On adaptive stratification

Pierre Etoré, Gersende Fort, Benjamin Jourdain, Éric Moulines

Noname manuscript No.
(will be inserted by the editor)

On adaptive stratification

Pierre Etore · Gersende Fort · Benjamin Jourdain · Eric Moulines

the date of receipt and acceptance should be inserted later

Abstract This paper investigates the use of stratified sampling as a variance reduction technique for approximating integrals over large dimensional spaces. The accuracy of this method critically depends on the choice of the space partition, the *strata*, which should be ideally fitted to the subsets where the functions to integrate is nearly constant, and on the allocation of the number of samples within each strata. When the dimension is large and the function to integrate is complex, finding such partitions and allocating the sample is a highly non-trivial problem. In this work, we investigate a novel method to improve the efficiency of the estimator "on the fly", by jointly sampling and adapting the strata and the allocation within the strata. The accuracy of estimators when this method is used is examined in detail, in the so-called asymptotic regime (*i.e.* when both the number of samples and the number of strata are large). We illustrate the use of the method for the computation of the price of path-dependent options in models with both constant and stochastic volatility. The use of this adaptive technique yields variance reduction by factors sometimes larger than 1000 compared to classical Monte Carlo estimators.

Acknowledgements This work has been written in honour of R. Rubinstein, for his 70th birthday. Most of the ideas used in this paper to reduce the variance of Monte Carlo estimator have been inspired by the pioneering work of R. Rubinstein on coupling simulation and stochastic optimization. These very fruitful ideas were a constant source of inspiration during this work.

This work is supported by the french National Research Agency (ANR) under the program ANR-05-BLAN-0299.

P. Etore
CMAP, École Polytechnique,
Route de Saclay, 91128 Palaiseau Cedex
Tel.: +33 (0)1 69 33 45 67
E-mail: etore@cmap.polytechnique.fr

G. Fort and E. Moulines
Institut des Télécoms, Télécom ParisTech,
46 Rue Barrault, 75634 Paris Cedex 13, France
Tel.: +33 (0)1 45 81 77 82
E-mail: surname.name@telecom-paristech.fr

B. Jourdain
Université Paris-Est, CERMICS, Projet MathFi ENPC-INRIA-UMLV,
6 et 8 avenue Blaise Pascal, 77455 Marne La Vallée, Cedex 2, France
Tel.: +33 (0)1 64 15 35 67 E-mail: benjamin.jourdain@enpc.fr

1 Introduction

A number of problems in statistics, operation research and mathematical finance boils down to the evaluation of the expectation (or higher order moments) of a random variable $\phi(Y)$, known to be a complicated real valued function of a vector $Y = (Y_1, \ldots, Y_d)$ of independent random variables. In our applications, we will mainly focus on simulations driven by a sequence of independent standard normal random variables, in situations where the dimension d is very large. We have in particular in mind the computations of moments of functionals of diffusion processes; the dimension d can be very large if the mapping ϕ is *path-dependent* (ϕ a general functional defined on the space of continuous function) and the solutions of the diffusion process cannot be explicitly computed and simulated (*i.e.* should be approximated using for example the Euler or an higher-order model discretization scheme). Such problems arise in particular in computational finance for the pricing of path-dependent options, either when the number of underlying assets is large, or when additional source of randomness is present such as in stochastic volatility models, which in general preclude the existence of explicit solutions for the multi-dimensional diffusions modeling the price of the asset.

Since the distribution of $\phi(Y)$ is most often impossible to obtain in closed analytic form, then a classical approach is to resort to Monte Carlo integration. In its most elementary form, a random sample of points $Y_1, \ldots, Y_d$ is drawn from $\mathbb{R}^d$, ϕ is evaluated at each of these points and the moments of interest are estimated from these values. Intuitively, rather than calculate ϕ at independently sampled points, it seems to be a better option to dissect $\mathbb{R}^d$ into mutually exclusive subsets (or *strata*) and ensure that ϕ is evaluated for a prescribed and appropriate number of points in each stratum. This is referred to as *stratified sampling*. Good reviews of the method include Glasserman (2004) (with an emphasis on finance applications), Asmussen and Glynn (2007), Rubinstein and Kroese (2008).

The main purpose of this paper is to discuss a way of dissecting the space into strata and sampling from the strata, adapted to the case where Y is a (typically large-dimensional) standard Gaussian vector. We also address the accuracy of estimators when this method of sampling is used, and give conditions upon which the variance reduction is most effective.

Determining an efficient dissection in a large dimensional space is a highly non-trivial problem. We shall consider a computationally inexpensive way to overcome this difficulty, which uses a kind of dimensionality reduction. This method makes use of one or more orthogonal directions, to induce a dissection of $\mathbb{R}^d$ with the right property. These directions and the associated allocation are learnt adaptively, while the simulations are performed. The advantage of the adaptive method, similar to those introduced for importance sampling by Rubinstein and Kroese (2004) is that information is collected as the simulations are done, and computations of means and variance of $\phi(Y)$ in strata are used to update the choice of these strata and of the allocation. We investigate in some details the asymptotic regime *i.e.* where the number of simulations and the number of the strata both go to infinity. We show that the variance of the estimator critically depends on the relations of the strata to the regions of the space where ϕ is "nearly" constant which can be fairly complex for example when pricing path-dependent basket options or when the underlying model of the asset is a multi-dimensional non-linear diffusion process.

The method is illustrated for pricing path-dependent options driven by high-dimensional gaussian vectors, combining adaptive importance sampling based on a change of drift together with the suggested adaptive stratification. The combination of these two methods, already advocated in an earlier work by Glasserman et al (1999), is very effective; nevertheless, these examples show that, contrary to what is suggested in this work, the asymptotical optimal drift vector is not always the most effective direction of stratification.

The paper is organized as follows. In section 2, an introduction to the main ideas of the stratification is presented. Section 3 addresses the behavior of the stratified estimator in the asymptotic regime (*i.e.* when both the number of samples and the number of strata go to infinity). The roles of the stratification directions, the strata in each direction of stratification and of the allocation within each strata are evidenced. In section 4, an algorithm is proposed to adapt the directions of stratifications and the allocation of simulations within each stratum. In Section 5, the proposed adaptive stratification procedure is illustrated using applications for the pricing of path-dependent options.

2 An introduction to stratification

Suppose we want to compute an expectation of the form $\mathrm{E}\left[\phi(Y)\right]$ where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a measurable function and Y is a $\mathbb{R}^d$-valued random variable. We assume hereafter that

$$\mathrm{E}\left[\phi^2(Y)\right] < +\infty \,. \tag{1}$$

In the examples we have in mind, ϕ is the payoff of a path-dependent option and $Y = (Y_1, \ldots, Y_d)$ is a typically large-dimensional standard Gaussian vector.

Stratified sampling is a variance reduction method which produces an alternative estimator of $\mathrm{E}[\phi(Y)]$ having smaller variance than the crude Monte Carlo estimator. Fully stratifying a random vector is typically infeasible in high dimension. We therefore focus on methods where the stratification is applied to a low-dimensional projection of the random vector Y. In a simulation driven by arbitrary random vectors, stratifying on a linear combination would typically be impractical because of the difficulty of sampling from the distribution of the vector conditional on a given linear combination, but in the Gaussian case, this conditional distribution is itself Gaussian which makes this approach practical.

We therefore consider a *stratification variable* of the form $\mu^T Y$ where μ is an orthonormal $(d \times m)$ matrix with $m \leq d$; recall that μ is orthonormal if $\mu^T \mu = \mathrm{Id}_m$ where Id_m is the identity matrix in dimension m and $\mu\mu^T$ is the orthonormal projector onto the range of the matrix μ. In all our examples, m is equal to one or two. Given a partition $\{\mathsf{S}_\mathbf{i}, \mathbf{i} \in \mathcal{I}\}$ of $\mathbb{R}^m$, the sample space of $\mu^T Y$, the sample space $\mathbb{R}^d$ of Y is divided into *strata* defined by

$$\mathsf{S}_{\mu,\mathbf{i}} \stackrel{\mathrm{def}}{=} \left\{ x \in \mathbb{R}^d, \mu^T x \in \mathsf{S}_\mathbf{i} \right\} \,, \quad \mathbf{i} \in \mathcal{I} \,. \tag{2}$$

The strata $\mathsf{S}_\mathbf{i}$ need not be a connected region or might have a curve surface in full generality, but will typically be a an hyperrectangle in all our applications. It is assumed in the sequel that the probability of the strata $\{p_\mathbf{i}, \mathbf{i} \in \mathcal{I}\}$

$$p_\mathbf{i}(\mu) \stackrel{\mathrm{def}}{=} \mathbb{P}\left(Y \in \mathsf{S}_{\mu,\mathbf{i}}\right) = \mathbb{P}\left(\mu^T Y \in \mathsf{S}_\mathbf{i}\right) \,, \tag{3}$$

are known; the dependence of the probability $p_\mathbf{i}(\mu)$ on the strata $\{\mathsf{S}_\mathbf{i}, \mathbf{i} \in \mathcal{I}\}$ is implicit. If Y is a large dimensional standard Gaussian vector and if $S_\mathbf{i}$ is an hyperrectangle, computing (3) is easy since in such case $\mu^T Y$ also is a standard Gaussian vector. Glasserman (2004, section 4.3, p. 223) (see also Section 5.1) presents a simple algorithm to sample according to the conditional distribution of Y given $\mu^T Y \in \mathsf{S}_\mathbf{i}$. Up to removing some strata, we may assume without loss of generality that $p_\mathbf{i}(\mu) > 0$, for any $\mathbf{i} \in \mathcal{I}$. For the special case where all the $p_\mathbf{i}(\mu)$ are equal, *i.e.* $p_\mathbf{i}(\mu) = |\mathcal{I}|^{-1}$, $\mathbf{i} \in \mathcal{I}$, we shall say that the strata are *equiprobable*.

Let M be the total number of draws and $\mathcal{Q} = \{q_\mathbf{i}, \mathbf{i} \in \mathcal{I}\}$ be an allocation vector (i.e. $q_\mathbf{i} \geq 0$ and $\sum_{\mathbf{i} \in \mathcal{I}} q_\mathbf{i} = 1$) : the number $M_\mathbf{i}$ of samples allocated to the $\mathbf{i}$-th stratum is given by

$$M_\mathbf{i} \stackrel{\mathrm{def}}{=} \left\lfloor M \sum_{\mathbf{j} \leq \mathbf{i}} q_\mathbf{j} \right\rfloor - \left\lfloor M \sum_{\mathbf{j} < \mathbf{i}} q_\mathbf{j} \right\rfloor \,, \quad \mathbf{i} \in \mathcal{I} \,, \tag{4}$$

where $\lfloor \cdot \rfloor$ denotes the lower integer part and by convention, $\sum_\emptyset q_\mathbf{j} = 0$ (it is assumed that the set of indices $\mathcal{I}$ is totally ordered: e.g. if $\mathcal{I}$ is a cartesian product of a set of totally ordered sets indexed by an ordinal, the order on $\mathcal{I}$ is the lexicographical one). If the number of points in each stratum is chosen to be proportional to the probability of the strata, the allocation is said to be *proportional*. Given the strata $\{\mathsf{S}_\mathbf{i}, \mathbf{i} \in \mathcal{I}\}$ and the allocation $\mathcal{Q}$, the *stratified estimator* with M draws is defined by

$$\sum_{\mathbf{i} \in \mathcal{I} : M_\mathbf{i} > 0} p_\mathbf{i}(\mu) \left\{ \frac{1}{M_\mathbf{i}} \sum_{j=1}^{M_\mathbf{i}} \phi(Y_{\mathbf{i},j}) \right\} \,, \tag{5}$$

where for each $\mathbf{i} \in \mathcal{I}$, $\{Y_{\mathbf{i},j}, j \leq M_\mathbf{i}\}$ are i.i.d. random variables distributed according to the conditional distribution of the vector Y given the strata, $\mathrm{P}\left[Y \in \cdot \,|\, \mu^T Y \in \mathsf{S}_\mathbf{i}\right]$. In addition, the random variables $\{Y_{\mathbf{i},j}, j \leq M_\mathbf{i}, \mathbf{i} \in \mathcal{I}\}$ are independent.

The stratified estimator is an unbiased estimator of $\mathrm{E}[\phi(Y)]$ if the $M_\mathbf{i}$'s are all positive (a sufficient condition is $M \geq \{\min_\mathbf{i} \rho_\mathbf{i}(\mu)\}^{-1}$). Its variance is given by

$$\sum_{\mathbf{i} \in \mathcal{I} : M_\mathbf{i} > 0} M_\mathbf{i}^{-1} p_\mathbf{i}^2(\mu) \sigma_\mathbf{i}^2(\mu) \tag{6}$$

where $\sigma^2_{\mathbf{i}}(\mu)$ is the conditional variance of the random vector $\phi(Y)$ given $\mu^T Y \in \mathsf{S}_{\mathbf{i}}$,

$$\sigma^2_{\mathbf{i}}(\mu) \stackrel{\text{def}}{=} \mathrm{E}\left[\phi^2(Y) \,\middle|\, \mu^T Y \in \mathsf{S}_{\mathbf{i}}\right] - \left(\mathrm{E}\left[\phi(Y) \,|\, \mu^T Y \in \mathsf{S}_{\mathbf{i}}\right]\right)^2 . \tag{7}$$

When M goes to infinity and the number of strata is either fixed or goes to infinity slowly enough, the variance of the stratified estimator is equivalent to

$$M^{-1} \sum_{\mathbf{i} \in \mathcal{I}: q_{\mathbf{i}} > 0} q_{\mathbf{i}}^{-1}\, p^2_{\mathbf{i}}(\mu) \sigma^2_{\mathbf{i}}(\mu) \tag{8}$$

(see e.g. Lemma 1 in Section 6.1 for a proof of this assertion).

The two key questions that arise in every application of the stratified sampling method are (i) the choice of the dissection of the space and (ii) for a fixed M, the determination of the number of samples $M_{\mathbf{i}}$ to be generated in each stratum $\mathbf{i}$. It is well-know (see *e.g.* (Fishman, 1996, Theorem4.15)) that, whatever the choice of the strata $\mathsf{S}_{\mu,\mathbf{i}}$ is, the stratification with proportional allocation always produces a variance reduction compared to the crude Monte Carlo. More ambitiously than just considering proportional allocation, the optimal allocation (in the sense of variance minimization) is obtained by minimizing the asymptotic variance (8) subject to the constraint $\sum_{\mathbf{i}\in\mathcal{I}} q_{\mathbf{i}} = 1$. The solution of this problem is given by:

$$q^\star_{\mathbf{i}}(\mu) \stackrel{\text{def}}{=} \frac{p_{\mathbf{i}}(\mu)\, \sigma_{\mathbf{i}}(\mu)}{\sum_{\mathbf{j}\in\mathcal{I}} p_{\mathbf{j}}(\mu)\sigma_{\mathbf{j}}(\mu)} . \tag{9}$$

Note that Eq. 6 reveals that the magnitude of variance reduction depends crucially on how widely dispersed the strata means $\mathrm{E}\left[\phi(Y) \,|\, Y \in \mathsf{S}_{\mu,\mathbf{i}}\right]$ are around the population mean $\mathrm{E}[\phi(Y)]$.

For a given stratification matrix μ, we refer to $\mathcal{Q}^\star(\mu) = \{q^\star_{\mathbf{i}}(\mu), \mathbf{i} \in \mathcal{I}\}$ as the *optimal stratification vector*. Of course, contrary to the proportions $p_{\mathbf{i}}(\mu)$, the conditional expectations $\mathrm{E}\left[\phi(Y) \,|\, Y \in \mathsf{S}_{\mu,\mathbf{i}}\right]$ are unknown and so are the conditional variances $\sigma^2_{\mathbf{i}}(\mu)$. Because the stratification matrix is also unknown, an adaptive procedure is required.

The simplest approach would be to estimate these conditional variance in a pilot run, to determine the optimal allocation vector from these estimates, and then to use this allocation vector in a second stage to determine the stratified estimator. Such a procedure is clearly suboptimal, since the results obtained in the pilot step are not fully exploited. This calls for a more sophisticated procedure, in the spirit of those used for adaptive importance sampling; see for example, Rubinstein and Kroese (2004) and Rubinstein and Kroese (2008). In these algorithms, the estimate of conditional variance and the stratification directions is gradually improved while computing the stratified estimator and estimating its variance. Of course, the limiting behavior of such estimators is more complex, because of the dependence between the successive draws and the definition of the strata themselves. Such algorithm extends the procedure by Etore and Jourdain (2007), who proposed to adaptively learn the optimal allocation vector for a set of given strata and derived a central limit theorem for the adaptive estimator (with the optimal asymptotic variance).

3 Asymptotic analysis of the stratification performance

We derive in this Section the asymptotic variance of the stratified estimator when both the total number of draws M and the number of strata (possibly depending upon M) tend to $+\infty$. The variance of the estimator depends on the stratification matrix μ, on the partition $\{\mathsf{S}_{\mathbf{i}}, \mathbf{i} \in \mathcal{I}\}$ of the sample space of $\mu^T Y$ and on the allocation $\mathcal{Q}$.

