Radon Series Comp. Appl. Math **8**, 1–18

Adaptive variance reduction techniques in finance

Benjamin Jourdain

Abstract. This paper gives an overview of adaptive variance reduction techniques recently developed for financial applications. More precisely, we explain how information available in the random drawings made to compute the expectation of interest may be used at the same time to optimize control variates, importance sampling or stratified sampling.

Key words. Variance reduction techniques, control variates, importance sampling, stratification, sample average optimization

AMS classification. 65C05 90C15 91-08

Introduction

In mathematical finance, the price of a European option is expressed as the expectation under the risk neutral probability measure of the discounted payoff of the option. Sensitivities of the price with respect to various parameters, the so-called greeks, and in particular the delta which is of paramount importance for hedging purposes, may also be expressed as expectations. The simplest and most natural numerical approach to compute these expectations, the Monte Carlo method, is widely used in banks. According to the central limit theorem, the precision of the empirical mean approximation of the expectation of a random variable is proportional to the standard deviation of this variable. Variance reduction techniques aim at improving this precision by computing the empirical mean of independent copies of a random variable with the same expectation as the original one but with a lower variance. These techniques may be classified into two categories :

- the ones which guarantee that the variance of the new variable will be lower than the variance of the original one : antithetic variables and conditioning. In general, the variance reduction ratio obtained with these techniques is not very large.
- the ones which may lead to a more significant variance reduction ratio but may also increase the variance depending on whether they are properly implemented : control variates and importance sampling.

Stratified sampling is at the boundary between these two classes : when the allocation of the random drawings into the strata is made proportionally to their probabilities, variance reduction is guaranteed. Nevertheless, to improve efficiency, one should try other allocation rules but then the variance may increase.

This research benefited from the support of the French National Research Agency (ANR) under the program ANR-05-BLAN-0299 and of the “Chair Risques Financiers”, Fondation du Risque

Adaptive methods have been developed to ensure a proper implementation of the second category of variance reduction techniques : information available in the random drawings made to compute the expectation of interest is used to optimize the variance reduction technique at the same time. In general, they save computation time in comparison to their more natural and earlier investigated alternative : optimize the variance reduction technique on a first pilot set of random drawings and then compute the empirical mean of the resulting random variable on a second set of independent drawings. Such two stages procedures lead to unbiased estimators whereas, in general, adaptive estimators are only asymptotically unbiased.

Sections 2, 3 and 4 are respectively devoted to adaptive control variates, adaptive importance sampling and adaptive stratified sampling. Since we are interested in financial applications, we will pay in what follows particular attention to the computation of $\mathbb{E}(f(G))$ where G is a standard d-dimensional normal random vector and $f : \mathbb{R}^d \to \mathbb{R}$. Indeed, the price and hedging ratios of European options written on underlying assets evolving according to a multi-dimensional Black Scholes model may be expressed in this way. When the underlyings evolve according to a more general stochastic differential equation, Euler discretization of this equation leads to approximations of the price and hedging ratios by expectations of the previous form, for a possibly high dimensional normal vector G and a complicated function f. Notice that in the present volume, Giles and Waterhouse [11] present an interesting multilevel path simulation technique which enables to reduce the time-discretization bias by computing the expectation corresponding to a refined time-grid. In order to reduce the computation time necessary to obtain a balanced statistical error, they suggest to combine results using different time-steps numbers. In the end, their method consists in computing $\mathbb{E}(f(G))$ for an even higher-dimensional and more complicated function f than the one derived from standard Euler discretization.

0.1 Adaptive control variates

Let us first illustrate the basic ideas of adaptive variance reduction on the simple example of linearly parametrized control variates (see for instace [21], [24] or Section 4.1 in [12]) before dealing with general parametrization.

0.1.1 Linearly parametrized control variates

Suppose that we want to compute the expectation $\mathbb{E}(Y)$ of a real random variable Y and that $Z = (Z^1, \dots, Z^d)^*$ is a related $\mathbb{R}^d$-valued centered random vector with Y and Z both square-integrable. We also assume, up to removing some coordinates of Z, that the covariance matrix $\mathrm{Cov}(Z)$ of Z is non-singular and we denote by $\mathrm{Cov}(Y, Z) = \mathbb{E}(YZ)$ the covariance between Y and Z . In finance, typically $Y = e^{-rT} f(X_T^1, \dots, X_T^d)$ where f is the payoff of a European option with maturity T written on d underlying assets $X^1, \dots, X^d$ with respective initial prices $x^1, \dots, x^d$ and since the discounted price of each asset is a martingale under the risk neutral measure, one may choose

$$Z = (X_T^1 - e^{rT} x^1, \dots, X_T^d - e^{rT} x^d)^*.$$

For $\theta \in \mathbb{R}^d$, since $\mathbb{E}(Y-\theta.Z) = \mathbb{E}(Y)$, one may approximate the expectation of interest $\mathbb{E}(Y)$ by the empirical mean $M_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n (Y_j - \theta.Z_j)$ where $((Y_j, Z_j))_{j\geq 1}$ are independent copies of (Y, Z). The classical estimator $\frac{1}{n}\sum_{j=1}^n Y_j$ corresponds to the choice $\theta = 0$. The variance of $M_n(\theta)$, equal to $\frac{v(\theta)}{n}$ where

$$v(\theta) \stackrel{\text{def}}{=} \mathrm{Var}(Y - \theta.Z) = \mathrm{Var}(Y) - 2\theta.\mathrm{Cov}(Y, Z) + \theta.\mathrm{Cov}(Z)\theta,$$

is minimal for $\theta_\star = \mathrm{Cov}(Z)^{-1}\mathrm{Cov}(Y, Z)$. Of course, when $\mathbb{E}(Y)$ is unknown, so is $\theta_\star$. But one may estimate the covariances $\mathrm{Cov}(Z)$ and $\mathrm{Cov}(Y, Z)$, respectively, by

$$C_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n Z_j Z_j^* - \left(\frac{1}{n}\sum_{j=1}^n Z_j\right)\left(\frac{1}{n}\sum_{j=1}^n Z_j^*\right)$$

$$\text{and } D_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n Y_j Z_j - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)\left(\frac{1}{n}\sum_{j=1}^n Z_j\right).$$

Let N be the smallest index n such that no strict affine subspace of $\mathbb{R}^d$ contains $\{Z_1, \ldots, Z_n\}$. Since $\mathrm{Cov}(Z)$ is non-singular, N is a.s. finite. Moreover C_n is non-singular if and only if $n \geq N$. For $n \geq N$, one may approximate $\theta_\star$ by the estimator $\theta_n \stackrel{\text{def}}{=} C_n^{-1} D_n$ which converges a.s. to $\theta_\star$ when $n \to \infty$. The derived adaptive control variate estimator $M_n(\theta_n) = \frac{1}{n}\sum_{j=1}^n (Y_j - \theta_n.Z_j)$ of $\mathbb{E}(Y)$ is biased in general (but not when (Y, Z) is a Gaussian vector or more generally when $\mathbb{E}(Y|Z) = \mathbb{E}(Y) + \theta_\star.Z$). Nevertheless, $M_n(\theta_n)$ is a.s. convergent to $\mathbb{E}(Y)$. Moreover, writing

$$\sqrt{n}(M_n(\theta_n) - \mathbb{E}(Y)) = \begin{pmatrix} 1 \\ \theta_n \end{pmatrix} . \frac{1}{\sqrt{n}} \sum_{j=1}^n \begin{pmatrix} Y_j - \mathbb{E}(Y) \\ Z_j \end{pmatrix},$$

one deduces from the central limit theorem governing the convergence in law of the second term in the product and Slutsky's theorem that $M_n(\theta_n)$ is asymptotically normal with optimal asymptotic variance $v(\theta_\star)$. To sum up,

Proposition 0.1 *The vector $(\theta_n, M_n(\theta_n))$ converges a.s. to $(\theta_\star, \mathbb{E}(Y))$ and*

$$\sqrt{n}(M_n(\theta_n) - \mathbb{E}(Y)) \stackrel{\mathcal{L}}{\to} \mathcal{N}_1(0, v(\theta_\star)).$$

Variance reduction is guaranteed in the limit since $v(\theta_\star) \leq v(0) = \mathrm{Var}(Y)$, the inequality being an equality only when Y and Z are uncorrelated. When $v(\theta_\star) = 0$ i.e. when $Y = \mathbb{E}(Y) + \theta_\star.Z$ then for all $n \geq N$, $\theta_n = \theta_\star$ and $M_n(\theta_n) = \mathbb{E}(Y)$ (see [19]). This situation is not likely to occur in financial applications but an example in the context of Markov chains is given in [14] which also discusses the asymptotic properties of other adaptive estimators of $\mathbb{E}(Y)$.

