Journal of Computational and Applied Mathematics 319 (2017) 440–459

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Adaptive importance sampling Monte Carlo simulation for general multivariate probability laws

Reiichiro Kawai *

School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form 17 January 2017

MSC:
65C05
93E35
60F05
60E05

Keywords:
Bypass distribution
Central limit theorem
Exponential family
Stochastic approximation
Variance reduction

ABSTRACT

We establish a parametric adaptive importance sampling variance reduction method for general multivariate probability laws. Employing the principle of bypass distributions makes it possible to develop adaptive algorithms without relying on particular properties of the target and proposal laws, both of which in the proposed framework are as general as the uniform law on the unit hypercube, without changing the sampling distribution at each iteration. We establish the asymptotic normality of the estimator of the desired mean and of the importance sampling parameter as the number of observations tends to infinity. Although implementation of the proposed methodology requires a small amount of initial work, it has the potential to yield substantial improvements in estimator efficiency in various general problem settings. To illustrate the applicability and effectiveness, we provide numerical results throughout, in which we apply exponential and normal bypass distributions, as well as demonstrate that well-known adaptive importance sampling formulations in the literature can be easily rewritten in the proposed framework.

1. Introduction

With significant increases in computational demands of Monte Carlo simulation, methods for increasing efficiency have attracted substantial interest, in particular for routines to be run frequently. It is thus worthwhile to carry out some initial analysis yielding more efficient simulation. In this spirit, in standard parametrized variance reduction methods, the parameter value is chosen in advance in the way to reduce, hopefully minimize, the estimator variance with differences in computing costs put into consideration. This parameter choice can be done, for instance, by taking a pure guess, or a more plausible guess from the past experience, or by running a pilot simulation for a while, etc., all without strong confidence of optimality. In particular, running a sufficiently long pilot simulation is a little contradictory from a practical point of view, in the sense that we might have been better off spending this effort running the crude Monte Carlo simulation longer, as well as, in the first place, the sufficient length of the pilot run is not obvious in advance.

Adaptive Monte Carlo variance reduction methods aim at concurrently running a Monte Carlo simulation and a search algorithm for optimal variance reduction parameters, where common replications can be used for both procedures without changing the sampling distribution at each iteration, rather than generating two independent sets of replications. Adaptive Monte Carlo methods require a certain amount of such initial work for its implementation, while it has the potential to provide significant variance reduction as a result. Equally importantly, the adaptive framework avoids the need for frequent

* Fax: +61 (0) 2 9351 4534.
E-mail address: reiichiro.kawai@sydney.edu.au.

http://dx.doi.org/10.1016/j.cam.2017.01.029
0377-0427/

recalibration of the parameters of the variance reduction techniques when changes occur in the experimental conditions governing system performance. The idea of adaptive Monte Carlo simulation and its practical use has been studied for a long time in various formulations and problem settings; for example, parametrized distributions [1–6], mixture densities [7–11], and (combinations of) control variates, importance sampling and stratified sampling [12–17], to mention just a few.

This paper is concerned with the construction and analysis of adaptive importance sampling variance reduction methods, for general multivariate probability laws, more precisely, the uniform law on the unit hypercube. Focusing on the uniform law is not a restriction but a generalization, in the sense that the expected value of a functional of a multivariate random vector can be rewritten with the standard uniform random vector in the same dimension with a suitable change of variables or the principle of inverse transform sampling. In the proposed framework, both target and proposal laws are simply the same uniform law on the unit hypercube. In particular, one class of parametric adaptive importance sampling methods is designed to adaptively reform the proposal law to make it closer to the target law (e.g. [10,17]), while the other class relies largely on special properties of the target distribution. For instance, a Gaussian random vector after an exponential change of measure is identical in law to the original Gaussian random vector with a suitable mean under the original probability measure (e.g. [1]), or a gamma random variable after an exponential change of measure is identical in law to a scaled original gamma random variable under the original probability measure (e.g. [14]).

The proposed framework is built upon a principle different from all those: it relies on no particular properties of the target and proposal laws. The key principle can be summarized as follows; first rewrite the expected value on the uniform law by a suitable parametric probability law with the principle of inverse transform sampling, secondly change the probability measure on the parametric law, thirdly rewrite the parametric law back to the uniform law under the original probability measure again with the principle of inverse transform sampling, and finally find optimal parameters introduced in the second step in the way to reduce the estimator variance. As is clear, the parametric law is employed in the second step solely to inject a parametrization into the expression of the expected value, whereas it does neither appear in the expected value in the final form nor change the underlying probability measure. For those reasons, we give the name of the *bypass* distribution to this parametric law in the second step. As a Monte Carlo simulation and a search algorithm for optimal variance reduction parameters are both expressed with common random elements under the original probability measure, the existing optimal parameter search techniques can be applied, such as the stochastic approximation [1,13–15,3,4,6] as well as the sample average approximation [2,18], under suitable technical conditions.

The rest of this paper is organized as follows. In Section 2, we begin with general notation and then briefly summarize background material on adaptive Monte Carlo variance reduction methods. In Section 3, we introduce the principle of bypass distributions so as to induce importance sampling. The choice of bypass distributions is quite flexible as far as technical conditions are satisfied, whereas theoretical over-complications often do not contribute to the practical effectiveness. In Section 4, we thus pay particular attention to the continuous canonical exponential family, a relatively simple yet wide class of continuous distributions. The derivation of the results entails rather lengthy proofs of somewhat routine nature. To avoid overloading the paper, we omit non-essential details in some instances. We provide numerical examples to demonstrate in Section 5 the procedure of the proposed method, as well as to illustrate the performance relative to the choice of bypass distributions. We also show that well-known adaptive importance sampling formulations in the literature can be easily rewritten in the proposed framework. In Section 6, we formulate the adaptive Monte Carlo simulation, along with convergence results and numerical illustrations of searching the parameter by the stochastic approximation and the sample average approximation. Finally, Section 7 concludes this study and highlights future research directions.

2. Problem setup

We begin with general notation which will be used throughout the paper. We use the notation $\mathbb{N} := \{1, 2, \ldots\}$ and denote by $|\cdot|$ and $\|\cdot\|$, respectively, the magnitude and the Euclidean norm. As usual, for a square matrix A, we denote by $|A|$, $\|A\|$, $A^\top$, $A^{\otimes 2}$ and $\sqrt{A}$, respectively, the determinant, a suitable matrix norm, the transpose, the outer product and a lower triangular matrix of the Cholesky decomposition of A, provided that those are well defined. We denote by ϕ, Φ and Φ^{-1}, respectively, the standard normal density function, the standard normal cumulative distribution function and its inverse. We denote by $\mathrm{Leb}(D)$, $\mathrm{int}(D)$, ∂D, $\overline{D}$ and $\mathscr{B}(D)$, respectively, the Lebesgue area, the interior, the boundary, the closure and the Borel σ-field of a domain D. We denote by $\mathbb{1}_D(\mathbf{x})$ the indicator function of a set D at $\mathbf{x}$. We let $\overset{\mathscr{L}}{=}$ and $\overset{\mathscr{L}}{\to}$ denote the identity and convergence in law. For the sake of simplicity, we use the notation ∂_x^q for the qth partial derivative with respect to the univariate variable x, as well as $\nabla_{\mathbf{x}}$ and $\mathrm{Hess}_{\mathbf{x}}$ indicate the gradient and the Hessian matrix with respect to the multivariate variable $\mathbf{x}$.

Throughout this paper, we are interested in constructing a fairly general framework of parametric adaptive importance sampling Monte Carlo methods for the integral

$$C := \int_{(0,1)^d} \Psi(\mathbf{u})\, d\mathbf{u} = \mathbb{E}_{\mathbb{P}}[\Psi(U)], \tag{2.1}$$

where Ψ is a function mapping from $(0, 1)^d$ to $\mathbb{R}$, and where U is a uniform random variables on $(0, 1)^d$ under the probability measure $\mathbb{P}$. We reserve the capital "C" for this integral value throughout the paper. To avoid triviality, we impose a finite

second moment;

$$\int_{(0,1)^d} |\Psi(\mathbf{u})|^2 \, d\mathbf{u} = \mathbb{E}_{\mathbb{P}}\left[|\Psi(U)|^2\right] < +\infty, \tag{2.2}$$

as well as non-degeneracy of the integrand $\mathbb{P}(|\Psi(U)| > 0) > 0$.

We reserve the notation $\{U_k\}_{k\in\mathbb{N}}$ for a sequence of i.i.d. uniform random variables on $(0,1)^d$ and define the natural filtration $(\mathscr{F}_k)_{k\in\mathbb{N}}$ generated by the sequence $\{U_k\}_{k\in\mathbb{N}}$ of i.i.d. uniform random variables on $(0,1)^d$, that is, for each $n \in \mathbb{N}$, $\mathscr{F}_n = \sigma(\{U_k\}_{k=1,\dots,n})$ is the σ-field generated by the i.i.d. uniform random vectors $U_1, \dots, U_n$. Throughout the paper, there is no need to specify under what probability measure the expectation $\mathbb{E}$ is taken, since we end up with taking expectations under $\mathbb{P}$ all the time, although we do change the probability measure in the middle of derivations. We thus set $(\Omega, \mathscr{F}, (\mathscr{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$ as our underlying filtered probability space throughout.

We now define the family of probability distributions, which we will shortly name the bypass distribution.

