Adaptive Importance Sampling via Stochastic Convex Programming

Ernest K. Ryu[1] and Stephen P. Boyd[1]

[1]Institute for Computational and Mathematical Engineering, Stanford University

January 8, 2015

Abstract

We show that the variance of the Monte Carlo estimator that is importance sampled from an exponential family is a convex function of the natural parameter of the distribution. With this insight, we propose an adaptive importance sampling algorithm that simultaneously improves the choice of sampling distribution while accumulating a Monte Carlo estimate. Exploiting convexity, we prove that the method's unbiased estimator has variance that is asymptotically optimal over the exponential family.

1 Introduction

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

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

where $X \sim f$ is a random variable on $\mathbf{R}^k$ and $\phi : \mathbf{R}^k \to \mathbf{R}$.

The standard Monte Carlo method estimates I by taking independent identically distributed (IID) samples $X_1, X_2, \ldots, X_n \sim f$ and using

$$\hat{I}_n^{\mathrm{MC}} = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i).$$

This estimator is *unbiased*, *i.e.*, $\mathbf{E}\hat{I}_n^{\mathrm{MC}} = I$, and has variance

$$\mathbf{Var}(\hat{I}_n^{\mathrm{MC}}) = \frac{1}{n}\mathbf{Var}_{X\sim f}\left[\phi(X)\right] = \frac{1}{n}\left(\int \phi^2(x) f(x)\, dx - I^2\right).$$

To reduce the variance of the estimator, the standard figure of merit, one can use importance sampling: choose a *sampling* (*importance*) distribution $\tilde{f}$ satisfying $\tilde{f}(x) > 0$ whenever

$\phi(x)f(x) \neq 0$, take IID samples $X_1, X_2, \ldots, X_n \sim \tilde{f}$ (as opposed to sampling from f, the *nominal* distribution) and use

$$\hat{I}_n^{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{f(X_i)}{\tilde{f}(X_i)}.$$

Again, the estimator is unbaised, *i.e.*, $\mathbf{E}\hat{I}_n^{\text{IS}} = I$, and has variance

$$\mathbf{Var}(\hat{I}_n^{\text{IS}}) = \frac{1}{n}\mathbf{Var}_{X\sim\tilde{f}}\left[\frac{\phi(X)f(X)}{\tilde{f}(X)}\right] = \frac{1}{n}\left(\int \frac{\phi^2(x)f^2(x)}{\tilde{f}(x)}\,dx - I^2\right).$$

When $\tilde{f} = f$, importance sampling reduces to standard Monte Carlo. Choosing $\tilde{f}$ wisely can reduce the variance, but this can be difficult in general. One approach is to use a priori information on the integrand $\phi(x)f(x)$ to manually find an appropriate sampling distribution $\tilde{f}$, perhaps through several informal iterations [14, 20, 24, 30, 34, 36, 37]. Another approach is to automate the process of finding the sampling distribution through adaptive importance sampling.

In adaptive importance sampling, one adaptively improves the sampling distribution while simultaneously accumulating the estimate for I. A particular form of importance sampling generates a sequence of sampling distributions $\tilde{f}_1, \tilde{f}_2, \ldots$ and a series of samples $X_1 \sim \tilde{f}_1$, $X_2 \sim \tilde{f}_2$, $\ldots$ and forms the estimate

$$\hat{I}_n^{\text{AIS}} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{f(X_i)}{\tilde{f}_i(X_i)}.$$

At each iteration n, the sampling distribution $\tilde{f}_n$, which is itself random, is adaptively determined based on the past data, $\tilde{f}_n, \ldots, \tilde{f}_{n-1}$ and $X_1, \ldots, X_{n-1}$. Again, $\hat{I}_n^{\text{AIS}}$ is unbiased, *i.e.*, $\mathbf{E}\hat{I}_n^{\text{AIS}} = I$, and

$$\mathbf{Var}(\hat{I}_n^{\text{AIS}}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}_{\tilde{f}_i}\mathbf{Var}_{X_i\sim\tilde{f}_i}\left[\frac{\phi(X_i)f(X_i)}{\tilde{f}_i(X_i)}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}_{\tilde{f}_i}\left(\int \frac{\phi^2(x)f^2(x)}{\tilde{f}^2(x)}\,dx - I^2\right),$$

where $\mathbf{E}_{\tilde{f}_i}$ denotes the expectation over the random sampling distribution $\tilde{f}_i$. Again, when $\tilde{f}_i = \tilde{f}$ for all i, adaptive importance sampling reduces to standard (non-adaptive) importance sampling. Now determining how to choose $\tilde{f}_n$ at each iteration fully specifies the method.