One could also approximate $\mathbb{E}(Y)$ by the unbiased estimator $M_n(\tilde{\theta}_m)$ with

$$\tilde{\theta}_m = \left(\sum_{k=1}^m \tilde{Z}_k \tilde{Z}_k^* - \frac{1}{m}\sum_{k=1}^m \tilde{Z}_k \sum_{k=1}^m \tilde{Z}_k^*\right)^{-1} \left(\sum_{k=1}^m \tilde{Y}_k \tilde{Z}_k - \frac{1}{m}\sum_{k=1}^m \tilde{Y}_k \sum_{k=1}^m \tilde{Z}_k\right)$$

where $((\tilde{Y}_k,\tilde{Z}_k))_{k\geq 1}$ are i.i.d. copies of (Y,Z) independent of $((Y_j,Z_j))_{j\geq 1}$. This is an example of the two stages procedure mentioned in the introduction. But it is a pity not to use the drawings $((\tilde{Y}_k,\tilde{Z}_k))_{1\leq k\leq m}$ made to compute $\tilde{\theta}_m$ also in the computation of the expectation of interest.
Let us finally mention that θ_n introduced above as a sample average approximation of the optimal parameter $\theta_\star$ also has another interpretation. The vector θ_n minimizes the sample approximation $v_n(\theta) = \frac{1}{n}\sum_{j=1}^n (Y_j-\theta.Z_j)^2 - \left(\frac{1}{n}\sum_{j=1}^n (Y_j-\theta.Z_j)\right)^2$ of $v(\theta)$. For more complex variance reduction techniques involving a parameter, no explicit expression of the optimal parameter $\theta_\star$ is in general available. So defining θ_n as an estimator of $\theta_\star$ is no longer possible. But the alternative definition of θ_n as the parameter minimizing the sample average approximation of the variance remains possible. We will see applications to generally parametrized control variates in the next paragraph and to importance sampling for normal random vectors in Section 0.2.

0.1.2 General parametrization

General parametrization of control variates for the computation of the expectation $\mathbb{E}(Y)$ of a square-integrable random variable Y is addressed by Kim and Henderson [19, 20]. Let $\Theta \subset U \subset \mathbb{R}^p$ with Θ compact and U bounded open, Z be a d-dimensional random vector related to Y, $h : U\times\mathbb{R}^d \mapsto \mathbb{R}$ be such that

$$\forall \theta\in U,\ \mathbb{E}(h^2(\theta,Z)) < +\infty \text{ and } \mathbb{E}(h(\theta,Z)) = 0,$$

and $((Y_j,Z_j))_{j\geq 1}$ be a sequence of independent copies of (Y,Z).
For any $\theta\in U$, $M_n(\theta) \overset{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n (Y_j - h(\theta,Z_j))$ is an unbiased and a.s. convergent estimator of the expectation of interest $\mathbb{E}(Y)$. Moreover $\mathrm{Var}(M_n(\theta)) = \frac{v(\theta)}{n}$ where $v(\theta) \overset{\text{def}}{=} \mathrm{Var}(Y-h(\theta,Z))$.

Let $m\geq 2$. When for all $z\in\mathbb{R}^d$, $U\ni\theta\mapsto h(\theta,z)$ is C^1, the unbiased estimator $\frac{1}{m-1}\sum_{j=1}^m (Y_j-h(\theta,Z_j)-M_m(\theta))^2$ of $v(\theta)$ is differentiable on U with respect to θ with gradient equal to $\frac{2}{m-1}\sum_{j=1}^m (Y_j-h(\theta,Z_j)-M_m(\theta))\nabla_\theta\left[h(\theta,Z_j)-\frac{1}{m}\sum_{k=1}^m h(\theta,Z_k)\right]$.

Let $(\gamma_l)_{l\geq 0}$ be a sequence of positive steps such that $\sum_l \gamma_l = \infty$ and $\sum_l \gamma_l^2 < \infty$. Starting from $\theta_0\in\Theta$, Kim and Henderson [19, 20] suggest to optimize $v(\theta)$ with respect to θ using the following gradient-based stochastic approximation procedure :

$$\begin{cases} A_{l+1} = \frac{1}{m}\sum_{j=lm+1}^{(l+1)m}(Y_j - h(\theta_l,Z_j)) \\ \theta_{l+1} = \Pi_\Theta\Bigg(\theta_l - \frac{2\gamma_l}{m-1}\sum_{j=lm+1}^{(l+1)m}(Y_j-h(\theta_l,Z_j)-A_{l+1}) \\ \qquad\qquad \times\nabla_\theta\left[h(\theta,Z_j)-\frac{1}{m}\sum_{k=lm+1}^{(l+1)m}h(\theta_l,Z_k)\right]\Big|_{\theta=\theta_l}\Bigg) \end{cases}$$

where Π_Θ denotes a projection of points outside Θ back into Θ. Using the law of large numbers and the central limit theorem for martingales, they study the asymptotic behaviour of the associated estimator $\mu_k \overset{\text{def}}{=} \frac{1}{k}\sum_{l=1}^k A_l$ of $\mathbb{E}(Y)$.

Theorem 0.2 *Assume that for all $z \in \mathbb{R}^d$, $U \ni \theta \mapsto h(\theta, z)$ is C^1 and that*

$$\mathbb{E}\left(\sup_{\theta\in U}|\nabla_\theta(\theta, Z)|\left(1+\sup_{\theta\in U}|Y-h(\theta, Z)|\right)\right) < +\infty.$$

Then μ_k converges a.s. to $\mathbb{E}(Y)$ as $k \to \infty$. If moreover θ_k converges a.s. to a random variable θ_∞, then $\sqrt{km}(\mu_k - \mathbb{E}(Y)) \xrightarrow{\mathcal{L}} \sqrt{v(\theta_\infty)} \times G$ where $G \sim \mathcal{N}_1(0,1)$ is independent from θ_∞ and $\frac{1}{k(m-1)}\sum_{l=1}^{k}\sum_{j=(l-1)m+1}^{lm}(Y_j - h(\theta_{l-1}, Z_j) - A_l)^2$ converges a.s. to $v(\theta_\infty)$.
Last, if Θ is a box i.e. $\Theta = \prod_{i=1}^{p}[a_i, b_i]$ and $\exists\theta_0 \in \Theta$ such that

$$\mathbb{E}\left(Y^4 + \sup_{\theta\in U}|\nabla_\theta(\theta, Z)|^4 + h^4(\theta_0, Z)\right) < +\infty,$$

then the distance of θ_k to the set S of first order critical points of v on Θ converges a.s. to 0 and, when S is discrete, θ_k converges a.s. to an S-valued random variable θ_∞.

Kim and Henderson also study in [19, 20] the estimator $M_n(\tilde{\theta}_m)$ obtained by a two stages procedures where $\tilde{\theta}_m$ is obtained as a first order critical point of the sample average estimator of the variance $\frac{1}{m-1}\sum_{k=1}^{m}(\tilde{Y}_k - h(\theta, \tilde{Z}_k) - \frac{1}{m}\sum_{j=1}^{m}(\tilde{Y}_j - h(\theta, \tilde{Z}_j)))^2$ computed on a sequence $((\tilde{Y}_k, \tilde{Z}_k))_{k\geq 1}$ of independent copies of (Y, Z) independent from $((Y_j, Z_j))_{j\geq 1}$.