Assumption 2.1. We choose in advance an open set $\Theta_0 \subseteq \mathbb{R}^d$ with $\mathrm{Leb}(\Theta_0) > 0$, a family $\{g(\cdot;\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta_0\}$ of probability density functions on $\mathbb{R}^d$ and a family $\{G(\cdot;\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta_0\}$ of functions on $\mathbb{R}^d$ in such a way that

(a) The support D of the probability density function $g(\cdot;\boldsymbol{\theta})$ is open and independent of the parameter $\boldsymbol{\theta}$;
(b) For almost every $\mathbf{z} \in D$ (with respect to $d\mathbf{z}$), the probability density function $g(\mathbf{z};\boldsymbol{\theta})$ is twice continuously differentiable at $\boldsymbol{\theta} \in \Theta_0$;
(c) For each $\boldsymbol{\theta} \in \Theta_0$ and $B \in \mathscr{B}(D)$, it holds that $\int_D \mathbb{1}(G(\mathbf{z};\boldsymbol{\theta}) \in B) g(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} = \mathrm{Leb}(B)$;
(d) For each $\boldsymbol{\theta} \in \Theta_0$, the inverse $G^{-1}(\mathbf{u};\boldsymbol{\theta})$ (with respect to $\mathbf{u}$) is well defined and continuous in $\mathbf{u}$ on $(0,1)^d$;
(e) For each $\boldsymbol{\theta} \in \Theta_0$ and $B \in \mathscr{B}(D)$, it holds that $\int_{(0,1)^d} \mathbb{1}(G^{-1}(\mathbf{u};\boldsymbol{\theta}) \in B) d\mathbf{u} = \int_B g(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z}$;
(f) For almost every $\mathbf{z} \in D$ (with respect to $d\mathbf{z}$), it holds that $\lim_{n\uparrow+\infty} \sup_{\boldsymbol{\theta}\in\partial K_n} g(\mathbf{z};\boldsymbol{\theta}) = 0$, where $\{K_n\}_{n\in\mathbb{N}}$ is an increasing sequence of compact subsets of the open set Θ_0, satisfying $\cup_{n=1}^{+\infty} K_n = \Theta_0$ and $K_n \subsetneq \mathrm{int}(K_{n+1})$.

Assumption 2.1(c), (d) and (e) indicate that if Z is a random vector in $\mathbb{R}^d$ with density $g(\mathbf{z};\boldsymbol{\theta})$ and $U \sim U(0,1)^d$, then

$$G(Z;\boldsymbol{\theta}) \overset{\mathscr{L}}{=} U, \quad Z \overset{\mathscr{L}}{=} G^{-1}(U;\boldsymbol{\theta}). \tag{2.3}$$

Assumption 2.1(f) will serve as an important requirement for convexity of the estimator variance later in Theorem 3.3.

Example 2.2. Although we have set up Assumption 2.1 in a somewhat abstract manner for theoretical purposes, it will soon turn out to be important in practice that the function G and its inverse G^{-1} can be written in closed (at least, computable) form. That is, in practice, we wish to focus on independent components

$$g(\mathbf{z};\boldsymbol{\theta}) = \prod_{k=1}^{d} g_k(z_k;\boldsymbol{\theta}), \quad \mathbf{z} = (z_1,\dots,z_d)^\top \in D, \tag{2.4}$$

where each $g_k(z_k;\boldsymbol{\theta})$ is a probability density function on D_k for a suitable $D_k \subseteq \mathbb{R}$ with $\mathrm{Leb}(D_k) > 0$. Then, the function G can also be written componentwise, as $G(\mathbf{z};\boldsymbol{\theta}) := (G_1(z_1;\boldsymbol{\theta}), \dots, G_d(z_d;\boldsymbol{\theta}))^\top$, where each $G_k(z;\boldsymbol{\theta})$ is defined by either

$$G_k(z_k;\boldsymbol{\theta}) := \int_{D_k} \mathbb{1}(x \le z_k) g_k(x;\boldsymbol{\theta}) dx, \quad \text{or} \quad G_k(z_k;\boldsymbol{\theta}) := \int_{D_k} \mathbb{1}(x > z_k) g_k(x;\boldsymbol{\theta}) dx. \tag{2.5}$$

Clearly, $G_k(Z;\boldsymbol{\theta}) \sim U(0,1)$ if Z is a random variable with density $g_k(z_k;\boldsymbol{\theta})$, and thus the multivariate version (2.3) holds true on the whole. For illustration purposes, we consider exponential and standard normal distributions for the first and second components, that is, $d = 2$, $D = (0,+\infty)\times\mathbb{R}$, and

$$g(\mathbf{z};\boldsymbol{\theta}) = \theta_1 e^{-\theta_1 z_1} \phi(z_2 - \theta_2), \quad \boldsymbol{\theta} = (\theta_1,\theta_2)^\top. \tag{2.6}$$

The support $D = (0,+\infty)\times\mathbb{R}$ is independent of the parameter $\boldsymbol{\theta}$. The function $g(\mathbf{z};\boldsymbol{\theta})$ is well defined as a joint probability density function on D, as long as $\boldsymbol{\theta} \in (0,+\infty)\times\mathbb{R}$. For each $\mathbf{z} \in D$, the joint probability density function $g(\mathbf{z};\boldsymbol{\theta})$ is twice continuously differentiable at $\boldsymbol{\theta}$ in $(0,+\infty)\times\mathbb{R}$. If $\boldsymbol{\theta}$ in $(0,+\infty)\times\mathbb{R}$, then the random vector $Z := (Z_1, Z_2)^\top$ with density $g(\mathbf{z};\boldsymbol{\theta})$ has independent components, where $Z_1 \sim \mathrm{Exp}(\theta_1)$ and $Z_2 \sim \mathscr{N}(\theta_2, 1)$. Then, we have

$$G(\mathbf{z};\boldsymbol{\theta}) = \left(e^{-\theta_1 z_1}, \Phi(z_2 - \theta_2)\right)^\top, \qquad G^{-1}(\mathbf{u};\boldsymbol{\theta}) = \left(-\theta_1^{-1}\ln(u_1), \theta_2 + \Phi^{-1}(u_2)\right)^\top. \tag{2.7}$$

Note that the first components of $G(\mathbf{z};\boldsymbol{\theta})$ and $G^{-1}(\mathbf{u};\boldsymbol{\theta})$ are obtained based on the latter definition in (2.5). As a matter of course, the former definition in (2.5) can be used as well, then we obtain $1 - e^{-\theta_1 z_1}$ and $-\theta_1^{-1}\ln(1-u_1)$ instead. Finally, for almost every $\mathbf{z} \in D$ (with respect to $d\mathbf{z}$), $g(\mathbf{z};\boldsymbol{\theta})$ tends to zero, whenever $\theta_1 \to \{0,+\infty\}$ or $|\theta_2| \uparrow +\infty$. Hence, we can set $\Theta_0 = (0,+\infty)\times\mathbb{R}$, which is open and independent of $\boldsymbol{\theta}$. We will continue this problem setting in Section 5.1.1. □

The handiest adaptive Monte Carlo variance reduction method is perhaps the one based on control variates. For instance, with the centered variates $U - \mathbb{E}_{\mathbb{P}}[U]$, one can construct an adaptive Monte Carlo simulation

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\Psi\left(U_k\right) - \langle \boldsymbol{\lambda}_{k-1}, U_k - \mathbb{E}_{\mathbb{P}}[U]\rangle\right) - \mathbb{E}_{\mathbb{P}}\left[\Psi(U)\right]\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, \mathrm{Var}_{\mathbb{P}}\left(\Psi(U)\right) - \frac{1}{12}\left\|\boldsymbol{\lambda}^*\right\|^2\right), \tag{2.8}$$

with $\boldsymbol{\lambda}_0 := 0$ and

$$\begin{aligned}\boldsymbol{\lambda}_{k-1} &:= \frac{12}{k-1}\sum_{l=1}^{k-1}\left(\Psi\left(U_l\right) - \frac{1}{k-1}\sum_{m=1}^{k-1}\Psi\left(U_m\right)\right)\left(U_l - \mathbb{E}_{\mathbb{P}}[U]\right)\\ &\to \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}^d}\mathrm{Var}_{\mathbb{P}}\left(\Psi\left(U\right) - \langle\boldsymbol{\lambda}, U - \mathbb{E}_{\mathbb{P}}[U]\rangle\right) =: \boldsymbol{\lambda}^*,\end{aligned}$$

under suitable technical conditions, where $\{U_k\}_{k\in\mathbb{N}}$ is a sequence of i.i.d. uniform random vectors on $(0,1)^d$. (We refer the reader to [15,16] for details.) This control variates method is often effective when the magnitude $\|\boldsymbol{\lambda}^*\|^2$ is large. If, however, $\|\boldsymbol{\lambda}^*\|^2$ is very small relative to the crude estimator variance $\mathrm{Var}_{\mathbb{P}}(\Psi(U))$, that is, the random variable $\Psi(U)$ is not strongly correlated to the uniform random vector U, then this method fails to reduce variance much. A typical situation is the rare event simulation, in which the integrand $\Psi(U)$ returns the value of exactly zero with a relatively high probability. The main scope of the present paper lies in such situations where control variates (2.8) are not very effective.

3. Change of measure and importance sampling through bypass distributions

We are now in a position to construct our theoretical base. We use the family $\{g(\cdot;\boldsymbol{\theta});\boldsymbol{\theta}\in\Theta_0\}$ of probability density functions to introduce a parametrization so that the estimator variance may be reduced.

