white Gaussian noise (AWGN). We simply provide a terse description and refer interested readers to [5, 70, 35, 45].

The digital modulation scheme 8-PSK sends 3 bits of information (so cases $n = 1, 2, \ldots, 8$) via sending

$$s_n(t) = \sqrt{\frac{2E}{T}} \cos\left(2\pi f_c t + n\frac{\pi}{4}\right),$$

from time $t = 0$ to $t = T$. Here E is the energy per symbol (i.e., energy per 3 bits),

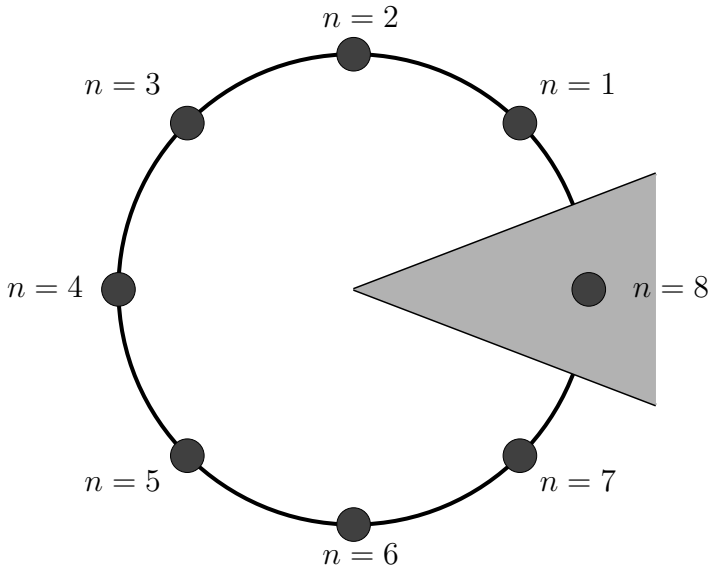


Figure 3.4: Constellation diagram for 8PSK. When $n = 8$ is sent, the decoding succeeds if the random variable X lands within the shaded region.

T is the symbol duration, and f_c is the carrier wave frequency (such that f_cT is a positive integer). This signal is sent through a channel and is corrupted by additive white Gaussian noise:

$$d\tilde{s}_n(t) = ds_n + \sqrt{\frac{N_0}{2}}dW_t,$$

where dW_t is white noise or Brownian motion and $N_0/2$ is the noise level. The decoder receives this corrupted signal and performs integration (say via an integrator circuit) to obtain

$$X(1) = \sqrt{\frac{E}{2}}\int_0^T \sqrt{\frac{2}{T}}\cos(2\pi f_c t)\tilde{s}(t)\,dt \stackrel{\mathcal{D}}{=} \cos\left(n\frac{\pi}{4}\right) + \sqrt{\frac{N_0}{E}}Z_1$$
$$X(2) = \sqrt{\frac{E}{2}}\int_0^T \sqrt{\frac{2}{T}}\sin(2\pi f_c t)\tilde{s}(t)\,dt \stackrel{\mathcal{D}}{=} \sin\left(n\frac{\pi}{4}\right) + \sqrt{\frac{N_0}{E}}Z_2,$$

where Z_1 and Z_2 are independent standard normals. Finally, the decoder uses the *nearest neighbor decoding rule*, i.e., it reports the n for which the angle $n\pi/8$ is closest to the angle of the coordinate $(X(1), X(2))$.

Assume the signal $n = 8$ is transmitted. Then the decoder receives a 2D Gaussian

$$X \sim \mathcal{N}\left((1,0), \sqrt{N_0/E}I\right).$$

Let

$$\phi(X) = \begin{cases} 0 & \text{if } -\pi/8 \leq \textbf{angle}(X) \leq \pi/8 \text{ (decoding correct)} \\ 1 & \text{otherwise (decoding fail).} \end{cases}$$

In other words, $\phi(X) = 1$ if the random variable X lies outside the shaded region of Figure 3.4. We define $I = \mathbb{E}\phi(X)$ as the error rate. (The error rates when the signal $n = 1, 2, \ldots, 7$ are sent are the same by symmetry.)

We consider the adaptive importance sampling algorithm we get when we choose multivariate Gaussians for the family $\mathcal{F}$, the per-sample variance, $D_2(F_\star \| F_\theta)$, for the objective to minimize, and stochastic mirror descent for the stochastic optimization

algorithm. The adaptive importance sampling method with batch size m is

$$
\begin{aligned}
&L_n L_n^T = \Sigma_n \quad \text{(Cholesky factorization)}\\
&Y_{n1}, Y_{n2}, \ldots, Y_{nm} \sim \mathcal{N}(0, I)\\
&X_{nj} = L_n^T Y_{nj} + \mu_n, \quad j = 1, \ldots, m\\
&w_{nj} = \frac{\phi(X_{nj})}{\sqrt{\det(S_n)}} \exp(-\|X_{nj} - (1,0)\|_2^2/2 + (X_{nj} - \mu_n)^T S_n (X_{nj} - \mu_n)/2) \quad j = 1, \ldots, m\\
&\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m}\sum_{j=1}^{m} w_{ij}\\
&g_n^S = \frac{1}{m}\sum_{j=1}^{m} w_{ij}^2 (X_{nj} X_{nj}^T - \mu_n \mu_n^T - \Sigma_n)\\
&g_n^b = \frac{1}{m}\sum_{j=1}^{m} w_{ij}^2 (\mu_n - X_{nj})\\
&S_{n+1}^* = S_n^* - (C_1/\sqrt{n}) g_n^S\\
&S_{n+1} = \exp S_{n+1}^*\\
&b_{n+1} = \Pi_{C_b}(b_n - (C_2/\sqrt{n}) g_n^b)\\
&\Sigma_{n+1} = S_{n+1}^{-1}\\
&\mu_{n+1} = S_{n+1}^{-1} b_{n+1},
\end{aligned}
$$

with $C_b = [-10, 10]^k$.

We run this simulation with $n = 10^3$, $m = 10^3$, $N_0/E = 0.01$, and $C_1 = C_2 = 10$. For comparison, we also run plain Monte Carlo. We get $I \approx 1.2866 \times 10^{-4}$. The results are shown in Figures 3.5 and 3.6.

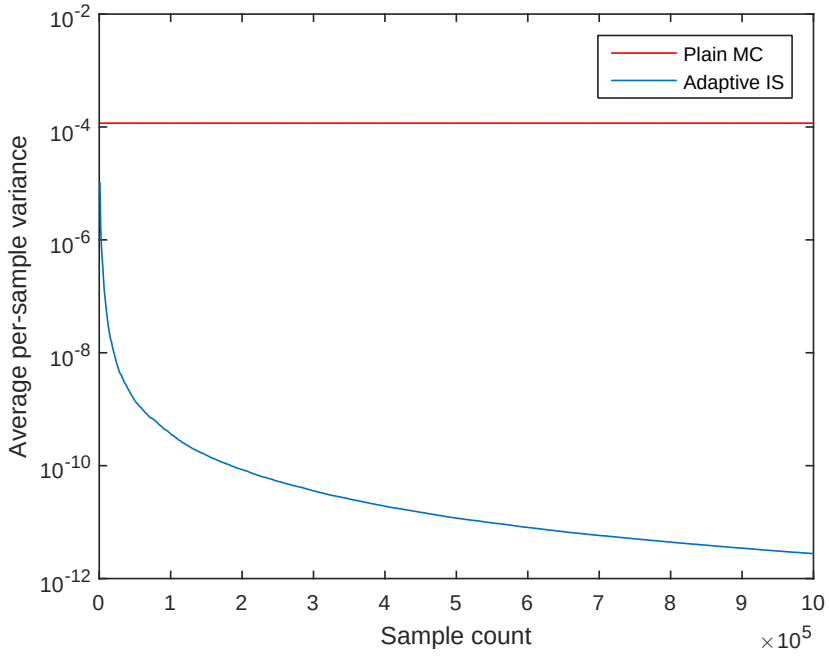


Figure 3.5: Average per-sample variance for the error rate problem.

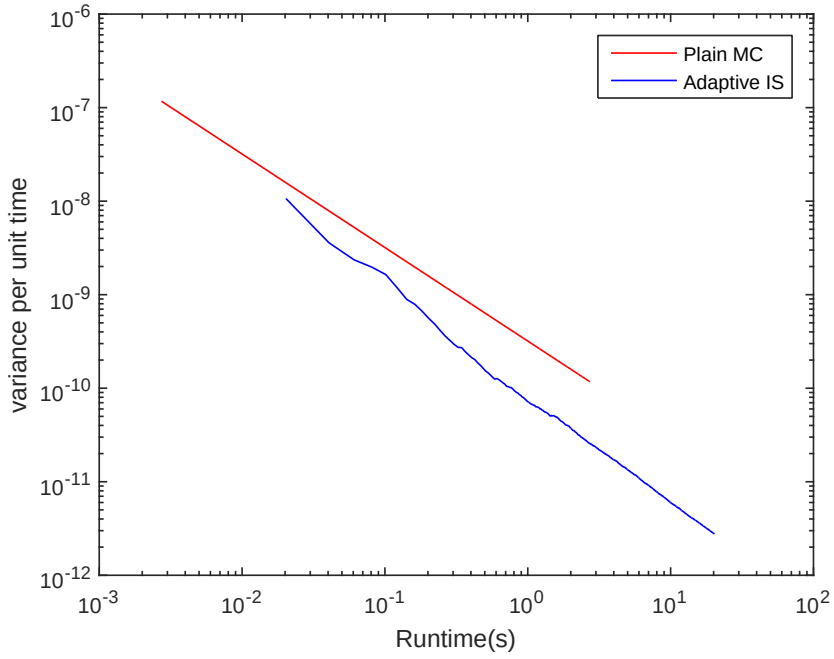


Figure 3.6: Variance of the estimator as a function of time for the error rate problem.

Chapter 4

Self-normalized importance sampling

Let F be an *unnormalized* probability measure, i.e., the *normalizing factor* $0 < \int dF < \infty$ is unknown and not necessarily 1. (F is a nonnegative measure.) Consider the problem of computing

$$I = \mathbb{E}_{F/\int dF}\phi(X) = \frac{\int \phi \, dF}{\int dF},$$

where $X \sim F/\int dF$ is a random variable on $\mathcal{X}$ and $\phi : \mathcal{X} \to \mathbf{R}$. Again we assume $0 < \mathbb{E}|\phi(X)| < \infty$.