In [20], the behaviour of both estimators is illustrated on the example of the pricing of barrier options.

0.2 Importance sampling for normal random vectors

Adaptive importance sampling techniques have been developed to approximate multidimensional integrals over the unit hypercube (see [25] and the reference therein) or in the context of Markov chains (see for instance [3] [8]). But research on this topic in view of financial applications was centered on normal random vectors due to the importance of this specific case for models given by stochastic differential equations. That is why the present section is devoted to the computation of $\mathbb{E}(f(G))$ where G is distributed according to the standard d-dimensional normal law $\mathcal{N}_d(0, I_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$.

We assume that

$$\mathbb{P}(f(G) \neq 0) > 0 \text{ and } \forall\theta \in \mathbb{R}^d,\ \mathbb{E}(f^2(G)e^{-\theta.G}) < +\infty. \tag{0.1}$$

The second hypothesis is implied for instance by the existence of a finite moment of order $2+\varepsilon$ with $\varepsilon > 0$ for $|f(G)|$. Let $(G_j)_{j\geq 1}$ be i.i.d. copies of G. For $\theta \in \mathbb{R}^d$, since

$$\mathbb{E}\left(f(G+\theta)e^{-\theta.G-\frac{|\theta|^2}{2}}\right) = \mathbb{E}(f(G)), \tag{0.2}$$

$M_n(\theta) \overset{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n f(G_j+\theta)e^{-\theta.G_j-\frac{|\theta|^2}{2}}$ is an a.s. convergent and asymptotically normal estimator of $\mathbb{E}(f(G))$ with variance $\mathrm{Var}(M_n(\theta)) = \frac{v(\theta)-\mathbb{E}^2(f(G))}{n}$, where

$$v(\theta) \overset{\text{def}}{=} \mathbb{E}\left(f^2(G+\theta)e^{-2\theta.G-|\theta|^2}\right) = \mathbb{E}\left(f^2(G+\theta)e^{-\theta.(G+\theta)+\frac{|\theta|^2}{2}}e^{-\theta.G-\frac{|\theta|^2}{2}}\right)$$
$$= \mathbb{E}\left(f^2(G)e^{-\theta.G+\frac{|\theta|^2}{2}}\right). \tag{0.3}$$

Notice that the translated normal variable $G+\theta$ has the density $p_\theta(x) = (2\pi)^{-\frac{d}{2}}e^{-\frac{|x-\theta|^2}{2}}$ and that the importance sampling ratio $\frac{p_0}{p_\theta}(G+\theta) = e^{-\theta.G-\frac{|\theta|^2}{2}}$ appears as a factor in the left-hand-side of (0.2). The interest of the class of importance sampling estimators $M_n(\theta)$ parametrized by the translation vector $\theta \in \mathbb{R}^d$ is that a very simple analytic mapping (addition of θ) permits to transform an i.i.d. sample of the standard normal law $\mathcal{N}_d(0,I_d)$ into an i.i.d. sample of $\mathcal{N}_d(\theta,I_d)$. This feature is particularly convenient to compute and study adaptive estimators in which the parameter evolves during the simulation.

Under (0.1) the function v is

1. C^∞ with derivatives obtained by differentiation under the expectation (0.3) :

$$\nabla_\theta v^f(\theta) = \mathbb{E}\left((\theta-G)f^2(G)e^{-\theta.G+\frac{|\theta|^2}{2}}\right)$$
$$\nabla^2_\theta v^f(\theta) = \mathbb{E}\left((I_d+(\theta-G)(\theta-G)^*)f^2(G)e^{-\theta.G+\frac{|\theta|^2}{2}}\right).$$

2. strongly convex.

Therefore

$$\exists!\theta_\star \in \mathbb{R}^d \ : \ v(\theta_\star) = \inf_{\theta\in\mathbb{R}^d} v(\theta).$$

This suggests to approximate $\mathbb{E}(f(G))$ by $M_n(\theta_\star)$ but $\theta_\star$ is unknown. Unlike in the analogous example of linear control variates developed in Section 0.1, no explicit expression is available for $\theta_\star$. Methods aimed at approximating $\theta_\star$ have been developed in the literature. These methods are based

- either on determistic optimization : in [13], the authors suggest to choose θ maximizing $\mathbb{R}^d \ni x \mapsto \log|f(x)| - \frac{|x|^2}{2}$ and justify this approximation by a large deviations asymptotics.
- or on stochastic optimization procedures analogous to the ones presented in Section 0.1.2 : gradient-based stochastic approximation ([27] [26]), adaptive Robbins-Monro procedures [2, 1, 16, 23], robust optimization of the sample average approximation of v by Newton's algorithm [15].

Let us now describe those stochastic optimization procedures more precisely.

0.2.1 Gradient based stochastic approximation and adaptive Robbins-Monro algorithms

In [27] and [26], the authors suggest to minimize $v(\theta)$ over a compact convex subset Θ of $\mathbb{R}^d$ by the following iterative procedure using an integer $m \in \mathbb{N}^*$, a sequence $(\tilde{G}_k)_{k\geq 1}$ of independent copies of G (possibly equal to $(G_j)_{j\geq 1}$) and a sequence of positive steps $(\gamma_l)_{l\geq 0}$ s.t. $\sum_l \gamma_l = \infty$ and $\sum_l \gamma_l^2 < \infty$:

- start with $\theta_0 \in \Theta$,
- at step $l \geq 0$ compute $g_l = \frac{1}{m}\sum_{lm+1}^{(l+1)m}(\theta_l - \tilde{G}_k)f^2(\tilde{G}_k)e^{-\theta_l.\tilde{G}_k+\frac{|\theta_l|^2}{2}}$ approximating $\nabla_\theta v(\theta_l)$, then define θ_{l+1} as the projection $\theta_l - \gamma_l g_l$ on Θ.

Proposition 0.3 *Under* (0.1)*, the sequence* $(\theta_l)_{l\geq 1}$ *converges a.s. to the unique* $\theta_\Theta \in \Theta$ *such that* $v(\theta_\Theta) = \inf_{\theta\in\Theta} v(\theta)$.

The papers [27, 26] do not deal with asymptotic properties of the estimators $M_n(\theta_l)$ as $n, l \to \infty$. These questions are adressed by Arouna [2, 1] who also gets rid of the compact Θ. More precisely, he obtains a sequence $(\theta_n)_{n\geq 1}$ adapted to the filtration $(\sigma(G_1, \ldots, G_n))_{n\geq 1}$ by stabilizing the Robbins-Monro algorithm corresponding to the choice $m = 1$ and $(\tilde{G}_k)_{k\geq 1} = (G_j)_{j\geq 1}$ with Chen's projection technique [6, 5]. Let $\theta_0 \in \mathbb{R}^d$, $\sigma_0 = 0$ and $(s_n)_{n\geq 0}$ be an increasing sequence of positive numbers tending to infinity with n and s.t. $s_0 \geq |\theta_0|$. The sequence (θ_n, σ_n) is defined inductively by

$$\forall n \in \mathbb{N},\ \begin{cases} \theta_{n+\frac{1}{2}} = \theta_n - \gamma_n(\theta_n - G_{n+1})f^2(G_{n+1})e^{-\theta_n.G_{n+1}+\frac{|\theta_n|^2}{2}} \\ \text{if } |\theta_{n+\frac{1}{2}}| \leq s_{\sigma_n} \text{ then } \theta_{n+1} = \theta_{n+\frac{1}{2}} \text{ and } \sigma_{n+1} = \sigma_n \\ \text{if } |\theta_{n+\frac{1}{2}}| > s_{\sigma_n} \text{ then } \theta_{n+1} = \theta_0 \text{ and } \sigma_{n+1} = \sigma_n + 1 \end{cases} .$$

Here σ_n is the number of projections made during the n first iterations.