3.1. Change of measure through bypass distributions

Fix a point $\boldsymbol{\theta}_0 \in \Theta_0$. By changing variables $\mathbf{u} = G(\mathbf{z};\boldsymbol{\theta}_0)$ thanks to Assumption 2.1, the desired integral (2.1) can be rewritten as

$$\int_{(0,1)^d}\Psi\left(\mathbf{u}\right)d\mathbf{u} = \int_D \Psi\left(G\left(\mathbf{z};\boldsymbol{\theta}_0\right)\right)g\left(\mathbf{z};\boldsymbol{\theta}_0\right)d\mathbf{z},$$

where the right-hand side can be interpreted as an integral of the integrand $\Psi(G(\mathbf{z};\boldsymbol{\theta}_0))$, with respect to the probability measure $g(\mathbf{z};\boldsymbol{\theta}_0)d\mathbf{z}$ on the support D. Next, pick a $\boldsymbol{\theta}\in\Theta_0$, which can be distinct from $\boldsymbol{\theta}_0$. Note that the definition of Θ_0 and the independence of the support D from the parameter $\boldsymbol{\theta}$ ensure that the probability measure $g(\mathbf{z};\boldsymbol{\theta}_0)d\mathbf{z}$ is equivalent to the probability measure $g(\mathbf{z};\boldsymbol{\theta})d\mathbf{z}$ on the support D. Hence, by changing variables $\mathbf{z} = G^{-1}(\mathbf{u};\boldsymbol{\theta})$, we obtain

$$\int_{(0,1)^d}\Psi\left(\mathbf{u}\right)d\mathbf{u} = \int_D \Psi\left(G\left(\mathbf{z};\boldsymbol{\theta}_0\right)\right)g\left(\mathbf{z};\boldsymbol{\theta}_0\right)d\mathbf{z} = \int_D \frac{g(\mathbf{z};\boldsymbol{\theta}_0)}{g(\mathbf{z};\boldsymbol{\theta})}\Psi\left(G\left(\mathbf{z};\boldsymbol{\theta}_0\right)\right)g\left(\mathbf{z};\boldsymbol{\theta}\right)d\mathbf{z} \tag{3.1}$$

$$= \int_{(0,1)^d}\frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta})}\Psi\left(G\left(G^{-1}\left(\mathbf{u};\boldsymbol{\theta}\right);\boldsymbol{\theta}_0\right)\right)d\mathbf{u}. \tag{3.2}$$

The expressions (3.1) and (3.2) can be written in a probabilistic manner. Recall that $\mathbb{P}$ is the original probability measure under which $U\sim U(0,1)^d$, as originally appeared in the desired expectation (2.1). Define the probability measure $\mathbb{Q}_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta}\in\Theta_0$, under which the random vector Z has the joint probability density function $g(\mathbf{z};\boldsymbol{\theta})$ on the support D. Then, we obtain

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\Psi\left(U\right)\right] &= \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\theta}_0}}\left[\Psi\left(G\left(Z;\boldsymbol{\theta}_0\right)\right)\right]\\ &= \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\theta}}}\left[\frac{g(Z;\boldsymbol{\theta}_0)}{g(Z;\boldsymbol{\theta})}\Psi\left(G\left(Z;\boldsymbol{\theta}_0\right)\right)\right]\\ &= \mathbb{E}_{\mathbb{P}}\left[\frac{g(G^{-1}(U;\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(U;\boldsymbol{\theta});\boldsymbol{\theta})}\Psi\left(G\left(G^{-1}\left(U;\boldsymbol{\theta}\right);\boldsymbol{\theta}_0\right)\right)\right].\end{aligned} \tag{3.3}$$

The first progression is concerned with the transform $U \leftarrow G(Z;\boldsymbol{\theta}_0)$ under the probability measure $\mathbb{Q}_{\boldsymbol{\theta}_0}$, whereas the last progression employs inverse transform sampling $Z\leftarrow G^{-1}(U;\boldsymbol{\theta})$ under a different probability measure $\mathbb{Q}_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$. In the other words, the first two estimators $\Psi(U)$ and $\Psi(G(Z;\boldsymbol{\theta}_0))$ are identical in law, but written in different forms. Similarly, the last two expressions are identical in law, again written in different forms. Moreover, if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then even the second and third expressions are identical. With $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$, however, the second and third estimators follow distinct laws, without changing the expected value $C(=\mathbb{E}_{\mathbb{P}}[\Psi(U)])$. Hence, by wisely choosing the parameter $\boldsymbol{\theta}$, we may achieve a smaller variance under $\mathbb{Q}_{\boldsymbol{\theta}}$, compared to the original variance under $\mathbb{Q}_{\boldsymbol{\theta}_0}$.

3.2. Importance sampling through bypass distributions

The estimator variance of (3.3) is defined as the $L^2(D)$-distance of the integrand from the integral value C with respect to the probability measure $g(\mathbf{z};\boldsymbol{\theta})d\mathbf{z}$, that is,

$$\begin{aligned}\int_D \left[\frac{g(\mathbf{z};\boldsymbol{\theta}_0)}{g(\mathbf{z};\boldsymbol{\theta})}\Psi\left(G\left(\mathbf{z};\boldsymbol{\theta}_0\right)\right)-C\right]^2 g\left(\mathbf{z};\boldsymbol{\theta}\right)d\mathbf{z} &= \int_D \left(\frac{g(\mathbf{z};\boldsymbol{\theta}_0)}{g(\mathbf{z};\boldsymbol{\theta})}\right)^2 \left|\Psi\left(G\left(\mathbf{z};\boldsymbol{\theta}_0\right)\right)\right|^2 g\left(\mathbf{z};\boldsymbol{\theta}\right)d\mathbf{z}-C^2\\ &= \int_D \frac{g(\mathbf{z};\boldsymbol{\theta}_0)}{g(\mathbf{z};\boldsymbol{\theta})}\left|\Psi\left(G\left(\mathbf{z};\boldsymbol{\theta}_0\right)\right)\right|^2 g\left(\mathbf{z};\boldsymbol{\theta}_0\right)d\mathbf{z}-C^2\\ &= \int_{(0,1)^d} \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta})}\left|\Psi\left(\mathbf{u}\right)\right|^2 d\mathbf{u}-C^2,\end{aligned}$$

provided that the integrals are all well defined. The second term C^2 is obviously independent of the parameter $\boldsymbol{\theta}$. Hence, we wish to choose $\boldsymbol{\theta}$ in such a way that the first term is minimized. To this end, we define

$$V\left(\boldsymbol{\theta}\right) := \int_{(0,1)^d} \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta})}\left|\Psi\left(\mathbf{u}\right)\right|^2 d\mathbf{u} = \mathbb{E}_{\mathbb{P}}\left[\frac{g(G^{-1}(U;\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)}{g(G^{-1}(U;\boldsymbol{\theta}_0);\boldsymbol{\theta})}\left|\Psi\left(U\right)\right|^2\right].$$

This expression is particularly useful, in the sense that the underlying probability measure $\mathbb{P}$ is independent of the parameter $\boldsymbol{\theta}$, as well as that the random element involved is identical in law to the uniform random vector U in the original expectation (2.1).

Hereafter, we reserve the notation $\boldsymbol{\theta}_0$ for a fixed point in Θ_0, and introduce the following notations

$$H(\mathbf{u};\boldsymbol{\theta}) := \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta})}, \qquad M(\mathbf{u};\boldsymbol{\theta}) := \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta})}, \tag{3.4}$$

so that

$$\mathbb{E}_{\mathbb{P}}\left[\Psi\left(U\right)\right] = \mathbb{E}_{\mathbb{P}}\left[M(U;\boldsymbol{\theta})\Psi\left(G\left(G^{-1}\left(U;\boldsymbol{\theta}\right);\boldsymbol{\theta}_0\right)\right)\right], \tag{3.5}$$

$$V\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\mathbb{P}}\left[H(U;\boldsymbol{\theta})\left|\Psi\left(U\right)\right|^2\right]. \tag{3.6}$$

Note a slight difference between $H(\mathbf{u};\boldsymbol{\theta})$ and $M(\mathbf{u};\boldsymbol{\theta})$.

Remark 3.1. Those simplified notations (3.4) are helpful not only to simplify the mathematical expressions, but also for ease of programming on an implementation level. That is to say, declaring $H(\mathbf{u};\boldsymbol{\theta})$ and $M(\mathbf{u};\boldsymbol{\theta})$ as user functions in advance simplifies the programming to a large extent. In fact, we may introduce a further simplification, such as

$$Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) := \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta})}, \qquad H(\mathbf{u};\boldsymbol{\theta}) = Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0), \qquad M(\mathbf{u};\boldsymbol{\theta}) = Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}), \tag{3.7}$$

for programming purposes. Nevertheless, we do not use the notation $Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})$ in the present paper, as it turns out to be an oversimplification causing rather unnecessary confusion. □

In order to discuss the estimator variance indexed by the parameter $\boldsymbol{\theta}$, we restrict our attention to the following set

$$\Theta_1 := \mathrm{int}\left\{\boldsymbol{\theta}\in\Theta_0 : \int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\theta})\left|\Psi\left(\mathbf{u}\right)\right|^2 d\mathbf{u} < +\infty\right\}. \tag{3.8}$$

Clearly, Θ_1 is (the interior) of the set of the parameter $\boldsymbol{\theta}$ with which the second moment $V(\boldsymbol{\theta})$ is well defined, whereas smoothness of $V(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ is not guaranteed yet. In order to avoid triviality, we assume the following.

Assumption 3.2. Throughout the paper, we assume

(i) $\mathrm{Leb}(\Theta_1) > 0$;
(ii) $\inf_{\boldsymbol{\theta}\in\Theta_1} V(\boldsymbol{\theta}) > C^2$.

Since $\boldsymbol{\theta}_0$ is chosen from the set Θ_0 in Section 3 and due to the original integrability condition (2.2), the set Θ_1 contains, at least, $\boldsymbol{\theta}_0$. If the set Θ_1 were a singleton against Assumption 3.2(i), then there would be no room for a variance reduction. Assumption 3.2(ii) indicates that the estimator variance cannot be reduced to zero in any way, that is, perfect importance sampling is impossible. This assumption is required to avoid technical issues later on the convergence of the parameter search phase in Theorem 6.1. In principle, however, this assumption is automatically satisfied due to parametricity of the bypass distribution.

We define the set

$$\Theta_2 := \operatorname{int} \bigcup_{B\subseteq\Theta_1} \left\{ B : \int_{(0,1)^d} \sup_{\boldsymbol{\theta}\in\overline{B}} \max\left\{1, \left\| \frac{\nabla_{\boldsymbol{\theta}} H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})} \right\|, \left\| \frac{\mathrm{Hess}_{\boldsymbol{\theta}}(H(\mathbf{u};\boldsymbol{\theta}))}{H(\mathbf{u};\boldsymbol{\theta})} \right\|\right\} H(\mathbf{u};\boldsymbol{\theta})\,|\Psi(\mathbf{u})|^2\,d\mathbf{u} < +\infty, \right.$$
$$\left. \text{and for almost every } \mathbf{z}\in D,\ (g(\mathbf{z};\boldsymbol{\theta}))^{-1} \text{is strictly convex in } \boldsymbol{\theta} \text{ on } \overline{B} \right\}, \tag{3.9}$$

which enables us to take a close look at the structure of the function $V(\boldsymbol{\theta})$, as follows. Note that in view of (3.4) and (3.7), the convexity requirement (3.9) on $(g(\mathbf{z};\boldsymbol{\theta}))^{-1}$ in $\boldsymbol{\theta}$ for almost every $\mathbf{z}$ ensures the convexity of $H(\mathbf{u};\boldsymbol{\theta})$ and $Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})$ in $\boldsymbol{\theta}$ for almost every $\mathbf{u}$ as well as for each $\boldsymbol{\lambda}$, provided that the inverse $G^{-1}(\mathbf{u};\boldsymbol{\lambda})$ is well defined.