In this paper, we propose an instance of adaptive importance sampling, which we call Convex Adaptive Monte Carlo (Convex AdaMC). First, we choose an exponential family of distributions $\mathcal{F}$ as the set of candidate sampling distributions. Define $T : \mathbf{R}^k \to \mathbf{R}^p$ and $h : \mathbf{R}^k \to \mathbf{R}_+$. Then our density function is

$$f_\theta(x) = \exp\left(\theta^T T(x) - A(\theta)\right) h(x),$$

where $A : \mathbf{R}^p \to \mathbf{R} \cup \{\infty\}$, defined as

$$A(\theta) = \log \int \exp(\theta^T T(x))h(x)\,dx,$$

serves as a normalizing factor. (When $A(\theta) = \infty$, we define $f_\theta = 0$ and remember that this does not define a distribution.) Finally, let $\Theta \subseteq \mathbf{R}^p$ be a convex set, and our exponential family is $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$, where θ is called the *natural parameter* of $\mathcal{F}$. Note that the choice of T, h, and Θ fully specifies our family $\mathcal{F}$.

Next, define $V : \mathbf{R}^p \to \mathbf{R} \cup \{\infty\}$ to be the per-sample variance of the importance sampled estimator with sampling distribution f_θ,

$$V(\theta) = \mathbf{Var}_{X \sim f_\theta}\left[\frac{\phi(X)f(X)}{f_\theta(X)}\right] = \int \frac{\phi^2(x)f^2(x)}{f_\theta(x)}\,dx - I^2.$$

(So the importance sampled estimator using n IID samples from f_θ has variance $V(\theta)/n$.) A natural approach is to solve

$$\begin{array}{ll} \text{minimize} & V(\theta) \\ \text{subject to} & \theta \in \Theta, \end{array} \tag{1}$$

where θ is the optimization variable, as this will give us the best sampling distribution among $\mathcal{F}$ to importance sample from. We write $V^\star$ to denote the optimal value, *i.e.*, the optimal per-sample variance over the family.

The first key insight of this paper is that V is a convex function, a consequence of $\mathcal{F}$ being an exponential family. Roughly speaking, one can efficiently find a *global* minimum of a convex functions through standard methods if one can compute the function value and its gradient [22]. This fact, however, is not directly applicable to our setting as evaluating $V(\theta)$ or $\nabla V(\theta)$ for any given θ is in general as hard as evaluating I itself.

The second key insight is that we can minimize the convex function V through a standard algorithm of stochastic optimization, *stochastic gradient descent*, while simuiltaneously accumulating an estimate for I. Because of convexity, we can prove theoretical guarantees.

In Convex AdaMC, we generate a sequence of sampling distribution parameters $\theta_1, \theta_2, \ldots$ and a series of samples $X_1 \sim f_{\theta_1}$, $X_2 \sim f_{\theta_2}, \ldots$, with which we form the estimate

$$\hat{I}_n^{\text{AMC}} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{f(X_i)}{f_{\theta_i}(X_i)}.$$

Again, $\hat{I}_n^{\text{AMC}}$ is unbiased, *i.e.*, $\mathbf{E}\hat{I}_n^{\text{AMC}} = I$. Furthermore, we show that

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

This shows that the per-sample variance of Convex AdaMC converges to the optimal per sample variance over our family $\mathcal{F}$; *i.e.*, our estimator Convex AdaMC asymptotically performs as well as any (adaptive or non-adaptive) importance sampling estimator using sampling distributions from $\mathcal{F}$. In particular, Convex AdaMC does not suffer from becoming trapped in (non-optimal) local minima, a problem other adaptive importance sampling methods can have.

2 Convexity of the variance

Let's establish a few important properties of our variance function

$$V(\theta) = \int \phi^2(x) f^2(x) \exp(A(\theta) - \theta^T T(x)) h(x)\, dx - I^2.$$

When $A(\theta) = \infty$, we define $V(\theta) = \infty$. Not only is this definition natural but is also convenient since $V(\theta) = \infty$ now indicates that θ is invalid either because the variance is infinite or because θ doesn't define a sampling distribution.

Recall that a function $V : \mathbf{R}^p \to \mathbf{R} \cup \{\infty\}$ is convex if

$$V(\eta\theta_1 + (1-\eta)\theta_2) \le \eta V(\theta_1) + (1-\eta) V(\theta_2)$$

holds for any $\eta \in [0,1]$ and $\theta_1, \theta_2 \in \Theta$. Convexity is important because it allows us to prove a theoretical guarantee.

Theorem 1. *The variance of the importance sampling estimator $V(\theta)$ is a convex function of θ, the natural parameter of the exponential family.*

Proof. We first show $A(\theta)$ is convex. By Hölder's inequality, we have

$$\begin{aligned}\exp(A(\eta\theta_1 + (1-\eta)\theta_2)) &= \int \exp((\eta\theta_1 + (1-\eta)\theta_2)^T T(x)) h(x)\, dx \\ &\le \left(\int \exp(\theta_1^T T(x)) h(x)\, dx\right)^{\eta} \left(\int \exp(\theta_2^T T(x)) h(x)\, dx\right)^{1-\eta},\end{aligned}$$

and by taking the log on both sides we get

$$A(\eta\theta_1 + (1-\eta)\theta_2) \le \eta A(\theta_1) + (1-\eta) A(\theta_2).$$

Since $\exp(\cdot)$ is an increasing convex function and $A(\theta) - \theta^T T(x)$ is convex in θ, the composition $\exp(A(\theta) - \theta^T T(x))$ is convex in θ. Finally, $V(\theta)$ is convex as it is an integral of the convex functions $\exp(A(\theta) - \theta^T T(x)) h(x)$ over x; see [16, §B.2], [31, §5], or [5, §3.2]. □

We note in passing that $\log V(\theta)$ is also a convex function of θ, which is a stronger statement than convexity of $V(\theta)$. This fact, however, is not useful for us since we do not have a simple way to obtain a stochastic gradient for $\log V(\theta)$, whereas, as we will see later, we do for $V(\theta)$.