Theorem 0.4 *Under* (0.1)*, the total number of projections* $\lim_{n\to\infty}\sigma_n$ *is finite and* θ_n *converges a.s. to* $\theta_\star$ *as* $n \to \infty$.
If moreover $\mathbb{E}(f^{4+\varepsilon}(G)) < +\infty$*, then as* $n \to \infty$,

$$\begin{pmatrix} M_n \\ S_n \end{pmatrix} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} f(G_j+\theta_{j-1})e^{-\theta_{j-1}.G_j-\frac{|\theta_{j-1}|^2}{2}} \\ f^2(G_j+\theta_{j-1})e^{-2\theta_{j-1}.G_j-|\theta_{j-1}|^2} \end{pmatrix} \stackrel{\text{a.s.}}{\longrightarrow} \begin{pmatrix} \mathbb{E}(f(G)) \\ v(\theta_\star) \end{pmatrix},$$

and $\sqrt{n}(M_n - \mathbb{E}(f(G))) \stackrel{\mathcal{L}}{\longrightarrow} \mathcal{N}_1(0, v(\theta_*) - \mathbb{E}^2(f(G)))$.

As a consequence, $\sqrt{\frac{n}{S_n - M_n^2}}(M_n - \mathbb{E}(f(G))) \stackrel{\mathcal{L}}{\longrightarrow} \mathcal{N}_1(0,1)$ which enables to construct confidence intervals for the expectation of interest $\mathbb{E}(f(G))$. The first statement follows from the verifiable sufficient conditions given by Lelong [22] for the convergence of randomly truncated stochastic algorithms. Originally, Arouna [2] checked the a.s. convergence of θ_n to $\theta_\star$ only under some explicit restrictive growth assumption on the sequence $(s_n)_n$. In [1], remarking that

$$\mathbb{E}\left(f(G_n+\theta_{n-1})e^{-\theta_{n-1}.G_n-\frac{|\theta_{n-1}|^2}{2}}\middle|\sigma(G_1,\ldots,G_{n-1})\right) = \mathbb{E}(f(G)),$$

he derived the second statement using the law of large numbers and the central limit theorem for martingales

The previous algorithm takes advantage of the characterization of $\theta_\star$ as the unique root of the equation $\mathbb{E}\left((\theta-G)f^2(G)e^{-\theta.G+\frac{|\theta|^2}{2}}\right)=0$. Remarking that for all $\theta\in\mathbb{R}^d$, $\mathbb{E}\left((\theta-G)f^2(G)e^{-\theta.G+\frac{|\theta|^2}{2}}\right)=e^{|\theta|^2}\mathbb{E}\left((2\theta-G)f^2(G-\theta)\right)$, Lemaire and Pagès [23] characterize $\theta_\star$ as the unique root of $\mathbb{E}\left((2\theta-G)f^2(G-\theta)\right)=0$. When

$$\exists c,\alpha>0,\ \exists\beta\in[0,2),\ \forall x\in\mathbb{R}^d, |f(x)|\le ce^{\alpha|x|^\beta}$$

then the Robbins-Monro procedure

$$\forall n\in\mathbb{N},\ \theta_{n+1}=\theta_n-\gamma_n e^{-2^\beta\alpha|\theta_n|^\beta}(2\theta_n-G_{n+1})f^2(G_{n+1}-\theta_n)$$

is stable without projections and Theorem 0.4 still holds with this new definition for the sequence $(\theta_n)_{n\ge1}$. In particular, when f is bounded, α may be chosen equal to 0 and the factor $e^{-2^\beta\alpha|\theta_n|^\beta}$ is then equal to 1.

In [16], Kawai combines importance sampling with control variates remarking that for $\theta,\lambda\in\mathbb{R}^d$, the expectation and variance of the random variable

$$[f(G+\theta)-\lambda.(G+\theta)]e^{-\theta.G-\frac{|\theta|^2}{2}}$$

are respectively equal to $\mathbb{E}(f(G))$ and $v(\theta,\lambda)-\mathbb{E}^2(f(G))$ where

$$v(\theta,\lambda)\stackrel{\text{def}}{=}\mathbb{E}\left((f(G)-\lambda.G)^2e^{-\theta.G+\frac{|\theta|^2}{2}}\right).$$

The function v is strictly convex in θ for fixed λ and strictly convex in λ for fixed θ. Let $g(\theta)$ (resp. $h(\lambda)$) denote the unique vector in $\mathbb{R}^d$ s.t. $v(\theta,g(\theta))=\inf_{\lambda\in\mathbb{R}^d}v(\theta,\lambda)$ (resp. $v(h(\lambda),\lambda)=\inf_{\theta\in\mathbb{R}^d}v(\theta,\lambda)$). According to Kawai [16], the functions $v(\theta,g(\theta))$ and $v(h(\lambda),\lambda)$ are still strictly convex (but the proof of this statement does not seem correct) and there exists a unique $\theta_\star\in\mathbb{R}^d$ (resp. $\lambda_\star\in\mathbb{R}^d$) s.t. $v(\theta_\star,g(\theta_\star))=\inf_{\theta\in\mathbb{R}^d}v(\theta,g(\theta))$ (resp. $v(h(\lambda_\star),\lambda_\star)=\inf_{\lambda\in\mathbb{R}^d}v(h(\lambda),\lambda)$). He proposes for (θ_n,λ_n) a two-scale Robbins Monro procedure with Chen's projection technique and increment

$$\begin{pmatrix}-\gamma_n(\theta_n-G_{n+1})(f(G_{n+1})-\lambda_n.G_{n+1})^2e^{-\theta_n.G_{n+1}+\frac{|\theta_n|^2}{2}}\\ 2\tilde\gamma_n(f(G_{n+1})-\lambda_n.G_{n+1})G_{n+1}e^{-\theta_n.G_{n+1}+\frac{|\theta_n|^2}{2}}\end{pmatrix},$$

where $\tilde\gamma_n$ is another sequence of positive steps s.t. $\sum_n\tilde\gamma_n=+\infty$ and $\sum_n\tilde\gamma_n^2<+\infty$. The sequence (θ_n,λ_n) converges a.s. to $(\theta_\star,g(\theta_\star))$ or $(h(\lambda_\star),\lambda_\star)$ depending on whether $\lim_{n\to\infty}\frac{\gamma_n}{\tilde\gamma_n}$ is equal to 0 or $+\infty$. Moereover the analogue of Theorem 0.4 holds in this setting, the estimator of $\mathbb{E}(f(G))$ being defined as

$$M_n=\frac{1}{n}\sum_{j=1}^n[f(G+\theta_{j-1})-\lambda_{j-1}.(G+\theta_{j-1})]e^{-\theta_{j-1}.G_j-\frac{|\theta_{j-1}|^2}{2}}.$$

In [17], Kawai adapts the previous algorithm when the Gaussian random vector G is replaced by an infinitely divisible random vector (stochastic approximation by Robbins-Monro procedures of the parameter θ only is treated in [18]). In finance, problems involving such vectors arise for instance when the Brownian motion driving continuous time models is replaced by a Lévy process. Kawai pays particular attention to the case of independent gamma distributed components. This particular distribution has the following nice property: after the exponential change of measure (also called Esscher transform) considered in the present section, the law of a gamma random variable is the same as the law of this random variable multiplied by a constant under the original probability measure. In comparison with the Gaussian case, addition is replaced by multiplication.