3.1 Notations and Assumptions

For any integer k, we denote by λ the Lebesgue measure on $\mathbb{R}^k$, equipped with its Borel sigma-field (the dependence in the dimension k is implicit). For a probability density h w.r.t the Lebesgue measure on $\mathbb{R}$, we denote by H the cumulative distribution function, and H^{-1} the *quantile* function, defined as the generalized inverse of H,

$$H^{-1}(u) = \inf\{x \in \{H > 0\} : H(x) \geq u\} , \quad \text{for any } u \in [0,1] ,$$

where, by convention, $\inf \emptyset = +\infty$. Let I be a positive integer. The choice of the strata boundaries is parameterized by an m-uplet $(g_1, \ldots, g_m)$ of probability densities on $\mathbb{R}$ in the following sense: for all m-uplet $\mathbf{i} = (i_1, \ldots, i_m) \in \{1, \cdots, I\}^m$,

$$\mathsf{S}_{\mathbf{i}} \stackrel{\text{def}}{=} \left\{ (x_1, \ldots, x_m) \in \mathbb{R}^m : G_k^{-1}\left(\frac{i_k - 1}{I}\right) < x_k \leq G_k^{-1}\left(\frac{i_k}{I}\right) \right\} . \tag{10}$$

We denote by $g(x_1, \ldots, x_m)$ the associated joint density:

$$g(x_1, \ldots, x_m) \stackrel{\text{def}}{=} \prod_{k=1}^{m} g_k(x_k) . \tag{11}$$

We consider allocations parameterized by a probability density $\chi : \mathbb{R}^m \to \mathbb{R}_+$ with respect to the Lebesgue measure by setting for all $\mathbf{i} \in \{1, \cdots, I\}^m$

$$q_{\mathbf{i}}(\chi) \stackrel{\text{def}}{=} \int_{\mathsf{S}_{\mathbf{i}}} \chi \, d\lambda \, ;$$

denote by $\mathcal{Q}_\chi = \{q_{\mathbf{i}}(\chi) \, , \mathbf{i} \in \{1, \ldots, I\}^m\}$ the associated allocation. Let μ be a $d \times m$ orthonormal matrix. We consider the stratification $\mathcal{S}(\mu) = \left\{\mathsf{S}_{\mu,\mathbf{i}}, \mathbf{i} \in \{1, \ldots, I\}^m\right\}$ of the space $\mathbb{R}^d$. Denote by $\varsigma^2_{I,M}(\mu, g, \mathcal{Q})$ the asymptotic variance of the stratified estimator, given by

$$\varsigma^2_{I,M}(\mu, g, \mathcal{Q}) \stackrel{\text{def}}{=} \sum_{\mathbf{i} \in \{1,\ldots,I\}^m : M_{\mathbf{i}} > 0} M_{\mathbf{i}}^{-1} \; p^2_{\mathbf{i}}(\mu) \sigma^2_{\mathbf{i}}(\mu) \, , \tag{12}$$

where the number of draws $M_{\mathbf{i}}$ is given by (4) and $p_{\mathbf{i}}(\mu), \sigma^2_{\mathbf{i}}(\mu)$, the probability and the conditional variance are given by (3), and (7), respectively. The dependence w.r.t. g and $\mathcal{Q}$ of $M_{\mathbf{i}}$, $p_{\mathbf{i}}(\mu)$ and $\sigma^2_{\mathbf{i}}(\mu)$ is implicit. We assume that the random variable $\mu^T Y$ possesses a density f_μ w.r.t. the Lebesgue measure (on $\mathbb{R}^m$). We consider the functions (taking a regular version of the conditional expectation)

$$\psi_\mu(x) \stackrel{\text{def}}{=} \mathrm{E}\left[\phi \; (Y) \,\middle|\, \mu^T Y = x\right] \, , \quad \text{and} \quad \zeta_\mu(x) \stackrel{\text{def}}{=} \mathrm{E}\left[\phi^2(Y) \,\middle|\, \mu^T Y = x\right] \, .$$

Using these notations, the asymptotic variance of the stratified estimator may be rewritten as

$$\varsigma^2_{I,M}(\mu, g, \mathcal{Q}_\chi) = \sum_{\mathbf{i} \in \{1,\ldots,I\}^m : M_{\mathbf{i}} > 0} M_{\mathbf{i}}^{-1} \; \left\{ \left(\int_{\mathsf{S}_{\mathbf{i}}} f_\mu \, d\lambda \right) \left(\int_{\mathsf{S}_{\mathbf{i}}} \zeta_\mu f_\mu \, d\lambda \right) - \left(\int_{\mathsf{S}_{\mathbf{i}}} \psi_\mu f_\mu \, d\lambda \right)^2 \right\} \, .$$

We will investigate the limiting behavior of asymptotic variance $\varsigma^2_{I,M}(\mu, g, \mathcal{Q}_\chi)$ when the total number of samples M and the number of strata I both tend to $+\infty$. For that purpose, some technical conditions are required. For ν a measure on $\mathbb{R}^n$ and h a real-valued measurable function on $\mathbb{R}^n$, we denote by $\mathrm{essinf}_\nu(h)$ and $\mathrm{esssup}_\nu(h)$ the essential infimum and supremum w.r.t. the measure ν. From now on we use the following convention : $z/0$ is equal to $+\infty$ if $z > 0$ and to 0 if $z = 0$.

A1 $\int_{\mathbb{R}^m} \chi^2 \, g^{-1} \, d\lambda < +\infty$ and $\mathrm{essinf}_{g \cdot \lambda} \left(\chi \, g^{-1} \right) > 0$.

A2 for $h \in \{f_\mu, \zeta_\mu f_\mu, \psi_\mu f_\mu\}$, $\int_{\mathbb{R}^m} h^2 \; g^{-1} \, d\lambda < +\infty$.

Under A2, λ-a.e. , $g = 0$ implies that $f_\mu = 0$. Finally, a reinforced integrability condition is needed

A3 $\int_{\mathbb{R}^m} \; f_\mu^4 (\zeta_\mu - \psi_\mu^2)^2 \; [\chi^2 g]^{-1} \, d\lambda < +\infty$.

Not surprisingly, the behavior of the asymptotic variance of the stratified estimator behaves differently if $m < d$ or if $m = d$. In the first case, the leading term of the variance remains proportional to the inverse of the number of samples and is asymptotically dominated by the variance in the subspace which is orthogonal to the stratification subspace. In the second case, the rate of convergence is faster, but depends on the choice of the strata in a more complex way.

3.2 The case $m < d$

Our main result is the following proposition which establishes the expression of the limit as the number of strata I goes to $+\infty$ of the limiting variance (as the number of simulations M goes to $+\infty$) of the stratified estimator. Define

$$\varsigma^2_\infty(\mu,\chi) \stackrel{\text{def}}{=} \int_{\mathbb{R}^m} f_\mu^2(\zeta_\mu - \psi_\mu^2)\chi^{-1}\, d\lambda\,. \tag{13}$$

Proposition 1 *Let m be an integer such that $m < d$, $g_1, \cdots, g_m$ be probability density functions (pdf) w.r.t. to the Lebesgue measure of $\mathbb{R}$, μ be a $d \times m$ orthonormal matrix, and χ be a pdf w.r.t. the Lebesgue measure on $\mathbb{R}^m$. Assume that g defined by (11) and χ satisfy assumptions A1-A3. Then*

$$\lim_{I\to+\infty}\lim_{M\to+\infty} M\varsigma^2_{I,M}(\mu,g,\mathcal{Q}_\chi) = \varsigma^2_\infty(\mu,\chi)\,.$$

Assume in addition one of the following conditions

(i) $\operatorname{esssup}_{\chi\cdot\lambda}\left(f_\mu\chi^{-1}\right) < +\infty$ *and* $\{I_M, M \geq 1\}$ *is an integer-valued sequence such that* $I_M^{-1} + I_M^m M^{-1} \to 0$ *as M goes to infinity.*

(ii) $\{I_M, M \geq 1\}$ *is an integer-valued sequence such that* $I_M^{-1} + I_M^{2m} M^{-1} \to 0$ *as M goes to infinity.*

Then,

$$\lim_{M\to+\infty} M\varsigma^2_{I_M,M}(\mu,g,\mathcal{Q}_\chi) = \varsigma^2_\infty(\mu,\chi)\,.$$

It is worthwhile to note that the limiting variance of the stratified estimator $\varsigma^2_\infty(\mu,\chi)$ does not depend on the densities $(g_1,\ldots,g_m)$ that define the strata. This might seem counter-intuitive because it means that only the directions of stratification μ and the allocation distribution $\mathcal{Q}_\chi$ enters in the limit. The contribution to the variance of the randomness in the directions orthogonal to the rows of μ dominates at the first order. Therefore, it is not required to optimize the choice of the distributions $g_1,\ldots,g_m$ which define the positions of the strata in each direction. In practice, this means that asymptotically, once the stratification direction is chosen, the choice of the strata is irrelevant (which is of course not true for any given finite sample); a convenient choice is to set g_i as the distribution of the i-th component of the random vector $\mu^T Y$, $i \in \{1,\ldots,m\}$. When Y is a standard normal random vector, then the components of the vector $\mu^T Y$ are standard gaussian variables, and the strata are simply chosen according to the quantiles of standard gaussian random variables (the distributions g_i, $i \in \{1,\ldots,m\}$ are in such case independent from μ).

On the contrary, the limiting variance $\varsigma^2_\infty(\mu,\chi)$ depends on the allocation density χ. For a given value of the stratification directions μ, it is possible to minimize the function $\chi \mapsto \varsigma^2_\infty(\mu,\chi)$. Assume that $\int_{\mathbb{R}^m} f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}\, d\lambda > 0$. Since

$$\int_{\mathbb{R}^m} f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}\, d\lambda = \mathrm{E}\left[\sqrt{\mathrm{Var}\left[\phi(Y)|\mu^T Y\right]}\right] \leq \sqrt{\mathrm{Var}(\phi(Y))},$$

the integral is finite by (1) and it is possible to define a density $\chi^\star_\mu$ by

$$\chi^\star_\mu \stackrel{\text{def}}{=} \frac{f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}}{\int_{\mathbb{R}^m} f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}\, d\lambda}\,. \tag{14}$$

Then $\chi^\star_\mu$ is the minimum of $\chi \mapsto \varsigma^2_\infty(\mu,\chi)$ and the minimal limiting variance is

$$\varsigma^2_\infty(\mu,\chi^\star_\mu) = \left(\int_{\mathbb{R}^m} f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}\, d\lambda\right)^2 = \left(\mathrm{E}\left[\sqrt{\mathrm{Var}\left[\phi(Y)|\mu^T Y\right]}\right]\right)^2\,.$$

Provided $\chi^\star_\mu$ satisfies assumptions A1-2 (note that in that case, A3 is automatically satisfied), the choice $\chi = \chi^\star_\mu$ for the allocation of the drawings in the strata is asymptotically optimal.

Remark 1 *An expression of the limiting variance $\varsigma^2_\infty(\mu,\chi)$ has been obtained in (Glasserman et al, 1999, Lemma 4.1) in the case $m = 1$ and for the proportional allocation rule which corresponds to $\chi = f_\mu$. It is shown by these authors that the limiting variance is* $\mathrm{E}\left(\mathrm{Var}\left[\,\phi(Y)\,|\,\mu^T Y\right]\right)$ *which is equal to* $\varsigma^2_\infty(\mu, f_\mu)$ *(note that in this case the stratification density $g = f_\mu$, satisfies the assumptions A1-3 provided that* $\mathrm{E}[\phi^4(Y)] < \infty$*). Unless* $\mathrm{Var}\left[\,\phi(Y)\,|\,\mu^T Y\right]$ *is a.s. constant, the asymptotic variance is strictly smaller for the optimal choice of the allocation density.*

The optimal allocation density $\chi_\mu^\star$ cannot in general be computed explicitly but, as shown in the following Proposition, can be approximated by computing the optimal allocation within each stratum.

Proposition 2 *Let $m < d$ be an integer and μ be an $(d \times m)$ orthonormal matrix. Assume that there exist p.d.f $g_1, \ldots, g_m$ such that assumptions A2 is satisfied with g given by (11). Then,*

$$\lim_{I \to +\infty} \sum_{\mathbf{i} \in \{1,\ldots,I\}^m} \left| q_{\mathbf{i}}^\star(\mu) - \int_{\mathsf{S}_\mathbf{i}} \chi_\mu^\star \, d\lambda \right| = 0 \, ,$$

where $\mathcal{Q}^\star(\mu) \stackrel{\text{def}}{=} \{q_{\mathbf{i}}^\star(\mu), \mathbf{i} \in \{1,\ldots,I\}^m\}$ is given by (9). Let $\{I_M, M \geq 1\}$ be an integer-valued sequence such that $I_M^{-1} + I_M^m M^{-1} \to 0$ as M goes to infinity. Then,

$$\lim_{M \to +\infty} M \varsigma_{I_M,M}^2(\mu, g, \mathcal{Q}^\star(\mu)) = \varsigma_\infty^2(\mu, \chi_\mu^\star) \, .$$

The proof is given in Section 6.1. As the number of strata goes to infinity, the stratified estimator run with the optimal allocation $\mathcal{Q}^\star(\mu)$ has the same asymptotic variance as the stratified estimator run with the allocation deduced from the optimal density $\chi_\mu^\star$. In practice, of course, the optimal allocation $\mathcal{Q}^\star(\mu)$ is unknown, but it is possible to construct an estimator of this quantity by estimating the conditional variance of $\mathrm{Var}[\phi(Y)|\mu^T Y \in \mathsf{S}_\mathbf{i}]$ within each stratum (Etore and Jourdain, 2007).

3.3 Case $m = d$

We will consider the case where $m = d$, the number of stratified directions is equal to the dimension of the space. Of course, the results obtained in that setting are markedly different, because this time, the accuracy of the stratified estimator will depend crucially on the definition of the strata along each direction (using the optimal allocation alone is no longer sufficient to reach the optimal asymptotic variance). Let $\phi_\mu(x) \stackrel{\text{def}}{=} \phi(\mu^T x)$ and for $k \in \{1,\ldots,d\}$, $\partial_k \phi_\mu$ denote the partial derivative of ϕ_μ w.r.t. its k-th coordinate. By a slight abuse of notation, we still denote by g_k the function $x = (x_1, \ldots, x_d) \in \mathbb{R}^d \mapsto g_k(x_k)$. When $m = d$, $\psi_\mu = \phi_\mu$, $\zeta_\mu = \phi_\mu^2 = \psi_\mu^2$ and the limits obtained in Propositions 1 and 2 are zero. The number M of random drawings is no longer the appropriate normalization to get a non-trivial limit for the asymptotic variance $\varsigma_{I,M}^2$. According to the following proposition, the right multiplicative factor is $I^2 M$.

Proposition 3 *Assume A1. Assume in addition that* $\mathrm{esssup}_\lambda (f_\mu/g) < +\infty$ *and that ϕ is C^1 and satisfies* $\mathrm{esssup}_\lambda \left(\sum_{k=1}^d \frac{|\partial_k \phi_\mu|}{g_k} \right) < +\infty$. *Finally, let $\{I_M, M \geq 1\}$ be an integer-valued sequence such that* $\lim_{M \to \infty} \left(I_M^{-1} + I_M^{d+2} M^{-1} \right) = 0$. *Then,*

$$\lim_{M \to +\infty} M I_M^2 \varsigma_{I_M,M}^2(\mu, g, \mathcal{Q}_\chi) = \varsigma_\infty^2(\mu, g, \chi)$$

where

$$\varsigma_\infty^2(\mu, g, \chi) \stackrel{\text{def}}{=} \frac{1}{12} \int_{\mathbb{R}^d} \frac{f_\mu^2}{\chi} \sum_{k=1}^d \left(\frac{\partial_k \phi_\mu}{g_k} \right)^2 d\lambda \, . \tag{15}$$

Notice that under the assumptions of the proposition, $\varsigma_\infty^2(\mu, g, \chi) < +\infty$ since $d\lambda$ a.e.,

$$\frac{f_\mu^2}{\chi} \sum_{k=1}^d \left(\frac{\partial_k \phi_\mu}{g_k} \right)^2 \leq \frac{\left(\mathrm{esssup}_\lambda \left(\frac{f_\mu}{g} \right) \mathrm{esssup}_\lambda \left(\sum_{k=1}^d \frac{|\partial_k \phi_\mu|}{g_k} \right) \right)^2}{\mathrm{essinf}_{g.\lambda} \left(\frac{\chi}{g} \right)} g.$$

In the same way, $\int_{\mathbb{R}^d} f_\mu \sqrt{\sum_{k=1}^d \left(\frac{\partial_k \phi_\mu}{g_k} \right)^2} d\lambda < +\infty$. If this last integral is positive, it is possible to define a density $\chi_{\mu,g}^\star$ by

$$\chi_{\mu,g}^\star \stackrel{\text{def}}{=} \frac{f_\mu \sqrt{\sum_{k=1}^d \left(\frac{\partial_k \phi_\mu}{g_k} \right)^2}}{\int_{\mathbb{R}^d} f_\mu \sqrt{\sum_{k=1}^d \left(\frac{\partial_k \phi_\mu}{g_k} \right)^2} d\lambda} .$$

Then the minimum of $\chi \mapsto \varsigma^2_\infty(\mu, g, \chi)$is attained at $\chi^\star_{\mu,g}$ and

$$\varsigma^2_\infty(\mu, g, \chi^\star_{\mu,g}) = \frac{1}{12}\left(\int_{\mathbb{R}^d} f_\mu \sqrt{\sum_{k=1}^d \left(\frac{\partial_k \phi_\mu}{g_k}\right)^2} d\lambda\right)^2 .$$

Remark 2 *When $d > 1$, the optimization of the stratified estimator asymptotic variance $\varsigma^2_\infty(\mu, g, \chi^\star_{\mu,g})$ w.r.t. g is not obvious, because of the restrictive choice of the stratification.*

Remark 3 *When $d = 1$, Proposition 3 continue to hold under a weakened assumption on ϕ, consisting in supposing that ϕ is locally bounded on $\mathbb{R}$ with a locally integrable distribution derivative ϕ' such that* $\operatorname{esssup}_\lambda\left(\frac{|\phi'|}{g}\right) < +\infty$. *Moreover, one has*

$$\begin{aligned}\varsigma^2_\infty(\mu, g, \chi) &= \int_{\mathbb{R}} \frac{(f_\mu \phi'_\mu)^2}{g^2 \chi} d\lambda = \int_{\mathbb{R}} \left(\frac{f_\mu |\phi'_\mu|}{g\chi}\right)^2 \chi d\lambda \\ &\geq \left(\int_{\mathbb{R}} \left(\frac{\sqrt{f_\mu |\phi'_\mu|}}{g}\right)^2 g d\lambda\right)^2 \geq \left(\int_{\mathbb{R}} \sqrt{f_\mu |\phi'_\mu|} d\lambda\right)^4\end{aligned}$$

where both inequalities are equalities for the choice $\chi^\star_{\mu,g} \propto \frac{f_\mu |\phi'_\mu|}{g}$ and $g^\star_\mu \propto \sqrt{f_\mu |\phi'_\mu|}$ which leads to $\chi^\star_{\mu,g^\star_\mu} = g^\star_\mu \propto \sqrt{f_\mu |\phi'_\mu|}$. For this choice the allocation is uniform ($q_i = \int_{G^{-1}(\frac{i-1}{I})}^{G^{-1}(\frac{i}{I})} g(x)dx = \frac{1}{I}$) and the asymptotic result given in the proposition is preserved for $I_M = M$:

$$\lim_{M\to+\infty} M^3 \varsigma^2_{M,M}(\mu, g^\star_\mu, \mathcal{Q}_{\chi^\star_{\mu,g^\star_\mu}}) = \left(\int_{\mathbb{R}} \sqrt{f_\mu |\phi'_\mu|} d\lambda\right)^4 .$$

Indeed, for this choice, there is no rounding error in the allocation of the drawings in the strata : $M_i = 1 = Mq_i$ for all $i \in \{1, \cdots, M\}$.