As we will see soon, stochastic gradient descent hinges on evaluating the derivative of V under the integral. The following lemma is a consequence of Theorem 2.7.1 of [18].

Lemma 1. *V is differentiable and its gradient can be evaluated under the integral on* $\mathbf{int}\,\{\theta \mid V(\theta) < \infty\}$*, where* **int** *denotes the interior.*

In particular, we have

$$\begin{aligned}\nabla V(\theta) &= \int \nabla_\theta \frac{\phi^2(x) f^2(x)}{f_\theta(x)}\, dx \\ &= \int (\nabla A(\theta) - T(x)) \frac{\phi^2(x) f^2(x)}{f_\theta^2(x)} f_\theta(x)\, dx \\ &= \mathbf{E}_{X \sim f_\theta}\left[(\nabla A(\theta) - T(X)) \frac{\phi^2(X) f^2(X)}{f_\theta^2(X)}\right].\end{aligned}$$

So when we take a sample $X \sim f_\theta$, the random vector

$$g = (\nabla A(\theta) - T(X)) \frac{\phi^2(X) f^2(X)}{f_\theta^2(X)}$$

satisfies $\mathbf{E} g = \nabla V(\theta)$.

3 The method

Stochastic gradient descent is a standard method for solving

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

using the algorithm

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

where Π is (Euclidean) projection onto Θ, the *step size* $\alpha_n > 0$ is an appropriately chosen sequence, and the *stochastic gradient* g_n is a random variable satisfying

$$\mathbf{E}\left[g_n \mid \theta_n\right] = \nabla V(\theta_n).$$

The intuition is that $-g_n$, although noisy, generally points towards a descent direction of V at θ_n, and therefore each step reduces the function value of V in expectation [17, 27, 29, 35].

Our algorithm, which we call Convex AdaMC, is

$$\begin{aligned} X_n &\sim f_{\theta_n} \\ \hat{I}_n^{\mathrm{AMC}} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(X_i) f(X_i)}{f_{\theta_i}(X_i)} \\ g_n &= (\nabla A(\theta_n) - T(X_n)) \frac{\phi^2(X_n) f^2(X_n)}{f_{\theta_n}^2(X_n)} \\ \theta_{n+1} &= \Pi\left(\theta_n - \frac{C}{\sqrt{n}} g_n\right), \end{aligned}$$

where $C > 0$ and $\theta_1 \in \Theta$. As mentioned in the introduction, the estimator $\hat{I}_n^{\mathrm{AMC}}$ is unbiased and has variance given by (2) under a technical condition to be presented in §4.

We can view Convex AdaMC as an adaptive importance sampling method where the third and fourth line of the algorithm updates the sampling distribution. Alternatively, we can view Convex AdaMC as stochastic gradient descent on the convex function V with an additional step, the second line of the algorithm, that accumulates the estimate of I but does not otherwise affect the iteration.

The computational cost of Convex AdaMC is cheap, of course, if all of its operations are cheap. This is the case if ϕ and f are functions we can easily evaluate, if our family of distributions, $\mathcal{F}$, is one of the well-known exponential families, and if Θ is a set we can easily project onto.

4 Analysis

Before we present our convergence results, we discuss the choice of Θ. For any convex domain Θ, our variance function V is convex and minimizing $V(\theta)$ over $\theta \in \Theta$ is a mathematically well-defined problem. However, for our method to be well-defined and for the proof of convergence to work out, we need further restrictions on Θ.

Define

$$K(\theta) = \mathbf{E}_{X \sim f_\theta}\left[\frac{\phi^4(X) f^4(X)}{f_\theta^4(X)}\right] = \int \frac{\phi^4(x) f^4(x)}{f_\theta^3(x)}\, dx.$$

We require that Θ is convex and compact and that $\Theta \subseteq \mathbf{int}\,\{\theta \mid K(\theta) < \infty\}$. In other words, Θ must be a convex compact subset of the interior of the set of natural parameters for which their importance sampled estimates have finite 4th moment. Since

$$\{\theta \mid K(\theta) < \infty\} \subseteq \{\theta \mid V(\theta) < \infty\} \subseteq \{\theta \mid A(\theta) < \infty\}$$

it follows that any $\theta \in \Theta$ defines a sampling distribution f_θ that produces an importance sampled estimate of finite variance.