Let us finally mention that an adaptive simulated annealing procedure has been recently developed by del Baño Rollin and Lázaro-Camí [7] to optimize antithetic variates. More precisely, using appropriate coordinates on the orthogonal group, the authors propose a Robbins-Monro procedure with an additional noise to compute a sequence $(O_n)_{n\geq 1}$ of orthogonal matrices converging to $O_\star$ minimizing $\mathbb{E}(f(G)f(OG))$ other all orthogonal matrices O. The additional noise, obtained from a sequence $(\tilde{G}_j)_{j\geq 1}$ of random vectors i.i.d. according to $\mathcal{N}(0, I_d)$ independent of $(G_j)_{j\geq 1}$, vanishes when n tends to infinity and avoids that the algorithm remains trapped in a critical point at which $\mathbb{E}(f(G)f(OG))$ is not minimal. The derived estimator

$$M_n = \frac{1}{4n}\sum_{j=1}^{n}\Big(f(G_j) + f(O_jG_j) + f(\tilde{G}_j) + f(O_j\tilde{G}_j)\Big)$$

of $\mathbb{E}(f(G))$ is then a.s. convergent and asymptotically normal with asymptotic variance $\frac{1}{4}\left(\mathrm{Var}(f(G)) + \mathrm{Cov}(f(G), f(OG))\right)$.

0.2.2 Robust sample average optimization

In order to save computation time, we introduce in [15] a parameter reduction. Indeed, numerical simulations show that, for a model driven by a Brownian motion, it is not useful to use different parameters for the increments of a single Brownian component. Let $A \in \mathbb{R}^{d\times d'}$ be a matrix with rank $d' \leq d$. We define $\tau_\star$ as the unique minimizer of the strongly convex and continuous function $\mathbb{R}^{d'} \ni \tau \mapsto v(A\tau)$. The sample average approximation of $v(A\tau)$ is given by $v_n(A\tau)$, where the C^∞ function

$$v_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} f^2(G_j)e^{-\theta.G_j + \frac{|\theta|^2}{2}}$$

is strongly convex as soon as $f(G_j) \neq 0$ for some $j \in \{1, \ldots, n\}$ which holds a.s. for n large enough by (0.1). The unique minimizer τ_n of $\tau \mapsto v_n(A\tau)$ is characterized by

the equality $\nabla_\tau v_n(A\tau)=0$, which also writes $\nabla_\tau u_n(\tau)=0$, where

$$u_n(\tau)=\frac{|A\tau|^2}{2}+\log\left(\sum_{j=1}^n f^2(G_j)e^{-A\tau\cdot G_j}\right)$$

$$\nabla_\tau u_n(\tau)=A^*A\tau-\frac{\sum_{j=1}^n A^*G_j f^2(G_j)e^{-A\tau\cdot G_j}}{\sum_{j=1}^n f^2(G_j)e^{-A\tau\cdot G_j}}$$

$$\nabla^2_\tau u_n(\tau)=A^*A+\frac{\sum_{j=1}^n A^*G_jG_j^*Af^2(G_j)e^{-A\tau\cdot G_j}}{\sum_{j=1}^n f^2(G_j)e^{-A\tau\cdot G_j}}$$
$$-\frac{\left(\sum_{j=1}^n A^*G_j f^2(G_j)e^{-A\tau\cdot G_j}\right)\left(\sum_{j=1}^n A^*G_j f^2(G_j)e^{-A\tau\cdot G_j}\right)^*}{\left(\sum_{j=1}^n f^2(G_j)e^{-A\tau\cdot G_j}\right)^2}.$$

The lowest eigenvalue of the Hessian matrix $\nabla^2_\tau u_n$ is always larger than the one of A^*A. Therefore τ_n can easily and precisely be computed by a few iterations of Newton's algorithm using the above explicit expressions of $\nabla_\tau u_n$ and $\nabla^2_\tau u_n$. Notice that the computation of the gradient and the Hessian of u_n is not too time-consuming since the points G_i, at which the payoff function f is evaluated, remain constant during the optimization procedure.

Convergence of τ_n to $\tau_\star$ is a consequence of classical results concerning M-estimators.

Proposition 0.5 *1. Under* (0.1), τ_n *and* $v_n(A\tau_n)$ *converge a.s. to* $\tau_\star$ *and* $v(A\tau_\star)$.

2. If moreover $\forall\theta\in\mathbb{R}^d, \mathbb{E}\left(f^4(G)e^{-\theta.G}\right)<+\infty$, *then* $\sqrt{n}(\tau_n-\tau_\star)\xrightarrow{\mathcal{L}}\mathcal{N}_{d'}(0,B^{-1}CB^{-1})$ *where* $B=A^*\nabla^2_\theta v(A\tau_\star)A$ *and* $C=\mathrm{Cov}\left(A^*(A\tau_\star-G)f^2(G)e^{-A\tau_\star\cdot G+\frac{|A\tau_\star|^2}{2}}\right)$.

In [15], we obtain convergence of $M_n(A\tau_n)$ to the expectation $\mathbb{E}(f(G))$ assuming that f is continuous and satisfies some growth assumption (see Theorem 0.7 below). When $d'=1$, continuity may be replaced by a monotonicity assumption introduced in the next definition.

Definition 0.6 We say that a function $h:\mathbb{R}^d\to\mathbb{R}$

- is A-nondecreasing (resp. A-nonincreasing) if
$$\forall x\in\mathbb{R}^d,\ \tau\in\mathbb{R}\mapsto h(x+A\tau) \text{ is nondecreasing (resp. nonincreasing),}$$
- is A-monotonic if it is either A-nondecreasing or A-nonincreasing,
- belongs to $\mathcal{V}_A$ if h may be decomposed as the sum of two A-monotonic functions h_1 and h_2 such that
$$\exists\lambda>0,\ \exists\beta\in[0,2),\ \forall x\in\mathbb{R},\ |h_i(x)|\le\lambda e^{|x|^\beta} \text{ for } i=1,2.$$

When $d=1$, $\mathcal{V}_1$ simply consists of functions with finite variation satisfying the previous growth assumption. The asymptotic properties of $M_n(A\tau_n)$ stated in the next theorem are proved in [15].

Theorem 0.7 *Assume* (0.1) *and that f admits a decomposition $f = f_1 + 1_{\{d'=1\}} f_2$ with*

1. *f_1 a continuous function s.t. $\forall M > 0, \mathbb{E}\left(\sup_{|\theta| \leq M} |f_1(G+\theta)| \right) < +\infty$,*
2. *$f_2 \in \mathcal{V}_A$ defined above.*

Then, for any deterministic integer-valued sequence $(\nu_n)_n$ going to ∞ with n, $M_n(A\tau_{\nu_n})$ converges a.s. to $\mathbb{E}(f(G))$.
Assume (0.1), *$\forall \theta \in \mathbb{R}^d$, $\mathbb{E}\left(f^4(G)e^{-\theta.G}\right) < +\infty$ and that f admits a decomposition $f = f_1 + f_2 + 1_{\{d'=1\}} f_3$ with*

1. *f_1 a C^1 function s.t.*

$$\forall M > 0,\ \mathbb{E}\left(\sup_{|\theta| \leq M} |f_1(G+\theta)| + \sup_{|\theta| \leq M} |\nabla f_1(G+\theta)| \right) < +\infty,$$

2. *$\exists \alpha \in \left((\sqrt{d'^2 + 8d'} - d')/4, 1 \right], \beta \in [0,2), \lambda > 0$,*

$$\forall x, y \in \mathbb{R}^d, |f_2(x) - f_2(y)| \leq \lambda e^{|x|^\beta \vee |y|^\beta} |x-y|^\alpha,$$

3. *$f_3 \in \mathcal{V}_A$.*

Then $\sqrt{n}(M_n(A\tau_n) - \mathbb{E}(f(G))) \xrightarrow{\mathcal{L}} \mathcal{N}_1\left(0, v(A\tau_\star) - \mathbb{E}^2(f(G))\right)$.