Theorem 3.3. (i) *It holds that*

$$\lim_{n\uparrow+\infty} \inf_{\boldsymbol{\theta}\in\partial K_n} V(\boldsymbol{\theta}) = +\infty,$$

where $\{K_n\}_{n\in\mathbb{N}}$ *is an increasing sequence of compact subsets of* Θ_1, *satisfying* $\cup_{n=1}^{+\infty} K_n = \Theta_1$ *and* $K_n \subsetneq \mathrm{int}(K_{n+1})$.

(ii) *If* $\mathrm{Leb}(\Theta_2) > 0$, *then* $V(\boldsymbol{\theta})$ *is twice continuously differentiable and strictly convex on* Θ_2, *with*

$$\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}\left[\nabla_{\boldsymbol{\theta}} H(U;\boldsymbol{\theta})\,|\Psi(U)|^2\right], \tag{3.10}$$

$$\mathrm{Hess}_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})) = \mathbb{E}_{\mathbb{P}}\left[\mathrm{Hess}_{\boldsymbol{\theta}}(H(U;\boldsymbol{\theta}))\,|\Psi(U)|^2\right]. \tag{3.11}$$

In particular, if $\Theta_1 = \Theta_2$, *then*

$$\boldsymbol{\theta}^* := \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta_1} V(\boldsymbol{\theta}), \tag{3.12}$$

is a unique interior point of Θ_2 *satisfying* $\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}^*) = 0$.

It is ideal to find the minimizer $\boldsymbol{\theta}^*$ in the set Θ_2 of the second moment $V(\boldsymbol{\theta})$ for Monte Carlo simulations with the least estimator variance. The results above indicate that there exists at least one $\boldsymbol{\theta}^*$ in the interior of Θ_1, not located towards the boundary $\partial\Theta_1$ (or at infinity if unbounded), while the minimizer is unique if $\Theta_1 = \Theta_2$.

Proof. (i) By the definition (3.8) of the set Θ_1, the claim holds true for every point $\boldsymbol{\theta}$ in $\partial\Theta_1 \setminus \partial\Theta_0$. Next, the integral in (3.8) explodes as $\boldsymbol{\theta} \to \partial\Theta_1 \cap \partial\Theta_0$, since the denominator $g(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta})$ of the likelihood in (3.8) tends to zero for almost every $\mathbf{u}\in(0,1)^d$, due to Assumption 2.1(f).

(ii) For ease of notation, we denote by $\nabla_{\boldsymbol{\theta}}^q$ the qth derivatives. Fix $\boldsymbol{\theta}\in\Theta_2$ and let $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ be a sequence in Θ_2 converging to $\boldsymbol{\theta}$. Since then $\boldsymbol{\theta}\in\Theta_1 \subseteq \Theta_0$, $H(\mathbf{u};\boldsymbol{\theta})$ is twice continuously differentiable for almost every $\mathbf{u}$ with respect to $|\Psi(\mathbf{u})|^2 d\mathbf{u}$. The integrability conditions in Θ_2 ensures that the families $\{\nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta});\ \boldsymbol{\theta}\in\Theta_2\}$ are dominated by a nonnegative function in $\mathbf{u}$, independent of $\boldsymbol{\theta}$, integrable with respect to $|\Psi(\mathbf{u})|^2 d\mathbf{u}$. It thus holds by the dominated convergence theorem that for each $q = 0, 1, 2$,

$$\lim_{k\uparrow+\infty} \mathbb{E}_{\mathbb{P}}\left[\nabla_{\boldsymbol{\theta}}^q (H(U;\boldsymbol{\theta}_k))\,|\Psi(U)|^2\right] = \mathbb{E}_{\mathbb{P}}\left[\lim_{k\uparrow+\infty} \nabla_{\boldsymbol{\theta}}^q (H(U;\boldsymbol{\theta}_k))\,|\Psi(U)|^2\right] = \mathbb{E}_{\mathbb{P}}\left[\nabla_{\boldsymbol{\theta}}^q (H(U;\boldsymbol{\theta}))\,|\Psi(U)|^2\right],$$

which shows the continuity of $V(\boldsymbol{\theta})$ and the right-hand sides of (3.10) and (3.11). For $q = 0, 1$ and for every $\mathbf{h}\in\mathbb{R}^d$ with $\|\mathbf{h}\| = 1$, there exists $\varepsilon > 0$ satisfying $\boldsymbol{\theta} + \varepsilon\mathbf{h} \in \Theta_2$ and

$$\int_{(0,1)^d} \left\| \frac{\nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta}+\varepsilon\mathbf{h}) - \nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta})}{\varepsilon} \right\| |\Psi(\mathbf{u})|^2\,d\mathbf{u} = \int_{(0,1)^d} \left\| \nabla_{\boldsymbol{\theta}}^{q+1} H\left(\mathbf{u};\boldsymbol{\theta}_{\varepsilon,\mathbf{h}}\right) \right\| |\Psi(\mathbf{u})|^2\,d\mathbf{u} < +\infty,$$

where $\boldsymbol{\theta}_{\varepsilon,\mathbf{h}}$ is an intermediate point on the line segment joining two points $\boldsymbol{\theta}+\varepsilon\mathbf{h}$ and $\boldsymbol{\theta}$, again due to twice continuous differentiability of $H(\mathbf{u};\boldsymbol{\theta})$, and the last inequality holds by the integrability condition in the domain Θ_2. Hence, it holds by the dominated convergence theorem that for $q = 0, 1$,

$$\lim_{\varepsilon\downarrow 0} \int_{(0,1)^d} \frac{\nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta}+\varepsilon\mathbf{h}) - \nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta})}{\varepsilon} |\Psi(\mathbf{u})|^2\,d\mathbf{u} = \int_{(0,1)^d} \lim_{\varepsilon\downarrow 0} \frac{\nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta}+\varepsilon\mathbf{h}) - \nabla_{\boldsymbol{\theta}}^q H(\mathbf{u};\boldsymbol{\theta})}{\varepsilon} |\Psi(\mathbf{u})|^2\,d\mathbf{u},$$

which gives the identities (3.10) and (3.11). Finally, the strict convexity of $V(\boldsymbol{\theta})$ on Θ_2 follows from the last condition in (3.9). The uniqueness and the first-order necessary optimality condition at $\boldsymbol{\theta}^*$ are obvious when $\Theta_1 = \Theta_2$, due to Theorem 3.3(i) and the strict convexity. □

4. Continuous exponential family

The choice of the bypass distribution is quite flexible as long as Assumption 2.1 is satisfied, whereas as discussed in Example 2.2, we wish to apply a somewhat approachable family of bypass distributions for implementation purposes. In this paper, we would rather adopt one of the simplest parametric families of continuous distributions, that is, the continuous exponential family. (In our context under Assumption 2.1, it suffices to focus on continuous distributions.) Recall that a continuous distribution is said to be in the exponential family if its probability density function can be written in the form

$$g(\mathbf{z};\boldsymbol{\theta}) = \exp\left[\langle \eta(\boldsymbol{\theta}), T(\mathbf{z})\rangle + \kappa(\boldsymbol{\theta}) + S(\mathbf{z})\right] \mathbb{1}_D(\mathbf{z}),$$

where $\eta : \Theta_0 \to \mathbb{R}^d, T : D \to \mathbb{R}^d, \kappa : \Theta_0 \to \mathbb{R}, S : D \to \mathbb{R}$, and the support D is independent of the parameter $\boldsymbol{\theta}$. Restriction to the exponential family simplifies the likelihood ratios (3.4) and derivatives, to a large extent, as

$$\begin{cases} M(\mathbf{u};\boldsymbol{\theta}) = \exp\left[\left\langle \eta(\boldsymbol{\theta}_0) - \eta(\boldsymbol{\theta}), T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta})\right)\right\rangle + \kappa(\boldsymbol{\theta}_0) - \kappa(\boldsymbol{\theta})\right], \\ H(\mathbf{u};\boldsymbol{\theta}) = \exp\left[\left\langle \eta(\boldsymbol{\theta}_0) - \eta(\boldsymbol{\theta}), T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)\right)\right\rangle + \kappa(\boldsymbol{\theta}_0) - \kappa(\boldsymbol{\theta})\right], \\ \dfrac{\nabla_{\boldsymbol{\theta}} H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})} = -\nabla_{\boldsymbol{\theta}}^{\top}\eta(\boldsymbol{\theta}) T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)\right) - \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}), \\ \dfrac{\mathrm{Hess}_{\boldsymbol{\theta}}(H(\mathbf{u};\boldsymbol{\theta}))}{H(\mathbf{u};\boldsymbol{\theta})} = \left(\dfrac{\nabla_{\boldsymbol{\theta}} H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})}\right)^{\otimes 2} - \displaystyle\sum_{j=1}^{m} T_j\left(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)\right) \mathrm{Hess}_{\boldsymbol{\theta}}\left(\eta_j(\boldsymbol{\theta})\right) - \mathrm{Hess}_{\boldsymbol{\theta}}\left(\kappa(\boldsymbol{\theta})\right), \end{cases}$$