In this setting, generating random samples $X \sim F/\int dF$ is often inefficient. So although one could employ the plain Monte Carlo method to compute I, its computational cost per iteration may be unappealing.

Self-normalized importance sampling circumvents this difficulty, in addition to possibly providing variance reduction. The method computes I with

$$\begin{aligned} X_n &\sim \tilde{F} \\ \hat{I}_n &= \left(\sum_{i=1}^{n} \phi(X_i)\frac{dF}{d\tilde{F}}(X_i)\right) / \left(\sum_{i=1}^{n} \frac{dF}{d\tilde{F}}(X_i)\right), \end{aligned}$$

where the sampling distribution $\tilde{F}$ satisfies $dF \ll d\tilde{F}$. The idea is that the numerator of $\hat{I}_n$ converges to $n \int \phi \, dF$ while the denominator converges to $n \int dF$. The estimator $\hat{I}_n$ is no longer unbiased, i.e., $\mathbb{E}\hat{I}_n \neq I$, but it is asymptotically consistent and has asymptotic variance

$$\mathbb{E}(\hat{I}_n - I)^2 \approx \frac{1}{n} \frac{J^2 \exp D_2(F_\star \| \tilde{F})}{(\int dF)^2}. \tag{4.1}$$

Like before, we call

$$\frac{J^2 \exp D_2(F_\star \| \tilde{F})}{(\int dF)^2}$$

the *asymptotic per-sample variance* of $\tilde{F}$. Here $F_\star$ is the distribution such that

$$dF_\star = \frac{1}{J} |\phi - I| \, dF$$

with normalizing factor

$$J = \int |\phi(x) - I| \, dF(x).$$

Asymptotic variance. Since

$$\frac{1}{n} \sum_{i=1}^{n} \frac{dF}{d\tilde{F}}(X_i) \to \int dF$$

almost surely by the law of large numbers, we have

$$\begin{aligned}
\mathbb{E}(\hat{I}_n - I)^2 &= \frac{1}{n^2 (\int dF)^2} \mathbb{E}\left(\frac{\sum_{i=1}^n (\phi(X_i) - I) \frac{dF}{d\tilde{F}}(X_i)}{1 + \left(\frac{1}{n \int dF} \sum_{i=1}^n \frac{dF}{d\tilde{F}}(X_i) - 1 \right)} \right)^2 \\
&\approx \frac{1}{n^2 (\int dF)^2} \mathbb{E}\left(\sum_{i=1}^{n} (\phi(X_i) - I) \frac{dF}{d\tilde{F}}(X_i) \right)^2 \\
&= \frac{1}{n (\int dF)^2} \mathbb{E}_{\tilde{F}}\left((\phi(X) - I) \frac{dF}{d\tilde{F}}(X) \right)^2 \\
&= \frac{1}{n (\int dF)^2} J^2 \exp D_2(F_\star \| \tilde{F}).
\end{aligned}$$

Although this argument is not rigorous, it does illustrate the main idea and should be good enough in practice.

Central limit theorem. If $dF \ll d\tilde{F}$,

$$\mathbf{Var}_{\tilde{F}}\left(\frac{dF}{d\tilde{F}}(X)\right) < \infty, \quad \mathbf{Var}_{\tilde{F}}\left(\phi(X)\frac{dF}{d\tilde{F}}(X)\right) < \infty,$$

then one can rigorously establish

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{(\int dF)^2} J^2 \exp D_2(F_\star \| \tilde{F})\right)$$

via the delta method.

Outline. In this section we show how the adaptive importance sampling method presented in Section 3 extends to adaptive self-normalized importance sampling. The material is presented in a way to emphasize the parallels between the two setups.

In this section, however, much of the theoretical analysis, such as the derivation of (4.1), is not rigorous. Rather, the analysis should be intepreted as heuristic justifications of why the methods of this section work as well as the methods of Section 3.

4.1 Adaptive self-normalized importance sampling

Adaptive self-normalized importance sampling computes I with

$$\begin{aligned} X_n &\sim \tilde{F}_i \\ \hat{I}_n &= \left(\sum_{i=1}^n \phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i)\right) / \left(\sum_{i=1}^n \frac{dF}{d\tilde{F}_i}(X_i)\right) \\ \tilde{F}_{n+1} &= \text{update with } X_1, \ldots, X_n, \tilde{F}_1, \ldots, \tilde{F}_n, \end{aligned}$$

where the sampling distributions satisfy $dF \ll d\tilde{F}_n$ for $n = 1, 2, \ldots$.

As in the non-adaptive case, we can argue that $\hat{I}_n$ is asymptotically consistent

with asymptotic variance

$$\mathbb{E}(\hat{I}_n - I)^2 \approx \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{J^2 \exp D_2(F_\star \| \tilde{F}_i)}{(\int dF)^2}\right).$$

Asymptotic consistency. Assume

$$\mathbf{Var}_{\tilde{F}_n}\left(\frac{dF}{d\tilde{F}_n}(X_n)\right), \quad \mathbf{Var}_{\tilde{F}_n}\left(\phi(X_n)\frac{dF}{d\tilde{F}_n}(X_n)\right)$$

are bounded for $n = 1, 2, \ldots$. This means the sampling distributions $\tilde{F}_1, \tilde{F}_2, \ldots$ cannot get worse indefinitely.

Then the strong law of large numbers for Martingales [29, §VII.9 Theorem 3] give us

$$\frac{1}{n}\sum_{i=1}^{n}\frac{dF}{d\tilde{F}_i}(X_i) \to \int dF$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)\frac{dF}{d\tilde{F}_i}(X_i) \to \int \phi \, dF$$

almost surely. This gives us $\hat{I}_n \to I$ almost surely.

Asymptotic variance. Assume

$$\mathbf{Var}_{\tilde{F}_n}\left(\frac{dF}{d\tilde{F}_n}(X_n)\right)$$

is bounded for $n = 1, 2, \ldots$. Then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{dF}{d\tilde{F}_i}(X_i) \to \int dF,$$

almost surely, as discussed before.

Then we can say

$$\begin{aligned}
\mathbb{E}(\hat{I}_n - I)^2 &= \frac{1}{n^2(\int dF)^2}\mathbb{E}\left(\frac{\sum_{i=1}^n(\phi(X_i)-I)\frac{dF}{d\tilde{F}_i}(X_i)}{1+\left(\frac{1}{n\int dF}\sum_{i=1}^n\frac{dF}{d\tilde{F}_i}(X_i)-1\right)}\right)^2 \\
&\approx \frac{1}{n^2(\int dF)^2}\mathbb{E}\left(\sum_{i=1}^n(\phi(X_i)-I)\frac{dF}{d\tilde{F}_i}(X_i)\right)^2 \\
&= \frac{1}{n(\int dF)^2}\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}\mathbb{E}_{\tilde{F}_i}\left((\phi(X)-I)\frac{dF}{d\tilde{F}_i}(X)\right)^2\right) \\
&= \frac{1}{n(\int dF)^2}\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}J^2\exp D_2(F_\star\|\tilde{F}_i)\right),
\end{aligned}$$

where we use the same conditional dependency argument as in Section 3.2. Again, this argument is not rigorous, but is made to illustrate the main idea.

4.1.1 Main framework

As in Section 3, determining how to update $\tilde{F}_n$ fully specifies the method and its performance, and we will do so by making following 3 choices:

- Choose a family of distributions.
- Choose an objective to minimize.
- Choose a stochastic optimization method to perform the minimization.

Again, if we choose a family $\mathcal{F}$ with a log-concave parameterization, the asymptotic per-sample variance

$$\frac{J^2\exp D_2(F_\star\|F_\theta)}{(\int dF)^2}$$

is a convex function of θ. Using the parameterization, the algorithm will be of the form:

$$\begin{aligned} X_n &\sim F_{\theta_n} \\ \hat{I}_n &= \left(\sum_{i=1}^{n} \phi(X_i)\frac{dF}{dF_{\theta_i}}(X_i)\right) / \left(\sum_{i=1}^{n} \frac{dF}{dF_{\theta_i}}(X_i)\right) \\ \theta_{n+1} &= \text{Stochastic optimization with } X_1, \ldots, X_n, \theta_1, \ldots, \theta_n. \end{aligned} \tag{4.2}$$

Write $V_\star$ for the mimumum asymptotic per-sample variance for the family $\mathcal{F}$, i.e., $V_\star$ is the optimal value for

$$\begin{array}{ll} \text{minimize} & (J/\int dF)^2 \exp D_2(F_\star \| F_\theta) \\ \text{subject to} & \theta \in \Theta. \end{array}$$

If algorithm (4.2) is run with a stochastic optimization algorithm such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\frac{J^2 \exp D_2(F_\star \| F_{\theta_i})}{(\int dF)^2} = V_\star + o(1),$$

then we have

$$\frac{1}{n}V_\star \lessapprox \mathbf{Var}\hat{I}_n \lessapprox \frac{1}{n}V_\star + o\left(\frac{1}{n}\right).$$

So we can say

$$\mathbf{Var}\hat{I}_n \approx \frac{1}{n}V_\star$$

and the variance of $\hat{I}$ is asymptotically optimal with respect to the family $\mathcal{F}$.

As in Section 3, we can use stochastic subgradient descent or stochastic mirror descent for this. Unfortunately, however, there is a complication for self-normalized adaptive importance sampling: we do not have access to stochastic gradients the same way we did in adaptive importance sampling.

4.2 Biased stochastic subgradients

As discussed in Section 2.4, when $X_n \sim F_{\theta_n}$

$$g_n = -\frac{(\phi(X_n) - I)^2}{f_{\theta_n}(X)} \left(\frac{dF}{dF_{\theta_n}}(X) \right)^2 \nabla_\theta f_{\theta_n}(X)$$

is a (constant multiple of a) stochastic subgradient of the asymptotic per-sample variance at θ_n:

$$\frac{1}{\int dF} \mathbb{E} g \in \partial_\theta \left(\frac{J^2 \exp D_2(F_\star \| F_\theta)}{(\int dF)^2} \right) (\theta_n)$$

That we do not know $\int dF$ is not a problem; we can absorb the fixed unknown constant into the step size in our stochastic optimization methods.