Let us compute the asymptotic variance under the optimal allocation :

Proposition 4 *Assume* $\operatorname{esssup}_\lambda\left(\frac{f_\mu}{g}\right) < +\infty$, *that ϕ is C^1 and such that* $\operatorname{esssup}_\lambda\left(\sum_{k=1}^d \frac{|\partial_k \phi_\mu|}{g_k}\right) < +\infty$ *and $\{I_M, M \geq 1\}$ is an integer-valued sequence such that $I_M^{-1} + I_M^{d+2} M^{-1} \to 0$. Then*

$$\lim_{M\to+\infty} M I_M^2 \varsigma^2_{I_M,M}(\mu, g, \mathcal{Q}^\star(\mu)) = \varsigma^2_\infty(\mu, g, \chi^\star_{\mu,g}).$$

4 An adaptive stratification algorithm

As shown in the asymptotic theory presented above, under optimal allocation, it is more important to optimize the stratification matrix μ than the strata boundaries along each stratification direction [1]. Proposition 1 suggests the following strategy: the "optimal" matrix $\mu_\star$ is defined as a minimizer (which is not necessarily unique) of the limiting variance $\mu \mapsto \varsigma^2_\infty(\mu, \chi^\star_\mu)$. Of course, this optimization problem does not have a closed form expression because it is unrealistic to assume that the functions $x \mapsto \psi_\mu(x)$, $x \mapsto \zeta_\mu(x)$ are available.

We rather use the characterization of the optimal limiting variance of the stratified estimator given in Proposition 2, *i.e.* the problem boils down to search for a minimizer μ of the variance $\varsigma^2_{I,M}(\mu, g, \mathcal{Q}^\star(\mu))$. The choice of g is, as emphasized above, largely arbitrary. In our applications, Y is a d-dimensional standard normal vector, and $\mu^T Y$ is a m-dimensional standard Gaussian vector. In this case, we set g_i, $i = \{1, \ldots, I\}^m$ to be the standard Gaussian distribution so that the strata boundaries in each directions are the quantiles of the standard normal variable (independently from the direction matrix μ). This choice leads to equiprobable strata for the vector $\mu^T Y$.

Of course, the optimization of $\varsigma^2_{I,M}(\mu, g, \mathcal{Q}^\star(\mu))$ is a difficult task because in particular the definition of this function involves multidimensional integrals, which cannot be computed with high accuracy. Note

[1] Of course, this is an asymptotic result, but our numerical experiments suggest that optimizing the strata boundaries along each stratification direction does not lead to a significant reduction of the variance. This is why we concentrate on the optimization of the stratification matrix

also that, in most situations, the optimization should be done in parallel to the main objective, namely, the estimation of the quantity of interest $\mathrm{E}[\phi(Y)]$, which is obtained using a stratified estimator based on the adaptively defined directions of stratification μ (and thus on the adaptively defined strata $\mathsf{S}_{\mu,\mathbf{i}}, \mathbf{i} \in \mathcal{I}$). The adaptive stratification algorithm might be seen as an analog to the very popular adaptive importance sampling; see for example Rubinstein and Kroese (2004), Arouna (2004), Kawai (2007), and Rubinstein and Kroese (2008).

When the function to minimize is an expectation, the classical approaches to tackle this problem are based on some forms of Monte Carlo approximations for the integrals appearing in the expression of the objective function and its gradients. There are typically two approaches to Monte Carlo methods, the stochastic approximation procedure and the sample average approximation method; see for example Judistsky et al (2007) for an in-depth comparison of these procedures. None of these procedures can be directly applied in our context, but they can be more or less directly adapted to solve our problem. In the adaptive stratification context, these Monte Carlo estimators are based on the current fit of the stratification matrix and of the conditional variances within each stratum, the underlying idea being that the algorithm is able to progressively learn the optimal stratification, while the stratified estimator is constructed.

The algorithm described here is closely related to the sample average approximation method, the main difference with the classical approach being that, at every time a new search direction is computed, a new Monte Carlo sample (using the current fit of the strata and of the allocation) is drawn; this is due to the fact that we are not only willing to minimize the asymptotic variance of the stratified estimator but we also want to compute the stratified estimator "on the fly".

Denote by f the density of Y w.r.t. the Lebesgue measure. Define for $\mathbf{i} \in \{1, \cdots, I\}^m$, a function $h \in \{f, \phi f, \phi^2 f\}$, and an orthonormal $d \times m$ matrix μ,

$$\nu_{\mathbf{i}}(h,\mu) \stackrel{\text{def}}{=} \int_{\mathsf{S}_{\mu,\mathbf{i}}} h \; d\lambda = \int \prod_{k=1}^{m} \mathbb{1}_{\{y, G_k^{-1}((\mathbf{i}_k-1)/I) \le \langle \mu_k, y\rangle \le G_k^{-1}(\mathbf{i}_k/I)\}} h \; d\lambda \; , \tag{16}$$

where $\langle x, y\rangle$ denotes the scalar product of the vectors x and y. Using the definition of $\nu_{\mathbf{i}}$, the proportions $p_{\mathbf{i}}(\mu)$ and the conditional variances with each stratum $\sigma^2_{\mathbf{i}}(\mu)$ respectively given by (3) and (7) may be expressed as, when $\nu_{\mathbf{i}}(f,\mu) > 0$,

$$p_{\mathbf{i}}(\mu) = \nu_{\mathbf{i}}(f,\mu) \; , \quad \text{and} \quad \sigma^2_{\mathbf{i}}(\mu) = \frac{\nu_{\mathbf{i}}(f\phi^2,\mu)}{\nu_{\mathbf{i}}(f,\mu)} - \left(\frac{\nu_{\mathbf{i}}(f\phi,\mu)}{\nu_{\mathbf{i}}(f,\mu)}\right)^2 \; . \tag{17}$$

When M is large and I is fixed, minimizing the asymptotic variance of the stratified estimate with optimal allocation is equivalent to minimize $V(\mu)$ w.r.t. the stratification matrix μ where (see Lemma 1)

$$V(\mu) \stackrel{\text{def}}{=} \sum_{\mathbf{i}=1}^{\mathcal{I}} p_{\mathbf{i}}(\mu)\sigma_{\mathbf{i}}(\mu) = \sum_{\mathbf{i}=1}^{\mathcal{I}} \Big(\nu_{\mathbf{i}}(f,\mu)\nu_{\mathbf{i}}(f\phi^2,\mu) - \nu^2_{\mathbf{i}}(f\phi,\mu)\Big)^{1/2} \; .$$

Assuming that the functions $\mu \mapsto \nu_{\mathbf{i}}(h,\mu)$ are differentiable at μ for $h \in \{f, f\phi, f\phi^2\}$ (which we prove below, under appropriate technical conditions), the gradient may be expressed as

$$\nabla_\mu \, V(\mu) = \sum_{\mathbf{i}=1}^{\mathcal{I}} \frac{\nabla_\mu \nu_{\mathbf{i}}(1,\mu) \; \nu_{\mathbf{i}}(\phi^2,\mu) + p_{\mathbf{i}}(\mu) \; \nabla_\mu \nu_{\mathbf{i}}(\phi^2,\mu) - 2\nu_{\mathbf{i}}(\phi,\mu) \; \nabla_\mu \nu_{\mathbf{i}}(\phi,\mu)}{2 \; p_{\mathbf{i}}(\mu)\sigma_{\mathbf{i}}(\mu)} \mathbb{1}_{\{p_{\mathbf{i}}(\mu)\sigma_{\mathbf{i}}(\mu)\neq 0\}} \; . \tag{18}$$

The computation of this gradient thus requires to establish the differentiability and to compute the gradients $\nabla_\mu \, \nu_{\mathbf{i}}(h,\mu)$ for $h \in \{f, f\phi, f\phi^2\}$. For a vector $\nu \in \mathbb{R}^d$, $\nu \neq 0$, and $z \in \mathbb{R}$, define λ^ν_z, the restriction of the Lebesgue measure on the hyperplane $\{y \in \mathbb{R}^d, \langle \nu, y\rangle = z\}$.

Proposition 5 *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a locally bounded integrable real function and z be a real. Let $g_z : \mathbb{R}^d \to \mathbb{R}$ be the function*

$$g_z(\nu) \stackrel{\text{def}}{=} \int \mathbb{1}_{\{y, \langle \nu, y\rangle \le z\}} h(y) \; d\lambda(y) \; .$$

Let $\mu \in \mathbb{R}^d$ be a non-zero vector. Assume that h is continuous λ^μ_z almost everywhere and that there exists $\varepsilon > 0$ such that

$$\lim_{M \to +\infty} \sup_{|\nu - \mu| \le \varepsilon} \int |y| \mathbb{1}_{\{|y| \ge M\}} |h(y)| \; d\lambda^\nu_z(y) = 0 \; . \tag{19}$$

Then, the function $\nu \mapsto g_z(\nu)$ is differentiable at μ and

$$\nabla_\mu \; g_z(\mu) = -\int \frac{y}{|\mu|} \; h(y) \; d\lambda^\mu_z(y) \; .$$

It is worthwhile to note that the function $\nu \mapsto g_z(\nu)$ is differentiable whereas the integrand $\nu \mapsto \mathbb{1}_{\{y,\langle \nu,y\rangle \le z\}} h(y)$ is not even continuous. The situation is rather different to the classical case where the gradient is obtained as the empirical mean of the gradient of the estimate. The expression of the gradient involves the computation of the integral with respect to a measure located on an hyperplane (a surface integral).

Corollary 1 *Assume that h is a real locally bounded integrable function. Let m be an integer and $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$. Let $g_z : \mathbb{R}^{d\times m} \to \mathbb{R}$ be the function*

$$g_z(\nu_1, \ldots, \nu_m) \stackrel{\text{def}}{=} \int \prod_{k=1}^{m} \mathbb{1}_{\{y,\langle \nu_k,y\rangle \le z_k\}} h(y)\, d\lambda(y)\,.$$

Let $\mu = [\mu_1, \ldots, \mu_m] \in \mathbb{R}^{d\times m}$ be a full rank matrix. Assume that h is continuous $\sum_{k=1}^m \lambda_z^{\mu_k}$ almost everywhere and that there exists $\varepsilon > 0$ such that, for any $k \in \{1, \ldots, m\}$,

$$\lim_{M\to+\infty} \sup_{|\nu-\mu_k|\le\varepsilon} \int |y| \mathbb{1}_{\{|y|\ge M\}} |h(y)|\, d\lambda_z^{\nu}(y) = 0\,. \tag{20}$$

Then, g_z is differentiable at μ and the differential $\nabla_\mu g_z$ is given by $\nabla_\mu g_z = [\nabla_{\mu_1} g_z, \ldots, \nabla_{\mu_m} g_z]$, where

$$\nabla_{\mu_i} g_z(\mu) = -\int \frac{y}{|\mu_i|} \prod_{k\ne i} \mathbb{1}_{\{y,\langle \mu_k,y\rangle \le z_k\}} h(y)\, d\lambda_{z_i}^{\mu_i}(y)\,.$$

For $k \in \{1, \cdots, m\}$, (16) shows that $\nabla_{\mu_k} \nu_{\mathbf{i}}(h, \mu)$ may be expressed as

$$\begin{aligned}\nabla_{\mu_k} \int \mathbb{1}_{\{y,\langle \mu_k,y\rangle \le G_k^{-1}(\mathbf{i}_k/I)\}} \left[\prod_{j\ne k} \mathbb{1}_{\{y,G_j^{-1}((\mathbf{i}_j-1)/I)\le\langle \mu_j,y\rangle \le G_j^{-1}(\mathbf{i}_j/I)\}} h\right] d\lambda \\ - \nabla_{\mu_k} \int \mathbb{1}_{\{y,\langle \mu_k,y\rangle \le G_k^{-1}((\mathbf{i}_k-1)/I)\}} \left[\prod_{j\ne k} \mathbb{1}_{\{y,G_j^{-1}((\mathbf{i}_j-1)/I)\le\langle \mu_j,y\rangle \le G_j^{-1}(\mathbf{i}_j/I)\}} h\right] d\lambda\,.\end{aligned} \tag{21}$$

The algorithm goes as follows. Denote by $\{\gamma_t\}$ a sequence of stepsizes. Consider the strata $\{\mathsf{S}_{\mathbf{i}}, \mathbf{i} \in \{1, \cdots, I\}^m\}$ given by (10) for some product density g.

1. Initialization. Choose initial stratification directions $\mu^{(0)}$ and an initial number of draws in each statum $M^{(0)} \stackrel{\text{def}}{=} \{M_{\mathbf{i}}^{(0)}, \mathbf{i} \in \stackrel{\text{def}}{=} \{1, \ldots, I\}^m\}$ such that $\sum_{\mathbf{i}} M_{\mathbf{i}}^{(0)} = M$. Compute the probabilities $p_{\mathbf{i}}(\mu^{(0)})$ of each stratum.
2. Iteration. At iteration $t+1$, given $\mu^{(t)}$, $M^{(t)}$ and $\{p_{\mathbf{i}}(\mu^{(t)}), \mathbf{i} \in \{1, \cdots, I\}^m\}$,
   (a) *Compute $\widehat{\nabla V}(\mu^{(t)})$:*
      (i) for $\mathbf{i} \in \{1, \cdots, I\}^m$, draw $M_{\mathbf{i}}^{(t)}$ realizations of i.i.d. random variables $\{Y_{\mathbf{i},k}^{(t)}, k \le M_{\mathbf{i}}^{(t)}\}$ with distribution $\mathbb{P}(Y \in \cdot | Y \in \mathsf{S}_{\mu^{(t)},\mathbf{i}})$ and evaluate for $h \in \{\phi, \phi^2\}$

$$\hat{\nu}_{\mathbf{i}}^{(t+1)}(h) = \frac{p_{\mathbf{i}}(\mu^{(t)})}{M_{\mathbf{i}}^{(t)}} \sum_{k=1}^{M_{\mathbf{i}}^{(t)}} h\left(Y_{\mathbf{i},k}^{(t)}\right)$$

      which is a Monte Carlo estimate of $\nu_{\mathbf{i}}(h, \mu)$ with $\mu = \mu^{(t)}$.
      (ii) for $k \in \{1, \cdots, m\}$, $s \in \{G_k^{-1}(1/I), \cdots, G_k^{-1}((I-1)/I)\}$, draw $\tilde{M}_{k,s}^{(t)}$ realizations of i.i.d. random variables with distribution $\mathbb{P}(Y \in \cdot | [\mu_k^{(t)}]^T Y = s)$. Compute a Monte Carlo estimate of $\nabla_\mu \nu_{\mathbf{i}}(h, \mu^{(t)})$ for $h \in \{f, f\phi, f\phi^2\}$ based on (21) and Corollary 1.
      (iii) deduce from these Monte Carlo approximations, a Monte Carlo estimate of $\nabla V(\mu^{(t)})$ based on the expression (18).
   (b) *Update the direction of stratification:* Set $\tilde{\mu} = \mu^{(t)} - \gamma_t \widehat{\nabla V}(\mu^{(t)})$; define $\mu^{(t+1)}$ as the orthonormal matrix found by computing the singular value decomposition of $\tilde{\mu}$ and keeping the m left singular vectors.
   (c) *Update the allocation policy:*
      (i) compute an estimate $\hat{\sigma}_{\mathbf{i}}^{(t+1)}$ of the standard deviation within stratum $\mathbf{i}$

$$\hat{\sigma}_{\mathbf{i}}^{(t+1)} = \left(\frac{\hat{\nu}_{\mathbf{i}}^{(t+1)}(\phi^2)}{p_{\mathbf{i}}(\mu^{(t)})} - \left(\frac{\hat{\nu}_{\mathbf{i}}^{(t+1)}(\phi)}{p_{\mathbf{i}}(\mu^{(t)})}\right)^2\right)^{1/2}\,.$$

(ii) Update the allocation vector

$$q_{\mathbf{i}}^{(t+1)} = \frac{p_{\mathbf{i}}(\mu^{(t)})\ \hat{\sigma}_{\mathbf{i}}^{(t+1)}}{\sum_{\mathbf{j}\in\{1,\ldots,I\}^m} p_{\mathbf{j}}(\mu^{(t)})\ \hat{\sigma}_{\mathbf{j}}^{(t+1)}}\ ,$$

and the number of draws $\{M_{\mathbf{i}}^{(t+1)}, \mathbf{i} \in \{1,\ldots,I\}^m\}$ by applying the formula (4) with a total number of draws equal to M.

(d) *Update the probabilities* $p_{\mathbf{i}}(\mu^{(t+1)})$, $\mathbf{i} \in \{1,\cdots,I\}^m$.

(e) *Compute an averaged stratified estimate of the quantity of interest:* Estimate the Monte Carlo variance of the stratified estimator for the current fit of the strata and the optimal allocation

$$[\varsigma^2]^{(t+1)} = \frac{1}{M}\left(\sum_{\mathbf{i}\in\{1,\cdots,I\}^m} p_{\mathbf{i}}(\mu^{(t)})\ \hat{\sigma}_{\mathbf{i}}^{(t+1)}\right)^2 .$$

Compute the current fit of the stratified estimator by the following weighted average

$$\mathcal{E}^{(t+1)} = \left(\sum_{\tau=1}^{t+1}\frac{1}{[\varsigma^2]^{(\tau)}}\right)^{-1}\sum_{\tau=1}^{t+1}\frac{1}{[\varsigma^2]^{(\tau)}}\sum_{\mathbf{i}\in\{1,\cdots,I\}^m}\hat{\nu}_{\mathbf{i}}^{(\tau)}(\phi)\ . \tag{22}$$

There are two options to choose the stepsizes $\{\gamma_t, t \geq 0\}$. The traditional approach consists in taking a decreasing sequence satisfying the following conditions (see for example Pflug (1996); Kushner and Yin (2003))

$$\sum_{t\geq 0}\gamma_t = +\infty\ , \qquad \sum_{t\geq 0}\gamma_t^2 < +\infty\ .$$

If the number of simulations is fixed in advance, say equal to N, then one can use a constant stepsize strategy, *i.e.* choose $\gamma_t = \gamma$ for all $t \in \{1,\ldots,N\}$. As advocated in Juditsky et al (2007), a sensible choice in this setting is to take γ_t proportional to $N^{-1/2}$. This is a rather crude optimization algorithm but line-searching is computationally heavy and should therefore better be avoided in this context; the convergence of a crude gradient proved to be quite fast in all our applications, so it is not required to resort to computationally intensive alternatives.