In contrast to the estimator M_n constructed using Robbins-Monro procedures in the previous section, there is no martingale structure for $M_n(A\tau_{\nu_n})$. This explains why we need some regularity assumptions on the function f. Except for $d' = 1$, asymptotic normality with optimal asymptotic variance $v(A\tau_\star) - \mathbb{E}^2(f(G))$ requires more regularity on f than a.s. convergence. Note that $\frac{\sqrt{d'^2+8d'}-d'}{4}$ is increasing with d', equals $\frac{1}{2}$ for $d' = 1$ and converges to 1 as $d' \to \infty$. Therefore the choice $\alpha = 1$ is always possible for f_2. So all the financial payoffs except the discontinuous ones (barrier or digital options) satisfy the assumption made on f_2 to ensure the asymptotic normality of the adaptive estimator $M_n(A\tau_n)$. If $\mathrm{Var}(f(G)) > 0$, then the previous results imply that

$$\sqrt{\frac{n}{v_n(A\tau_n) - M_n^2(A\tau_n)}}(M_n(A\tau_n) - \mathbb{E}(f(G))) \xrightarrow{\mathcal{L}} \mathcal{N}_1(0,1),$$

and one may easily derive confidence intervals for $\mathbb{E}(f(G))$.

The numerical experiments performed in [15] suggest that strong convergence and asymptotic normality of $M_n(A\tau_n)$ still hold under less restrictive assumptions on f than those stated in the previous theorem.

0.3 Stratified sampling

We are interested in the computation of $c = \mathbb{E}(f(X))$ where X is an $\mathbb{R}^d$-valued random vector and $f : \mathbb{R}^d \to \mathbb{R}$ a measurable function such that $\mathbb{E}(f^2(X)) < \infty$. We suppose

that $(A_i)_{1\le i\le I}$ is a partition of $\mathbb{R}^d$ into I *strata* such that $p_i = \mathbb{P}[X \in A_i]$ is known explicitly for $i \in \{1,\dots,I\}$. Up to removing some strata, we assume from now on that p_i is positive for all $i \in \{1,\dots,I\}$. The stratified Monte-Carlo estimator of c (see [12, p.209-235] and the references therein for a detailed presentation) is based on the equality $\mathbb{E}(f(X)) = \sum_{i=1}^I p_i \mathbb{E}(f(X^i))$ where X^i denotes a random variable distributed according to the conditional law of X given $X \in A_i$. Indeed, when the variables X_i can be simulated, it is possible to estimate each expectation in the right-hand side using n_i i.i.d drawings of X^i. Let $n = \sum_{i=1}^I n_i$ be the total number of drawings (in all the strata) and $q_i = n_i/n$ denote the proportion of drawings made in stratum i.

Then $\widehat{c}$ is defined by

$$\widehat{c} = \sum_{i=1}^I \frac{p_i}{n_i} \sum_{j=1}^{n_i} f(X_j^i) = \frac{1}{n} \sum_{i=1}^I \frac{p_i}{q_i} \sum_{j=1}^{q_i n} f(X_j^i),$$

where for each i the X_j^i's, $1 \le j \le n_i$, are distributed like X^i, and all the X_j^i's, for $1 \le i \le I$, $1 \le j \le n_i$ are drawn independently. This stratified sampling estimator can be implemented for instance when X is distributed according to the standard normal law on $\mathbb{R}^d$, $A_i = \{x \in \mathbb{R}^d : y_{i-1} \le \theta.x < y_i\}$ where $-\infty = y_0 < y_1 < \ldots < y_{I-1} < y_I = +\infty$ and $\theta \in \mathbb{R}^d$ is such that $|\theta| = 1$. Indeed, then one has $p_i = N(y_i) - N(y_{i-1})$ with $N(.)$ denoting the cumulative distribution function of the one-dimensional normal law and when U is uniformly distributed on $[0,1]$ and independent from X, then

$$X + (N^{-1}[N(y_{i-1}) + U(N(y_i) - N(y_{i-1}))] - \theta.X)\theta$$

follows the conditional law of X given $y_{i-1} \le \theta.X < y_i$.

We have $\mathbb{E}(\widehat{c}) = c$ and

$$\mathrm{Var}(\widehat{c}) = \sum_{i=1}^I \frac{p_i^2\sigma_i^2}{n_i} = \frac{1}{n}\sum_{i=1}^I \frac{p_i^2\sigma_i^2}{q_i} = \frac{1}{n}\sum_{i=1}^I \Big(\frac{p_i\sigma_i}{q_i}\Big)^2 q_i \ge \frac{1}{n}\Big(\sum_{i=1}^I \frac{p_i\sigma_i}{q_i} q_i\Big)^2, \quad (0.4)$$

where $\sigma_i^2 = \mathrm{Var}(f(X^i)) = \mathrm{Var}(f(X)|X \in A_i)$ for all $1 \le i \le I$.

In the sequel, we assume $\sigma_{i_0} > 0$ for at least one index i_0.

Let $(X_j)_{j\ge1}$ be i.i.d. drawings of X. The variance of the crude Monte Carlo estimator $\frac{1}{n}\sum_{j=1}^n f(X_j)$ of $\mathbb{E}(f(X))$ is

$$\frac{1}{n}\left(\sum_{i=1}^I p_i(\sigma_i^2 + \mathbb{E}^2(f(X^i))) - \left(\sum_{i=1}^I p_i \mathbb{E}(f(X^i))\right)^2\right) \ge \frac{1}{n}\sum_{i=1}^I p_i\sigma_i^2.$$

For given strata, the stratified estimator achieves variance reduction if the allocations n_i or equivalently the proportions q_i are properly chosen. For instance, for the so-called proportional allocation $q \equiv p$, the variance of the stratified estimator is equal to the previous lower bound of the variance of the crude Monte Carlo estimator. For the *optimal allocation* $q_i^\star \stackrel{\text{def}}{=} p_i\sigma_i / \sum_{j=1}^I p_j\sigma_j$, $1 \le i \le I$, the lower-bound in (0.4) is attained. Then

$$\mathrm{Var}(\widehat{c}) = \frac{1}{n}\Big(\sum_{i=1}^I p_i\sigma_i\Big)^2 \stackrel{\text{def}}{=} \frac{\sigma_\star^2}{n}.$$

In general, when the conditional expectations $\mathbb{E}(f(X)|X \in A_i) = \mathbb{E}(f(X^i))$ are unknown, then so are the conditional variances σ_i^2. Therefore optimal allocation of the drawings is not feasible at once. One can of course estimate the conditional variances and the optimal proportions by a first Monte Carlo algorithm and run a second Monte Carlo procedure with drawings independent from the first one to compute the stratified estimator corresponding to these estimated proportions. But why not use the drawings made in the first Monte Carlo procedure also for the final computation of the conditional expectations?

Instead of running two successive Monte Carlo procedures, one can think to obtain a first estimation of the σ_i's, using the first drawings of the X^i's made to compute the stratified estimator. One could then estimate the optimal allocations before making further drawings allocated in the strata according to these estimated proportions. One can next obtain another estimation of the σ_i's, compute again the allocations and so on. This is the principle of the adaptive allocation procedure proposed in [10] and described in the next section. Then, we will present the adaptive algorithm proposed in [9] in order to optimize the strata themselves.

0.3.1 Adaptive optimal allocation

Let N^k (resp. N_i^k) denote the total number of random drawings X_j^i made in all the strata (resp. in stratum i) at the end of step k of the following algorithm :

1. At step 1, allocate the N^1 first drawings in the strata proportionally to the p_i and estimate $\mathbb{E}(f(X^i))$ and σ_i, $1 \leq i \leq I$,
2. At the beginning of step $k \geq 2$, compute the vector $(n_1, \ldots, n_I) \in \mathbb{R}_+^I$ obtained by allocating the $N^k - N^{k-1}$ new drawings

   - either proportionally to the estimations $p_i \widehat{\sigma}_i^{k-1} / \sum_{l=1}^I p_l \widehat{\sigma}_l^{k-1}$ of the $q_i^\star$ available at the end of step $k-1$,
   - or in order to minimize the estimated variance $\sum_{i=1}^I (p_i \widehat{\sigma}_i^{k-1})^2 / N_i^k$ of the stratified estimator after step k under the constraints $\sum_{i=1}^I N_i^k = N^k$ and $\forall i$, $N_i^k \geq N_i^{k-1}$. The explicit solution of this constrained optimization problem is given in [10].