However, it is a problem that g uses I, the unknown quantity of interest. So instead of g, we use

$$\tilde{g}_n = -\frac{(\phi(X_n) - \hat{I}_{n-1})^2}{f_{\theta_n}(X)} \left(\frac{dF}{dF_{\theta_n}}(X) \right)^2 \nabla_\theta f_{\theta_n}(X),$$

and we say $\tilde{g}_n$ is a *biased stochastic subgradient*.

Stochastic subgradient descent or stochastic mirror descent can use biased stochastic subgradients, as long as the bias or error diminishes to 0 as $n \to \infty$. For example, the algorithm

$$\theta_{n+1} = \Pi_\Theta(\theta_n - \alpha_n \tilde{g}_n) \tag{4.3}$$

with $\alpha_n = C/\sqrt{n}$ solves problem (2.1) with rate

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} U(\theta_i) \leq U(\theta_\star) + \mathcal{O}\left(\frac{1}{\sqrt{n}} \right) + \mathcal{O}\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E} \|\tilde{g}_i - g_i\|_2 \right).$$

So

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} U(\theta_i) \to U(\theta_\star)$$

provided that $\mathbb{E}\|\tilde{g}_n - g_n\|_2 \to 0$.

Convergence proof. With appropriate assumptions, we can establish $\hat{I}_n \to I$ almost surely, as discussed. However, this does not necessarily imply $\mathbb{E}\|\tilde{g}_n - g_n\| \to 0$. Nevertheless, the assumption $\mathbb{E}\|\tilde{g}_n - g_n\| \to 0$ seems reasonable, and the following lemma based on it should reasonably explain the behavior of the stochastic optimization we use in our methods.

Lemma 8. *Assume that Θ is nonempty compact convex, that U has a subgradient for all $\theta \in \Theta$, that $\mathbb{E}[\|g_n\|_2^2|\theta_n] \le G^2 < \infty$ for $n = 1, 2, \ldots$, If g_n is a stochastic subgradient of U at θ_n then (4.3) converges with rate*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U(\theta_i) \le U(\theta_\star) + \mathcal{O}(1/\sqrt{n}) + \frac{D}{n}\sum_{i=1}^{n}\mathbb{E}\|\tilde{g}_i - g_i\|_2.$$

So if $\mathbb{E}\|\tilde{g}_n - g_n\|_2 \to 0$ then

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U(\theta_i) \to U(\theta_\star).$$

Proof. Here we simply outline the key steps that are different from the analysis of Section 2.2.1. The initial inequalities change to

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{i+1} - \theta_\star\|_2^2 \mid \theta_i\right] &\le \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}G^2 - 2\frac{C}{\sqrt{i}}(U(\theta_i) - U(\theta_\star)) \\
&\quad + 2\frac{C}{\sqrt{i}}\mathbb{E}[(g_i - \tilde{g}_i)^T(\theta_i - \theta_\star) \mid \theta_i] \\
&\le \|\theta_i - \theta_\star\|_2^2 + \frac{CG^2}{i} - 2\frac{C}{\sqrt{i}}(U(\theta_i) - U(\theta_\star)) \\
&\quad + 2\frac{C}{\sqrt{i}}\mathbb{E}[\|\tilde{g}_i - g_i\|_2 \mid \theta_i]\|\theta_i - \theta_\star\|_2 \\
&\le \|\theta_i - \theta_\star\|_2^2 + \frac{CG^2}{i} - 2\frac{C}{\sqrt{i}}(U(\theta_i) - U(\theta_\star)) \\
&\quad + 2\frac{CD}{\sqrt{i}}\mathbb{E}[\|\tilde{g}_i - g_i\|_2 \mid \theta_i],
\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz. Following the same steps

as before we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U(\theta_i) \le U(\theta_\star) + \left(\frac{D^2}{2C} + CG^2\right)\frac{1}{\sqrt{n}} + \frac{D}{n}\sum_{i=1}^{n}\mathbb{E}\|\tilde{g}_i - g_i\|_2.$$

□

4.3 Central limit theorem

Lemma 9. *Assume algorithm* (4.2) *uses a stochastic optimization with performance*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\frac{J^2 \exp D_2(F_\star\|F_{\theta_i})}{(\int dF)^2} = V_\star + o(1).$$

Furthermore, assume $D_2(F/\int dF\|F_\theta)$ *and* $D_{2+\varepsilon}(F_\star\|F_\theta)$ *are bounded for all* $\theta \in \Theta$. *Then we have*

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

Proof. With the same Martingale CLT argument as in Section 3.3, we have

$$\frac{1}{\sqrt{n}\int dF}\sum_{i=1}^{n}(\phi(X_i) - I)\frac{dF}{dF_{\theta_i}}(X_i) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

Since the 2nd moments are bounded we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{dF}{dF_{\theta_i}}(X_i) \to \int dF$$

in probability. Since

$$\hat{I}_n - I = \frac{1}{n\int dF}\left(1 + \left(\frac{1}{n\int dF}\sum_{i=1}^{n}\frac{dF}{dF_{\theta_i}}(X_i) - 1\right)\right)^{-1}\sum_{i=1}^{n}(\phi(X_i) - I)\frac{dF}{dF_{\theta_i}}(X_i)$$

we apply Slutsky's lemma to get the desired result. □

4.4 Examples

4.4.1 Stochastic gradient descent with exponential family

In this section, we consider the adaptive importance sampling algorithm we get when we choose an exponential family for the family $\mathcal{F}$ (as defined in Section 2.3), $D_2(F_\star \| F_\theta)$, the asymptotic per-sample variance, for the objective to minimize, and stochastic subgradient descent (with biased stochastic subgradients) for the stochastic optimization algorithm.

The adaptive importance sampling method with these choices is

$$\begin{aligned}
X_n &\sim F_{\theta_n} \\
w_n &= \frac{dF}{dF_{\theta_n}}(X_n) \\
\hat{I}_n &= \left(\sum_{i=1}^{n} \phi(X_i) w_i\right) / \left(\sum_{i=1}^{n} w_i\right) \\
g_n &= (\phi(X_n) - \hat{I}_{n-1})^2 w_n^2 (\nabla A(\theta) - T(X_n)) \\
\theta_{n+1} &= \Pi_\Theta(\theta_n - (C/\sqrt{n}) g_n).
\end{aligned}$$

We expect $\hat{I}_n$ to have performance

$$\mathbb{E}(\hat{I}_n - I)^2 \approx \frac{1}{n} V_\star$$

and

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_\star).$$

4.4.2 Bayesian estimation

Consider the Bayesian setup where $Z_1, Z_2, \ldots, Z_\ell \sim \mathcal{N}(\mu, I)$ and μ is given the prior

$$\pi(\mu) \propto \begin{cases} \left(\prod_{i=1}^{k} \mu(i)\right) \exp\left(-\prod_{i=1}^{k} \mu(i)\right) & \text{if } \mu \succ 0 \\ 0 & \text{otherwise.} \end{cases}$$

Write

$$\bar{Z} = \frac{1}{\ell} \sum_{i=1}^{\ell} Z_i.$$

Then the unnormalized posterior distribution of μ is

$$f(\mu) = \frac{\ell^k}{(2\pi)^{k/2}} \exp(-\ell \|\mu - \bar{Z}\|_2^2/2)\pi(\mu).$$

We wish to estimate

$$I = \mathbb{E}[\mathbf{1}^T \mu | Z_1, \ldots, Z_\ell].$$

We consider the adaptive importance sampling algorithm we get when we choose affine transformations of the continuous random variable with density

$$p(y) = \frac{1}{2^k} \exp(-\|y\|_1)$$

for the family $\mathcal{F}$, $D_2(F_\star \| F_\theta)$, the asymptotic per-sample variance, for the objective to minimize, and stochastic mirror descent for the stochastic optimization algorithm.

The adaptive self-normalized importance sampling algorithm with batch size m is

$$
\begin{aligned}
&Y_{n1}, \ldots, Y_{nm} \sim p(x) \\
&X_{nj} \sim A_n^{-1}(Y_{nj} - b_n), \quad j = 1, \ldots, m \\
&w_{nj} = \frac{f(X_{nj})}{p(Y_{nj}) \det(A_n)} \\
&\hat{I}_n = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \phi(X_{ij}) w_{ij} \right) / \left(\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} \right) \\
&h_{nj} = \mathbf{sign}(Y_{nj}) \\
&g_n^A = \frac{1}{2m} \sum_{j=1}^{m} (\phi(X_{nj}) - \hat{I}_{n-1})^2 w_{nj}^2 (h_{nj} X_{nj}^T + X_{nj} h_{nj}^T - 2A_n^{-1}) \\
&g_n^b = \frac{1}{m} \sum_{j=1}^{m} (\phi(X_{nj}) - \hat{I}_{n-1})^2 w_{nj}^2 h_{nj} \\
&A_{n+1}^* = A_n^* - (C_1/\sqrt{n}) g_n^A \\
&A_{n+1} = \exp A_{n+1}^* \\
&b_{n+1} = b_n - (C_2/\sqrt{n}) g_n^b.
\end{aligned}
$$

For comparison, we also consider the non-adaptive importance sampling method where $X = Y + \bar{Z}$ and $Y \sim p(x)dx$.

We use the parameters $n = 10^3$, $m = 10^3$, $k = 3$, $\ell = 20$, and $C_1 = C_2 = 0.3$. The computations give us $I \approx 4.5203$. Figures 4.1 and 4.2 show the performance.