Step 2(a)ii is specific to the optimization problem to solve and is not related to the stratification sampler itself. The number of draws for the computation of the surface integral (see Corollary 1) can be chosen independently of the allocation $M^{(t)}$. When the samples in steps 2(a)i and 2(a)ii can be obtained by transforming the same set of variables (see Section 5 for such a situation), it is natural to choose $\tilde{M}^{(t)} = \{\tilde{M}_{k,s}^{(t)}, k \in \{1,\cdots,m\}, s \in \{G_k^{-1}(1/I),\cdots,G_k^{-1}((I-1)/I)\}\}$ such that $\sum_{k,s}\tilde{M}_{k,s}^{(t)} = M$.

When f_μ has a product form (which is the case e.g. when Y is a standard d-dimensional Gaussian distribution), we can set $g = f_\mu$. Then, the strata are equiprobable and $p_{\mathbf{i}}(\mu) = 1/I^m$ for any $(\mathbf{i}, \mu)$.

It is out of the scope of this paper to prove the convergence of this algorithm and we refer the reader to classical treatises on this subject. The above algorithm provides, at convergence, both *(i)* "optimal" directions of stratification and an estimate of the associated optimal allocation; *(ii)* an averaged stratified estimate $\mathcal{E}$. By omitting the step 2e, the algorithm might be seen as a mean for computing the stratification directions and the associated optimal allocation, and these quantities can then be plugged in a "usual" stratification procedure.

5 Applications in Financial Engineering

The pricing of an option under classical Black-Scholes assumptions amounts to compute the expectation $\mathrm{E}[\Xi(Y)]$ for some measurable non-negative function Ξ on $\mathbb{R}^d$, where Y is a standard d-multivariate Gaussian variable. Examples of such situations include the pricing of Asian options or Basket options when the underlying asset prices are described by geometric Brownian motions. The Cameron-Martin formula implies that for any $\nu \in \mathbb{R}^d$,

$$\mathrm{E}[\Xi(Y)] = \mathrm{E}\left[\Xi(Y+\nu)\ \exp(-\nu^T Y - 0.5\nu^T\nu)\right]\ , \tag{23}$$

Classical results on importance sampling show that the variance of the crude Monte Carlo estimate depends on ν. In the numerical applications below, we apply the adaptive stratification procedure introduced in

Section 4 (hereafter referred to as "`AdaptStr`") with $\phi(y) = \Xi(y+\nu_\star)\ \exp(-\nu_\star^T y - 0.5\nu_\star^T\nu_\star)$ where $\nu_\star$ is the solution of the optimization problem

$$\text{argmax}_{\{\nu\in\mathbb{R}^d, \Xi(\nu)>0\}} \quad \left\{\ln \Xi(\nu) - 0.5\nu^T\nu\right\}\ , \tag{24}$$

(case $\nu = \nu_\star$), and with $\phi(y) = \Xi(y)$ (case "no drift" or "ν is the null vector"). The motivations for this particular choice of the direction ν and procedures to solve this optimization problem are discussed in Glasserman et al (1999).

For comparison purposes, we also run the stratification procedure proposed in Glasserman et al (1999) (hereafter referred to as " `GHS`"): we implement the algorithm which combines *(i)* importance sampling with the drift $\nu_\star$ defined as above, and *(ii)* stratification with direction $\mu_{\mathtt{g}}$ defined as some eigenvector of some Hessian matrix (see (Glasserman et al, 1999, Section 4.2)). We also run for comparison the plain Monte Carlo estimator.

5.1 Practical implementations of the adaptive stratification procedure

The numerical results have been obtained by running `Matlab` codes [2] In the numerical applications below, $m = 1$. We choose $g = f_\mu$ so that the strata are equiprobable ($p_{\mathbf{i}}(\mu) = 1/I$). We choose $I = 100$ strata and $M = 10\,000$ draws per iterations.

The drift vector ν that solves (24) is obtained by running `solnp`, a nonlinear optimization program in Matlab freely available at `http://www.stanford.edu/~yyye/matlab/`. The direction $\mu^{(0)}$ is set to the unitary constant vector $(1,\cdots,1)/\sqrt{d}$ -except when specified-; and the initial allocation $M^{(0)}$ is the proportional one. Exact sampling under the conditional distributions $\mathbb{P}(Y \in \cdot | Y \in \mathsf{S}_{\mu^{(t)},\mathbf{i}})$ and $\mathbb{P}(Y \in \cdot | [\mu^{(t)}]^T Y = s)$ can be done by linear transformation of standard Gaussian vectors (see (Glasserman, 2004, section 4.3, p. 223)). For example, when $m = 1$, the procedure

(i) Draw independently $V \sim \mathcal{N}_d(0, \mathrm{Id})$ and $U \sim \mathcal{U}([0,1])$
(ii) Set $\tilde{U} = \Phi^{-1}(\Phi(s_{\mathbf{i}-1}) + U\{\Phi(s_{\mathbf{i}}) - \Phi(s_{\mathbf{i}-1}\})$, where Φ is the c.d.f. of a standard Gaussian random variable $\mathcal{N}(0,1)$
(iii) Set $Z = \tilde{U}\mu + (\mathrm{Id} - \mu\mu^T)V$

produces a r.v. Z with distribution $\mathbb{P}(Y \in \cdot | \mu^T Y \in [s_{\mathbf{i}-1}, s_{\mathbf{i}}])$ (by convention $\Phi(-\infty) = 1 - \Phi(+\infty) = 0$); and the procedure

(i) Draw $V \sim \mathcal{N}_d(0, \mathrm{Id})$
(ii) Set $Z = s\mu + (\mathrm{Id} - \mu\mu^T)V$

produces a r.v. Z with distribution $\mathbb{P}(Y \in \cdot | \mu^T Y = s)$. The draws in step 2(a)i and 2(a)ii can thus be obtained by transforming the same set of $M^{(t)}$ Gaussian random variables $\{V_j^{\mathbf{i}}, j \le M_{\mathbf{i}}^{(t)}, \mathbf{i} \in \{1,\cdots,I\}\}$. Therefore, the total number of draws by iteration is M (the estimates of $\nu_{\mathbf{i}}(h,\mu)$ and $\nabla_\mu \nu_{\mathbf{i}}(h,\mu)$ are not independent). The criterion is optimized using a fixed stepsize steepest descent algorithm (we take $\gamma_t = \gamma$ for some $\gamma \in [0.001, 0.01]$).

5.2 Assessing efficiency of the adaptive stratification procedure

We compare the averaged stratified estimate $\mathcal{E}^N$ obtained after $N = 200$ iterations, with the stratified estimate obtained by running `GHS`, and with the crude Monte Carlo estimate. For a fair comparison, the `GHS` algorithm and the Monte Carlo procedure are run with the *same* MN realizations of standard Gaussian vectors (in the present case, $MN = 2\ 10^6$). We report in the tables below the estimates of the option prices given by the stratification procedures (column "Price") and the estimates of the variance of the estimators. The column "`GHS`" is an estimate of $\sum_{\mathbf{i}} p_{\mathbf{i}} \sigma_{\mathbf{i}}^2(\mu_{\mathtt{g}})$ computed with MN samples; the column "`AdaptStr,` $\nu = \nu_\star$" is the limiting variance per sample of $\mathcal{E}^N$ which is equal to

$$N\left\{\sum_{t=1}^{N}\left(\left[\sum_{\mathbf{i}} p_{\mathbf{i}}\,\hat{\sigma}_{\mathbf{i}}^{(t)}\right]^2\right)^{-1}\right\}^{-1} \sim \left(\sum_{\mathbf{i}} p_{\mathbf{i}}\,\sigma_{\mathbf{i}}(\mu^{(+\infty)})\right)^2\ ,$$

when the objective function is given by (23) with $\nu = \nu_\star$. For comparison purposes, we also report in column "`Monte Carlo`", an estimate of the variance of the crude Monte-Carlo estimator computed with MN samples.

[2] These codes are freely available from the url `http://www.tsi.enst.fr/~gfort/`

5.3 Asian options

Consider the pricing of an arithmetic Asian option on a single underlying asset under standard Black-Scholes assumptions. The price of the asset is described by the stochastic differential equation

$$\frac{dS_t}{S_t} = r\,dt + \upsilon\,dW_t\ , \quad S_0 = s_0\ ,$$

where $\{W_t, t \geq 0\}$ is a standard Brownian motion, r is the risk-free mean rate of return, υ is the volatility and s_0 is the initial value. The asset price is discretized on a regular grid $0 = t_0 < t_1 < \cdots < t_d = T$, with $t_i \stackrel{\text{def}}{=} iT/d$. The increment of the Brownian motion on $[t_{i-1}, t_i)$ is simulated as $\sqrt{T/d}\,Y_i$ for $i \in \{1, \cdots, d\}$ where $Y = (Y_1, \cdots, Y_d) \sim \mathcal{N}_d(0, \mathrm{Id})$. The discounted payoff of a discretely monitored arithmetic average Asian option with strike price K is given by $\Xi(Y)$,

$$\Xi(y) = \exp(-rT)\left(\frac{s_0}{d}\sum_{k=1}^{d}\exp\left((r-0.5\upsilon^2)\frac{kT}{d} + \upsilon\sqrt{\frac{T}{d}}\sum_{j=1}^{k} y_j\right) - K\right)_+ , \quad y = (y_1, \cdots, y_d) \in \mathbb{R}^d\ ,$$

where for $x \in \mathbb{R}$, $x_+ = \max(x, 0)$. In the numerical applications, we take $s_0 = 50$, $r = 0.05$, $T = 1$, $(\upsilon, K) \in \{(0.1, 45), (0.1, 50), (0.1, 55), (0.3, 45), (0.3, 50), (0.3, 55)\}$ and $d = 16$. We run `AdaptStr` for $N = 100$ iterations, when $\upsilon = 0.1$ and $K = 45$. On Figure 1, the optimal drift vector $\nu_\star$, the direction $\mu^{(N)}$ obtained after N iterations of `AdaptStr`, and the direction of stratification $\mu_{\mathtt{g}}$ are plotted.

Insert Figure 1 about here

In Figure 2 , the successive directions $t \mapsto \mu^{(t)}$, the successive estimations of the quantity of interest $t \mapsto \mathcal{E}^{(t)}$ and of the variance $t \mapsto (\sum_{\mathbf{i}} p_{\mathbf{i}} \hat{\sigma}_{\mathbf{i}}^{(t)})^2$ are displayed. As shown on Figure 1,$\{\mu^{(t)}, t \geq 0\}$ converges to the direction $\mu_{\mathtt{g}}$, and the convergence takes place after about 30 iterations. We find the same pattern for a wide range of parameter values. The choice of the stratification direction has a major impact on the variance of the estimate $\mathcal{E}^{(t)}$ as shown on Figure 1[bottom right]. Along the 100 iterations of the algorithm, the variance decreases from 0.1862 to 0.0015. We also observed that the convergence of the algorithm and the limiting values were independent of the initial values $(\mu^{(0)}, M^{(0)})$ (these results are not reported for brevity). These initial values (and the choice of the sequence $\{\gamma^{(t)}, t \geq 1\}$) only influence the number of iterations required to converge.

Insert Figure 2 about here

`AdaptStr` can also be read as a procedure that computes a stratification direction and provides the associated optimal allocation. These quantities can then be used for running a (usual) stratification procedure with M draws and for the optimal allocation. By doing such with $M = 10\,000$, we obtain an estimate of the quantity $\mathrm{E}[\phi(Y)]$ equal to 6.05 and of the variance equal to $0.0015/M$. We can compare these results to the output of `GHS`: this yields the same estimator of $\mathrm{E}[\phi(Y)]$ and a larger standard deviation equal to $0.0070/M$. Observe that since $\mu^{(N)} = \mu_{\mathtt{g}}$, the two algorithms differ from the allocations in the strata (in `GHS`, an equal number of replications in each stratum is used).

We conclude this study of `AdaptStr` by illustrating the role of the drift vector ν (see Eq. 23). We report on Figure 3 the limiting direction $\mu^{(N)}$, the estimates $t \mapsto \mathcal{E}^{(t)}$ and the variance $t \mapsto (\sum_{\mathbf{i}} p_{\mathbf{i}} \hat{\sigma}_{\mathbf{i}}^{(t)})^2$ when the drift vector ν is the null vector. The limiting direction $\mu^{(N)}$ slightly differs from $\mu_{\mathtt{g}}$ and is close to $\nu_\star$. Moreover, the variance reduction is weaker: the limiting value of $t \mapsto (\sum_{\mathbf{i}} p_{\mathbf{i}} \hat{\sigma}_{\mathbf{i}}^{(t)})^2$ is 0.0035. The efficiency of the adaptive stratification procedure `AdaptStr` is thus related to the drift vector ν in (23); similar conclusions are reached in Glasserman et al (1999) (see also Glasserman (2004)).

Insert Figure 3 about here

We report in Table 1 the variance (per sample) of the plain Monte Carlo estimate, of `GHS` and of `AdaptStr`. For `AdaptStr`, we consider the cases $\nu = \nu_\star$ and ν equal to the null vector in the formula (23).

Insert Table 1 about here

5.4 Options with knock-out at expiration

A knock-out barrier option is a path-dependent option that expires worthless if the underlier reaches a specified barrier level. The payoff of this option is given by

$$\Xi(y) = \exp(-rT)\left(\frac{s_0}{d}\sum_{k=1}^{d}\exp\left((r-0.5\sigma^2)\frac{kT}{d}+\sigma\sqrt{\frac{T}{d}}\sum_{j=1}^{k}y_j\right)-K\right)_+ \mathbb{1}_{S_T(y)\le B}\ ,$$

where K is the strike price, B is the barrier and $S_T(y)$ is the underlier price modeled as

$$S_T(y) = s_0\exp\left((r-0.5\sigma^2)T+\sigma\sqrt{\frac{T}{d}}\sum_{j=1}^{d}y_j\right)\ .$$

In the numerical applications, we set $s_0 = 50$, $r = 0.05$, $T = 1$, $\sigma = 0.1$, $d = 16$ and $(K, B) \in \{(50, 60), (50, 70), (50, 80), (55, 70), (55, 80)\}$.

On Figure 4, we plot $\mu^{(N)}$ and $\mu_{\mathtt{g}}$ for different values of the parameters (K, B). In this case, the optimal importance sampling direction does not necessarily coincide with the optimal direction of stratification $\mu_{\mathtt{g}}$. In the case $(K, B) = (50, 60)$, we display in Figure 5, the successive directions $t \mapsto \mu^{(t)}$ when $\mu^{(0)}$ is proportional to the constant vector $(1, \cdots, 1)$, and $\mu^{(0)} = \mu_{\mathtt{g}}$: the limiting direction does not depend on the initial value and this is an example where the limiting direction differs from $\mu_{\mathtt{g}}$.

Insert Figure 4 and Figure 5 about here

We report in Table 2 the variance (per sample) of the plain Monte Carlo estimate, of `GHS` and of `AdaptStr`. For `AdaptStr`, we consider the cases $\nu = \nu_\star$ and ν equal to the null vector in the formula (23).

Insert Table 2 about here

5.5 Basket options

Consider a portfolio consisting of d assets. The portfolio contains a proportion α_k of asset k, $k \in \{1, \ldots, d\}$. The price of each asset is described by a geometric Brownian motion (under the risk neutral probability measure)

$$\frac{dS_t^{(k)}}{S_t^{(k)}} = r\,dt + \upsilon_k\,dW_t^{(k)}$$

but the standard Brownian motions $\{W_\cdot^{(k)}, k \in \{1, \ldots, d\}\}$ are not necessarily independent. For any $t \ge s$ and $k \in \{1, \ldots, d\}$

$$\ln S_t^{(k)} = \ln S_s^{(k)} + \left(r - 0.5\upsilon_k^2\right)(t-s) + \upsilon_k\sqrt{t-s}\tilde{Y}_k$$

where $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_d) \sim \mathcal{N}_d(0, \Sigma)$. The $d \times d$ matrix Σ is a positive semidefinite matrix with diagonal coefficients equal to 1. Therefore, the variance of the log-return on asset k in the time interval $[s, t]$ is $(t-s)\upsilon_k^2$, and the covariance between the log-returns i, j is $(t-s)\upsilon_i\upsilon_j\Sigma_{i,j}$. It follows that $\Sigma_{i,j}$ is the correlation between the log-returns. The price at time 0 of a European call option with strike price K and exercise time T is given by $\mathrm{E}[\Xi(Y)]$ where

$$\Xi(y) = \exp(-rT)\left(\sum_{k=1}^{d}\alpha_k s_0^{(k)}\exp\left((r-0.5\upsilon_k^2)T+\upsilon_k\sqrt{T}\tilde{y}_k\right)-K\right)_+$$

and $\tilde{y} = \sqrt{\Sigma}y$ ($\sqrt{\Sigma}$ denotes a square root of the matrix Σ i.e. solves $MM^T = \Sigma$). In the numerical applications, Σ is of the form $\Sigma_{i,i} = 1$, $\Sigma_{i,j} = c$, $\alpha_k = 1/d$, $r = 0.05$, $T = 1$, and $d = 40$. We consider $(c, K) \in \{(0.1, 45), (0.1, 60), (0.5, 45), (0.5, 60), (0.9, 45), (0.9, 60)\}$. The initial values $\{s_0^k, k \le d\}$ are realizations of uniform random draws in the range $[20, 80]$; the volatilities $\{\upsilon_k, k \le d\}$ are chosen linearly equally spaced in the set $[0.1, 0.4]$. The assets are sorted so that $\upsilon_1 \le \cdots \le \upsilon_d$.

On Figure 6, we observe the limiting direction $\mu^{(N)}$ which, here again is very close to $\mu_{\mathtt{g}}$. We also plot on Figure 7 a path of $t \mapsto \mu^{(t)}$ along one run of the algorithm `AdaptStr`.

Insert Figure 6 and Figure 7 about here

We report in Table 3 the variance (per sample) of the plain Monte Carlo estimate, of `GHS` and of `AdaptStr`. For `AdaptStr`, we consider the cases $\nu = \nu_\star$ and ν equal to the null vector in the formula (23).