where we write $\eta(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta}))^{\top}$ and $T(\mathbf{z}) = (T_1(\mathbf{z}), \ldots, T_d(\mathbf{z}))^{\top}$ and let $\nabla_{\boldsymbol{\theta}}^{\top}\eta(\boldsymbol{\theta})$ indicate the transpose of the Jacobian matrix of $\eta(\boldsymbol{\theta})$. In particular, the term $S(\mathbf{z})$ disappears. In our context, $\eta_j(\boldsymbol{\theta})$ and $T_j(\mathbf{z})$ depend, respectively, only on the jth component of $\boldsymbol{\theta}$ and $\mathbf{z}$, since we have restricted our attention to the case of $g(\mathbf{z};\boldsymbol{\theta})$ with independent components (2.4). Hence, as soon as we focus on the exponential family, our bypass distributions are not curved. Moreover, since every non-curved continuous distribution in the exponential family can be converted to canonical form, that is, $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$, it is not really a restriction to focus on the continuous canonical exponential family, with

$$\begin{cases} g(\mathbf{z};\boldsymbol{\theta}) = \exp\left[\langle \boldsymbol{\theta}, T(\mathbf{z})\rangle + \kappa(\boldsymbol{\theta}) + S(\mathbf{z})\right] \mathbb{1}_D(\mathbf{z}), \\ M(\mathbf{u};\boldsymbol{\theta}) = \exp\left[\left\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta})\right)\right\rangle + \kappa(\boldsymbol{\theta}_0) - \kappa(\boldsymbol{\theta})\right], \\ H(\mathbf{u};\boldsymbol{\theta}) = \exp\left[\left\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)\right)\right\rangle + \kappa(\boldsymbol{\theta}_0) - \kappa(\boldsymbol{\theta})\right], \\ \dfrac{\nabla_{\boldsymbol{\theta}} H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})} = -T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)\right) - \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}), \\ \dfrac{\mathrm{Hess}_{\boldsymbol{\theta}}(H(\mathbf{u};\boldsymbol{\theta}))}{H(\mathbf{u};\boldsymbol{\theta})} = \left(\dfrac{\nabla_{\boldsymbol{\theta}} H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})}\right)^{\otimes 2} - \mathrm{Hess}_{\boldsymbol{\theta}}\left(\kappa(\boldsymbol{\theta})\right) \end{cases} \tag{4.1}$$

so that (3.5) and (3.6) are as simple as

$$\mathbb{E}_{\mathbb{P}}\left[\Psi(U)\right] = e^{\kappa(\boldsymbol{\theta}_0)-\kappa(\boldsymbol{\theta})}\mathbb{E}_{\mathbb{P}}\left[\exp\left[\left\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, T\left(G^{-1}(U;\boldsymbol{\theta})\right)\right\rangle\right] \Psi\left(G\left(G^{-1}(U;\boldsymbol{\theta});\boldsymbol{\theta}_0\right)\right)\right], \tag{4.2}$$

$$V(\boldsymbol{\theta}) = e^{\kappa(\boldsymbol{\theta}_0)-\kappa(\boldsymbol{\theta})}\mathbb{E}_{\mathbb{P}}\left[\exp\left[\left\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, T\left(G^{-1}(U;\boldsymbol{\theta}_0)\right)\right\rangle\right] |\Psi(U)|^2\right]. \tag{4.3}$$

In light of the expressions (4.1), we impose the additional assumption of second differentiability of $\kappa(\boldsymbol{\theta})$, which is usually the case. Also, it is beneficial to simplify the definition of the set (3.9) as follows;

$$\Theta_2 = \mathrm{int} \bigcup_{B \subseteq \Theta_1} \left\{ B : \int_{(0,1)^d} \max\left\{1, \left\|T\left(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)\right)\right\|^2\right\} \sup_{\boldsymbol{\theta}\in\bar{B}} H(\mathbf{u};\boldsymbol{\theta}) |\Psi(\mathbf{u})|^2 \, d\mathbf{u} < +\infty, \right.$$
$$\left. \text{and } \mathrm{Hess}_{\boldsymbol{\theta}}\left(\kappa(\boldsymbol{\theta})\right) \prec 0, \ \boldsymbol{\theta} \in \bar{B} \right\}.$$

The concavity of $\kappa(\boldsymbol{\theta})$ holds true for our numerical experiments throughout the paper.

For later use, we summarize the behavior of the likelihood ratio $H(\mathbf{u};\boldsymbol{\theta})$ and its derivatives when exponential (i) and normal (ii) bypass distributions are applied. Since all cases are in the canonical form with $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$, the reduced expressions (4.1) can be applied. Moreover, all cases satisfy the convexity condition in Θ_2 (subject to the integrability conditions) since $\kappa(\boldsymbol{\theta})$ is concave.

Lemma 4.1. (i) *Let $g(z;\theta) = \theta e^{-\theta z}$, and fix $\theta_0 = 1$. Then, $D = (0, +\infty)$, $\eta(\theta) = \theta$, $T(z) = -z$, $S(z) = 0$, and $\kappa(\theta) = \ln(\theta)$, which is strictly concave on D.*

(i-a) *If* $G(z;\theta) = e^{-\theta z}$, *then it holds that for each* $\theta \in D$, $G^{-1}(u;\theta) = -\theta^{-1}\ln(u)$, $T(G^{-1}(u;\theta)) = \theta^{-1}\ln(u)$,

$$\begin{aligned} H(u;\theta) &= \exp\left[-\ln(\theta) + (1-\theta)\ln(u)\right], \\ |\partial_\theta H(u;\theta)| &= u^{1-\theta}\frac{|\theta\ln(u)+1|}{\theta^2} \sim \begin{cases} \theta^{-1}u^{1-\theta}\,|\ln(u)|, & \text{as } u \downarrow 0, \\ \theta^{-2}, & \text{as } u \uparrow 1, \end{cases} \\ \partial_\theta^2 H(u;\theta) &= u^{1-\theta}\frac{(\theta\ln(u)+1)^2+1}{\theta^3} \sim \begin{cases} \theta^{-1}u^{1-\theta}\,(\ln(u))^2, & \text{as } u \downarrow 0, \\ 2/\theta^3, & \text{as } u \uparrow 1. \end{cases} \end{aligned}$$

(i-b) *If* $G(z;\theta) = 1 - e^{-\theta z}$, *then it holds that for each* $\theta \in D$, $G^{-1}(u;\theta) = -\theta^{-1}\ln(1-u)$, $T(G^{-1}(u;\theta)) = \theta^{-1}\ln(1-u)$,

$$\begin{aligned} H(u;\theta) &= \exp\left[-\ln(\theta) + (1-\theta)\ln(1-u)\right], \\ |\partial_\theta H(u;\theta)| &= (1-u)^{1-\theta}\frac{|\theta\ln(1-u)+1|}{\theta^2} \sim \begin{cases} \theta^{-1}(1-u)^{1-\theta}\,|\ln(1-u)|, & \text{as } u \downarrow 0, \\ \theta^{-2}, & \text{as } u \uparrow 1, \end{cases} \\ \partial_\theta^2 H(u;\theta) &= (1-u)^{1-\theta}\frac{(\theta\ln(1-u)+1)^2+1}{\theta^3} \sim \begin{cases} 2/\theta^3, & \text{as } u \downarrow 0, \\ \theta^{-1}(1-u)^{1-\theta}\,(\ln(1-u))^2, & \text{as } u \uparrow 1. \end{cases} \end{aligned}$$

(ii) *Let* $g(z;\theta) = \phi(z-\theta)$, *and fix* $\theta_0 = 0$. *Then,* $D = \mathbb{R}$, $\eta(\theta) = \theta$, $T(z) = z$, $S(z) = -z^2/2 - \ln\sqrt{2\pi}$, *and* $\kappa(\theta) = -\theta^2/2$, *which is strictly concave on* D. *If* $G(z;\theta) = \Phi(z-\theta)$, *then it holds that for each* $\theta \in D$, $G^{-1}(u;\theta) = \theta + \Phi^{-1}(u)$, $T(G^{-1}(u;\theta)) = \theta + \Phi^{-1}(u)$,

$$\begin{aligned} H(u;\theta) &= e^{-\theta\Phi^{-1}(u)+\theta^2/2}, \\ |\partial_\theta H(u;\theta)| &= \left|\theta - \Phi^{-1}(u)\right| e^{-\theta\Phi^{-1}(u)+\theta^2/2} \sim \left|\Phi^{-1}(u)\right| e^{-\theta\Phi^{-1}(u)+\theta^2/2}, \\ \partial_\theta^2 H(u;\theta) &= \left(\left(\theta - \Phi^{-1}(u)\right)^2 + 1\right) e^{-\theta\Phi^{-1}(u)+\theta^2/2} \sim \left|\Phi^{-1}(u)\right|^2 e^{-\theta\Phi^{-1}(u)+\theta^2/2}, \end{aligned}$$

where the asymptotic equivalences hold true as either $u \downarrow 0$ *or* $u \uparrow 1$.

Remark 4.2. Apart from the exponential family, one might wonder whether the class of linear probability density functions can act as a more reasonable bypass distribution. This is not the case. Consider the simplest linear case $g(z;\theta) = \theta(z - 1/2) + 1$, $z \in (0,1)$ with $\theta \in (-2+2)$. Note that this density function is not in the exponential family. Although $G(z;\theta) = (\theta z^2 + (2-\theta)z)/2$ is still simple, its inverse is quite intricate. The case of piecewise linear density functions is even worse. Consider a triangular distribution on $(0,1)$ at mode $\theta \in (0,1)$, that is,

$$g(z;\theta) = \begin{cases} 2\dfrac{z}{\theta}, & z \in (0,\theta], \\ 2\dfrac{1-z}{1-\theta}, & z \in (\theta,1) \end{cases}, \qquad G(z;\theta) = \begin{cases} \dfrac{z^2}{\theta}, & z \in (0,\theta], \\ 1 - \dfrac{(1-z)^2}{1-\theta}, & z \in (\theta,1), \end{cases}$$

which are at most quadratic. First of all, this density function violates the condition in the set Θ_0 as it is not twice continuously differentiable with respect to $\boldsymbol{\theta}$. Moreover, the computation of, for example, the likelihood ratio $g(G^{-1}(u;\theta_0);\theta)/g(G^{-1}(u;\theta_0);\theta_0)$ turns out to be surprisingly involved. Two distinct break points θ_0 and θ (for instance, $\theta_0 < \theta$) induce the function $g(G^{-1}(u;\theta_0);\theta)$ piecewise-defined on three domains, that is, $(0,\theta_0]$, $(\theta_0,\theta]$ and $(\theta,1)$. The (non-triangular) trapezoidal distribution is even a worse choice. Two break points in the trapezoidal distribution result in either four or five pieces of the function $g(G^{-1}(u;\theta_0);\theta)$. □

5. Examples

In this section, we provide examples with numerical results to present the procedure, illustrate that the choice of bypass distributions affects the effectiveness of the proposed method (Section 5.1), and demonstrate that the proposed framework is general enough to cover problem settings under a variety of multivariate probability laws (Section 5.2).