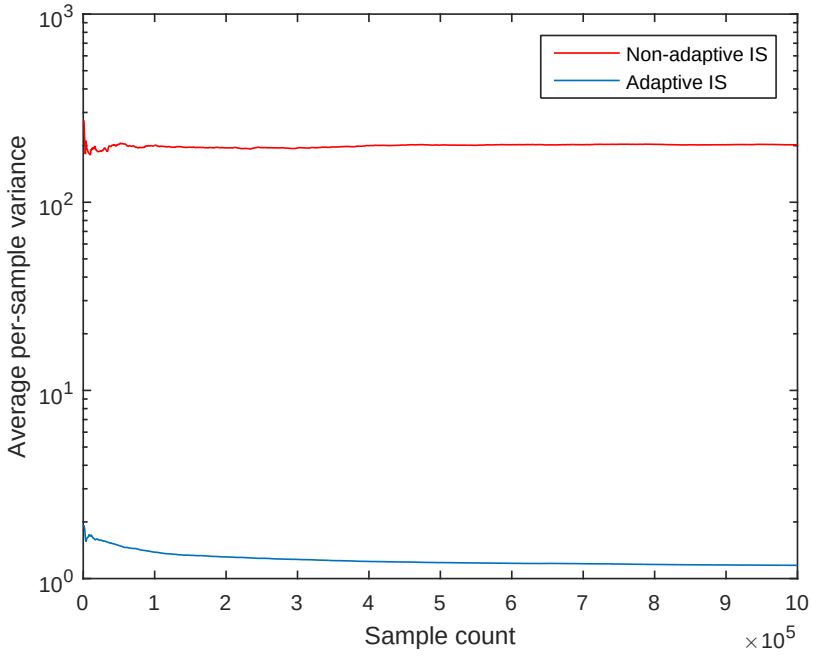


Figure 4.1: Asymptotic per-sample variance for the Bayes estimation problem.

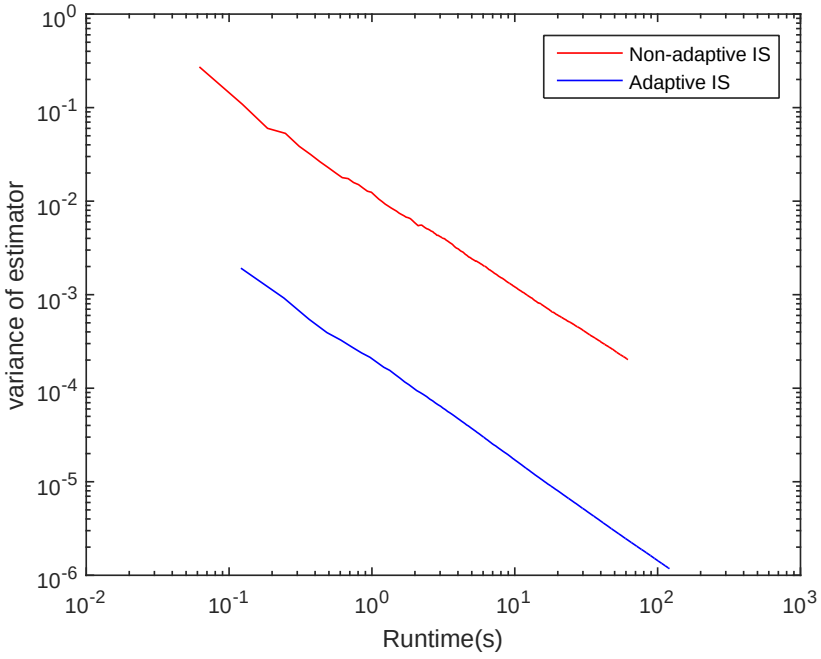


Figure 4.2: Variance of the estimator for the Bayes estimation problem.

Chapter 5

What-if simulation

Consider the problem of computing the function

$$I(a) = \mathbf{E}_{F(a)}\phi(X;a) = \int \phi(x;a)\, dF(x;a)$$

for $a \in \mathcal{A}$. Here $F(a)$ is a probability measure on $\mathcal{X}$ for each a and $X \sim F(a)$ is a random variable on $\mathcal{X}$. The integrand is $\phi : \mathcal{X} \times \mathcal{A} \to \mathbf{R}$. Again, we assume $0 < \mathbb{E}_{F(a)}|\phi(X;a)| < \infty$.

One approach is to perform separate Monte Carlo simulations on $I(a)$ for each $a \in \mathcal{A}$. However, it is often better to compute $I(a)$ for all $a \in \mathcal{A}$ simulatneously via importance sampling:

$$\begin{aligned} X_n &\sim \tilde{F}(a) \\ \hat{I}_n(a) &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i;a)\frac{dF(a)}{d\tilde{F}}(X_i). \end{aligned}$$

The sampling distribution has to satisfy $\phi(x;a)dF(x;a) \ll d\tilde{F}(x)$ for all $a \in \mathcal{A}$. As before, $\hat{I}_n(a)$ is unbiased, i.e., $I(a) = \mathbb{E}\hat{I}_n(a)$, and

$$\max_{a\in\mathcal{A}} \mathbf{Var}\hat{I}_n(a) = \frac{1}{n}\max_{a\in\mathcal{A}} \mathbf{Var}_{\tilde{F}}\left(\phi(X;a)\frac{dF(a)}{d\tilde{F}}(X)\right).$$

Adaptive importance sampling for what-if simulations has the form

$$\begin{aligned} X_n &\sim \tilde{F}_i \\ \hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n} \phi(X_i;a)\frac{dF(a)}{d\tilde{F}_i}(X_i) \\ \tilde{F}_{n+1} &= \text{update with } X_1,\ldots,X_n,\tilde{F}_1,\ldots,\tilde{F}_n, \end{aligned}$$

where $\phi(x;a)dF(a) \ll \tilde{F}_i$ for all $a \in \mathcal{A}$ and $i = 1,2,\ldots$. Again, $\hat{I}_n$ is unbiased, i.e., $\mathbb{E}\hat{I}_n(a) = I$, and

$$\max_{a\in\mathcal{A}} \mathbf{Var}\hat{I}_n(a) = \frac{1}{n}\max_{a\in\mathcal{A}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbf{Var}_{\tilde{F}_i}\left(\phi(X;a)\frac{dF(a)}{d\tilde{F}_i}(X)\right)\right).$$

In this section, we explore adaptive importance sampling methods that minimize the maximum variance of the estimator.

5.1 Primal-dual formulation

Often, $\mathcal{A}$, the set of parameters, is continuous, and therefore $|\mathcal{A}| = \infty$. For now, however, let us assume $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$.

We write Δ_ℓ for the ℓ-dimensional probability simplex, i.e.,

$$\Delta_\ell = \{\lambda \in \mathbf{R}^\ell \mid \mathbf{1}^T\lambda = 1,\ \lambda_1,\ldots,\lambda_\ell \geq 0\}.$$

Then it is simple to verify

$$\max_{a\in\mathcal{A}} \mathbf{Var}_{\tilde{F}}\left(\phi(X_i;a)\frac{dF(a)}{d\tilde{F}}(X)\right) = \max_{\lambda\in\Delta_\ell}\sum_{k=1}^{\ell}\lambda(k)\mathbf{Var}_{\tilde{F}}\left(\phi(X;a_k)\frac{dF(a_k)}{d\tilde{F}}(X)\right).$$

Now let $\mathcal{F}$ be a family with log-concave parameterization, where $\phi(x;a)dF(x;a) \ll$

$dF_\theta(x)$ for all $\theta \in \Theta$ and $a \in \mathcal{A}$. Then define

$$K(\theta, \lambda) = \sum_{k=1}^{\ell} \lambda(k) \mathbf{Var}_{F_\theta} \left(\phi(X; a_k) \frac{dF(a_k)}{dF_\theta}(X) \right),$$

which is convex in θ and concave in λ.

Finding the sampling distribution within $\mathcal{F}$ with minimum maximum variance is equivalent to finding a saddle point of K, i.e., the sampling dsitribution $F_{\theta_\star}$ has the minimum maximum variance if

$$\theta_\star = \operatorname*{argmin}_{\theta \in \Theta} \max_{\lambda \in \Delta_\ell} K(\theta, \lambda).$$

We write $K_\star$ for the minimum maximum variance, i.e.,

$$K_\star = \inf_{\theta \in \Theta} \max_{\lambda \in \Delta_\ell} K(\theta, \lambda) = \max_{\lambda \in \Delta_\ell} \inf_{\theta \in \Theta} K(\theta, \lambda).$$

The order of min and max can be swapped due to Sion's minimax theorem [69].

Stochastic subgradients. For a function $K(\theta, \lambda)$ that is convex in θ and concave in λ, we say (g, h) is a subgradient of K at (θ_0, λ_0) if

$$\begin{aligned} \mathbb{E}g &\in \partial_\theta K(\theta_0, \lambda_0) \\ \mathbb{E}h &\in -\partial_\lambda(-K)(\theta_0, \lambda_0). \end{aligned}$$

If K is differentiable with respect to λ at (θ_0, λ_0), then

$$\mathbb{E}h = \nabla_\lambda K(\theta_0, \lambda_0).$$

Let $X_1, \ldots, X_m \sim F_\theta$ be independent, and let the parameterization of $\mathcal{F}$ be log-concave. Then by a similar argument as Section 2.4.1,

$$\begin{aligned}
w_j(a) &= \phi(X_j; a)\frac{dF(a)}{dF_{\theta_0}}(X_j), \quad j = 1, \ldots, m \\
g &= \frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{\ell}\lambda(k)\gamma_k^2 w_j(a_k)^2\left(-\frac{1}{f_{\theta_n}(X_j)}\nabla_\theta f_{\theta_0}(X_j)\right) \\
h(k) &= \text{Sample variance of}\left\{\gamma_k w_j(a_k) \mid j = 1, \ldots, m\right\}, \quad k = 1, \ldots, \ell
\end{aligned}$$

provides stochastic subgradients of K at (θ_0, λ_0). We note that sample variance must use Bessel's correction, i.e., divide by $m - 1$, for it to be an unbiased estimate.

5.1.1 Main framework

As in Section 3, determining how to update $\tilde{F}_n$ fully specifies the adaptive importance sampling method and its performance, and we will do so by making following 3 choices:

- Choose a family of distributions.
- Choose an objective to minimize.
- Choose a stochastic optimization method to perform the minimization.

As mentioned, if we choose a family $\mathcal{F}$ with a log-concave parameterization, $K(\theta, \lambda)$ is convex in θ and concave in λ. Using the parameterization, the algorithm will be of the form:

$$\begin{aligned}
X_n &\sim F_{\theta_n} \\
\hat{I}_n &= \frac{1}{n}\sum_{i=1}^{n}\phi(X_i; a)\frac{dF(a)}{dF_{\theta_i}}(X_i) \\
(\theta_{n+1}, \lambda_{n+1}) &= \text{Stochastic optimization with } X_1, \ldots, X_n, \theta_1, \ldots, \theta_n,
\end{aligned}$$

where the stochastic optimization method seeks a saddle point of K.