Insert Table 3 about here

5.5.1 Stochastic volatility

We now want to test our method on the pricing of an Asian option in the Heston model, which is specified as follows

$$\begin{aligned} S_t &= S_0 + \int_0^t r S_s \, ds + \int_0^t \sqrt{\xi_s}\, S_s \, (\sqrt{1-\rho^2} dW_s^1 + \rho dW_s^2) \\ X_t &= \int_0^t S_s \, ds \\ \xi_t &= \xi_0 + k \int_0^t (\theta - \xi_s) ds + \sigma \int_0^t \sqrt{\xi_s}\, dW_s^2 \end{aligned}$$

where $\{W_t^1, t \geq 0\}$ and $\{W_t^2, t \geq 0\}$ are two independent Brownian motions, r is the risk free mean rate of return, $\sigma > 0$ is the volatility, $k \geq 0$ the mean reversion rate, $\theta \geq 0$ the long run average volatility, and $\rho \in [-1, 1]$ a correlation rate. The processes $\{S_t, t \geq 0\}$ and $\{\xi_t, t \geq 0\}$ are respectively the stock process and the volatility process, and $\{X_t, t \geq 0\}$ is the integral of the stock price.

The stock and the volatility are driven by SDEs correlated with correlation rate ρ. Indeed by construction $\left\{ \sqrt{1-\rho^2}\, W_t^1 + \rho\, W_t^2,\ t \geq 0 \right\}$ is a Brownian motion with $\left\langle \sqrt{1-\rho^2}\, W^1 + \rho\, W^2, W^2 \right\rangle_t = \rho t$. The price of an Asian Call option at time 0 with strike price K is

$$\mathrm{E}\left[\exp(-rT) \left(X_T - K \right)_+ \right]. \tag{25}$$

An Exact simulation method for the Heston model has recently been proposed in Broadie and Kaya (2006). However, it is computationally intensive especially for pathwise options, and practical numerical schemes for the Heston model are still a very active research field. In our tests we have chosen to use a variation of a scheme introduced in Ninomya and Victoir (2008) and refined in Alfonsi (2008). The weak error of this scheme is potentially of order two. We will not describe this scheme in full details, but will focus on the case where $\sigma < 4k\theta$. Define

$$\psi_k(t) = \frac{1 - e^{-kt}}{k},\ k \neq 0 \quad \text{and} \quad \psi_0(t) = t,$$

and

$$\varphi(\xi, t, y) = e^{-\frac{kt}{2}} \left(\sqrt{(k\theta - \frac{\sigma^2}{4})\psi_k(\frac{t}{2}) + e^{-\frac{kt}{2}}\xi} + \frac{\sigma}{2} y \right)^2 + (k\theta - \frac{\sigma^2}{4})\psi_k(\frac{t}{2}).$$

Consider a regular time grid $0 = t_0 < t_1 < \cdots < t_d = T$, with $t_i = iT/d$ and put $\Delta t = T/d$. At time t_i the scheme is in the state $\left(\hat{S}_i, \hat{X}_i, \hat{\xi}_i \right)$. The next state $\left(\hat{S}_{i+1}, \hat{X}_{i+1}, \hat{\xi}_{i+1} \right)$ is computed by applying:

1. Draw $B_{i+1} \sim \mathcal{U}([0,1])$
2. Draw independently Y_{i+1} and Y_{d+i+1} of law $\mathcal{N}(0,1)$ (independently from B_{i+1})
3. (a) If $B_{i+1} < 0.5$
   - i. Compute $\hat{S}_{i+1/2} = \hat{S}_i \exp\left(\sqrt{(1-\rho^2)\hat{\xi}_i \Delta t}\, Y_{i+1} \right)$
   - ii. Compute $\Delta\hat{\xi}_{i+1} = \varphi(\hat{\xi}_i, \Delta t, \sqrt{\Delta t}\, Y_{d+i+1}) - \hat{\xi}_i$
   - iii. Compute $\hat{X}_{i+1/2} = \hat{X}_i + 0.5 \hat{S}_{i+1/2} \Delta t$.
   - iv. Compute $\hat{S}_{i+1} = \hat{S}_{i+1/2} \exp[(r - \rho k\theta/\sigma)\Delta t + \rho \Delta\hat{\xi}_{i+1}/\sigma + (\rho k/\sigma - 0.5)(\hat{\xi}_i + 0.5\Delta\hat{\xi}_{i+1})\Delta t]$
   - v. Compute $\hat{X}_{i+1} = \hat{X}_{i+1/2} + 0.5 \hat{S}_{i+1} \Delta t$
   - vi. Compute $\hat{\xi}_{i+1} = \hat{\xi}_i + \Delta\hat{\xi}_{i+1}$

   (b) If $B_{i+1} > 0.5$
   - i. Compute $\Delta\hat{\xi}_{i+1} = \varphi(\hat{\xi}_i, \Delta t, \sqrt{\Delta t}\, Y_{d+i+1}) - \hat{\xi}_i$
   - ii. Compute $\hat{X}_{i+1/2} = \hat{X}_i + 0.5 \hat{S}_i \Delta t$
   - iii. Compute $\hat{S}_{i+1/2} = \hat{S}_i \exp[(r - \rho k\theta/\sigma)\Delta t + \rho \Delta\hat{\xi}_{i+1}/\sigma + (\rho k/\sigma - 0.5)(\hat{\xi}_i + 0.5\Delta\hat{\xi}_{i+1})\Delta t]$
   - iv. Compute $\hat{X}_{i+1} = \hat{X}_{i+1/2} + 0.5 \hat{S}_{i+1/2} \Delta t$
   - v. Compute $\hat{\xi}_{i+1} = \hat{\xi}_i + \Delta\hat{\xi}_{i+1}$
   - vi. Compute $\hat{S}_{i+1} = \hat{S}_{i+1/2} \exp\left(\sqrt{(1-\rho^2)\hat{\xi}_{i+1} \Delta t}\, Y_{i+1} \right)$

At each time step i the random variables B_{i+1}, Y_{i+1} and Y_{i+1+d} are drawn independently from the past. The price (25) can be approximated by $\mathrm{E}[\exp(-rT)\left(\hat{X}_d - K\right)_+]$. With the scheme described above we have, in the case $\sigma < 4k\theta$,

$$\mathrm{E}\left[\exp(-rT)\left(\hat{X}_d - K\right)_+\right] = \mathrm{E}\left[\Xi(Y,B)\right],$$

with $Y = (Y_1, \ldots, Y_d, Y_{d+1}, \ldots, Y_{2d}) \sim \mathcal{N}_{2d}(0, \mathrm{Id})$, and $B = (B_1, \ldots, B_d)$ being a vector of independent random variables with law $\mathcal{U}([0,1])$. The vector Y represents the increments of the two Brownian motions W^1 and W^2.

In the case $\sigma > 4k\theta$, the scheme is more complicated and we have,

$$\mathrm{E}\left[\exp(-rT)\left(\hat{X}_d - K\right)_+\right] = \mathrm{E}\left[\Xi(Y,B)\right],$$

with $Y \sim \mathcal{N}_{2d}(0, \mathrm{Id})$, and $B \sim \mathcal{U}([0,1]^{2d})$, Y and B being again independent.

We want now to use our algorithm for the estimation of $\mathrm{E}\left[\Xi(Y,B)\right]$, stratifying only the gaussian vector Y. Our procedure is very easy to adapt to this situation. We consider the case $m = 1$. As Y and B are independent, it is easy to sample under the law $\mathbb{P}(\,(Y,B) \in \cdot | [\mu^{(t)}]^T Y = s)$, with $\mu^{(t)} \in \mathbb{R}^{2d}$. For the computation of the gradient, set $\mu = (\mu_1, \cdots, \mu_{2d})$, and denote by $f_{Y,B}(y,b)$ the density of (Y,B). Using the proof of Proposition 5, we can write

$$\partial_\mu \left(\int \mathbb{1}_{\{y, \langle \mu, y \rangle \leq z\}} \phi(y,b) f_{Y,B}(y,b)\, d\lambda(y,b) \right) = - \int \mathbb{1}_{\{y, \langle \mu, y \rangle = z\}} \frac{y}{|\mu|}\, \phi(y,b) f_{Y,B}(y,b)\, d\lambda(y,b),$$

under mild assumptions on the function ϕ. This allows the computation of an estimated gradient $\widehat{\nabla V}(\mu^{(t)}) \in \mathbb{R}^{2d}$ to update at each time step the current direction $\mu^{(t)} \in \mathbb{R}^{2d}$.

Note that, as in the case $\sigma < 4k\theta$ the vector B is only here to draw Bernoulli variables, we could artificially use standard normal variables to draw these Bernoulli samples (testing positivity). We thus would have to estimate $\mathrm{E}\left[\Xi(Y)\right]$ with $Y \sim \mathcal{N}_{3d}(0, \mathrm{Id})$ and could stratify the whole vector Y. This is not the case for $\sigma > 4k\theta$. We could also think to stratify the hypercube $[0,1]^d$ (or $[0,1]^{2d}$). This was not done in the presented tests.

In the following tests we do not do any previous importance sampling; ν is the null vector in (23). Indeed the additive randomness introduced by B somehow complicates the setting.

We choose $m = 1$, $I = 50$ and $N = 40$. The total amount of drawings done till the end of iteration N is $M\,N = 100\,000$. The parameters of the model are fixed to $S_0 = 100$, $r = 0.1$, $T = 1.0$ and $\sigma = 0.2$. On Figure 8 we plot the evolution of the cost function $t \mapsto (\sum_i p_i \hat{\sigma}_i^{(t)})$, for $K = 100$, $\theta = 0.01$, $k = 2$ and $\rho = 0.5$. The discretization step of the scheme is $d = 100$, and the initial volatility $\xi_0 = 0.01$. The initial direction was arbitrary set to $\mu^{(0)} = (-1, 0, \ldots, 0)$.

Insert Figure 8 about here

We plot on Figure 9[left] the components of $\mu^{(N)}$ with respect to the component index i. Note that on this example the correlation was positive and the two parts of $\mu^{(N)}$, each one corresponding to the increments of respectively W^1 and W^2 are similar. Note that if we take $\rho = -0.5$ (keeping the other parameters unchanged) the algorithm converge to $\mu^{(N)}$, whose components are displayed on Figure 9[right]. This time there is a difference of sign between the components of the first and second half of the vector.

In this example the variance is divided by 25 compared to initial direction, and by 20 compared to plain Monte Carlo. We can wonder on the effect of the moneyness and the volatility of the model on the reduction variance. The results are shown in Table 4. The variance ratio indicated in Table 4 has been computed by dividing an estimation of the variance per sample of the plain Monte Carlo estimator by $(\sum_{\mathbf{i}} p_{\mathbf{i}} \hat{\sigma}_{\mathbf{i}}^{(N)})^2$. We observed indeed that the empirical variance of the estimator $\mathcal{E}^N$ based on the output of 300 independent runs of our procedure, is close to $(\sum_{\mathbf{i}} p_{\mathbf{i}} \hat{\sigma}_{\mathbf{i}}^{(N)})^2/(MN)$. In general the achieved variance reduction is larger when the option is out of the money.

Insert Figure 9 about here

Insert Table 4 about here

5.6 Conclusions

The results show that `AdaptStr` and `GHS` provide similar variance reduction when compared to the crude Monte Carlo procedure. In many applications, $\mu^{(\infty)} \sim \mu_{\mathtt{g}}$; in these cases, in the long time behavior, `AdaptStr` - applied with $\nu = \nu_\star$ in (23) - and `GHS` may be seen as stratification procedures for the estimate of the same target quantity, with the same direction of stratification $\mu_{\mathtt{g}}$ but different allocations (resp. the optimal one, and the proportional one).

In complex applications, the optimization problem (24) is not easy to solve and the `GHS` procedure can not be applied. In that case, the procedure `AdaptStr` can be implemented with ν equal to the null vector in (23). This yields to a significant variance reduction when compared to plain Monte Carlo.

`AdaptStr` is thus an efficient stratification procedure, that learns "on the fly" the direction of stratification and the optimal allocation. It can be combined with importance sampling (choice of ν in (23)) and the direction $\nu_\star$ that solves (24) is an efficient drift vector. Even when the stratification procedure is not combined with importance sampling, `AdaptStr` still strongly reduces the variance w.r.t. the crude Monte Carlo procedure.

6 Proofs

6.1 Proofs of Sections 3

In the sequel, we denote $\mathcal{I}_m \stackrel{\text{def}}{=} \{1, \ldots, I\}^m$.

6.2 Proofs of Section 3

Lemma 1 *Let $m < d$, μ be a $d \times m$ orthonormal matrix, p.d.f. densities $g_1, \cdots, g_m$ on $\mathbb{R}$ and χ be a density on $\mathbb{R}^m$. Let g be given by (11) and $\{\mathsf{S}_{\mathbf{i}}, \mathbf{i} \in \mathcal{I}_m\}$ be the strata given by (2) and define*

$$p_{\mathbf{i}}\sigma_{\mathbf{i}} \stackrel{\text{def}}{=} \sqrt{\left(\int_{\mathsf{S}_{\mathbf{i}}} f_\mu \, d\lambda\right)\left(\int_{\mathsf{S}_{\mathbf{i}}} \zeta_\mu f_\mu \, d\lambda\right) - \left(\int_{\mathsf{S}_{\mathbf{i}}} \psi_\mu f_\mu \, d\lambda\right)^2} \, .$$

(i) Let $\epsilon > 0$. For any $M > \epsilon^{-1}$,

$$\sup_{\mathcal{Q}:\inf_{\mathbf{i}\in\mathcal{I}_m} q_{\mathbf{i}} \geq \epsilon} \left| M\varsigma^2_{I,M}(\mu, g, \mathcal{Q}) - \sum_{\mathbf{i}\in\mathcal{I}_m} q_{\mathbf{i}}^{-1} \; p_{\mathbf{i}}^2 \, \sigma_{\mathbf{i}}^2 \right| \leq \frac{1}{M\epsilon(\epsilon - M^{-1})} \; \mathrm{Var}[\phi(Y)] \, .$$

(ii) Assume that $\mathrm{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right) > 0$ and $\mathrm{esssup}_{\chi\cdot\lambda}\left(f_\mu \chi^{-1}\right) < +\infty$. Let $\epsilon > 0$. For any (I, M) such that $MI^{-m}\,\mathrm{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right) \geq 1 + \epsilon$

$$\left| M\varsigma^2_{I,M}(\mu, g, \mathcal{Q}_\chi) - \sum_{\mathbf{i}\in\mathcal{I}_m} [q_{\mathbf{i}}(\chi)]^{-1} \; p_{\mathbf{i}}^2 \, \sigma_{\mathbf{i}}^2 \right| \leq \frac{(1+\epsilon^{-1})\mathrm{Var}[\phi(Y)]}{\mathrm{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right)} \; \frac{I^m}{M} \; \left(\operatorname*{esssup}_{\chi\cdot\lambda}\left(\frac{f_\mu}{\chi}\right) \wedge \frac{I^m}{\mathrm{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right)} \right) \, .$$

(iii) For any positive integers M, I and real $\epsilon > 1$,

$$\left| M\varsigma^2_{I,M}(\mu, g, \mathcal{Q}^\star(\mu)) - \sum_{\mathbf{i}\in\mathcal{I}_m} \left(q^\star_{\mathbf{i}}[\mathcal{S}(\mu)]\right)^{-1} \; p_{\mathbf{i}}^2 \, \sigma_{\mathbf{i}}^2 \right| \leq \mathrm{Var}[\phi(Y)] \; \left((1+\epsilon)\frac{I^m}{M} + \frac{1}{\epsilon - 1} \right) \, ,$$

where $\mathcal{Q}^\star[\mathcal{S}(\mu) = \{q^\star_{\mathbf{i}}[\mathcal{S}(\mu)], \mathbf{i} \in \mathcal{I}_m\}$ is the optimal allocation defined by (9).

Proof It is easily shown that $M\ \varsigma^2_{I,M}(\mu, g, \mathcal{Q}_\chi) = \sum_{\mathbf{i}\in\mathcal{I}_m:M_\mathbf{i}>0} M M_\mathbf{i}^{-1} p_\mathbf{i}^2\sigma_\mathbf{i}^2$. By definition of $M_\mathbf{i}$ (see Eq. 4), $M_\mathbf{i} = 0$ when $q_\mathbf{i} = 0$ and $M_\mathbf{i} \geq 1$ when $q_\mathbf{i} \geq M^{-1}$. One may have $M_\mathbf{i} = 1$ when $q_\mathbf{i} \in (0, M^{-1})$ but then $MM_\mathbf{i}^{-1} \leq q_\mathbf{i}^{-1}$. Hence,

$$\left| M\ \varsigma^2_{I,M}(\mu, g, \mathcal{Q}_\chi) - \sum_{\mathbf{i}\in\mathcal{I}_m:q_\mathbf{i}>0} \frac{p_\mathbf{i}^2\sigma_\mathbf{i}^2}{q_\mathbf{i}} \right| \leq \sum_{\mathbf{i}\in\mathcal{I}_m,q_\mathbf{i}\geq 1/M} \left|\frac{Mq_\mathbf{i} - M_\mathbf{i}}{M_\mathbf{i}}\right| \frac{p_\mathbf{i}^2\sigma_\mathbf{i}^2}{q_\mathbf{i}} + \sum_{\mathbf{i}\in\mathcal{I}_m:0<q_\mathbf{i}<1/M} \frac{p_\mathbf{i}^2\sigma_\mathbf{i}^2}{q_\mathbf{i}} \quad (26)$$

(i) Since $q_\mathbf{i} \geq \epsilon > M^{-1}$, the second term in the rhs of (26) is null and since by (4), $Mq_\mathbf{i} - 1 < M_\mathbf{i} < Mq_\mathbf{i} + 1$, the first term is upper bounded by

$$M^{-1} \left(\sup_{\mathbf{i}\in\mathcal{I}_m:q_\mathbf{i}\geq M^{-1}} p_\mathbf{i} q_\mathbf{i}^{-1} \right) \sum_{\mathbf{i}\in\mathcal{I}_m,q_\mathbf{i}\geq 1/M} (q_\mathbf{i} - M^{-1})^{-1}\ p_\mathbf{i}\sigma_\mathbf{i}^2$$

which yields the desired result upon noting that $p_\mathbf{i} q_\mathbf{i}^{-1} \leq q_\mathbf{i}^{-1} \leq \epsilon^{-1}$ and $\sum_\mathbf{i} p_\mathbf{i}\sigma_\mathbf{i}^2 \leq \mathrm{Var}[\phi(Y)]$.