5.1. Choice of bypass distributions

We start with a toy example of a Monte Carlo simulation of an area on the unit square $(0,1)^2$. To be precise, we examine an estimation of the area of the indicator function

$$\Psi(\mathbf{u}) = \mathbb{1}\left(u_2 \le c_2 - \frac{c_2}{c_1}u_1\right), \quad 0 \le c_1 \le c_2 \le 1, \ \mathbf{u} := (u_1,u_2)^\top.$$

Throughout this numerical experiment, we fix $(c_1,c_2) = (0.1, 0.2)$, and thus

$$C = \mathbb{E}_{\mathbb{P}}[\Psi(U)] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\left(U_2 \le c_2 - \frac{c_2}{c_1}U_1\right)\right] = \frac{c_1c_2}{2} = 0.01, \quad U := (U_1,U_2)^\top,$$

with the crude estimator variance

$$\mathrm{Var}_{\mathbb{P}}\left(\Psi\left(U\right)\right) = \frac{c_1c_2}{2} - \left(\frac{c_1c_2}{2}\right)^2 = 0.0099.$$

Suppose first we apply the adaptive control variates (2.8), which is available in closed form as

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\left[\Psi\left(U_k\right) - \left\langle \boldsymbol{\lambda}_{n-1}, U_k - \mathbb{E}_{\mathbb{P}}[U]\right\rangle\right] - C\right) \overset{\mathscr{L}}{\to} \mathscr{N}\left(0, \mathrm{Var}_{\mathbb{P}}\left(\Psi\left(U\right)\right) - \frac{1}{12}\left\|\boldsymbol{\lambda}^*\right\|^2\right),$$

$$\boldsymbol{\lambda}_{n-1} := \frac{12}{n-1}\sum_{k=1}^{n-1}\left(\Psi\left(U_k\right) - \frac{1}{n-1}\sum_{m=1}^{n-1}\Psi\left(U_m\right)\right)\left(U_k - \mathbb{E}_{\mathbb{P}}[U]\right) \to c_1c_2\begin{bmatrix}2c_1-3\\2c_2-3\end{bmatrix} =: \boldsymbol{\lambda}^*,$$

where $\{U_k\}_{k\in\mathbb{N}}$ here is a sequence of i.i.d. uniform random vectors on $(0,1)^2$. The optimal parameter and limiting minimized variance when $(c_1, c_2) = (0.1, 0.2)$ are given by

$$\boldsymbol{\lambda}^* = (-0.056, -0.052)^\top, \quad \mathrm{Var}_{\mathbb{P}}\left(\Psi\left(U\right)\right) - \frac{1}{12}\left\|\boldsymbol{\lambda}^*\right\|^2 = 0.0094133.$$

This adaptive control variates method (5.1) accelerates the convergence of a Monte Carlo simulation in the long run by only a variance factor of 1.052 ($\approx 0.0099/0.0094133$), which indicates effectively no improvement. (This control variates is however not always useless at all. For instance, the variance factor increases monotonically up to 3, as c_1 and c_2 increase.)

5.1.1. Exponential–normal bypass

We first examine the exponential and normal distributions for the first and second dimensions with (2.6) and (2.7) of Example 2.2. For further simplicity, we fix $\boldsymbol{\theta}_0 = (1,0)^\top$. That is, in the form of the exponential family in canonical form, we have

$$\eta(\boldsymbol{\theta}) = \boldsymbol{\theta} := (\theta_1, \theta_2)^\top, \qquad T(\mathbf{z}) = (-z_1, z_2)^\top, \qquad \kappa(\boldsymbol{\theta}) = \ln(\theta_1) - \frac{\theta_2^2}{2} - \ln\sqrt{2\pi}, \qquad S(\mathbf{z}) = -\frac{z_2^2}{2}.$$

Clearly, $\Theta_0 = (0,+\infty)\times\mathbb{R}$. With the expressions

$$\begin{aligned}
M(\mathbf{u};\boldsymbol{\theta}) &= \exp\left[-\ln(\theta_1) + \frac{1-\theta_1}{\theta_1}\ln(u_1) - \frac{1}{2}\theta_2^2 - \theta_2\Phi^{-1}(u_2)\right],\\
H(\mathbf{u};\boldsymbol{\theta}) &= \exp\left[-\ln(\theta_1) + (1-\theta_1)\ln(u_1) + \frac{1}{2}\theta_2^2 - \theta_2\Phi^{-1}(u_2)\right],\\
G\left(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0\right) &= \left(u_1^{1/\theta_1},\ \Phi\left(\theta_2 + \Phi^{-1}(u_2)\right)\right)^\top,
\end{aligned} \tag{5.1}$$

the probabilistic representations (4.2) and (4.3) can be written as

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\Psi\left(U\right)\right] &= \mathbb{E}_{\mathbb{P}}\left[\exp\left[-\ln(\theta_1) + \frac{1-\theta_1}{\theta_1}\ln(U_1) - \frac{1}{2}\theta_2^2 - \theta_2\Phi^{-1}(U_2)\right]\Psi\left(U_1^{1/\theta_1},\ \Phi\left(\theta_2+\Phi^{-1}(U_2)\right)\right)\right],\\
V(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbb{P}}\left[\exp\left[-\ln(\theta_1) + (1-\theta_1)\ln(U_1) + \frac{1}{2}\theta_2^2 - \theta_2\Phi^{-1}(U_2)\right]|\Psi\left(U\right)|^2\right].
\end{aligned} \tag{5.2}$$

As illustrated in Fig. 2, the function $\Psi(\mathbf{u})$ is 1, supported only around the origin, and vanishes whenever $u_1 \uparrow 1$ and $u_2 \uparrow 1$. Hence, it suffices to check the behavior around the origin $\mathbf{u} = (0,0)^\top$. It is straightforward to check that $V(\boldsymbol{\theta})$ is well defined in $\Theta_1 = (0,2)\times\mathbb{R}(\subset\Theta_0)$, where we have used $\int_0^1 e^{-\theta\Phi^{-1}(u)}du = \int_{\mathbb{R}} e^{-\theta y}\phi(y)dy = e^{\theta^2/2}$. With the aid of the reduced expression (4.1), the score function and Hessian matrix are given by

$$\frac{\nabla_{\boldsymbol{\theta}}H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})} = \begin{bmatrix}-1/\theta_1 - \ln(u_1)\\ \theta_2 - \Phi^{-1}(u_2)\end{bmatrix}, \tag{5.3}$$

$$\frac{\mathrm{Hess}_{\boldsymbol{\theta}}(H(\mathbf{u};\boldsymbol{\theta}))}{H(\mathbf{u};\boldsymbol{\theta})} = \left(\frac{\nabla_{\boldsymbol{\theta}}H(\mathbf{u};\boldsymbol{\theta})}{H(\mathbf{u};\boldsymbol{\theta})}\right)^{\otimes 2} + \begin{bmatrix}1/\theta_1^2 & 0\\ 0 & 1\end{bmatrix}. \tag{5.4}$$

For the integrability condition in Θ_2, it suffices to check componentwise. With the aid of Lemma 4.1, it is easy to show that for each $k\in\mathbb{N}$,

$$\begin{aligned}
&\int_{(0,\varepsilon)}\max\left\{1, |\ln(u)|^2\right\}\sup_{\theta\in[(k+1)^{-1},2-(k+1)^{-1}]}\theta^{-1}u^{1-\theta}du < +\infty,\\
&\int_{(0,\varepsilon)}\max\left\{1, \left|\Phi^{-1}(u)\right|^2\right\}\sup_{\theta\in[-k,+k]}e^{-\theta\Phi^{-1}(u)+\theta^2/2}du < +\infty,
\end{aligned}$$

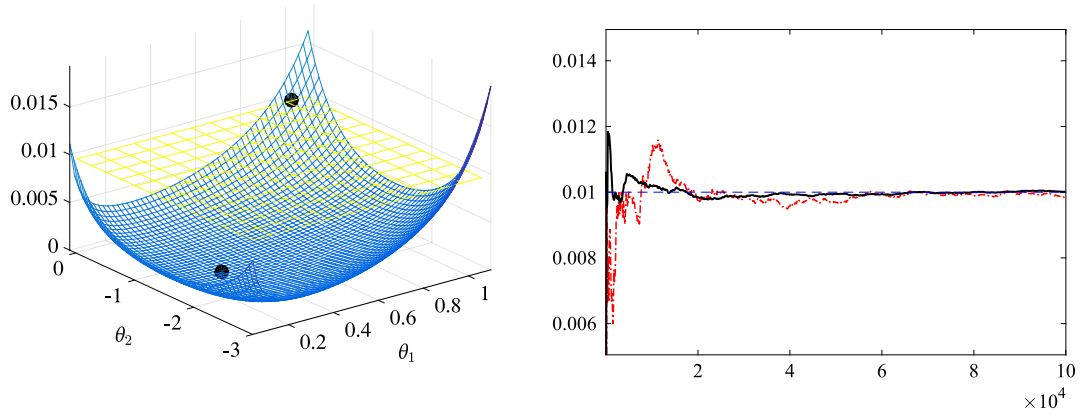


Fig. 1. The left figure plots the estimator variance $V(\boldsymbol{\theta}) - C^2$, along with two filled circles indicating the crude Monte Carlo case $\boldsymbol{\theta}_0$, as well as the minimizer $\boldsymbol{\theta}^*$. The right figure plots a typical convergence of the minimum variance Monte Carlo simulation with the minimizer $\boldsymbol{\theta}^*$ (solid), and that of the crude Monte Carlo simulation (dash–dot), along with the true mean $\mathbb{E}_{\mathbb{P}}[\Psi(U)]$ (dash).