5.2 Stochastic optimization for convex-concave saddle functions

The stochastic optimization method *stochastic saddle point subgradient descent* solves

$$\min_{\theta\in\Theta}\max_{\lambda\in\Delta_\ell} K(\theta,\lambda)$$

with

$$\begin{aligned}\theta_{n+1} &= \Pi_\Theta(\theta_n - (C_1/\sqrt{n})g_n) \\ \lambda_{n+1} &= \Pi_{\Delta_\ell}(\lambda_n + (C_2/\sqrt{n})h_n),\end{aligned} \tag{5.1}$$

where $C_1$and C_2 are positive constants, Π_Θ and Π_{Δ_ℓ} are the projections onto Θ and Δ_ℓ, respectively, and g_n and h_n are subgradients of K at (θ_n, λ_n).

Convergence proof.

Lemma 10. *Assume K has a subgradient for all $\theta \in \Theta$ and $\lambda \in \Delta_\ell$. Assume $\mathbb{E}[\|g_n\|_2^2|\theta_n] \le G_1^2$ and $\mathbb{E}[\|h_n\|_2^2|\theta_n] \le G_2$. Also assume Θ is nonempty compact convex. Then* (5.1) *converges with rate*

$$\max_{\lambda\in\Delta_\ell}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}K(\theta_i,\lambda_i) \le K_\star + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Write D_1 and D_2 for the diameters of Θ and Δ_ℓ, respectively. As discussed in Section 2.2.1 that K has a subgradient on all of $\Theta\times\Delta_\ell$ implies $-\infty < K(\theta,\lambda) < \infty$ on $\Theta\times\Delta_\ell$.

By convexity and concavity of K, we have

$$\begin{aligned}K(\theta_i,\lambda_i) - K(\theta,\lambda_i) &\le g_i^T(\theta_i-\theta) \\ -K(\theta_i,\lambda_i) + K(\theta_i,\lambda) &\le h_i^T(\lambda-\lambda_i).\end{aligned}$$

Adding these two inequalities we get

$$K(\theta_i, \lambda) - K(\theta, \lambda_i) \leq g_i^T(\theta_i - \theta) - h_i^T(\lambda_i - \lambda).$$

We sum and take the expectation to get

$$\sum_{i=1}^{n} \mathbb{E}K(\theta_i, \lambda) - \sum_{i=1}^{n} \mathbb{E}K(\theta, \lambda_i) \leq \sum_{i=1}^{n} \mathbb{E}g_i^T(\lambda - \lambda_i) - \sum_{i=1}^{n} \mathbb{E}h_i^T(\theta_i - \theta).$$

Using the nonexpansivity of the projection (c.f. Lemma 13 of the appendix), we have

$$\frac{1}{2}\|\theta_{i+1} - \theta\|_2^2 \leq \frac{1}{2}\|\theta_i - \theta\|_2^2 - \frac{C_1}{\sqrt{i}} g_i^T(\theta_i - \theta) + \frac{C_1^2}{2i}\|g_i\|_2^2.$$

Take the full expectation and reorganize to get

$$\mathbb{E}g_i^T(\theta_i - \theta) \leq \frac{\sqrt{i}}{2C_1}\mathbb{E}\|\theta_i - \theta\|_2^2 - \frac{\sqrt{i}}{2C_1}\mathbb{E}\|\theta_{i+1} - \theta\|_2^2 + \frac{C_1}{2\sqrt{i}}G_1^2.$$

With the same "almost telescoping" series argument as before, we get

$$\sum_{i=1}^{n} \mathbb{E}g_i^T(\theta_i - \theta) \leq \frac{D_1^2}{2C_1}\sqrt{n} + C_1 G_1^2 \sqrt{n}.$$

Repeating the same argument with λ, we get

$$-\sum_{i=1}^{n} \mathbb{E}h_i^T(\theta_i - \theta) \leq \frac{D_2^2}{2C_2}\sqrt{n} + C_2 G_2^2 \sqrt{n}.$$

We put these inequalities together to get

$$\sum_{i=1}^{n} \mathbb{E}K(\theta_i, \lambda) - \sum_{i=1}^{n} \mathbb{E}K(\theta, \lambda_i) \leq \left(\frac{D_1^2}{2C_1} + \frac{D_2^2}{2C_2} + C_1 G_1^2 + C_2 G_2^2 \right) \sqrt{n}.$$

Since $K(\theta, \lambda)$ is linear in λ, we have

$$\min_{\theta\in\Theta} \sum_{k=1}^{\ell} \mathbb{E}K(\theta, \lambda_k) = n \min_{\theta\in\Theta} K\left(\theta, \mathbb{E}\frac{1}{n}\sum_{k=1}^{\ell} \lambda_k\right) \leq nK_\star,$$

we we conclude

$$\begin{aligned}\max_{\lambda\in\Delta_\ell} \sum_{i=1}^{n} (\mathbb{E}K(\theta_i, \lambda) - K^\star) &\leq \max_{\lambda\in\Delta_\ell} \sum_{i=1}^{n} \mathbb{E}K(\theta_i, \lambda) - \min_{\theta\in\Theta} \sum_{i=1}^{n} \mathbb{E}K(\theta, \lambda_i) \\ &\leq \left(\frac{D_1^2}{2C_1} + \frac{D_2^2}{2C_2} + C_1G_1^2 + C_2G_2^2\right)\sqrt{n}.\end{aligned}$$

□

5.3 Examples

5.3.1 Stochastic saddle point subgradient descent with exponential family

In this section, we consider the algorithm we get when we choose an exponential family for the family $\mathcal{F}$ (as defined in Section 2.3), the maximum per-sample variance for the objective to minimize, and stochastic saddle point subgradient descent for the stochastic optimization algorithm.

Assume Θ is nonempty convex compact and

$$\Theta \subset \mathbf{int}\left\{\theta \;\middle|\; \int \left(\phi(x; a)\frac{dF(a)}{dF_\theta}\right)^4 \, dF_\theta < \infty, \text{ for all } a \in \mathcal{A}\right\},$$

So Θ is in the interior of the set for which the estimators have finite 4th moments for all $a \in \mathcal{A}$.

The adaptive importance sampling method with these choices is

$$
\begin{aligned}
&X_{n1}, X_{n2}, \ldots, X_{nm} \sim F_{\theta_n} \\
&w_{nj}(a) = \phi(X_{nj}; a)\frac{dF(a)}{dF_{\theta_n}}(X_{nj}), \quad j = 1, \ldots, m \\
&\hat{I}_n(a) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{nj}(a) \\
&g_n = \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{k=1}^{\ell}\lambda_n(k)w_{nj}^2(a_k)\right)(\nabla A(\theta) - T(X_{nj})) \\
&h_n(k) = \text{Sample variance of}\left\{w_{nj}(a_k) \mid j = 1, \ldots, m\right\}, \quad k = 1, \ldots, \ell \\
&\theta_{n+1} = \Pi_\Theta(\theta_n - (C_1/\sqrt{n})g_n) \\
&\lambda_{n+1} = \Pi_{\Delta_\ell}(\lambda_n + (C_2/\sqrt{n})h_n)
\end{aligned}
$$

The estimator is unbiased, i.e., $\mathbb{E}\hat{I}_n(a) = I(a)$, and has maximum variance

$$\frac{1}{nm}K_\star \leq \max_{a\in\mathcal{A}} \mathbf{Var}\hat{I}_n(a) \leq \frac{1}{nm}K_\star + \mathcal{O}\left(\frac{1}{n^{3/2}m}\right).$$

We note that batching is necessary as the sample variance is undefined when $m = 1$.

Discussion of assumptions. The assumptions do in fact imply the assumptions necessary to ensure stochastic saddle point subgradient descent converges. The justification is essentially the same as that of Section 3.4.1, so we omit it.

5.3.2 Stochastic saddle point mirror descent

Just as stochastic subgradient descent generalizes to stochastic mirror descent, stochastic saddle point subgradient descent generalizes to stochastic saddle point mirror descent. Instead of explaining the generalization, we merely point out that replacing the update

$$\lambda_{n+1} = \Pi_{\Delta_\ell}(\lambda_n + (C_2/\sqrt{n})h_n)$$

with

$$\lambda^*_{n+1} = \lambda^*_n + (C_2/\sqrt{n})h_n$$
$$\lambda_{n+1} \propto \exp(\lambda^*_{n+1})$$

yields an instance of stochastic saddle point mirror descent.

5.3.3 Order statistic

Let $X \in \mathbf{R}^k$ be IID standard normals. We sort the entries of X to get its order statistic

$$X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(k)}.$$

We define $\phi(X; j) = X^{(j)}$ for $j = 1, \ldots, k$. The goal is to compute

$$I(j) = \mathbb{E}\phi(X; j),$$

i.e., we wish to compute the means of the order statistic.

We consider the adaptive importance sampling algorithm we get when we choose multivariate Gaussians with zero mean for the family $\mathcal{F}$, the maximum per-sample variance, for the objective to minimize, and stochastic saddle point mirror descent for the stochastic optimization algorithm.