(ii) Under the stated assumptions, $q_\mathbf{i}(\chi) = \int_{\mathsf{S}_\mathbf{i}} \chi d\lambda \geq \mathrm{essinf}_{g\cdot\lambda} \left(\chi g^{-1}\right)\ \ I^{-m}$. Hence $Mq_\mathbf{i} \geq 1 + \epsilon$ which implies that the second term in the rhs of (26) is null. This also implies that

$$q_\mathbf{i} - M^{-1} \geq \left(1 - \frac{1}{1+\epsilon}\right) \operatorname*{essinf}_{g\cdot\lambda} \left(\chi g^{-1}\right)\ I^{-m}$$

To conclude the proof,

$$\frac{p_\mathbf{i}}{q_\mathbf{i}(\chi)} = \frac{\int_{\mathsf{S}_\mathbf{i}} f_\mu\ d\lambda}{\int_{\mathsf{S}_\mathbf{i}} \chi\ d\lambda} \leq \operatorname*{esssup}_{\chi\cdot\lambda} \left(\frac{f_\mu}{\chi}\right) \wedge \frac{1}{q_\mathbf{i}(\chi)} \leq \operatorname*{esssup}_{\chi\cdot\lambda} \left(\frac{f_\mu}{\chi}\right) \wedge \frac{I^m}{\mathrm{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right)}\ .$$

(iii) Note that by convention, $p_\mathbf{i}^2\sigma_\mathbf{i}^2/q_\mathbf{i}^\star[\mathcal{S}(\mu)] = 0$ when $q_\mathbf{i}^\star[\mathcal{S}(\mu)] = 0$. By definition of the optimal allocation (see Eq. 9),

$$\left(q_\mathbf{i}^\star[\mathcal{S}(\mu)]\right)^{-1}\ p_\mathbf{i}^2\sigma_\mathbf{i}^2 \leq q_\mathbf{i}^\star[\mathcal{S}(\mu)] \left(\sum_\mathbf{j} p_\mathbf{j}\sigma_\mathbf{j}\right)^2 \leq q_\mathbf{i}^\star[\mathcal{S}(\mu)]\ \mathrm{Var}[\phi(Y)]\ .$$

The second term in the rhs of (26) is upper bounded by $I^m M^{-1}\ \mathrm{Var}[\phi(Y)]$. For the first term,

$$\begin{aligned}
&[\mathrm{Var}[\phi(Y)]]^{-1} \sum_{\mathbf{i}\in\mathcal{I}_n,q_\mathbf{i}^\star[\mathcal{S}(\mu)]\geq 1/M} \left|\frac{Mq_\mathbf{i} - M_\mathbf{i}}{M_\mathbf{i}}\right| \frac{p_\mathbf{i}^2\sigma_\mathbf{i}^2}{q_\mathbf{i}^\star[\mathcal{S}(\mu)]} \\
&\quad\leq \sum_{\mathbf{i}\in\mathcal{I}_m,1/M\leq q_\mathbf{i}^\star[\mathcal{S}(\mu)]\leq\epsilon/M} \left|\frac{Mq_\mathbf{i} - M_\mathbf{i}}{M_\mathbf{i}}\right|\ q_\mathbf{i}^\star[\mathcal{S}(\mu)] + \sum_{\mathbf{i}\in\mathcal{I}_m,q_\mathbf{i}^\star[\mathcal{S}(\mu)]\geq\epsilon/M} \left|\frac{Mq_\mathbf{i} - M_\mathbf{i}}{M_\mathbf{i}}\right|\ q_\mathbf{i}^\star[\mathcal{S}(\mu)]\ .
\end{aligned}$$

For all $\mathbf{i}$ such that $q_\mathbf{i}^\star[\mathcal{S}(\mu)] \geq 1/M$, $M_\mathbf{i}^{-1}|Mq_\mathbf{i} - M_\mathbf{i}| \leq 1$ which implies that

$$\sum_{\mathbf{i}\in\mathcal{I}_m,1/M\leq q_\mathbf{i}^\star[\mathcal{S}(\mu)]\leq\epsilon/M} \left|\frac{Mq_\mathbf{i} - M_\mathbf{i}}{M_\mathbf{i}}\right| q_\mathbf{i}^\star[\mathcal{S}(\mu)] \leq \frac{\epsilon I^m}{M}\ .$$

For all $\mathbf{i}$ such that $q_\mathbf{i}^\star[\mathcal{S}(\mu)] \geq \epsilon/M$, $M_\mathbf{i}^{-1}|Mq_\mathbf{i} - M_\mathbf{i}| \leq M_\mathbf{i}^{-1} \leq (Mq_\mathbf{i}^\star[\mathcal{S}(\mu)] - 1)^{-1} \leq (\epsilon - 1)^{-1}$ which implies that

$$\sum_{\mathbf{i}\in\mathcal{I}_m,q_\mathbf{i}^\star[\mathcal{S}(\mu)]\geq\epsilon/M} \left|\frac{Mq_\mathbf{i} - M_\mathbf{i}}{M_\mathbf{i}}\right| q_\mathbf{i}^\star[\mathcal{S}(\mu)] \leq (\epsilon - 1)^{-1}\ .$$

Proof of Proposition 1 To prove the Proposition 1, we need the two following Lemmas. Define

$$G^{-1}(x_1, \ldots, x_m) \stackrel{\mathrm{def}}{=} (G_1^{-1}(x_1), \ldots, G_m^{-1}(x_m))\ , \quad (27)$$

where G_k is the c.d.f. associated to the density g_k on $\mathbb{R}$. The first is a standard change of variables formula (see for example, (Dudley, 2002, Theorem 4.1.11)).

Lemma 2 *Let $h : \mathbb{R}^m \to \mathbb{R}$ be a measurable function. Assume that h is nonnegative or is such that $\int_{\mathbb{R}^m} |h| 1_{\{g>0\}}\ d\lambda < +\infty$. Then, for all $0 \le v_k \le w_k \le 1$, $k \in \{1, \dots, I\}$*

$$\int_{\prod_{k=1}^m [G_k^{-1}(v_k), G_k^{-1}(w_k)]} h 1_{\{g>0\}}\ d\lambda = \int_{\prod_{k=1}^m [v_k, w_k]} \frac{h}{g} \circ G^{-1}\ d\lambda\,. \tag{28}$$

The second technical Lemma is our key approximation result.

Lemma 3 *Let $h, \gamma : \mathbb{R}^m \to \mathbb{R}$ be functions such that $\int_{\mathbb{R}^m} \left(h^2 + \gamma^2\right) g^{-1}\ d\lambda < +\infty$. Define for $\mathbf{i} \in \mathcal{I}_m$,*

$$R_{\mathbf{i}}[h, \gamma] \stackrel{\text{def}}{=} \int_{\mathsf{S}_{\mathbf{i}}} h\gamma g^{-1}\ d\lambda - I^m \left(\int_{\mathsf{S}_{\mathbf{i}}} h\ d\lambda\right)\left(\int_{\mathsf{S}_{\mathbf{i}}} \gamma\ d\lambda\right)\,. \tag{29}$$

Then $\lim_{I\to+\infty} \sum_{\mathbf{i}\in\mathcal{I}_m} |R_{\mathbf{i}}[h, \gamma]| = 0$.

Proof By polarization, it is enough to prove the result when $\gamma = h$ with $\int_{\mathbb{R}^m} h^2 g^{-1}\ d\lambda < +\infty$. This integrability condition ensures that λ-a.e. , $g = 0$ implies $h = 0$ and by (28), one has

$$R_{\mathbf{i}}[h, h] = \int_{\prod_{k=1}^m [(i_k-1)/I, i_k/I]} \frac{h^2}{g^2} \circ G^{-1}\ d\lambda - I^m \left(\int_{\prod_{k=1}^m [(i_k-1)/I, i_k/I]} \frac{h}{g} \circ G^{-1}\ d\lambda\right)^2\,,$$

where the right-hand-side is non-negative by Cauchy-Schwarz inequality. Set $\tilde{h}(u) \stackrel{\text{def}}{=} \frac{h}{g}(G^{-1}(u))$ if $u \in (0,1)^m$ and 0 otherwise. By (28) and the integrability assumption made on h, the function $\tilde{h}$ is square integrable on $\mathbb{R}^m$. Using the definition of $\tilde{h}$ for the first equality and symmetry for the second one, one has

$$\begin{aligned}\sum_{\mathbf{i}\in\mathcal{I}_m} R_{\mathbf{i}}[h, h] &= I^m \sum_{\mathbf{i}\in\mathcal{I}_m} \int_{\mathcal{J}_{\mathbf{i}}^2} \tilde{h}(u)\{\tilde{h}(u) - \tilde{h}(v)\} du dv = \frac{I^m}{2} \sum_{\mathbf{i}\in\mathcal{I}_m} \int_{\mathcal{J}_{\mathbf{i}}^2} \{\tilde{h}(u) - \tilde{h}(v)\}^2\ du dv \\ &= \frac{I^m}{2} \sum_{\mathbf{i}\in\mathcal{I}_m} \int_{\mathcal{J}_{\mathbf{i}}} \int_{\mathcal{J}_{\mathbf{i}} - u_k} \{\tilde{h}(u) - \tilde{h}(u+w)\}^2\ du dw \le \frac{1}{2} \int_{[0,1]^m} \int_{[-1,1]^m} (\tilde{h}(u) - \tilde{h}(u + z/I))^2\ du dz\,.\end{aligned}$$

where we have set, for $\mathbf{i} \in \{1, \dots, m\}$, $\mathcal{J}_{\mathbf{i}} = \prod_{k=1}^m [(i_k - 1)/I, i_k/I]$. By continuity of the translations in $L^2(\mathbb{R}^m, du)$ and the dominated convergence Theorem, one obtains that the right-hand-side converges to 0 as $I \to \infty$.

We now proceed to the proof of Proposition 1. Under A1, it holds that

$$q_{\mathbf{i}}(\chi) \ge \left(\operatorname*{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right)\right) \int_{\mathsf{S}_{\mathbf{i}}} g\ d\lambda = I^{-m} \operatorname*{essinf}_{g\cdot\lambda}\left(\chi g^{-1}\right)\,. \tag{30}$$

Hence, by Lemma 1(i), to prove the first assertion, it is enough to check that $\lim_{I\to+\infty} \sum_{\mathbf{i}\in\{1,\dots,I\}^m} \frac{p_{\mathbf{i}}^2\ \sigma_{\mathbf{i}}^2}{q_{\mathbf{i}}(\chi)} = \varsigma_\infty^2(\mu, \chi)$. By definition of $R_{\mathbf{i}}$ (see Eq. (29)),

$$\begin{aligned}\frac{p_{\mathbf{i}}^2\ \sigma_{\mathbf{i}}^2}{q_{\mathbf{i}}(\chi)} &= \frac{\left(\int_{\mathsf{S}_{\mathbf{i}}} f_\mu\ d\lambda\right)\left(\int_{\mathsf{S}_{\mathbf{i}}} [\zeta_\mu f_\mu]\ d\lambda\right) - \left(\int_{\mathsf{S}_{\mathbf{i}}} [\psi_\mu f_\mu]\ d\lambda\right)^2}{\int_{\mathsf{S}_{\mathbf{i}}} \chi\ d\lambda} \\ &= \frac{\int_{\mathsf{S}_{\mathbf{i}}} f_\mu^2(\zeta_\mu - \psi_\mu^2) g^{-1}\ d\lambda - R_{\mathbf{i}}[f_\mu, \zeta_\mu f_\mu] + R_{\mathbf{i}}[\psi_\mu f_\mu, \psi_\mu f_\mu]}{I^m\ \int_{\mathsf{S}_{\mathbf{i}}} \chi\ d\lambda}\,,\end{aligned}$$

and

$$\varsigma_\infty^2(\mu, \chi) = \frac{\int_{\mathsf{S}_{\mathbf{i}}} f_\mu^2(\zeta_\mu - \psi_\mu^2) g^{-1}\ d\lambda - R_{\mathbf{i}}[\chi, f_\mu^2(\zeta_\mu - \psi_\mu^2)\chi^{-1}]}{I^m\ \int_{\mathsf{S}_{\mathbf{i}}} \chi\ d\lambda}\,.$$

Therefore

$$\sum_{\mathbf{i}\in\mathcal{I}_m} \frac{p_{\mathbf{i}}^2\ \sigma_{\mathbf{i}}^2}{q_{\mathbf{i}}(\chi)} - \varsigma_\infty^2(\mu, \chi) = \sum_{\mathbf{i}\in\mathcal{I}_m} \frac{R_{\mathbf{i}}[\chi, f_\mu^2(\zeta_\mu - \psi_\mu^2)\chi^{-1}] + R_{\mathbf{i}}[\psi_\mu f_\mu, \psi_\mu f_\mu] - R_{\mathbf{i}}[f_\mu, \zeta_\mu f_\mu]}{I^m\ \int_{\mathsf{S}_{\mathbf{i}}} \chi\ d\lambda},$$

and one easily concludes with (30) and Lemma 3 (which applies under A2 and A3). The second assertion is a consequence of Lemma 1(ii).

Proof of Proposition 2 For ease of notations, in this proof, the dependence upon μ and the strata $\{\mathsf{S}_\mathbf{i}, \mathbf{i} \in \mathcal{I}_m$ is omitted. We denote $p_\mathbf{i}\sigma_\mathbf{i} \stackrel{\text{def}}{=} \sqrt{\left(\int_{\mathsf{S}_\mathbf{i}} f_\mu \, d\lambda\right)\left(\int_{\mathsf{S}_\mathbf{i}} \zeta_\mu f_\mu \, d\lambda\right) - \left(\int_{\mathsf{S}_\mathbf{i}} \psi_\mu f_\mu \, d\lambda\right)^2}$. Since for $a, b \geq 0$, $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$, one has

$$\begin{aligned}
&\sum_{\mathbf{i}\in\mathcal{I}_m} \left| p_\mathbf{i}\sigma_\mathbf{i} - \int_{\mathsf{S}_\mathbf{i}} \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda \right| \\
&\qquad \leq \sum_{\mathbf{i}\in\mathcal{I}_m} \left| \int_{\mathsf{S}_\mathbf{i}} f_\mu \, d\lambda \int_{\mathsf{S}_\mathbf{i}} \zeta_\mu f_\mu \, d\lambda - \left(\int_{\mathsf{S}_\mathbf{i}} \psi_\mu f_\mu \, d\lambda\right)^2 - \left(\int_{\mathsf{S}_\mathbf{i}} \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda\right)^2 \right|^{1/2} \\
&\qquad = \sum_{\mathbf{i}\in\mathcal{I}_m} \sqrt{\frac{1}{I^m} \left| -R_\mathbf{i}[f_\mu, \zeta_\mu f_\mu] + R_\mathbf{i}[\psi_\mu f_\mu, \psi_\mu f_\mu] + R_\mathbf{i}[f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}, f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}] \right|} \\
&\qquad \leq \left(\sum_{\mathbf{i}\in\{1,\ldots,I\}^m} \left| -R_\mathbf{i}[f_\mu, \zeta_\mu f_\mu] + R_\mathbf{i}[\psi_\mu f_\mu, \psi_\mu f_\mu] + R_\mathbf{i}[f_\mu\sqrt{\zeta_\mu - \psi_\mu^2}, f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}] \right| \right)^{1/2} .
\end{aligned}$$

Under A2, $\int f_\mu^2(\zeta_\mu - \psi_\mu^2) g^{-1} \, d\lambda < +\infty$, and by Lemma 3, the right-hand-side converges to 0 as $I \to +\infty$. Therefore,

$$\lim_{I\to+\infty} \sum_{\mathbf{i}\in\mathcal{I}_m} \left| p_\mathbf{i}\sigma_\mathbf{i} - \int_{\mathsf{S}_\mathbf{i}} \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda \right| = 0 \, . \tag{31}$$

We write

$$\begin{aligned}
&\left(\int \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda\right) \sum_{\mathbf{i}\in\mathcal{I}_m} \left| q_\mathbf{i}(\chi_\mu^\star) - q_\mathbf{i}^\star[\mathcal{S}(\mu)] \right| \\
&\qquad \leq \sum_{\mathbf{i}\in\mathcal{I}_m} q_\mathbf{i}^\star[\mathcal{S}(\mu)] \left| \sum_{\mathbf{j}\in\mathcal{I}_m} p_\mathbf{j}\sigma_\mathbf{j} - \int \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda \right| + \sum_{\mathbf{i}\in\mathcal{I}_m} \left| p_\mathbf{i}\sigma_\mathbf{i} - \int_{\mathsf{S}_\mathbf{i}} \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda \right| .
\end{aligned}$$

By Eq.(31), the rhs tend to zero as $I \to +\infty$. The second assertion is a consequence of Lemma 1(iii) applied with $\epsilon = \sqrt{M/I^m}$ and of Eq. (31) upon noting that

$$\left| \sum_{\mathbf{i}\in\mathcal{I}_m} p_\mathbf{i}\sigma_\mathbf{i} - \int_{\mathbb{R}^m} \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda \right| \leq \sum_{\mathbf{i}\in\mathcal{I}_m} \left| p_\mathbf{i}\sigma_\mathbf{i} - \int_{\mathsf{S}_\mathbf{i}} \left[f_\mu \sqrt{\zeta_\mu - \psi_\mu^2}\right] \, d\lambda \right| .$$

Proof of Proposition 3 Since $\operatorname{esssup}_{\chi.\lambda}\left(f_\mu \chi^{-1}\right) \leq \operatorname{esssup}_\lambda\left(f_\mu g^{-1}\right) / \operatorname{essinf}_{g.\lambda}\left(\chi g^{-1}\right) < +\infty$, Lemma 1(ii) ensures that it is enough to check that

$$\lim_{I\to+\infty} I^2 \sum_{\mathbf{i}\in\{1,\ldots,I\}^d} \frac{p_\mathbf{i}^2(\xi)\sigma_\mathbf{i}^2(\xi)}{q_\mathbf{i}} = \varsigma_\infty^2(\mu, g, \chi).$$

In the sequel, for $\mathbf{i} = (i_1, \ldots, i_d) \in \{1, \ldots, I\}^d$, we denote $\mathcal{J}_\mathbf{i} \stackrel{\text{def}}{=} \prod_{j=1}^d [(i_j - 1)/I, i_j/I]$. Set $\tilde{f}(u) \stackrel{\text{def}}{=} [f_\mu]/g]G^{-1}(u)$ if $u \in (0,1)^d$ and 0 otherwise and similarly, $\tilde{h}_k(u) \stackrel{\text{def}}{=} [\partial_k\phi_\mu/g_k](G^{-1}(u))$ if $u \in (0,1)^d$ and 0 otherwise. Using symmetry and (28) one obtains

$$p_\mathbf{i}^2\sigma_\mathbf{i}^2 = \int_{\mathsf{S}_\mathbf{i}}\int_{\mathsf{S}_\mathbf{i}} f_\mu(x) f_\mu(y) \phi_\mu(y)(\phi_\mu(y) - \phi_\mu(x)) dxdy = \frac{1}{2} \int_{\mathcal{J}_\mathbf{i}^2} \tilde{f}(u)\tilde{f}(v) \left(\phi_\mu(G^{-1}(v)) - \phi_\mu(G^{-1}(u))\right)^2 dvdu.$$