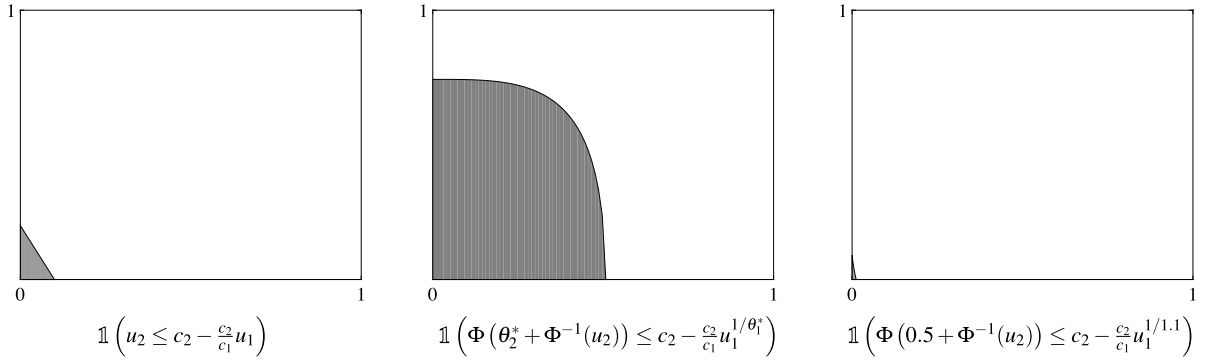


Fig. 2. Domain $\{\mathbf{u} \in (0,1)^2 : \Phi(\theta_2 + \Phi^{-1}(u_2)) \le c_2 - \frac{c_2}{c_1}u_1^{1/\theta_1}\}$, with $\boldsymbol{\theta}_0 = (1,0)^\top$ (left), with the minimizer $\boldsymbol{\theta}^*$ (middle), and with a wrong choice $\boldsymbol{\theta} = (1.1, 0.5)^\top$.

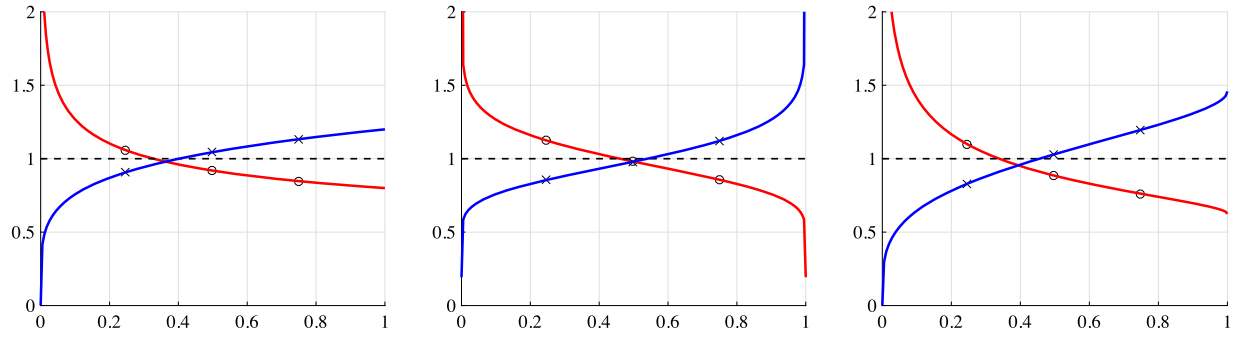
where we have used $\int_{(0,1/2)} |\Phi^{-1}(u)|^2 e^{-\theta\Phi^{-1}(u)} du = \int_{-\infty}^{0} |y|^2 e^{-\theta y}\phi(y)dy < +\infty$. Moreover, $\kappa(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$. Hence, $\Theta_2 = (0,2) \times \mathbb{R}$, which means $\Theta_1 = \Theta_2$. Theorem 3.3 ensures that the function $V(\boldsymbol{\theta})$ is well defined and strictly convex on $(0,2) \times \mathbb{R}$, and thus there exists a unique minimizer $\boldsymbol{\theta}^*$ in its interior.

In this numerical experiment, we find the minimizer $\boldsymbol{\theta}^* \approx (0.288600, -1.499630)^\top$ by a numerical approximation. The estimator variance attains the minimum $V(\boldsymbol{\theta}^*) - C^2 \approx 6.869 \times 10^{-4}$. Hence, the optimal variance reduction in this framework reduces the estimator variance by a factor of $14.41 (\approx 0.0099/0.0006869)$. We plot the estimator variance $V(\boldsymbol{\theta}) - C^2$ in the left of Fig. 1, along with the transparent flat plane indicating the crude Monte Carlo estimator variance. For further illustration, we give two filled circles to indicate the crude Monte Carlo case $\boldsymbol{\theta}_0 = (1,0)^\top$, as well as the minimizer $\boldsymbol{\theta}^*$. In the right figure, we plot a typical convergence of the minimum variance Monte Carlo simulation with the minimizer $\boldsymbol{\theta}^*$ (solid), and that of the crude Monte Carlo simulation (dash–dot), along with the true mean $\mathbb{E}_{\mathbb{P}}[\Psi(U)] = 0.01$ (dash), to illustrate the effectiveness of the achieved variance reduction.

We plot in Fig. 2 the support of the indicator function

$$\Psi\left(u_1^{1/\theta_1},\ \Phi\left(\theta_2 + \Phi^{-1}(u_2)\right)\right) = \mathbb{1}\left(\Phi\left(\theta_2 + \Phi^{-1}(u_2)\right) \le c_2 - \frac{c_2}{c_1}u_1^{1/\theta_1}\right), \qquad \mathbf{u} \in (0,1)^2,$$

in the expectation (5.2). The leftmost figure corresponds to the case when importance sampling is not applied at all, that is, $\boldsymbol{\theta}_0 = (1,0)^\top$. From the original formulation $\mathbb{E}_{\mathbb{P}}[\Psi(U)] = 0.01$, it follows that, on average, 99% of i.i.d. realizations return the value of zero since they fall into the unfilled section. The middle figure indicates the support with the minimizer $\boldsymbol{\theta}^*$, that is substantially expanded to a Lebesgue area of 0.3408. Rather than throwing out 99%, this importance sampling allows us to make use of 34.08% of i.i.d. realizations of non-zero integrands for Monte Carlo estimation. A wrong choice of the parameter $\boldsymbol{\theta}$ may deteriorate the simulation, as illustrated in the rightmost figure. That is, this choice yields a Lebesgue area of 8.34×10^{-4}, that is, around 99.92% of i.i.d. realizations will return zero.



Exponential: $g(z;\theta) = \theta e^{-\theta z}$
$\theta = \{0.8\,(\text{-}\circ\text{-}),\ 1.0\,(\text{- -}),\ 1.2\,(\text{-}\times\text{-})\}$,
$\theta_0 = 1.0$

Gaussian: $g(z;\theta) = \phi(z-\theta)$
$\theta = \{-0.2\,(\text{-}\circ\text{-}),\ 0.0\,(\text{- -}),\ +0.2\,(\text{-}\times\text{-})\}$,
$\theta_0 = 0.0$

Beta: $g(z;\theta) = \frac{z^{\theta-1}(1-z)^{2-1}}{B(\theta,2)}$
$\theta = \{1.5\,(\text{-}\circ\text{-}),\ 2.0\,(\text{- -}),\ 2.5\,(\text{-}\times\text{-})\}$
$\theta_0 = 2.0$

Fig. 3. The density function of the transformed random variable $G(G^{-1}(U;\theta);\theta_0)$ through the exponential bypass (left), the normal bypass (middle) and the beta bypass (right).

We can show that the probability density function of those two components is given, respectively, by

$$\partial_x \mathbb{P}\left(U_1^{1/\theta_1} \le x\right) = \theta_1 x^{\theta_1 - 1}, \quad x \in (0,1),$$

$$\partial_x \mathbb{P}\left(\Phi\left(\theta_2 + \Phi^{-1}(U_2)\right) \le x\right) = \frac{\phi(\Phi^{-1}(x) - \theta_2)}{\phi(\Phi^{-1}(x))}, \quad x \in (0,1),$$

which are illustrated in the leftmost and middle figures in Fig. 3, for two representing values of the parameter θ around the crude case θ_0. As is now obvious in comparison with Fig. 2, the integrand $\Psi(u_1^{1/\theta_1}, \Phi(\theta_2 + \Phi^{-1}(u_2)))$ is more likely to return the value of 1 than the original integrand $\Psi(\mathbf{u})$ thanks to more mass on values towards $\mathbf{u} = (0,0)^\top$ (and thus less mass towards $\mathbf{u} = (1,1)^\top$), by suitably choosing $\theta_1 < 1$ and $\theta_2 < 0$, both illustrated by the density functions with (- ∘ -) in Fig. 3. In comparison between the exponential and normal bypass distributions, the normal bypass can symmetrically weight more on an end and less on the opposite end.

In addition, we plot in the rightmost figure in Fig. 3 the density function of the transformed random variable $G(G^{-1}(U;\theta);\theta_0)$, through the beta bypass $g(z;\theta) = z^{\theta-1}(1-z)^{b-1}/B(\theta,b)$, which is given by

$$\partial_x \mathbb{P}\left(G(G^{-1}(U;\theta);\theta_0) \le x\right) = \frac{B(\theta_0,b)}{B(\theta,b)}\left(G^{-1}(x;\theta_0)\right)^{\theta-\theta_0}, \quad x \in (0,1). \tag{5.5}$$

Indeed, what this beta bypass offers does not seem significantly different from what the exponential distribution does (the leftmost figure). An important point here is that the computation of $G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)$ with the beta bypass is far more expensive (although some mathematical tools have preset functions for beta cumulative distribution function and its inverse as default). Hence, it does not seem very convincing to employ the beta bypass, in particular in place of the exponential bypass.