   Then convert $(n_1, \ldots, n_I)$ to $\mathbb{N}^I$ by the following rounding procedure preserving the sum : $n_i^k = \lfloor \sum_{l=1}^i n_l \rfloor - \lfloor \sum_{l=1}^{i-1} n_l \rfloor$ and allocate n_i^k new drawings in stratum i. Refine the estimations $\hat{c}_i^k$ and $\widehat{\sigma}_i^k$ of $\mathbb{E}(f(X^i))$ and σ_i using these new drawings.

In fact, one has to modify this algorithm in order to enforce at least one drawing in each stratum at each step. Indeed, if $\widehat{\sigma}_{i_0}^1 = 0$ whereas $\sigma_{i_0} > 0$, then no drawings are made after step $k = 1$ in the stratum i_0 and $\frac{1}{N_{i_0}^k} \sum_{j=1}^{N_{i_0}^k} f(X_j^{i_0}) = \frac{1}{N_{i_0}^1} \sum_{j=1}^{N_{i_0}^1} f(X_j^{i_0})$ does not converges to $\mathbb{E}(f(X^{i_0}))$ when $k \to +\infty$ which prevents the stratified estimator $\sum_{i=1}^I \frac{p_i}{N_i^k} \sum_{j=1}^{N_i^k} f(X_j^i)$ from converging to $\mathbb{E}(f(X))$. Choosing the sequence $(N^k)_{k \geq 1}$ so that $N^k \geq N^{k-1} + I$ for all $k \geq 2$, enforcing one drawing in each stratum at each step k, and allocating the remaining $N^k - N^{k-1} - I$ drawings according the previous

procedure permits to overcome this difficulty. Then $\forall 1 \leq i \leq I,\ \forall k \geq 1,\ N_i^k \geq k$ and the following result is proved in [10] by first checking that the proportions $\frac{N_i^k}{N^k}$ converge a.s. to the optimal ones $q_i^\star$ as $k \to \infty$ and then applying the central limit theorem for martingales :

Theorem 0.8

$$\mathbb{P}\left(\sum_{i=1}^{I} \frac{p_i}{N_i^k} \sum_{j=1}^{N_i^k} f(X_j^i) \xrightarrow[k\to\infty]{} \mathbb{E}(f(X))\right) = 1.$$

If, moreover, $\sigma_{i_0} > 0$ *for some* $i_0 \in \{1, \ldots, I\}$ *and* $\lim_{k\to+\infty} \frac{k}{N^k} = 0$*, then*

$$\sqrt{N^k}\left(\sum_{i=1}^{I} \frac{p_i}{N_i^k} \sum_{j=1}^{N_i^k} f(X_j^i) - \mathbb{E}(f(X))\right) \xrightarrow[k\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma_\star^2\right)$$

with $\sigma_\star^2 = \left(\sum_{i=1}^{I} p_i\sigma_i\right)^2$ *the asymptotic variance for the optimal allocation.*

As a consequence, $\frac{\sqrt{N^k}}{\sum_{i=1}^{I} p_i\widehat{\sigma}_i^k}\left(\sum_{i=1}^{I} \frac{p_i}{N_i^k} \sum_{j=1}^{N_i^k} f(X_j^i) - \mathbb{E}(f(X))\right) \xrightarrow[k\to\infty]{\mathcal{L}} \mathcal{N}_1(0,1)$ and one may easily construct confidence intervals for $\mathbb{E}(f(X))$. Numerical experiments performed in [10] on the pricing of arithmetic average Asian options in the Black-Scholes model show that adaptive allocation permits to divide the variance obtained with proportional allocation by a factor up to 50.

Another stratified sampling algorithm in which the optimal proportions and the conditional expectations are estimated using the same drawings has been proposed in [4] for quantile estimation. More precisely, for a total number of drawings equal to N, the authors suggest to allocate the N^γ with $0 < \gamma < 1$ first ones proportionally to the probabilities of the strata and then use the estimation of the optimal proportions obtained from these first drawings to allocate the $N - N^\gamma$ remaining ones. Their stratified estimator is also asymptotically normal with asymptotic variance equal to the optimal one. In practice, N is finite and it seems better to take advantage of all the drawings and not only the N^γ first ones to modify adaptively the allocation between the strata.

0.3.2 Adaptive optimization of the strata for normal random vectors

Let us now consider the problem of optimally designing the strata when they are parametrized in the following way : for $1 \leq i \leq I$, $A_i = \left\{x \in \mathbb{R}^d : \theta.x \in [y_{i-1}, y_i)\right\}$ where

$$-\infty = y_0 < y_1 < \cdots < y_{I-1} < y_I = +\infty \text{ and } \theta \in \mathbb{R}^d \text{ is s.t. } |\theta| = 1.$$

In [9], we address a more general parametrization where the strata are defined by hyperrectangles but the present section is devoted to the particular case of a single stratification direction.

Our aim is to approximate the parameters $(\theta, y_1, \ldots, y_{I-1})$ defining the strata which minimize the standard deviation $\sigma_\star = \sum_{i=1}^{I} p_i\sigma_i$ obtained either by optimal allocation

or with the adaptive allocation algorithm described above. This standard deviation $\sigma_\star$ is equal to

$$\sum_{i=1}^{I} \sqrt{(\nu_\theta(1,y_i)-\nu_\theta(1,y_{i-1}))(\nu_\theta(f^2,y_i)-\nu_\theta(f^2,y_{i-1}))-(\nu_\theta(f,y_i)-\nu_\theta(f,y_{i-1}))^2}.$$

where $\nu_\theta(h,y) \stackrel{\text{def}}{=} \mathbb{E}(h(X)1_{\{\theta.X\le y\}})$ for $y\in\mathbb{R}$ and $h:\mathbb{R}^d\to\mathbb{R}$ such that $h(X)$ is integrable. According to the following Lemma proved in [9] it is possible to express the gradient of $\nu_\theta(h,y)$ in terms of conditional expectations.

Lemma 0.9 *When $\theta.X$ admits a density p_θ w.r.t. the Lebesgue measure on the real line and under further technical regularity assumptions not precised here,*

$$\partial_y\nu_\theta(h,y) = p_\theta(y)\mathbb{E}(h(X)|\theta.X=y)$$
$$\nabla_\theta\nu_\theta(h,y) = -p_\theta(y)\mathbb{E}(Xh(X)|\theta.X=y).$$

We suppose from now on that $X\sim\mathcal{N}_d(0,I_d)$ is a standard normal random vector. Then $p_\theta(y)=\frac{e^{-y^2/2}}{\sqrt{2\pi}}$ and

$$\forall i\in\{1,\dots,I\},\ \mathbb{E}(h(X)|\theta.X=y)=\mathbb{E}[h(X^i+(y-\theta.X^i)\theta)].$$

At each step k of the above optimal allocation algorithm, this enables us

1. to estimate the gradient of $\sigma_\star$ w.r.t. $(y_1,\dots,y_{I-1})$ and θ using the orthogonal projections on the boundaries of the random drawings X_j^i made at this step in the strata,
2. to perform a gradient descent step to update the stratification direction and boundaries.