Since ϕ is continuously differentiable, Eq. (28) implies that $\phi(G^{-1}(v)) - \phi(G^{-1}(u)) = \sum_{k=1}^d \int_{u_k}^{v_k} \tilde{h}_k(uv_k(t))dt$ where $uv_k(t) \stackrel{\text{def}}{=} (u_1, \ldots, u_{k-1}, t, v_{k+1}, \ldots, v_d)$. Therefore,

$$p_\mathbf{i}^2\sigma_\mathbf{i}^2 = \frac{1}{2} \sum_{k,l=1}^d \iint_{\mathcal{J}_\mathbf{i}\times\mathcal{J}_\mathbf{i}} \int_{u_k}^{v_k} \int_{u_l}^{v_l} \tilde{f}(u)\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s)) dtdsdvdu. \tag{32}$$

We may similarly obtain

$$\int_{\mathsf{S}_\mathbf{i}} \frac{(f_\mu \partial_k \phi_\mu)^2}{g_k^2 \chi} d\lambda = \frac{\int_{S_\mathbf{i}} \frac{(f_\mu \partial_k \phi_\mu)^2}{g g_k^2} d\lambda - R_\mathbf{i}[\frac{(f_\mu \partial_k \phi_\mu)^2}{g_k^2 \chi}, \chi]}{I^d \int_{S_\mathbf{i}} \chi d\lambda} = \frac{\int_{\mathcal{J}_\mathbf{i}} (\tilde{f}\tilde{h}_k)^2 d\lambda - R_\mathbf{i}[\frac{(f_\mu \partial_k \phi_\mu)^2}{g_k^2 \chi}, \chi]}{I^d \int_{S_\mathbf{i}} \chi d\lambda}.$$

Noting that

$$\frac{1}{2} \int_{\mathcal{J}_\mathbf{i}^2} \int_{u_k}^{v_k} \int_{u_l}^{v_l} dt ds dv du = \frac{1_{\{k=l\}}}{12 I^{2d+2}}, \tag{33}$$

one deduces that

$$\begin{aligned} I^2 \sum_{\mathbf{i} \in \{1,\dots,I\}^d} \frac{p_\mathbf{i}^2 \sigma_\mathbf{i}^2}{q_\mathbf{i}} - \varsigma_\infty^2(\mu, g, \chi) &= \frac{1}{12} \sum_{\mathbf{i} \in \{1,\dots,I\}^d} \frac{R_\mathbf{i}[\sum_{k=1}^d \frac{(f_\mu \partial_k \phi_\mu)^2}{g_k^2 \chi}, \chi]}{I^d \int_{S_\mathbf{i}} \chi d\lambda} \\ + \sum_{k,l=1}^d \sum_{\mathbf{i} \in \{1,\dots,I\}^d} & \frac{I^{2d+2} \int_{\mathcal{J}_\mathbf{i}^3} \int_{u_k}^{v_k} \int_{u_l}^{v_l} \left(\tilde{f}(u)\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s)) - \tilde{f}^2 \tilde{h}_k \tilde{h}_l(w) \right) dt ds dv du dw}{2 I^d \int_{S_\mathbf{i}} \chi d\lambda} \end{aligned} \tag{34}$$

Since $\int_{\mathbb{R}^d} \frac{(f_\mu \partial_k \phi_\mu)^4}{g g_k^4 \chi^2} d\lambda \leq \frac{\left(\mathrm{esssup}_\lambda \left(\frac{f_\mu}{g}\right) \mathrm{esssup}_\lambda \left(\frac{|\partial_k \phi_\mu|}{g_k}\right)\right)^4}{\left(\mathrm{essinf}_{g \cdot \lambda} \left(\frac{\chi}{g}\right)\right)^2} < +\infty$, by Lemma 3 and (30), the first term of the right-hand-side converges to 0 as $I \to +\infty$. Let us now prove that for fixed k and l in $\{1, \dots, d\}$, the corresponding sum of ratios over $\mathbf{i}$ in the second term also converges to 0. As the denominators are bounded from below away from 0 by (30), it is enough to check that the sum of the numerators tends to 0. For $u, v \in \mathbb{R}^d$ $uv_p \stackrel{\text{def}}{=} (u_1, \dots, u_p, v_{p+1}, \dots, v_d)$ if $p \in \{1, \dots, d-1\}$ and $uv_0 \stackrel{\text{def}}{=} v$, $uv_d \stackrel{\text{def}}{=} u$. One has

$$\begin{aligned} & \left| \int_{\mathcal{J}_\mathbf{i}^3} \int_{u_k}^{v_k} \int_{u_l}^{v_l} \left(\tilde{f}(u)\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s)) - \tilde{f}^2 \tilde{h}_k \tilde{h}_l(w) \right) dt ds dv du dw \right| \\ & \leq \int_{\mathcal{J}_\mathbf{i}^3} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} \left| \tilde{f}(u)\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s)) - \tilde{f}^2 \tilde{h}_k \tilde{h}_l(w) \right| dt ds dv du dw \\ & \leq \sum_{p=1}^d \Big(\int_{\mathcal{J}_\mathbf{i}^3} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} |\tilde{f}(uw_p) - \tilde{f}(uw_{p-1})||\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s))| dt ds dv du dw \\ & \qquad + \int_{\mathcal{J}_\mathbf{i}^3} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} |\tilde{f}(vw_p) - \tilde{f}(vw_{p-1})||\tilde{f}(w)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s))| dt ds dv du dw \\ & \qquad + \int_{\mathcal{J}_\mathbf{i}^3} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} |\tilde{h}_k([uv_k(t)]w_p) - \tilde{h}_k([uv_k(t)]w_{p-1})||\tilde{f}^2(w)\tilde{h}_l(uv_l(s))| dt ds dv du dw \\ & \qquad + \int_{\mathcal{J}_\mathbf{i}^3} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} |\tilde{h}_l([uv_l(s)]w_p) - \tilde{h}_l([uv_l(s)]w_{p-1})||\tilde{f}^2(w)\tilde{h}_k(w)| dt ds dv du dw \Big). \end{aligned}$$

In each of the $4d$ integrals in the right-hand-side, only $d+1$ of the $3d+2$ integration variables are involved in the difference which appears in the integrand. Their domain of integration is $\prod_{j=1}^d [(i_j - 1)/I, i_j/I] \times [(i_p - 1)/I, i_p/I]$. Integrating first the absolute value of the product of three functions with respect to the $2d+1$ remaining variables one obtains a function of these $d+1$ variables smaller than CI^{2d+1} with $C = \left(\mathrm{esssup}_\lambda \left(\frac{f_\mu}{g}\right) \vee \mathrm{esssup}_\lambda \left(\sum_{k=1}^d \frac{|\partial_k \phi_\mu|}{g_k} \right) \right)^3$. Dealing for instance with the p-th integral of the first kind, one has

$$\begin{aligned} & \sum_{\mathbf{i} \in \{1,\dots,I\}^d} I^{2d+2} \int_{\mathcal{J}_\mathbf{i}^3} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} |\tilde{f}(uw_p) - \tilde{f}(uw_{p-1})||\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s))| dt ds dv du dw \\ & \leq C \sum_{\mathbf{i} \in \{1,\dots,I\}^d} I \int_{\prod_{j=1}^d [(i_j-1)/I, i_j/I] \times [(i_p-1)/I, i_p/I]} |\tilde{f}(uu_p(t)) - \tilde{f}(u)| dt du \\ & \leq C \int_{-1}^1 \int_{\mathbb{R}^d} |\tilde{f}(u + \frac{s}{I} e_p) - \tilde{f}(u)| du ds \end{aligned}$$

where e_p denotes the p-th element of the canonical basis on $\mathbb{R}^d$. By continuity of the translations in $L^1(\mathbb{R}^d, du)$, one concludes that the second term of the right-hand-side of (34) tends to 0.

Proof of Proposition 4 By Lemma 1(iii), it is enough to check that

$$\lim_{I\to+\infty} I \sum_{\mathbf{i}\in\{1,\dots,I\}^d} p_{\mathbf{i}}\sigma_{\mathbf{i}} = \varsigma_\infty(\mu, g, \chi^\star_{\mu,g}).$$

Since for $a, b \geq 0$, $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$, the relations (32), (33) together with the Cauchy-Schwarz inequality imply

$$\begin{aligned}
&\left| I \sum_{\mathbf{i}\in\{1,\dots,I\}^d} p_{\mathbf{i}}\sigma_{\mathbf{i}} - \varsigma_\infty(\mu, g, \chi^\star_{\mu,g}) \right| \\
&\leq \sum_{\mathbf{i}\in\{1,\dots,I\}^d} \left| \frac{I^2}{2} \sum_{k,l=1}^d \int_{\mathcal{J}_{\mathbf{i}}^2} \int_{u_k}^{v_k} \int_{u_l}^{v_l} \tilde{f}(u)\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s))dtdsdvdu - \frac{1}{12}\sum_{k=1}^d \int_{\mathcal{J}_{\mathbf{i}}} \tilde{f}\tilde{h}_k(w)\tilde{f}\tilde{h}_k(r)drdw \right|^{1/2} \\
&\leq \Bigg(\sum_{k,l=1}^d \sum_{\mathbf{i}\in\{1,\dots,I\}^d} I^{3d+2} \int_{\mathcal{J}_{\mathbf{i}}^4} \int_{\frac{i_k-1}{I}}^{\frac{i_k}{I}} \\
&\qquad\qquad \int_{\frac{i_l-1}{I}}^{\frac{i_l}{I}} \left| \tilde{f}(u)\tilde{f}(v)\tilde{h}_k(uv_k(t))\tilde{h}_l(uv_l(s)) - \tilde{f}\tilde{h}_k(w)\tilde{f}\tilde{h}_k(r) \right| dtdsdvdudwdr \Bigg)^{1/2}.
\end{aligned}$$

Reasoning like in the end of the proof of Proposition 3, one concludes that the right-hand-side converges to 0 as $I \to +\infty$.

6.3 Proofs of Section 4

Proof of Proposition 5 Let $H \in \mathbb{R}^d$ be such that $|H| < |\mu|$, $e_1 = \frac{\mu}{|\mu|}$, $a = \langle H, e_1\rangle$, $b = |H - ae_1|$ and e_2 be equal to $\frac{H-ae_1}{b}$ if $b \neq 0$ and to any vector with norm 1 orthogonal to e_1 otherwise. We complete (e_1, e_2) with $(e_3, \dots, e_d)$ to obtain an orthonormal basis of $\mathbb{R}^d$. For $\alpha \in \mathbb{R}^d$, $\alpha_k = \langle \alpha, e_k\rangle$.

$$\begin{aligned}
g_z(\mu+H) - g_z(\mu) &= \int_{\{\alpha,\alpha_1 \leq \frac{z-\alpha_2 b}{|\mu|+a}\}} h(\alpha)\, d\alpha - \int_{\{\alpha,\alpha_1\leq\frac{z}{|\mu|}\}} h(\alpha)\, d\alpha \\
&= \int_{\mathbb{R}^{d-1}} \int_{\frac{z}{|\mu|}}^{\frac{z-\alpha_2 b}{|\mu|+a}} h(\alpha) d\alpha_1 d\alpha_{2:d} \\
&= -\int_{\mathbb{R}^{d-1}} \int_0^1 h\left(\frac{z-\alpha_2 bs}{|\mu|+as} e_1 + \sum_{k=2}^d \alpha_k e_k \right) \frac{az + \alpha_2 b|\mu|}{(|\mu|+as)^2} ds d\alpha_{2:d} \\
&= -\int_0^1 \int_{\mathbb{R}^{d-1}} h\Bigg(z\frac{(|\mu|+as)e_1 + bse_2}{(|\mu|+as)^2 + (bs)^2} + \sum_{k=3}^d \alpha_k e_k \\
&\qquad + \left(\alpha_2 - \frac{zbs}{(|\mu|+as)^2+(bs)^2}\right) \frac{(|\mu|+as)e_2 - bse_1}{|\mu|+as} \Bigg) \frac{az+\alpha_2 b|\mu|}{(|\mu|+as)^2} d\alpha_{2:d} ds \\
&= -\int_0^1 \int h(y) \frac{\langle y, H\rangle}{|\mu+sH|} d\lambda_z^{\mu+sH} ds\,, \qquad (35)
\end{aligned}$$

where, for the last equality, we made the change of variable

$$\beta_2 = \frac{\sqrt{(|\mu|+as)^2+(bs)^2}}{|\mu|+as}\alpha_2 - \frac{zbs}{(|\mu|+as)\sqrt{(|\mu|+as)^2+(bs)^2}}\,,$$

used the equality $(|\mu|+as)e_1 + bse_2 = \mu + sH$ and remarked that $\langle \mu+sH, y\rangle = z$ implies that $az + \langle y, e_2\rangle\, b|\mu| = (|\mu|+as)\,\langle y, H\rangle$. Define, for $\nu \in \mathbb{R}^d_*$,

$$\gamma(h,\nu) \stackrel{\text{def}}{=} \int \frac{y}{|\nu|} h(y) d\lambda_z^\nu\,. \qquad (36)$$

We deduce that

$$g_z(\mu+H) - g_z(\mu) + \left\langle H, \int \frac{y}{|\mu|}\, h(y)\, d\lambda_z^\mu \right\rangle = \left\langle H, \int_0^1 \{\gamma(h,\mu) - \gamma(h,\mu+sH)\}\, ds \right\rangle\,.$$

Consider now the following decomposition

$$\gamma(h,\nu) = \gamma\left(h\mathbb{1}_{\{|\cdot|>M\}},\nu\right) + \gamma\left(h\mathbb{1}_{\{|\cdot|\le M\}},\nu\right) \ . \tag{37}$$

Under assumption 19, the first term is arbitrarily small as M goes to infinity uniformly in ν close to μ. When $\nu \to \mu$, the measure $\mathbb{1}_{\{|\cdot|\le M\}}\lambda_z^\nu$ converges weakly to $\mathbb{1}_{\{|\cdot|\le M\}}\lambda_z^\mu$; hence, the second term in the RHS of (37) converges to $\gamma\left(h\mathbb{1}_{\{|\cdot|\le M\}},\mu\right)$. Therefore, the function $\nu \mapsto \gamma(h,\nu)$ is continuous at μ and the conclusion follows easily.

Proof of Corollary 1 Let H be a $d \times m$ matrix with columns $(H_1,\ldots,H_m)$. Let $\{a_k, b_k, k \in \{1,\ldots,m\}\}$ be real numbers. We have

$$\prod_{k=1}^m a_k - \prod_{k=1}^m b_k = \sum_{k=1}^m (a_k - b_k)\left(\prod_{j=1}^{k-1} b_j\right)\left(\prod_{j=k+1}^m a_j\right) \ ,$$

where, by convention, $\prod_{k=j}^{\ell} c_k = 1$ for $j > \ell$. We deduce from the latter expression

$$\begin{aligned}\prod_{k=1}^m a_k - \prod_{k=1}^m b_k - \sum_{k=1}^m (a_k - b_k)\left(\prod_{j\neq k} a_j\right) \\ = -\sum_{k=1}^m (a_k - b_k)\left(\prod_{j=k+1}^m a_j\right)\left(\sum_{j=1}^{k-1}(a_j - b_j)\left\{\prod_{u=1}^{j-1} a_u\right\}\left\{\prod_{u=j+1}^{k-1} b_u\right\}\right) \ .\end{aligned}$$

We apply this equality with $a_k = \phi_k(y,0)$ and $b_k = \phi_k(y,H)$ where $\phi_k(y,\Delta) \stackrel{\text{def}}{=} \mathbb{1}_{\{y,\langle\mu_k+\Delta_k,y\rangle\le z_k\}}$, which yields

$$\begin{aligned}g_z(\mu+H) - g_z(\mu) - \sum_{k=1}^m \int \{\phi_k(y,H) - \phi_k(y,0)\}\left\{\prod_{j\neq k}\phi_j(y,H)\right\} h(y)d\lambda(y) \\ = \sum_{k=1}^m \sum_{j=1}^{k-1} \int \{\phi_k(y,H) - \phi_k(y,0)\}\, h_{j,k}(y;H)d\lambda(y) \ ,\end{aligned}$$

where the function $h_{j,k}(y,H)$ is defined as

$$h_{j,k}(y,H) \stackrel{\text{def}}{=} h(y)\{\phi_j(y,H) - \phi_j(y,0)\} \prod_{u=k+1}^m \phi_u(y,0) \prod_{u=1}^{j-1} \phi_u(y,0) \prod_{u=j+1}^{k-1} \phi_u(y,H) \ . \tag{38}$$

By the weak convergence argument used to conclude the proof of Proposition 5, we obtain

$$\lim_{|H|\to 0}\left|\sum_{k=1}^m \int \{\phi_k(y,H) - \phi_k(y,0)\}\prod_{j\neq k}\phi_j(y,0)h(y)d\lambda(y) - \sum_{k=1}^m \left\langle H_k, \int \frac{y}{|\mu_k|}\prod_{j\neq k}\phi_j(y,0)\, h(y)d\lambda_{z_k}^{\mu_k}\right\rangle\right| = 0 \ .$$

To conclude the proof, it is enough to check that for any $j < k$,

$$\mathcal{I}(H,h_{j,k}) \stackrel{\text{def}}{=} \int \{\phi_k(y,H) - \phi_k(y,0)\}\, h_{j,k}(y;H)d\lambda(y) = o(|H|) \quad \text{as } |H| \to 0 \ . \tag{39}$$

Using (35), the latter integral may be expressed as

$$\mathcal{I}(H,h_{j,k}) = -\left\langle H_k, \int_0^1 \int h_{j,k}(y;H)\frac{y}{|\mu_k + sH_k|}d\lambda_{z_k}^{\mu_k+sH_k}(dy)ds\right\rangle \ .$$

We write $\mathcal{I}(H,h_{j,k}) = \mathcal{I}(H,h_{j,k}\mathbb{1}_{\{|\cdot|>M\}}) + \mathcal{I}(H,h_{j,k}\mathbb{1}_{\{|\cdot|\le M\}})$. By (20), the first term is small uniformly in H for $|H| \le \varepsilon$, when M large enough.