Theorem 2. *Assume $\Theta \subseteq \mathbf{int}\,\{\theta \mid K(\theta) < \infty\}$ is nonempty, convex, and compact. Define $D = \max_{\theta_1,\theta_2 \in \Theta} \|\theta_1 - \theta_2\|_2$ and*

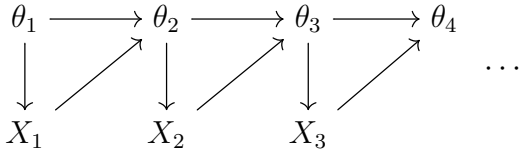
$$G^2 = \sup_{\theta \in \Theta} \mathbf{E}_{X \sim f_\theta} \left\| (\nabla A(\theta) - T(X)) \frac{\phi^2(X) f^2(X)}{f_\theta^2(X)} \right\|_2^2.$$

Then $G < \infty$, and $\hat{I}_n^{\mathrm{AMC}}$, the unbiased estimator of Convex AdaMC*, satisfies*

$$\frac{1}{n} V^* \leq \mathbf{Var}(\hat{I}_n^{\mathrm{AMC}}) \leq \frac{1}{n} V^* + \left(\frac{D^2}{2C} + CG^2\right) \frac{1}{n^{3/2}}.$$

Proof. We defer the proof of $G < \infty$ to the appendix.

Since the conditional dependency of our sequences $\theta_1, \theta_2, \ldots$ and $X_1, X_2, \ldots$ is



X_i is independent of the entire past conditioned on θ_i for all i. With this insight, we have

$$\mathbf{E}\hat{I}_n^{\mathrm{AMC}} = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\frac{\phi(X_i) f(X_i)}{f_{\theta_i}(X_i)} = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\mathbf{E}\left[\frac{\phi(X_i) f(X_i)}{f_{\theta_i}(X_i)} \mid \theta_i\right]\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\,[I] = I$$

and

$$\begin{aligned}
\mathbf{Var}(\hat{I}_n^{\mathrm{AMC}}) &= \mathbf{E}(\hat{I}_n^{\mathrm{AMC}} - I)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}\left(\frac{\phi(X_i)f(X_i)}{f_{\theta_i}(X_i)} - I\right)^2 \\
&\quad + \frac{2}{n^2}\sum_{1\le i<j\le n} \mathbf{E}\left(\frac{\phi(X_i)f(X_i)}{f_{\theta_i}(X_i)} - I\right)\left(\frac{\phi(X_j)f(X_j)}{f_{\theta_j}(X_j)} - I\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}\left[\mathbf{E}\left[\left(\frac{\phi(X_i)f(X_i)}{f_{\theta_i}(X_i)} - I\right)^2 \mid \theta_i\right]\right] \\
&\quad + \frac{2}{n^2}\sum_{1\le i<j\le n} \mathbf{E}\left[\mathbf{E}\left[\left(\frac{\phi(X_i)f(X_i)}{f_{\theta_i}(X_i)} - I\right)\left(\frac{\phi(X_j)f(X_j)}{f_{\theta_j}(X_j)} - I\right) \mid \theta_j\right]\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}V(\theta_i).
\end{aligned}$$

Since $V(\theta_i) \ge V^\star$ for any $\theta_i \in \Theta$, we conclude $\mathbf{Var}(\hat{I}_n^{\mathrm{AMC}}) \ge V^\star/n$.

Now let's prove the upper bound. Let $\theta_\star$ be a minimizer of V over Θ (which exists since V is continuous on the compact set Θ). Then we have

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

where the first inequality follows from nonexpansivity of Π (*i.e.*, $\|\Pi(u) - \Pi(v)\|_2 \le \|u - v\|_2$ for any u and v). We take expectation conditioned on θ_i on both sides to get

$$\begin{aligned}
\mathbf{E}\left[\|\theta_{i+1} - \theta_\star\|_2^2 \mid \theta_i\right] &\le \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}\mathbf{E}\left[\|g_i\|_2^2 \mid \theta_i\right] - 2\frac{C}{\sqrt{i}}\nabla V(\theta_i)^T(\theta_i - \theta_\star) \\
&\le \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}G^2 - 2\frac{C}{\sqrt{i}}\nabla V(\theta_i)^T(\theta_i - \theta_\star) \\
&\le \|\theta_i - \theta_\star\|_2^2 + \frac{C^2}{i}G_2^2 - 2\frac{C}{\sqrt{i}}(V(\theta_i) - V(\theta_\star)),
\end{aligned}$$

where the second inequality follows from the definition of G and the third inequality follows from re-arranging the following consequence of V's convexity

$$V(\theta_\star) \ge V(\theta_i) + \nabla V(\theta_i)^T(\theta_\star - \theta_i).$$

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

$$\mathbf{E}V(\theta_i) - V^\star \le \frac{\sqrt{i}}{2C}(\mathbf{E}\|\theta_i - \theta_\star\|_2^2 - \mathbf{E}\|\theta_{i+1} - \theta_\star\|_2^2) + \frac{C}{2\sqrt{i}}G^2.$$

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

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

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 $2n^2$ to get

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}V(\theta_i) \leq \frac{1}{n}V^\star + \left(\frac{D^2}{2C} + CG^2\right)\frac{1}{n^{3/2}}.$$

□

Not surprisingly, we have a central limit theorem (CLT) for our estimator. The proof of the following theorem is a straightforward application of a Martingale CLT, and is given in the appendix.