In practice, the differences N^k-N^{k-1} should be large enough not to increase significantly the computation time needed to calculate the crude Monte Carlo estimator. As a consequence, the Monte Carlo estimator of the gradient is precise and the optimization of the strata parameters is rather a noisy gradient descent than a stochastic algorithm. According to our numerical experiments, optimizing the direction θ works : the gradient procedure converges to some limit and this ensures effective variance reduction. On examples involving discontinuous payoffs such as barrier options, the optimal direction computed with our algorithm is significantly different and more efficient than the one derived analytically in [13] using some large deviations asymptotics. Numerical optimization of the strata boundaries was far less convincing. In [9], we explain this numerical observation by the following asymptotic analysis performed in the limit $I\to\infty$. We parametrize the boundaries by a positive probability density g on $\mathbb{R}$ with c.d.f. $G(y)=\int_{-\infty}^{y}g(z)dz$ and set $y_i=G^{-1}(\frac{i}{I})$ for $i\in\{0,\dots,I\}$.

Theorem 0.10 • *Let $d\ge 2$. If for $h\in\{p_\theta, p_\theta\times\mathbb{E}(f(X)|\theta.X=\cdot), p_\theta\times\mathbb{E}(f^2(X)|\theta.X=\cdot)\}$, $\int_{\mathbb{R}}\frac{h^2}{g}(y)dy<+\infty$, then* $\lim_{I\to\infty}\sigma_\star(I)=\mathbb{E}\left(\sqrt{\mathrm{Var}(f(X)|\theta.X)}\right)$.

- *When $d = 1$, and f is a locally bounded function on the real line with a locally integrable distribution derivative f' such that* $\mathrm{esssup}_{dy} \frac{p_\theta + |f'|}{g}(y) < +\infty$, *then* $\lim_{I\to\infty} I\sigma_\star(I) = \frac{1}{\sqrt{12}} \int_{\mathbb{R}} \frac{|f'|p_\theta}{g}(y)dy$.

The fact that, in the practical case $d \geq 2$, the limit does not depend on g means that under optimal or adaptive allocation, the choice of the boundaries of the strata is not important when the number of strata is large. So only the stratification direction θ should be optimized.

Note that the optimized direction θ computed by our algorithm can be used to design Latin hypercube or Quasi Monte Carlo (see [12]) estimators of $\mathbb{E}(f(X))$. When X is a standard normal random vector, for any orthogonal matrix $O \in \mathbb{R}^{d\times d}$, $\mathbb{E}(f(X)) = \mathbb{E}(f(OX))$, but the convergence properties of Latin hypercube or QMC estimators associated with the variable $f(OX)$ crucially depend on O. Unfortunately, it is very difficult to estimate these rates of convergence and adaptive optimization of the matrix O seems unreachable. As Latin hypercube or QMC methods somehow consist in stratifying each canonical direction, choosing the first column of O equal to θ should be effective.

Bibliography

[1] Bouhari Arouna, *Adaptative Monte Carlo method, a variance reduction technique*, Monte Carlo Methods Appl. 10 (2004), pp. 1–24. MR MR2054568 (2004m:62159)

[2] ———, *Robbins Monro algorithms and variance reduction in finance*, J. of Comput. Finance 7 (Winter 2003/04), pp. 35–61.

[3] Keith Baggerly, Dennis Cox, and Rick Picard, *Exponential convergence of adaptive importance sampling for Markov chains*, J. Appl. Probab. 37 (2000), pp. 342–358. MR MR1780995 (2001e:65008)

[4] Claire Cannamela, Josselin Garnier, and Bertrand Looss, *Controlled stratification for quantile estimation*, Ann. Appl. Stat. (To appear).

[5] Han Fu Chen, Guo Lei, and Ai Jun Gao, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds*, Stochastic Process. Appl. 27 (1988), pp. 217–231. MR MR931029 (89b:62180)

[6] Han Fu Chen and Yun Min Zhu, *Stochastic approximation procedures with randomly varying truncations*, Sci. Sinica Ser. A 29 (1986), pp. 914–926. MR MR869196 (88b:62158)

[7] Sebastian del Baño Rollin and Joan-Andreu Lázaro-Camí, *Antithetic variates in higher dimension*, Preprint ArXiv 0902.4211 (2009).

[8] Paul Dupuis and Hui Wang, *Dynamic importance sampling for uniformly recurrent Markov chains*, Ann. Appl. Probab. 15 (2005), pp. 1–38. MR MR2115034 (2006b:60042)

[9] Pierre Étoré, Gersende Fort, Benjamin Jourdain, and Éric Moulines, *On adaptive stratification*, Preprint ArXiv:0809.1135 (2008).

[10] Pierre Étoré and Benjamin Jourdain, *Adaptive optimal allocation in stratified sampling methods*, Methodol. Comput. Appl. Probab. (To appear).

[11] Michael B. Giles and Ben J. Waterhouse, *Multilevel quasi-Monte Carlo path simulation*, Radon Series Comp. Appl. Math. 8 (2009).

[12] Paul Glasserman, *Monte Carlo methods in financial engineering*, Applications of Mathematics (New York), vol. 53, Springer-Verlag, New York, 2004, Stochastic Modelling and Applied Probability. MR MR1999614 (2004g:65005)

[13] Paul Glasserman, Philip Heidelberger, and Perwez Shahabuddin, *Asymptotically optimal importance sampling and stratification for pricing path-dependent options*, Math. Finance 9 (1999), pp. 117–152. MR MR1849001 (2002m:91035)

[14] Shane G. Henderson and Burt Simon, *Adaptive simulation using perfect control variates*, J. Appl. Probab. 41 (2004), pp. 859–876. MR MR2074828 (2005h:65009)

[15] Benjamin Jourdain and Jérôme Lelong, *Robust adaptive importance sampling for normal random vectors*, Ann. Appl. Probab. (To appear).

[16] Reiichiro Kawai, *Adaptive Monte Carlo variance reduction with two-time-scale stochastic approximation*, Monte Carlo Methods Appl. 13 (2007), pp. 197–217. MR MR2349428 (2008h:62195)

[17] ———, *Adaptive Monte Carlo variance reduction for Lévy processes with two-time-scale stochastic approximation*, Methodol. Comput. Appl. Probab. 10 (2008), pp. 199–223. MR MR2399681

[18] ———, *Optimal importance sampling parameters search for Lévy processes via stochastic approximation*, SIAM J. Numer. Anal. 47 (2008), pp. 293–307.

[19] Sujin Kim and Shane G. Henderson, *Adaptive control variates*, Proceedings of the 2004 Winter Simulation Conference (2004), pp. 621–629.

[20] ———, *Adaptive control variates for finite-horizon simulation*, Math. Oper. Res. 32 (2007), pp. 508–527. MR MR2348231 (2008i:65005)

[21] Stephen Lavenberg, Thomas Moeller, and Peter Welch, *Statistical Results on Control Variables with Application to Queuing Network Simulation*, Oper. Res. 30 (1982), pp. 182–202.

[22] Jérôme Lelong, *Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions*, Stat. Probab. Letters 78 (2008), pp. 2632–2636.

[23] Vincent Lemaire and Gilles Pagès, *Unconstrained Recursive Importance Sampling*, Preprint ArXiv:0807.0762 (2008).

[24] Barry L. Nelson, *Control variate remedies*, Oper. Res. 38 (1990), pp. 974–992. MR MR1095954

[25] Teemu Pennanen and Matti Koivu, *An adaptive importance sampling technique*, Monte Carlo and quasi-Monte Carlo methods 2004, Springer, Berlin, 2006, pp. 443–455. MR MR2208724 (2006k:65065)

[26] Yi Su and Michael Fu, *Optimal importance sampling in securities pricing*, J. Comput. Finance 5 (2002), pp. 26–50.

[27] Felicia Vázquez-Abad and Daniel Dufresne, *Accelerated simulation for pricing Asian options*, Proceedings of the 1998 Winter Simulation Conference (1998), pp. 1493–1500.

Author information

Benjamin Jourdain, Université Paris-Est, CERMICS, Project team MathFi ENPC-INRIA-UMLV, 6 et 8 avenue Blaise Pascal, 77455 Marne La Vallée, Cedex 2, France .
Email: jourdain@cermics.enpc.fr