Let $[e_1, \ldots, e_d]$ be any (given) orthonormal basis of $\mathbb{R}^d$ such that $\langle \mu_k, e_d \rangle \neq 0$. Consider the following matrix $S_k(\mu) \stackrel{\text{def}}{=} [\Pi(\mu_k)e_1, \ldots, \Pi(\mu_k)e_{d-1}]$, where $\Pi(\mu_k)$ is the orthogonal projector on the orthogonal complement of the vector μ_k. By the change of variable formula, $\mathcal{I}(H, h_{j,k}\mathbb{1}_{\{|\cdot|\leq M\}})$ is equal to

$$-\left\langle H_k, \int_0^1 \det\left[\frac{\mu_k + sH_k}{|\mu_k + sH_k|^2}, S_k(\mu + sH)\right] \int_{\mathbb{R}^{d-1}} \tilde{h}_{j,k}\left(z_k \frac{\mu_k + sH_k}{|\mu_k + sH_k|^2} + S_k(\mu + sH)\tilde{y}; H\right) d\lambda(\tilde{y})ds \right\rangle ,$$

where $\tilde{h}_{j,k}(y; H) \stackrel{\text{def}}{=} h_{j,k}(y; H)\ y\mathbb{1}_{\{|y|\leq M\}}$.

We then conclude by the Lebesgue Theorem : by construction, $\tilde{h}_{j,k}$ is bounded and the integration domain is bounded; it is sufficient to check that the limit of the integrand is zero almost-everywhere w.r.t. the Lebesgue measure on $\mathbb{R}^{d-1}$. Note that $|\tilde{h}_{j,k}(y; H)| \leq C_M |\phi_j(y, H) - \phi_j(y, 0)|$ and that

$$\lim_{|H|\to 0} \left\{ \phi_j\left(z_k \frac{\mu_k + sH_k}{|\mu_k + sH_k|^2} + S_k(\mu + sH)\tilde{y}, H\right) - \phi_j\left(z_k \frac{\mu_k + sH_k}{|\mu_k + sH_k|^2} + S_k(\mu + sH)\tilde{y}, 0\right) \right\} = 0$$

except on the set $\{\tilde{y} \in \mathbb{R}^{d-1}, \left\langle \mu_j, z_k \frac{\mu_k}{|\mu_k|^2} + S_k(\mu)\tilde{y} \right\rangle = z_j\}$, which is of measure zero w.r.t. the Lebesgue measure on $\mathbb{R}^{d-1}$.

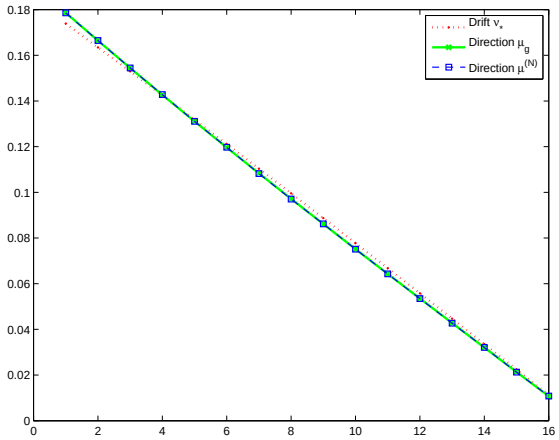

Fig. 1 Asian Option: Optimal drift vector $\nu_\star$, direction μ_{g} and direction $\mu^{(N)}$. The directions have been scaled to have the same norm as the drift $\nu_\star$.

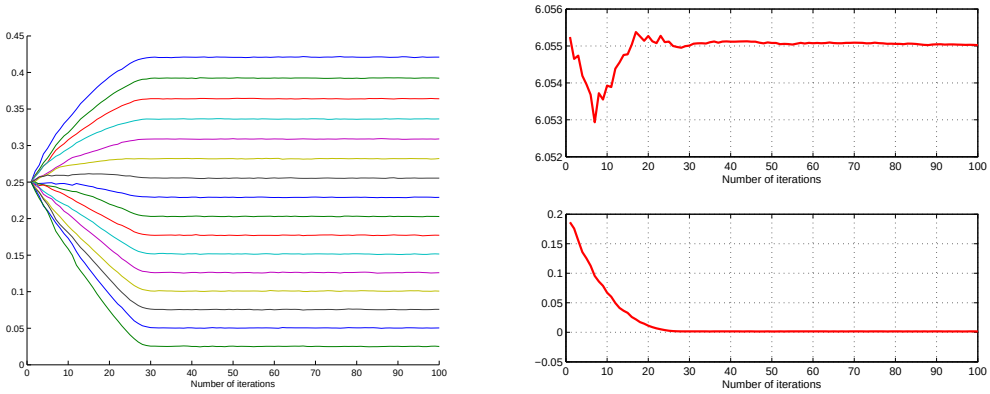

Fig. 2 Asian Option: [left] successive directions of stratification $t \mapsto \mu^{(t)}$. $\mu^{(0)}$ is proportional to the vector $(1, \cdots, 1)$ so that the d curves start from the same point $1/\sqrt{d}$. By convention, the first component of $\mu^{(t)}$ is positive. [top right] successive estimations of the quantity of interest $t \mapsto \mathcal{E}^{(t)}$. [bottom right] successive values of the variance $t \mapsto (\sum_{\mathbf{i}} p_{\mathbf{i}}\, \hat{\sigma}_{\mathbf{i}}^{(t)})^2$; the limiting value is 0.0015.

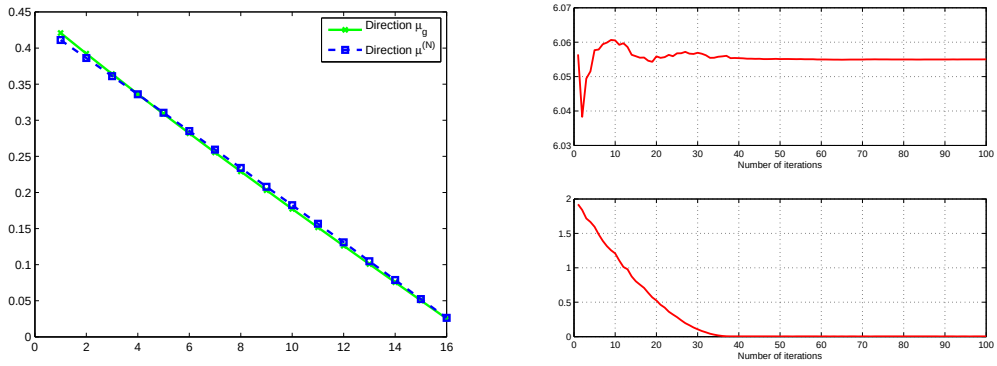

Fig. 3 Asian Option when ν is the null vector in (23): [left] the direction μ_{g} and the limiting direction $\mu^{(N)}$ when . [top right] successive estimations of the quantity of interest $t \mapsto \mathcal{E}^{(t)}$. [bottom right] successive values of the variance $t \mapsto (\sum_{\mathbf{i}} p_{\mathbf{i}}\, \hat{\sigma}_{\mathbf{i}}^{(t)})^2$; the limiting value is 0.0015.

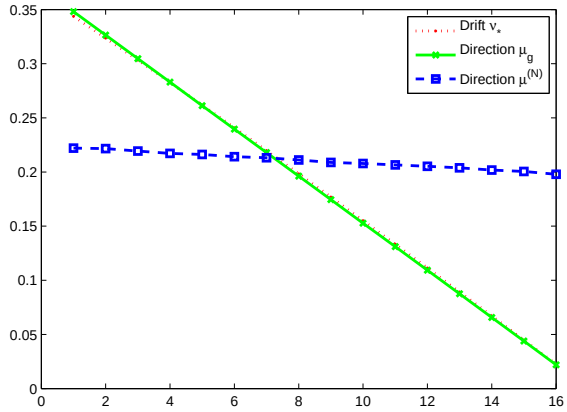

Fig. 4 Barrier Option: Optimal drift vector $\nu_\star$, direction μ_{g} and direction $\mu^{(N)}$. The directions have been scaled to have the same norm as the drift $\nu_\star$

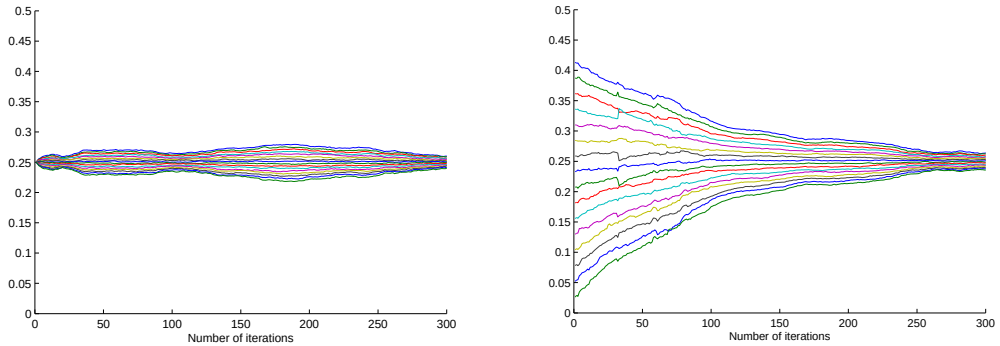

Fig. 5 Barrier Option: successive directions of stratification $t \mapsto \mu^{(t)}$. By convention, the first component of $\mu^{(t)}$ is positive. [left] $\mu^{(0)}$ is proportional to the vector $(1, \cdots, 1)$ so that the d curves start from the same point $1/\sqrt{d}$. [right] $\mu^{(0)} = \mu_{\mathrm{g}}$.

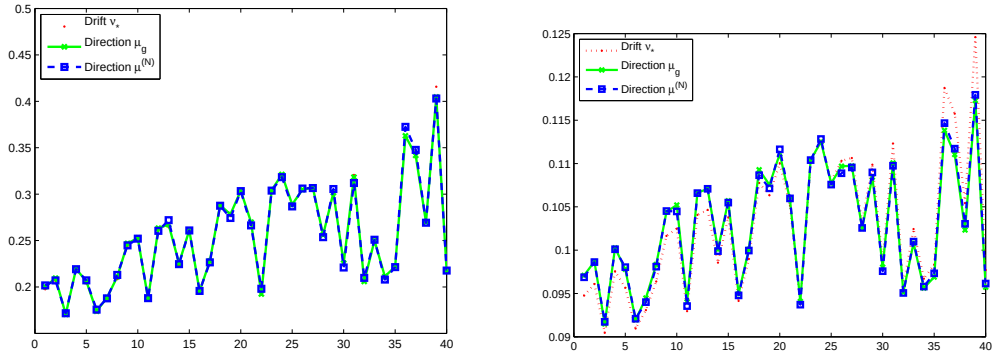

Fig. 6 Basket Option: Optimal drift vector $\nu_\star$, direction μ_{g} and direction $\mu^{(N)}$. The directions have been scaled to have the same norm as the drift $\nu_\star$ [left] when $(c, K) = (0.1, 60)$. [right] when $(c, K) = (0.5, 45)$

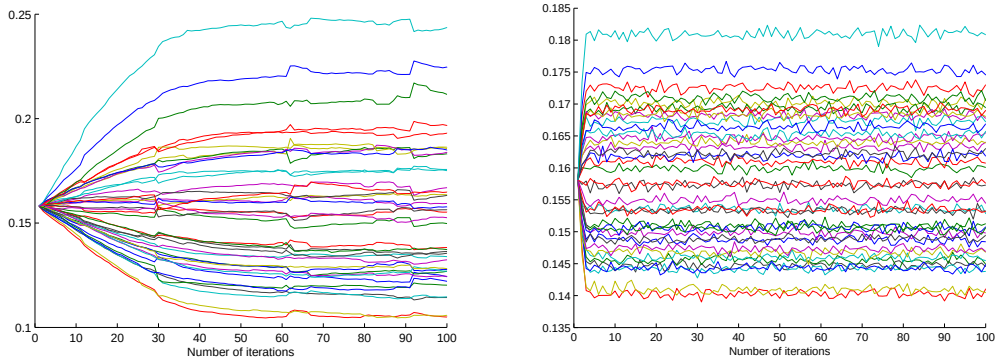

Fig. 7 Basket Option: successive directions of stratification $t \mapsto \mu^{(t)}$. By convention, the first component of $\mu^{(t)}$ is positive [left] when $(c, K) = (0.1, 60)$. [right] when $(c, K) = (0.5, 45)$

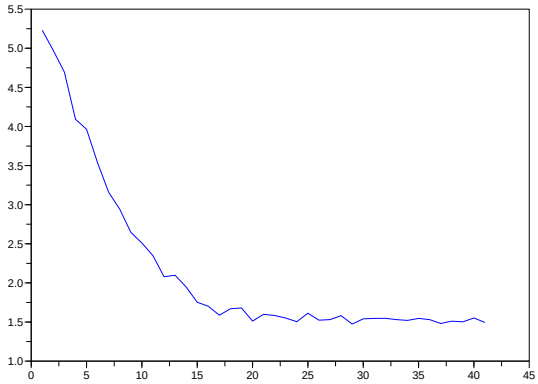

Fig. 8 Asian Option in Heston model: Value of $\sum_i p_i \hat{\sigma}_i^{(t)}$ as function of t for $K = 100$, $\theta = 0.01$, $k = 2$ and $\rho = 0.5$.

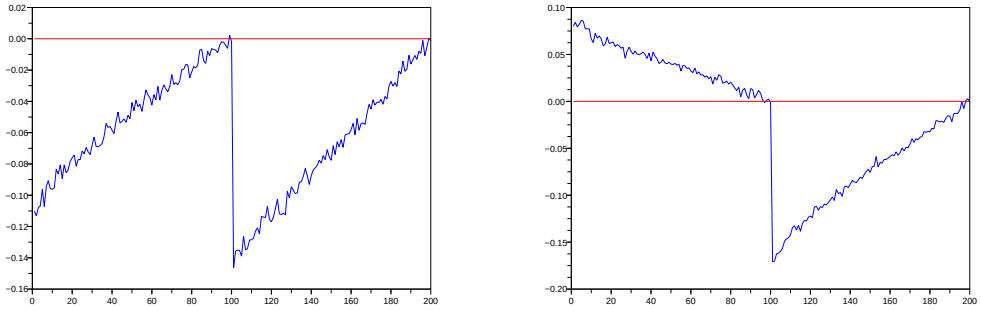

Fig. 9 Asian Option in Heston model: components of $\mu^{(N)}$ with respect to component number for $K = 100$, $\theta = 0.01$, $k = 2$ and [left] $\rho = 0.5$, [right]$\rho = -0.5$

Parameters		Price	Variance			
υ	K	-	`Monte Carlo`	`GHS`	`AdaptStr` ($\nu_\star$)	`AdaptStr` (no drift)
0.10	45	6.05	8.68	0.007	0.001	0.004
	50	1.92	4.93	0.0009	0.0004	0.0017
	55	0.20	0.55	0.00003	0.00002	0.00053
0.30	45	7.15	59.30	0.035	0.025	0.062
	50	4.17	40.11	0.021	0.013	0.039
	55	2.21	21.48	0.010	0.006	0.023

Table 1 Asian Option

Parameters		Price	Variance			
K	B	-	`Monte Carlo`	`GHS`	`AdaptStr` ($\nu_\star$)	`AdaptStr` (no drift)
50	60	1.38	2.99	0.494	0.130	0.106
	70	1.90	4.79	0.020	0.005	0.007
	80	1.92	4.92	0.0011	0.0005	0.0017
55	70	0.19	0.49	0.0014	0.0006	0.0012
	80	0.20	0.55	0.00004	0.00002	0.00053

Table 2 Barrier Option

Parameters		Price	Variance			
c	K	-	`Monte Carlo`	`GHS`	`AdaptStr` ($\nu_\star$)	`AdaptStr` (no drift)
0.1	45	11.20	22.18	0.256	0.206	0.215
	60	0.78	3.70	0.037	0.018	0.023
0.5	45	11.56	81.38	0.077	0.061	0.099
	60	2.54	27.00	0.021	0.012	0.032
0.9	45	12.09	134.31	0.022	0.008	0.053
	60	3.73	56.85	0.004	0.002	0.034

Table 3 Basket Option

Parameters		Price	Variance Ratio
ξ_0	K	-	`AdaptStr`
0.01	120	0.105	400
	100	4.93	25
	80	22.65	60
0.04	130	0.20	150
	120	0.63	18
	100	6.21	31
	80	22.65	19
	70	31.73	19.5

Table 4 Asian Option in Heston model

References

Alfonsi A (2008) High order discretization scheme for the cir process: application to affine term structure and heston model. URL `http://hal.archives-ouvertes.fr/docs/00/28/88/11/PDF/2nd_order_ATSM.pdf`

Arouna B (2004) Adaptative Monte Carlo method, a variance reduction technique. Monte Carlo Methods Appl 10(1):1–24

Asmussen S, Glynn PW (2007) Stochastic simulation: algorithms and analysis, Stochastic Modelling and Applied Probability, vol 57. Springer, New York

Broadie M, Kaya O (2006) Exact simulation of stochastic volatility and other affine jump diffusion processes. Operations Research 54(2):217–231

Dudley RM (2002) Real analysis and probability, Cambridge Studies in Advanced Mathematics, vol 74. Cambridge University Press, Cambridge, revised reprint of the 1989 original

Etore P, Jourdain B (2007) Adaptive optimal allocation in stratified sampling methods. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:0711.4514`

Fishman GS (1996) Monte Carlo. Springer Series in Operations Research, Springer-Verlag, New York, concepts, algorithms, and applications

Glasserman P (2004) Monte Carlo methods in financial engineering, Applications of Mathematics (New York), vol 53. Springer-Verlag, New York, stochastic Modelling and Applied Probability

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

Judistsky A, Lan G, Nemirovski A, Shapiro A (2007) Stochastic approximation approach to stochastic programming. URL `http://http://www2.isye.gatech.edu/~nemirovs/`

Kawai R (2007) Adaptive Monte Carlo variance reduction with two-time-scale stochastic approximation. Monte Carlo Methods Appl 13(3):197–217

Kushner HJ, Yin GG (2003) Stochastic approximation and recursive algorithms and applications, Applications of Mathematics (New York), vol 35, 2nd edn. Springer-Verlag, New York, stochastic Modelling and Applied Probability

Ninomya S, Victoir N (2008) Weak approximation of stochastic differential equations and application to derivatives pricing. Appl Math Finance 15(2):107–121

Pflug GC (1996) Optimization of stochastic models. The Kluwer International Series in Engineering and Computer Science, 373, Kluwer Academic Publishers, Boston, MA, the interface between simulation and optimization

Rubinstein RY, Kroese DP (2004) The cross-entropy method. Information Science and Statistics, Springer-Verlag, New York, a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning

Rubinstein RY, Kroese DP (2008) Simulation and the Monte Carlo method, 2nd edn. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