Theorem 3. *Under the assumptions of Theorem 2, we have*

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

Convex AdaMC has parameters C and θ_1 that must be chosen, but Theorem 2 or its proof does not give us insight on how to make this choice. Of course, the choice $C = D/\sqrt{2}G$ optimizes the bound of Theorem 2, but this is not very meaningful: The quantity G, in general, is unknown a priori, and the term $(D^2/2C + CG^2)/n^{3/2}$ is merely a bound that we suspect is not representative of the actual performance. In practice, one should vary C through several informal iterations to find what works well.

Likewise, the stated bound of Theorem 2 independent of θ_1, and the proof does not seem to reveal any significant dependence on θ_1. However, intuition and empirical experiments suggest that a θ_1 with a small value of $V(\theta_1)$ performs well.

Rather, the theoretical significance of Theorem 2 and 3 is that the leading order term of $\mathbf{Var}(\hat{I}_n^{\mathrm{AMC}})$ is $V^\star/n$, the optimum among the family $\mathcal{F}$, and that the following term is of order $\mathcal{O}(1/n^{3/2})$. In particular, this implies that Convex AdaMC cannot be trapped at a (non-optimal) local minimum.

5 Examples

Volume of a polytope. Consider the problem of computing the area of the quadrilateral Q with corners at $(0.05, 0.9)$, $(0.8, 0.9)$, $(1, 0.7)$, and $(0.15, 0.7)$. The answer is 0.16, which of course can be found with simple geometry.

First note that

$$I = \int_0^1 \int_0^1 1_Q \, dxdy,$$

where 1_Q is the indicator function that is 1 within the quadrilateral and 0 otherwise. Now let's see how to compute I with Convex AdaMC.

First, we choose bivariate Gaussians, which have the densities

$$f(x; \mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

as our candidate sampling distributions. To form these into an exponential family, we perform a change of variables. Loosely speaking, we say our natural parameter θ has two components: $m = \Sigma^{-1}\mu \in \mathbf{R}^2$ and $S = \Sigma^{-1} \in \mathbf{S}^2$ where $\mathbf{S}^2$ denotes the set of 2×2 symmetric matrices. Now our densities are

$$f_{m,S}(x) = \frac{1}{2\pi} \exp\left(m^T x - \frac{1}{2}\mathbf{Tr}(Sxx^T)\right) \exp\left(-\frac{1}{2}\left(m^T S^{-1} m - \log|S|\right)\right).$$

(Note that $\mathbf{Tr}(Sxx^T)$ is linear in S as it is the inner product between S and xx^T, interpreted as vectors of $\mathbf{R}^4$.) We choose our compact natural parameter set Θ to be