The algorithm with batch size m is

$$
\begin{aligned}
&L_n L_n^T = \Sigma_n \quad \text{(Cholesky factorization)} \\
&Y_{n1}, Y_{n2}, \ldots, Y_{nm} \sim \mathcal{N}(0, I) \\
&X_{nj} = L_n^T Y_{nj}, \quad j = 1, \ldots, m \\
&w_{nj}(a) = \phi(X_{nj}; a) \frac{dF(a)}{dF_{\theta_n}}(X_{nj}), \quad j = 1, \ldots, m \\
&\hat{I}_n(a) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{nj}(a) \\
&g_n = \frac{1}{m} \sum_{j=1}^{m} \left(\sum_{k=1}^{\ell} \lambda_n(k) w_{nj}^2(a_k) \right) (X_{nj} X_{nj}^T - \Sigma_n) \\
&h_n(k) = \text{Sample variance of} \left\{ w_{nj}(a_k) \mid j = 1, \ldots, m \right\}, \quad k = 1, \ldots, \ell \\
&S_{n+1}^* = S_n^* - (C_1/\sqrt{n}) g_n \\
&S_{n+1} = \exp S_{n+1}^* \\
&\lambda_{n+1}^* = \lambda_n^* + (C_2/\sqrt{n}) h_n \\
&\lambda_{n+1} \propto \exp(\lambda_{n+1}^*) \\
&\Sigma_{n+1} = S_{n+1}^{-1}.
\end{aligned}
$$

As a comparison, we also consider the what-if simulation via plain Monte Carlo. We run this simulation with $n = 10^3$, $m = 10^3$, $k = 10$, $C_1 = 0.3$, and $C_2 = 3$. We get

$$
\begin{aligned}
&(\mathbb{E}X^{(1)}, \mathbb{E}X^{(2)}, \ldots, \mathbb{E}X^{(10)}) \\
&\approx (-1.539, -1.002, -0.657, -0.376, -0.123, 0.122, 0.375, 0.655, 1.001, 1.538).
\end{aligned}
$$

Figure 5.1 and 5.2 shows the performance of the two methods. We can see that although the adaptive method has better maximum per-sample variance, but worse performance when measured computation time.

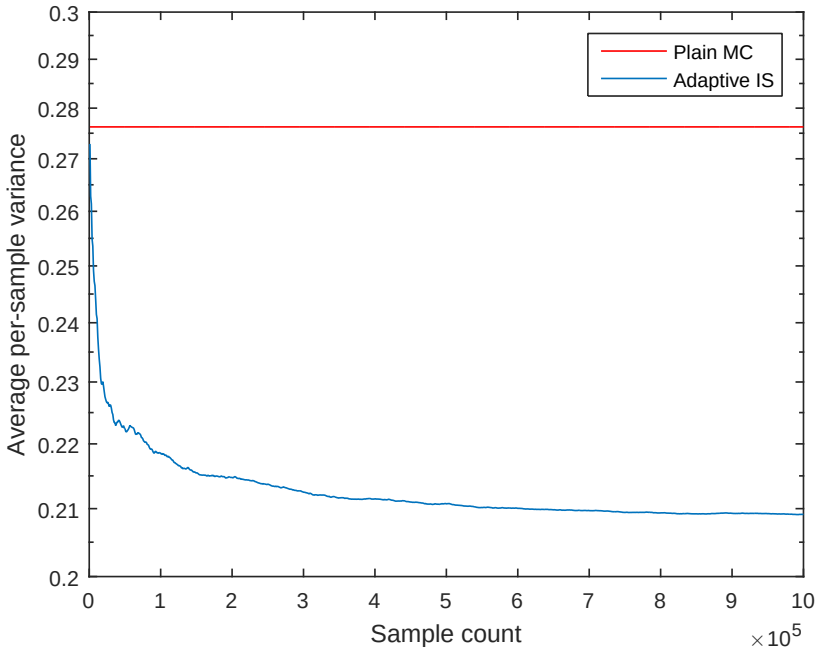


Figure 5.1: Maximum per-sample variance for the order statistic problem.

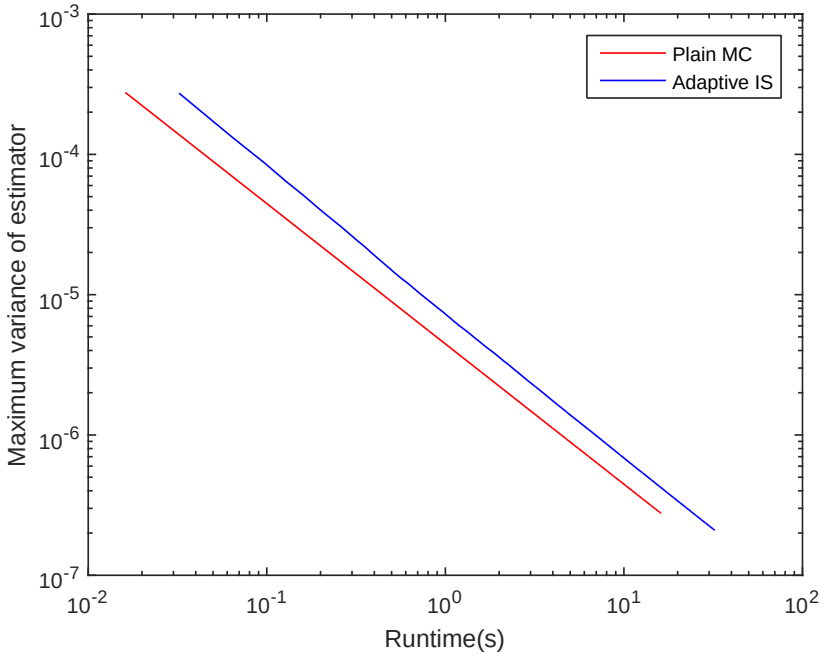


Figure 5.2: Maximum variance of the estimator for the order statistic problem.

5.3.4 Weighted maximum variance

So far, we considered the maximum per-sample variance for the objective to minimize. However, we can infact generalize this approach to consider the *weighted maximum variance*

$$\max_{k=1,\ldots,\ell} \gamma_k^2 \mathbf{Var} \hat{I}_n(a_k)$$

with $\gamma \in \mathbf{R}^k$, for the objective to minimize. If we can choose $\gamma_k \approx 1/I(a_k)$ then the objective becomes the approximate maximum coefficient of variation.

Without going into detail, we simply point out that the stochastic subgradients for the weighted per-sample variance becomes

$$\begin{aligned}
w_{nj}(a) &= \phi(X_{nj}; a) \frac{dF(a)}{dF_{\theta_n}}(X_{nj}), \quad j = 1, \ldots, m \\
g_n &= \frac{1}{m} \sum_{j=1}^{m} \sum_{k=1}^{\ell} \lambda_n(k) \gamma_k^2 w_{nj}(a_k)^2 \left(-\frac{1}{f_{\theta_n}(X_{nj})} \nabla_\theta f_{\theta_n}(X_{nj}) \right) \\
h_n(k) &= \text{Sample variance of} \left\{ \gamma_k w_{nj}(a_k) \mid j = 1, \ldots, m \right\}, \quad k = 1, \ldots, \ell.
\end{aligned}$$

Chapter 6

Other topics

6.1 When adaptive importance sampling fails

Importance sampling fails in a practical sense when the per-sample variance of $\tilde{F}$ is too large. Unfortunately, the adaptive importance sampling method presented in this work suffers from the same problem.

Consider the problem of estimating

$$I = \mathbb{E}\phi(X) = \mathbf{P}(X \geq 10^{10})$$

where X is an exponential random variable with mean 1 and $\phi(X) = \hat{I}_{\{X \geq 10^{10}\}}$. For our family with log-concave parameterization, we choose Gaussians with mean θ and $\Theta = [-10^{-100}, 10^{100}]$. The optimal parameter $\theta_\star$ will provide a decent per-sample variance. Moreover, one can verify that all the technical assumptions are met and thereby conclude that the adaptive importance sampling method (minimizing the per-sample variance with stochastic subgradient descent) will have the optimal asymptotic variance.

However, if we were to use the starting point $\theta_1 = 0$, the method will fail miserably. On average, the algorithm will make no progress for the first $e^{10^{10}}$ iterations, which of course is way too many. So even though the theory guarantees asymptotic optimality and even though we know the asymptotic optimum is good, the method is useless

under any practical criterion.

So in a sense, the adaptive importance sampling method inherits the same difficulties from importance sampling. If the starting point θ_1 is decent adaptive importance sampling will take you to the optimal parameter $\theta_\star$, but if the starting point is bad adaptive importance sampling does not offer much help.

We can make this statement more precise in the case of using exponential families for $\mathcal{F}$. In the convergence proofs, we saw that the constants of the higher order terms are related to $D_4(F_\star\|F_\theta)$. So the constants are reasonably small when $D_4(F_\star\|F_\theta)$ is small. However, when θ_1 is bad, meaning the per-sample variance $D_2(F_\star\|F_{\theta_1})$ is large, $D_4(F_\star\|F_{\theta_1}) \geq D_2(F_\star\|F_{\theta_1})$ is also large. So a starting point with reasonable per-sample variance is necessary for the adaptive method to make progress in a reasonable number of iterations.

In conclusion, we should view adaptive importance sampling as an approach to improve importance sampling, not an approach to fix importance sampling. In a setting where one cannot make importance sampling work, adaptive importance sampling will not work either.

6.2 Non-uniform weights

So far we have only considered uniform weights for our estimator $\hat{I}_n$, but this is not necessary. So instead of using

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{dF}{dF_{\theta_i}}(X_i),$$

we could use

$$\hat{I}_n = \left(\sum_{i=1}^{n} \eta_i\phi(X_i)\frac{dF}{dF_{\theta_i}}(X_i)\right) / \left(\sum_{i=1}^{n}\eta_i\right)$$

for a positive sequence of *weights* $\eta_1, \eta_2, \ldots$. It is easy to see $\hat{I}_n$ is unbiased. We can do likewise in self-normalized importance sampling and what-if simulations as well.

For why non-uniform weights might be better, one can argue that the initial iterates should have smaller weights as they have larger variance. Another argument

comes from [58, Theorem 4]. To analyze the performance of $\hat{I}_n$ with non-uniform weights, we would need a convergence rate on

$$\left(\sum_{i=1}^{n} \eta_i \mathbb{E}U(\theta_i)\right) / \left(\sum_{i=1}^{n} \eta_i\right) \to U(\theta_\star)$$

instead of (2.3). This can be done by modifying the convergence proof of Section 2.2.1.

6.3 Confidence intervals

It is well known that the sample variance of n IID random variables provide an unbiased estimate of their variance when Bessel's correction (dividing by $n-1$) is used. We can do the same to obtain an unbiased estimate of $\mathbf{Var}I_n$ in our adaptive importance sampling method.

For notational simplicity, write

$$Z_n = \phi(X_n)\frac{dF}{dF_{\theta_n}}(X_n), \quad \overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

Then

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} Z_n$$

and $\mathbb{E}Z_i = \mathbb{E}\overline{Z}_n = I$ for $i = 1, \ldots, n$.