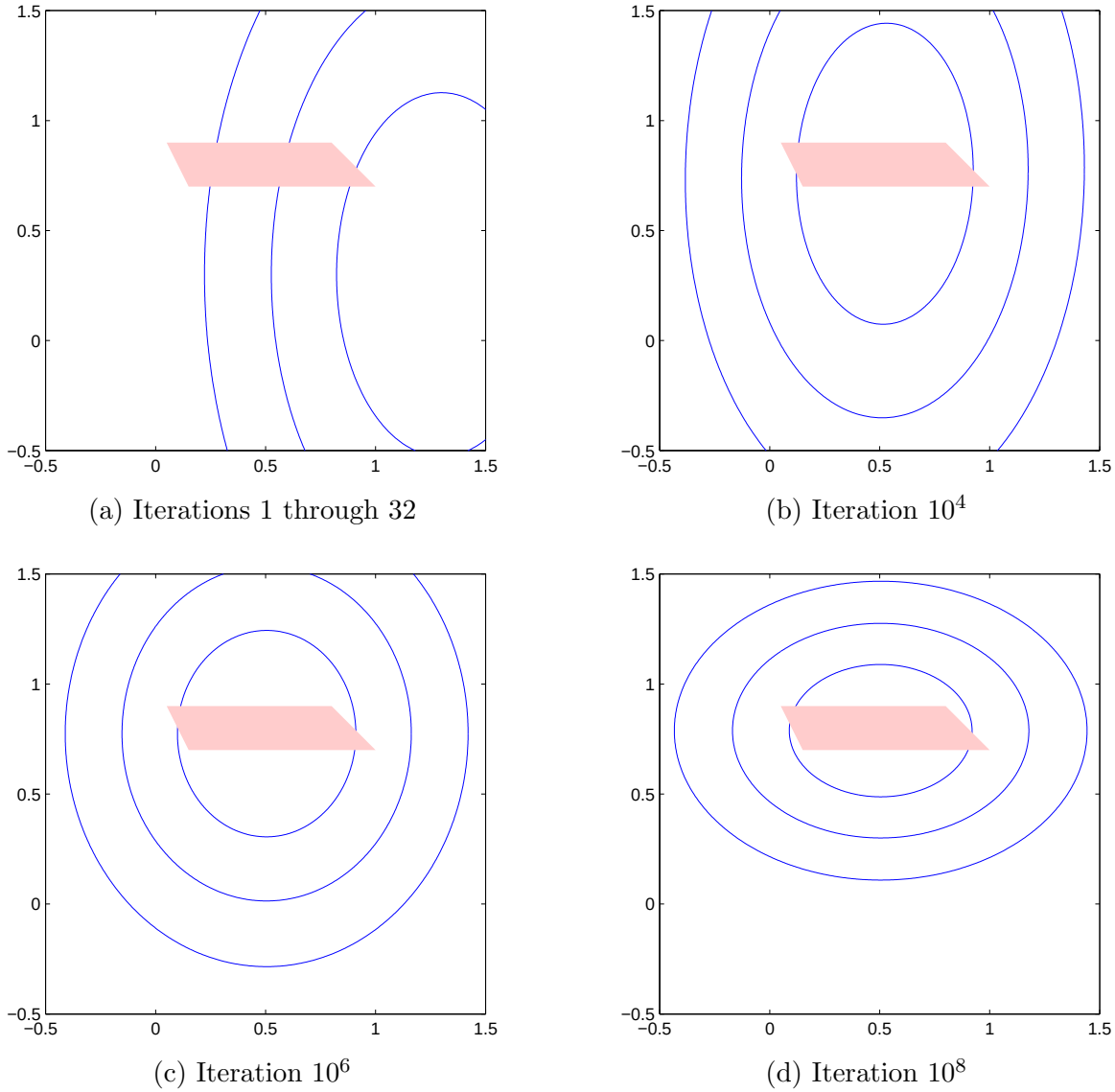
$$\Theta = [0, 25]^2 \times \left\{S \in \mathbf{S}^2 \mid I \preceq S \preceq 50I\right\}.$$

In other words, we restrict m_1 and m_2 to be within $[0, 25]$ and the eigenvalues of Θ to both be within $[1, 50]$. With this choice, the updates of Convex AdaMC are

$$\begin{aligned} m_{n+1} &= \Pi_{[0,25]^2}\left(m_n - \frac{C1_Q(X_n)}{f^2_{m,S}(X_n)\sqrt{n}}\left(S_n^{-1}m_n - X_n\right)\right), \\ S_{n+1} &= \Pi_{\{S\in\mathbf{S}^2 | I\preceq S\preceq 50I\}}\left(S_n - \frac{C1_Q(X_n)}{2f^2_{m,S}(X_n)\sqrt{n}}\left(X_nX_n^T - S_n^{-1}m_nm_n^TS_n^{-1} - S_n^{-1}\right)\right). \end{aligned}$$

Figure 1 shows the improvement of the sampling distributions for a particular run of this problem. Figure 1a gives the initial sampling distribution. Since the first sample to ever hit Q happens at iteration 33, the sampling distribution is identical for the first 32 iterations. As the algorithm progresses, we see that the density function of the Gaussian sampling distribution gradually matches the shape I_Q.

Option pricing. Consider the pricing of an arithmetic Asian call option on an underlying asset under standard Black-Scholes assumptions [14]. 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



(a) Iterations 1 through 32

(b) Iteration 10^4

(c) Iteration 10^6

(d) Iteration 10^8

Figure 1: Sampling distributions at different iterations. The red quadrilateral represents Q and the three ellipses denote the 68%, 95%, and 99.7% confidence ellipsoids.

T for the maturity time. Under the Black-Scholes model, the price of the asset at time jT/k is

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

for $t = 1, \ldots, k$, where $X \in \mathbf{R}^k$ is random with independent standard normal entries $X^1, \ldots, X^k$. (Here we will use superscripts to denote entries of a vector.) The discounted payoff of the option with strike K is given by

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

and we wish to compute $\mathbf{E}\phi(X)$.

To use Convex AdaMC, we choose the exponential family

$$f_\theta(x) = \frac{1}{(2\pi)^{k/2}} e^{-\|x-\theta\|_2^2} = \exp\left(\theta^T x - \frac{1}{2}\|\theta\|_2^2\right) \exp\left(-\frac{1}{2}\|x\|_2^2\right) / (2\pi)^{k/2},$$

where $\theta \in \mathbf{R}^k$ and $\Theta \in [-0.5, 0.5]^k$. In other words, $X \sim f_\theta$ contains independent standard normals with mean shifted by θ. So we have

$$\phi(X)\frac{f(X)}{f_\theta(X)} = \phi(X)\exp\left(\frac{1}{2}\|\theta\|_2^2 - X^T\theta\right)$$

and

$$\nabla A(\theta) = \theta \qquad T(X) = X.$$

We run Convex AdaMC with the parameters $S^0 = K = 50$, $r = 0.05$, $\sigma = 0.3$, $T = 1.0$, $k = 64$, $C = 0.01$, and $\theta_1 = 0$ for 10^7 iterations. Figure 2 shows the shifting at the end of the algorithm and the asset price estimate is 4.02.