We adapt the standard argument to our setup and use the conditional dependency argument of Section 3.2 to get

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} Z_i^2 - 2Z_i\overline{Z}_n + \overline{Z}_n^2\right] &= \mathbb{E}\sum_{i=1}^{n} Z_i^2 - n\mathbb{E}\overline{Z}_n^2 \\
&= \sum_{i=1}^{n} \mathbb{E}\mathbf{Var}_{F_{\theta_i}}\left(\phi(X_n)\frac{dF}{dF_{\theta_n}}(X_n)\right) + nI^2 \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbf{Var}_{F_{\theta_i}}\left(\phi(X_n)\frac{dF}{dF_{\theta_n}}(X_n)\right) - nI^2.
\end{aligned}$$

Finally, we reorganize to get

$$\begin{aligned}\mathbb{E}\frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i-\overline{Z}_n)^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbf{Var}_{F_{\theta_i}}\left(\phi(X_n)\frac{dF}{dF_{\theta_n}}(X_n)\right)\\ &= \mathbf{Var}\hat{I}_n.\end{aligned}$$

So we can establish a confidence interval around $\hat{I}_n$ with

$$\mathbf{Var}\hat{I}_n \approx \frac{1}{n}\,\text{Sample variance of}\left\{\phi(X_i)\frac{dF}{dF_{\theta_i}}(X_i)\middle| i=1,\ldots,n\right\}.$$

6.4 Optimal rates

Using step sizes $\alpha_n = C/\sqrt{n}$, we showed that stochastic subgradient descent converges with rate $\mathcal{O}(1/\sqrt{n})$. This translates to

$$\mathbf{Var}\hat{I}_n = \frac{1}{n}V_\star + \mathcal{O}\left(\frac{1}{n^{3/2}}\right).$$

As discussed, $V_\star/n$ is optimal with respect to the family $\mathcal{F}$. However, the higher-order term $1/n^{3/2}$ need not be optimal. Under certain assumptions, different step sizes can make stochastic subgradient descent to converge at a rate faster than $\mathcal{O}(1/\sqrt{n})$. This will translate to an improved higher-order term for $\mathbf{Var}\hat{I}_n$.

This discussion, however, goes beyond the scope of this work. The body of research investigating what assumptions and what step sizes yield what rates is vast and nuanced [13, 6]. In this work, we only considered step sizes $\alpha_n = C/\sqrt{n}$, because this choice is simple and is known to be robust, both in a theoretical and empirical sense [53].

Chapter 7

Conclusion

In this work, we presented a framework for performing adaptive importance sampling with stochastic convex optimization. When the family of sampling distributions in consideration has a log-concave parameterization, the objectives of interest become convex functions of the parameters. This allows us to use stochastic convex optimization methods that are computationally simple, and this allows us to use tools from convex analysis to analyze the performance of the methods. The resulting method performs importance sampling and stochastic optimization simultaneously, and has optimal asymptotic performance with respect to the family of sampling distributions in consideration. We also showed how this approach can be extended to self-normalized importance sampling and what-if simulations. For what-if simulations the method asymptotically achieves the minimum maximum variance, which, to the best of our knowledge, is a novel approach.

The proposed method, however, does have some practical drawbacks. First, the proposed adaptive importance sampling method will not be effective in setups where non-adaptive importance sampling is infeasible, even if the theory states asymptotic optimality. Also, the variance reduction may not be large enough to justify the additional cost of the adaptive update, compared to non-adaptive importance sampling. Finally, tuning the optimization parameters, an issue we did not explore, does add to cost in terms of human effort and computation time.

Nevertheless, there are setups where this method does bring a practical improvement, and the framework's generality allows a broad applicability. Furthermore, we believe this work is interesting from a theoretical standpoint.

In a borad sense, that adaptive importance sampling is an optimization problem that can benefit from convexity is a useful viewpoint that lead to this work. We are hopeful that future work in Monte Carlo simulations can also benefit from this approach.

Chapter 8

Appendix

8.1 Stochastic subgradients

Let $X(\omega)$ be a random variable under a probability space $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ is a σ-algebra of Ω, and P is a probability measure on $\mathcal{F}$. We say $f(\theta; X)$ is a random function of θ if $f(\theta; X)$ is a measurable function of X for all fixed θ.

Assume $f(\theta; x)$ is convex in θ on Θ for P-almost all x. Then $\mathbb{E}f(\theta; X)$ is a convex function of θ because

$$\mathbb{E}f(\eta\theta_1 + (1-\eta)\theta_2; X) \leq \eta\mathbb{E}f(\theta_1; X) + (1-\eta)\mathbb{E}f(\theta_2; X).$$

for any $\eta \in [0, 1]$ and $\theta_1, \theta_2 \in \Theta$. (We assume the expectation is never $-\infty$.)

If $f(\theta; x)$ is differentiable in θ for P-almost all x, then $\nabla f(\theta; x)$ is $\sigma(X)$-measurable and $\mathbb{E}\nabla_\theta f(\theta; X)$ is well-defined if $\mathbb{E}\|\nabla_\theta f(\theta; X)\|_1 < \infty$. When $f(\theta; x)$ is not necessarily differentiable in θ, one might be tempted to consider

$$\mathbb{E}\left[\partial_\theta f(\theta; X)\right]$$

in the discussion of stochastic subgradients. While making sense of this expectation of random sets is possible, we avoid this complication [9].

Rather, consider a $\mathcal{F}$-measurable random variable g such that

$$g(\omega) \in \partial_\theta f(\theta_0; X(\omega)).$$

We say g is a *measurable selection* of $\partial_\theta f(\theta_0; X)$. By definition $\mathbb{E}g(X)$ is well-defined provided that $\mathbb{E}\|g(X)\|_1 < \infty$. Furthermore,

$$\mathbb{E}g \in \partial \mathbb{E} f(\theta_0; X)$$

because

$$\begin{aligned} f(\theta; X) &\geq f(\theta_0; X) + g^T(\theta - \theta_0) \\ \mathbb{E}f(\theta; X) &\geq \mathbb{E}f(\theta_0; X) + \mathbb{E}g^T(\theta - \theta_0). \end{aligned}$$

This is convenient because the assertion

$$\nabla \mathbb{E} f(\theta; X) = \mathbb{E} \nabla f(\theta; X)$$

is, technically speaking, not always true. However, we can, in a sense, swap the order of ∂ and $\mathbb{E}$.

Throughout this work, the requirement that a stochastic subgradient should be measurable, i.e., that the selection $g \in \partial f(\theta; X)$ be measurable, is omitted. We consider this technical detail to be implied, since any real-world implementation of a stochastic subgradient will be measurable.

Lemma 11. *Assume $dF_\theta = f_\theta d\mu$ where $f_\theta(x)$ is log-concave in θ on Θ for μ-almost all x. Let $X \sim F_{\theta_0}$, and $h \in \partial(-\log f_{\theta_0}(X))$. Then*

$$g = \left(\frac{dF}{dF_{\theta_0}}(X) \right)^2 h$$

is a stochastic subgradient of the convex function

$$\int \frac{dF}{dF_\theta} \, dF$$

at θ_0.

Proof. We first show a result similar to the chain rule for derivatives. Let $U(\theta)$ be a convex function on Θ, and let $g \in \partial U(\theta_0)$. Then by definition of subgradients, we have

$$U(\theta) \geq U(\theta_0) + g^T(\theta - \theta_0)$$

for all $\theta \in \Theta$. Then

$$\begin{aligned}\exp U(\theta) &\geq \exp(U(\theta_0) + g^T(\theta - \theta_0)) \\ &\geq \exp U(\theta_0) + \exp U(\theta_0) g^T(\theta - \theta_0).\end{aligned}$$

The first and inequality follows from monotoncity and convexity of exp, respectively. (From this one could conclude $\exp(h(\theta))\partial h(\theta) \subseteq \partial \exp(h(\theta))$, but we just need the inequality.)

Now we have

$$\begin{aligned}\exp(-\log f_\theta)h &\in \partial \exp(-\log f_\theta) \\ \frac{1}{f_\theta}h &\in \partial \frac{1}{f_\theta} \\ \frac{dF}{d\mu}\frac{1}{f_\theta}h &\in \partial \frac{dF}{d\mu}\frac{1}{f_\theta} \\ \mathbb{E}_F\left[\frac{dF}{dF_\theta}h\right] &\in \partial \mathbb{E}_F\left[\frac{dF}{dF_\theta}\right]\end{aligned}$$

□

When differentiable, the random variable

$$-h(X) = \nabla_\theta \log f_\theta(X)$$

with $X \sim F_\theta$ is called the *score function* [22, 47]. Score functions are known to have zero mean under certain regularity conditions. Here we repeat the standard

argument:

$$\mathbb{E}_{F_\theta} - h(X) = \mathbb{E}_{F_\theta} \nabla_\theta (\log f_\theta(X)) = \int \frac{1}{f_\theta} (\nabla_\theta f_\theta) f_\theta \, d\mu$$
$$= \int \nabla_\theta f_\theta \, d\mu = \nabla_\theta \int f_\theta \, d\mu = \nabla 1 = 0.$$

Although this fact is not directly related to this work, it is often useful as a debugging tool.

Let us see Lemma 11 applied to the 3 families with log-concave parameterizations. With an exponential family, we get

$$g = \left(\frac{dF}{dF_\theta} \right)^2 (\nabla A(\theta) - T(x)).$$

With a mixture, we get

$$g = - \left(\frac{dF}{dF_\theta} \right)^2 \frac{1}{f_\theta} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_p \end{pmatrix}.$$

With affine transformations of a continuous log-concave variable, we get

$$\begin{aligned} Y &\sim p(x)dx \\ X &= A^{-1}(Y - b) \\ h &= \nabla(-\log p)(Y) \\ g_A &= \frac{1}{2} \left(\frac{dF(X)}{p(Y) \det(A) dx} \right)^2 \left(hX^T + Xh^T - 2A^{-1} \right) \\ g_b &= \left(\frac{dF(X)}{p(Y) \det(A) dx} \right)^2 h, \end{aligned}$$

where we use Lemma 12.

Lemma 12. *Let $Y \sim p(x)dx$. Then the convex function $-\log p$ is almost surely differentiable at Y, and we can write $\partial(-\log p)(Y) = \nabla(-\log p)(Y)$.*

Proof. A convex function on a finite dimensional Euclidean space is differentiable almost everywhere (with respect to the Lebesgue measure) on its domain [63, Theorem 25.5].

Since $p(x)dx \ll dx$ and since $-\log p(Y) < \infty$ (i.e., Y is in the domain of $-\log p$) almost surely, $-\log p$ is differentiable almost surely. □

8.2 Projection

The projection of x onto a nonempty closed convex set C is defined as

$$\Pi_C(x) = \operatorname*{argmin}_{z \in C} \|z - x\|_2,$$

which always exists and is unique. We can interpret Π_C as the point in C closest to x. In general, the projection onto any nonempty closed convex set is a convex optimization problem, and as such could be solved with a convex optimization solver. Usually, however, projection is a useful subroutine only when it has a analytical or semi-analytical solution.

Lemmas 13 and 14 are well-known, but we prove them anyways for the sake of completeness.

Lemma 13. *If C is a nonempty closed convex set, Π_C is nonexpansive.*

Proof. Reorganizing the optimality condition for the optimization problem [15, p. 139], we get that for any $u \in \mathbf{R}^n$ and $v \in C$ we have

$$(v - \Pi_C u)^T (\Pi_C u - u) \geq 0. \tag{8.1}$$

Now for any x, $y \in \mathbf{R}^n$, we get

$$(\Pi_C y - \Pi_C x)^T (\Pi_C x - x) \geq 0$$
$$(\Pi_C x - \Pi_C y)^T (\Pi_C y - y) \geq 0$$

using (8.1), and by adding these two we get

$$(\Pi_C x - \Pi_C y)^T (x - y) \geq \|\Pi_C x - \Pi_C y\|_2^2.$$

Finally, we apply the Cauchy-Schwartz inequality to conclude

$$\|\Pi_C x - \Pi_C y\|_2 \leq \|x - y\|_2.$$

□

Lemma 14. *Let*

$$\Delta_k = \{x \in \mathbf{R}^k \mid x_1, \ldots, x_k \geq 0,\ x_1 + \cdots + x_k = 1\},$$

define the function $(\cdot)_+ : \mathbf{R}^k \to \mathbf{R}^k$ *as* $((z)_+)_i = \max\{z_i, 0\}$ *for* $i = 1, \ldots, k$. *Then*

$$\Pi(z) = (z - \nu \mathbf{1})_+,$$

where ν *is a solution of the equation*

$$\mathbf{1}^T (z - \nu \mathbf{1})_{+,\varepsilon} = 1.$$

The left-hand side is a nonincreasing function of ν, *so* ν *can be obtained efficiently via bisection on with the starting interval* $[\max_i z_i - 1, \max_i z_i]$.

Proof. By definition of the projection, $x = \Pi(z)$ is the solution of

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|x - z\|_2^2 \\ \text{subject to} & \mathbf{1}^T x = 1 \\ & x \geq 0. \end{array}$$

This problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|x - z\|_2^2 + \nu(\mathbf{1}^T x - 1) \\ \text{subject to} & x \geq 0 \end{array}$$

for an optimal dual variable $\nu \in \mathbf{R}$. This follows from dualizing with respect to the constraint $\mathbf{1}^T x = 1$ (and not the others), applying strong Lagrange duality, and using fact that the objective is strictly convex [10]. This problem is in turn equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|x - (z - \nu\mathbf{1})\|_2^2 \\ \text{subject to} & x \geq 0, \end{array}$$

which has the analytic solution

$$x^\star = (z - \nu\mathbf{1})_+.$$

An optimal dual variable ν must satisfy

$$\mathbf{1}^T(z - \nu\mathbf{1})_+ = 1,$$

by the KKT conditions. Write $h(\nu) = (z - \nu\mathbf{1})_+$. Then $h(\nu)$ a continuous nonincreasing function with

$$h(\max_i z_i - 1) \geq 1, \quad h(\max_i z_i) = 0.$$

So a solution of $h(\nu) = 1$ is in the interval $[\max_i z_i - 1, \max_i z_i]$.

□

Bibliography

[1] W. A. Al-Qaq, M. Devetsikiotis, and J. K. Townsend. Stochastic gradient optimization of importance sampling for the efficient simulation of digital communication systems. *IEEE Transactions on Communications*, 43(12):2975–2985, 1995.

[2] B. Arouna. Robbins-Monro algorithms and variance reduction in finance. *The Journal of Computational Finance*, 7(2):35–61, 2003.

[3] B. Arouna. Adaptive Monte Carlo method, a variance reduction technique. *Monte Carlo Methods and Applications*, 10(1):1–24, 2004.

[4] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-Linear Programming*. 1972.

[5] E. Arthurs and H. Dym. On the optimum detection of digital signals in the presence of white Gaussian noise—a geometric interpretation and a study of three basic data transmission systems. *IRE Transactions on Communications Systems*, 10(4):336–372, 1962.

[6] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[7] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[8] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics, 2001.

[9] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

[10] D. P. Bertsekas. *Convex Optimization Theory*. 2009.

[11] D. P. Bertsekas. *Convex Optimization Algorithms*. 2015.

[12] P. Billingsley. *Probability and Measure*. Wiley, third edition, 1995.

[13] L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, 1998.

[14] L. Bottou. Stochastic gradient tricks. In G. Montavon, G. B. Orr, and K.-R. Müller, editors, *Neural Networks, Tricks of the Trade, Reloaded*, pages 430–445. Springer, 2012.

[15] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[16] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[17] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:1–279, 1986.

[18] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

[19] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

[20] M. A. Cauchy. Méthode générale pour la résolution des systémes d'équations simultanées. *Comptes Rendus Hebdomadaires des Séances de l'Acadmie des Sciences*, 25:536–538, 1847.

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

[22] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. 1974.

[23] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[24] P.-T. de Boer, D. P. Kroese, and R. Y. Rubinstein. A fast cross-entropy method for estimating buffer overflows in queueing networks. *Management Science*, 50(7):883–895, 2004.

[25] M. Devetsikiotis and J. K. Townsend. An algorithmic approach to the optimization of importance sampling parameters in digital communication system simulation. *IEEE Transactions on Communications*, 41(10):1464–1473, 1993.

[26] M. Devetsikiotis and J. K. Townsend. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM Transactions on Networking*, 1(3):293–305, 1993.

[27] R. Douc, R. Guillin, J.-M. Marin, and C. P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 35(1):420–448, 2007.

[28] D. Egloff and M. Leippold. Quantile estimation with adaptive importance sampling. *The Annals of Statistics*, 38(2):1244–1278, 2010.

[29] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. 2nd edition, 1966.

[30] W. Fenchel. On conjugate convex functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949.

[31] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. *CoRR*, 2014.

[32] R. M. Fung and K.-C. Chang. Weighing and integrating evidence for stochastic simulation in bayesian networks. In *Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*, pages 209–220, 1990.

[33] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Mathematical Finance*, 9(2):117–152, 1999.

[34] G. H. Golub and C. F. Van Loan. *Matrix Computations*. 4th edition, 2013.

[35] S. Haykin. *Communication Systems*. Wiley Publishing, fifth edition, 2009.

[36] H. Y. He and A. B. Owen. Optimal mixture weights in multiple importance sampling. 2014.

[37] T. C. Hesterberg. *Advances in Importance Sampling*. PhD thesis, Stanford University, 1988.

[38] T. C. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[39] R. A. Horn and C. R. Johnson, editors. *Matrix Analysis*. 1986.

[40] H. Kahn. Random sampling (Monte Carlo) techniques in neutron attenuation problems, i. *Nucleonics*, 6(5):27–33, 1950.

[41] H. Kahn. Random sampling (Monte Carlo) techniques in neutron attenuation problems, ii. *Nucleonics*, 6(6):60–65, 1950.

[42] H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

[43] R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

[44] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[45] A. Lapidoth. *A Foundation in Digital Communication*. Cambridge University Press, 2009.

[46] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, second edition, 1998.

[47] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, third edition, 2005.

[48] M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. *Proceedings of ACM SIGKDD*, pages 661–670, 2014.

[49] D. Lieber, R. Y. Rubinstein, and D. Elmakis. Quick estimation of rare events in stochastic networks. *IEEE Transactions on Reliability*, 46(2):254–265, 1997.

[50] D. L. McLeish. Bounded relative error importance sampling and rare event simulation. *ASTIN Bulletin*, 40(1), 2010.

[51] N. Metropolis. The beginning of the Monte Carlo method. *Annals of Operations Research*, 15:125–130, 1987.

[52] N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[53] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[54] A. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[55] A. S. Nemirovski and D. B. Yudin. On Cesari's convergence of the steepest descent method for approximating saddle points of convex-concave functions. *Doklady Akademii Nauk SSSR*, 239:1056–1059, 1978.

[56] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[57] M.-S. Oh and J. O. Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41:143–168, 1992.

[58] A. B. Owen and Y. Zhou. Adaptive importance sampling by mixtures of products of beta distributions. Technical report, Stanford University, 1999.

[59] T. Pennanen and M. Koivu. An adaptive importance sampling technique. In H. Niederreiter and D. Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 443–455. Springer, 2006.

[60] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

[61] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

[62] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[63] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[64] R. Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operations Research*, 99:89–112, 1997.

[65] R. Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1(2):127–190, 1999.

[66] R. D. Shachter and M. A. Peot. Simulation approaches to general probabilistic inference on belief networks. In *Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*, pages 221–234, 1990.

[67] A. Shapiro. Monte carlo sampling methods. In *Stochastic Programming*, pages 353–425. 2003.

[68] N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer, 1985.

[69] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[70] R. Srinivasan. *Importance Sampling: Applications in Communications and Detection*. Springer, 2002.

[71] H. K. van Dijk T. Kloek. Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica*, 46(1):1–19, 1978.

[72] H. F. Trotter and J. W. Tukey. Conditional Monte Carlo for normal samples. In *Symposium on Monte Carlo methods*, pages 64–79, 1956.

[73] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Keibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[74] W. H. Young. On classes of summable functions and their fourier series. *Proceedings of the Royal Society of London, Series A*, 87(594):225–229, 1912.