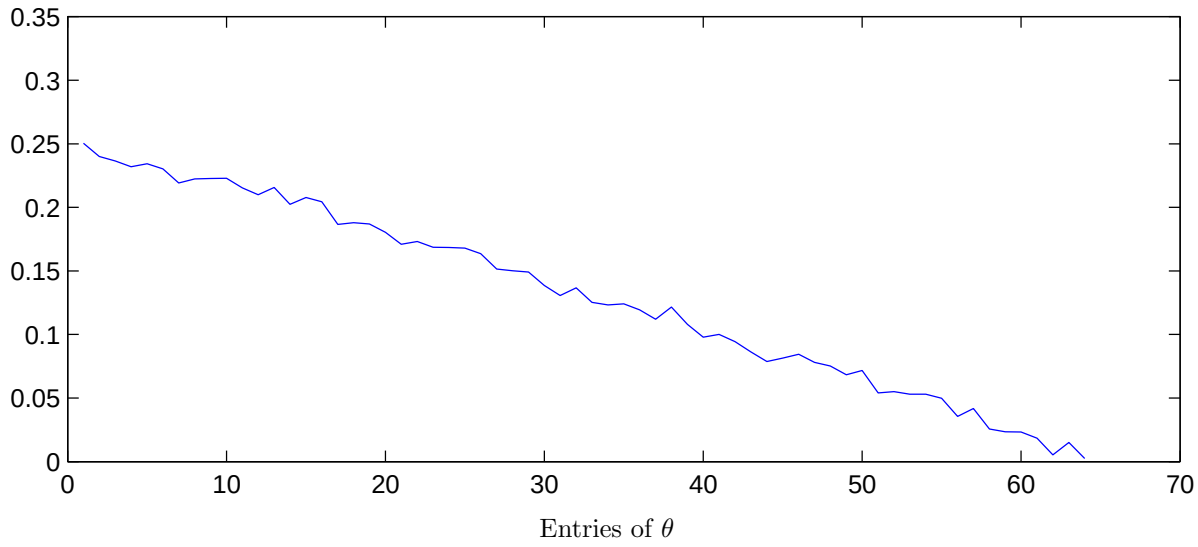


Figure 2: Importance sampling parameter θ for the Asian option pricing problem after 10^7 iterations.

6 Remarks, extensions, and variations

An iteration of Convex AdaMC is simultaneously an iteration of a convex optimization problem and an iteration of importance sampling. Because of this fact, each iteration of the method is computationally efficient, and we can prove convergence of the variance (and of course the estimator) as a function of the iteration count.

Some previous work on adaptive importance sampling have used stochastic gradient descent or similar stochastic approximation algorithms without a setup to make the variance

convex [1–3,13]. 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 as in general with nonconvex optimization problems, 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 [6–12, 15, 19, 23, 26, 32, 33]. 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.

We point out two straightforward generalizations that we omitted for the sake of simplicity. One is that when the nominal and importance distributions have densities with respect to any general measure, not the Lebesgue measure as assumed in our exposition, the same results apply.

Another is to adaptively minimize the Rényi generalized divergence with parameter $\alpha \geq 1$ of the sampling distribution to the "optimal" sampling distribution $|\phi(x)|f(x)$ [21,28]. What we did, minimizing the variance of the estimator, is the special with $\alpha = 2$. When $\alpha = 1$, the Rényi generalized divergence becomes the cross entropy and we get a method similar to the cross-entropy method [8, 33]. (The Rényi generalized divergence is convex for $\alpha \geq 1$.)

There are other not-so-straightforward generalizations worth pursuing. One is to try other stochastic optimization methods. In this paper, we used the most common and simplest stochastic optimization method, stochastic gradient descent with step size $C/\sqrt{n}$. However, there are many other methods to solve a given stochastic optimization problem, and these other methods could perform better under certain assumptions.

Another would be a different weighting scheme. In Convex AdaMC, we add a sequence of unbiased estimators with varying variance, which, loosely speaking, is decreasing in expectation. If we knew these variances in advance, then we can easily compute the optimal weighting, which is not a uniform weighting. Although we don't know the variances in advance, it would be interesting to know if there is a better weighting or to characterize the optimality of the uniform weighting in the spirit of Theorem 4 of [25].

Finally, it would be most interesting to understand Convex AdaMC's theoretical and empirical performance when used in conjunction with other variance reduction techniques such as control variates or mixture importance sampling.

Acknowledgement

We thank Art B. Owen for helpful discussions and many detailed suggestions. This research was supported by the Simons Foundation and DARPA X-DATA.

7 Appendix

The following lemma, like Lemma 1 follows from Theorem 2.7.1 of [18].

Lemma 2. *A, V, and K are infinitely differentiable and all derivatives can be evaluated under their integrals on the interiors of their respective domains.*

We are now ready to prove the part of the proof of Theorem 2 we omitted.

Proof of $G < \infty$ in Theorem 2. For any $i \in \{1, 2, \ldots, p\}$,

$$\frac{\partial}{\partial \theta_i} K(\theta) = 3 \int \left(\frac{\partial}{\partial \theta_i} A(\theta) - T_i(x) \right) \frac{\phi^4(x) f^4(x)}{f_\theta^3(x)} \, dx = 3K(\theta) \frac{\partial}{\partial \theta_i} A(\theta) - 3 \int T_i(x) \frac{\phi^4(x) f^4(x)}{f_\theta^3(x)} \, dx$$

exists and is and continuous on $\mathbf{int} \{\theta \mid K(\theta) < \infty\}$ by Lemma 2. As we already know the first term is continuous by Lemma 2, this tells us the second term is continuous.

Repeating this, we have

$$\frac{\partial^2}{\partial \theta_i^2} K(\theta) = \int \left(3 \frac{\partial^2}{\partial \theta_i^2} A(\theta) + 9 \left(\frac{\partial}{\partial \theta_i} A(\theta) \right)^2 - 18 T_i(x) \frac{\partial}{\partial \theta_i} A(\theta) + 9 T_i^2(x) \right) \frac{\phi^4(x) f^4(x)}{f_\theta^3(x)} \, dx.$$

We know the first 3 terms are continuous from what we just proved and Lemma 2. So we conclude that

$$\int T_i^2(x) \frac{\phi^4(x) f^4(x)}{f_\theta^3(x)} \, dx$$

is a continuous function of θ on $\mathbf{int} \{\theta \mid K(\theta) < \infty\}$.

Finally, we conclude that

$$\begin{aligned} &\mathbf{E}_{X \sim f_\theta} \left\| (\nabla A(\theta) - T(X)) \frac{\phi^2(X) f^2(X)}{f_\theta^2(X)} \right\|_2^2 \\ &= \|\nabla A(\theta)\|_2^2 K(\theta) - 2 \nabla A(\theta)^T \int T(X) \frac{\phi^4(X) f^4(X)}{f_\theta^3(X)} \, dx + \int \|T(X)\|_2^2 \frac{\phi^4(X) f^4(X)}{f_\theta^3(X)} \, dx \end{aligned}$$

is a continuous function on the compact set Θ and therefore the supremum, G^2, is finite. □

Finally, we prove the CLT.

Proof of Theorem 3. First define

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

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

$$\sigma_{ni}^2 = \mathbf{E}\left[Y_{ni}^2 \mid \mathcal{G}_{i-1}\right] = \begin{cases} \frac{1}{n} V(\theta_i) & \text{for } i \leq n \\ 0 & \text{otherwise.} \end{cases}$$

Then a form of the Martingale CLT, *c.f.*, Theorem 35.12 of [4], states that if

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

and

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

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

Since

$$\sum_{i=1}^{n} \sigma_{ni}^2 = \frac{1}{n} \sum_{i=1}^{n} V(\theta_i),$$

and since, by Theorem 2, we have

$$\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{E} V(\theta_i) - V^\star\right) = \mathbf{E} \left| \frac{1}{n} \sum_{i=1}^{n} V(\theta_i) - V^\star \right| = \mathcal{O}(1/\sqrt{n}) \to 0,$$

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

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

Finally, since $\Theta \subseteq \{\theta \mid K(\theta) < \infty\}$ is a compact set and $K(\theta)$ is a continuous function by Lemma 2, we have

$$B = \sup_{\theta \in \Theta} K(\theta) < \infty$$

and we conclude

$$\begin{aligned} \sum_{i=1}^{n} \mathbf{E} Y_{ni}^2 I_{\{|Y_{ni}| \geq \varepsilon\}} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} \frac{\phi^2(X_i) f^2(X_i)}{f_{\theta_i}^2(X_i)} I_{\left\{\phi^2(X_i) f^2(X_i) / f_{\theta_i}^2(X_i) \geq n\varepsilon^2\right\}} \\ &\leq \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^{n} \mathbf{E} \frac{\phi^4(X_i) f^4(X_i)}{f_{\theta_i}^4(X_i)} \leq \frac{B}{n\varepsilon^2} \to 0. \end{aligned}$$

Since this proves the conditions we need, applying the martingale CLT completes the proof. □

References

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

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

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

[19] 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.

[20] N. Madras. *Lectures on Monte Carlo Methods*. American Mathematical Society, 2002.

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

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

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

[24] A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

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

[26] 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.

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

[28] 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*. University of California Press, 1961.

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

[30] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

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

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

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

[34] J. S. Sadowsky and J. A. Bucklew. On large deviations theory and asymptotically efficient monte carlo estimation. *IEEE Transactions on Information Theory*, 36(3):579–588, 1990.

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

[36] P. J. Smith, M. Shafi, and H. Gao. Quick simulation: A review of importance sampling techniques in communication systems. *IEEE Journal on Selected Areas in Communications*, 15(4):597–613, 1997.

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