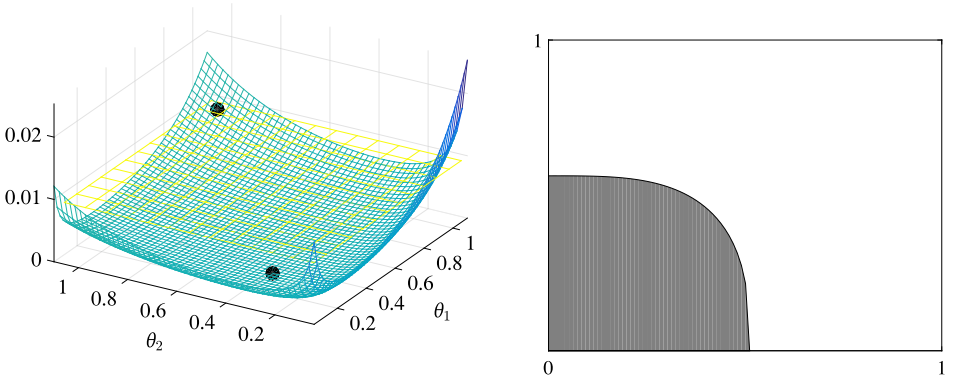
Remark 5.1. In the standard importance sampling method, the beta distribution is a very useful parametrized proposal distribution to directly control the weight of the uniform distribution, provided that all relevant integrability conditions are satisfied, as

$$\mathbb{E}_{\mathbb{P}}[\Psi(U)] = \mathbb{E}\left[\frac{B(a,b)}{Z^{a-1}(1-Z)^{b-1}}\Psi(Z)\right],$$

where the random variable Z on the right-hand side has the probability density function $z^{a-1}(1-z)^{b-1}/B(a,b)$ on $(0,1)$. The density function (5.5) indicates that the change of measure by the beta distribution above has nothing to do with the beta bypass in the proposed framework. □

Remark 5.2. The exponential bypass $g(z;\theta) = \theta e^{-\theta z}$ in our context should not be confused with the widely applied exponential tilting of the uniform distribution, that is,

$$\frac{e^{\theta u}}{\mathbb{E}[e^{\theta U}]} = \frac{\theta e^{\theta u}}{e^\theta - 1} \in \begin{cases} \left(\dfrac{\theta e^\theta}{e^\theta - 1}, \dfrac{\theta}{e^\theta - 1}\right), & \text{if } \theta < 0, \\ \left(\dfrac{\theta}{e^\theta - 1}, \dfrac{\theta e^\theta}{e^\theta - 1}\right), & \text{if } \theta > 0, \end{cases}$$



Estimator variance $V(\boldsymbol{\theta}) - C^2$ $\quad$ $\mathbb{1}\left(u_2^{1/\theta_2^*} \le c_2 - \frac{c_2}{c_1} u_1^{1/\theta_1^*}\right)$

Fig. 4. The left figure plots the estimator variance $V(\boldsymbol{\theta}) - C^2$, along with the locations of the crude parameter $\boldsymbol{\theta}_0$ and of the minimizer $\boldsymbol{\theta}^*$. The right figure plots the domain $\{\mathbf{u} \in (0,1)^2 : u_2^{1/\theta_2^*} \le c_2 - \frac{c_2}{c_1} u_1^{1/\theta_1^*}\}$, with the minimizer $\boldsymbol{\theta}^*$.

which is evidently different from the exponential bypass. This exponential tilting of the standard uniform distribution is however within our scope as well in that it is in the canonical exponential family with $g(z;\theta) = \theta e^{\theta z}/(e^\theta - 1) = \exp[\theta z + \ln(\theta/(e^\theta - 1))]\mathbb{1}_{(0,1)}(z)$ and $G^{-1}(z;\theta) = \theta^{-1}\ln(1 + (e^\theta - 1)z)$, which are somewhat unexpectedly not as simple as the exponential and normal bypass. □

5.1.2. Exponential–exponential bypass

We next examine the exponential bypass for both first and second components, with $d = 2$, $D = (0, +\infty)^2$, and $g(\mathbf{z};\boldsymbol{\theta}) = \theta_1 e^{-\theta_1 z_1}\theta_2 e^{-\theta_2 z_2}$, where $\Theta_0 = (0, +\infty)^2$. Then, the resulting Monte Carlo estimator is given by

$$\mathbb{E}_{\mathbb{P}}[\Psi(U)] = \mathbb{E}_{\mathbb{P}}\left[\frac{1}{\theta_1\theta_2} U_1^{\frac{1-\theta_1}{\theta_1}} U_2^{\frac{1-\theta_2}{\theta_2}} \Psi\left(U_1^{1/\theta_1}, U_2^{1/\theta_2}\right)\right],$$

whereas the second moment is given in closed form by

$$V(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}\left[\frac{1}{\theta_1\theta_2} U_1^{1-\theta_1} U_2^{1-\theta_2} |\Psi(U)|^2\right] = c_1^{2-\theta_1} c_2^{2-\theta_2} \frac{B(2-\theta_1, 2-\theta_2)}{\theta_1\theta_2(4-\theta_1-\theta_2)}.$$

With the minimizer $\boldsymbol{\theta}^* \approx (0.2912, 0.3569)^\top$, the estimator variance attains a minimum

$$V(\boldsymbol{\theta}^*) - C^2 = c_1^{2-\theta_1^*} c_2^{2-\theta_2^*} \frac{B(2-\theta_1^*, 2-\theta_2^*)}{\theta_1^*\theta_2^*(4-\theta_1^*-\theta_2^*)} - C^2 \approx 0.0010521.$$

Hence, we obtain a reduction of the estimator variance by a factor of 9.41($\approx 0.0099/0.0010521$). We plot the estimator variance $V(\boldsymbol{\theta}) - C^2$ in the left of Fig. 4, along with the crude Monte Carlo estimator variance indicated by the transparent flat plane. In the right figure, we plot the optimal domain $\{\mathbf{u} \in (0,1)^2 : u_2^{1/\theta_2^*} \le c_2 - \frac{c_2}{c_1} u_1^{1/\theta_1^*}\}$, with a Lebesgue area of 0.2561. This choice of bypass distributions (exponential–exponential) fails to reduce the estimator variance as much as the previous choice (exponential–normal) of Section 5.1.1.

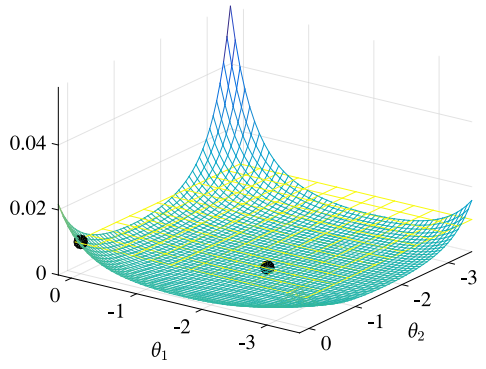
5.1.3. Normal–normal bypass

The next choice is the normal bypass for both first and second components, with $d = 2$, $D = \mathbb{R}^2$, and $g(\mathbf{z};\boldsymbol{\theta}) = \phi(z_1 - \theta_1)\phi(z_2 - \theta_2)$, where $\Theta_0 = \mathbb{R}^2$. Then, the resultant Monte Carlo estimator is given by

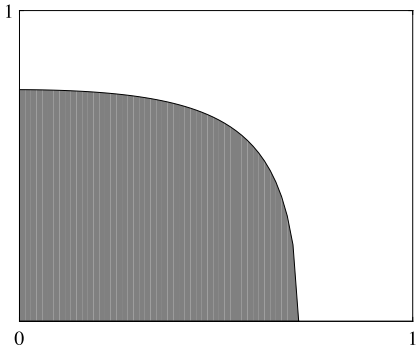
$$\mathbb{E}_{\mathbb{P}}[\Psi(U)] = \mathbb{E}_{\mathbb{P}}\left[\exp\left[-\frac{1}{2}\theta_1^2 - \theta_1\Phi^{-1}(U_1) - \frac{1}{2}\theta_2^2 - \theta_2\Phi^{-1}(U_2)\right] \times \Psi\left(\Phi\left(\theta_1 + \Phi^{-1}(U_1)\right), \Phi\left(\theta_2 + \Phi^{-1}(U_2)\right)\right)\right],$$

whereas the second moment is given by

$$V(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}\left[\exp\left[\frac{1}{2}\theta_1^2 - \theta_1\Phi^{-1}(U_1) + \frac{1}{2}\theta_2^2 - \theta_2\Phi^{-1}(U_2)\right]|\Psi(U)|^2\right].$$



Estimator variance $V(\boldsymbol{\theta}) - C^2$



$\mathbb{1}\left(\Phi(\theta_2^* + \Phi^{-1}(u_2)) \le c_2 - \frac{c_2}{c_1}\Phi(\theta_1^* + \Phi^{-1}(u_1))\right)$

Fig. 5. The left figure plots the estimator variance $V(\boldsymbol{\theta}) - C^2$, along with the locations of the crude parameter $\boldsymbol{\theta}_0$ and of the minimizer $\boldsymbol{\theta}^*$. The right figure plots the domain $\{\mathbf{u} \in (0,1)^2 : \Phi(\theta_2^* + \Phi^{-1}(u_2)) \le c_2 - \frac{c_2}{c_1}\Phi(\theta_1^* + \Phi^{-1}(u_1))\}$, with the minimizer $\boldsymbol{\theta}^*$.

With $\boldsymbol{\theta}^* \approx (-1.798, -1.502)^\top$, the estimator variance attains the minimum $V(\boldsymbol{\theta}^*) - C^2 \approx 0.00038966$. Hence, we obtain a reduction of variance by a factor of $25.41 (\approx 0.0099/0.00038966)$. We provide numerical results in Fig. 5. The Lebesgue area of the support is now 0.4563.

As observed in Sections 5.1.1–5.1.3, the choice of bypass distributions affects the effectiveness of the proposed method. In this problem setting, the normal bypass results in more reduction in estimator variance compared to the exponential bypass. Yet, it is difficult to draw a universally convincing conclusion, as the effectiveness is generally unforeseeable and depends largely on the integrand $\Psi(\mathbf{u})$ as well. This model selection problem is outside the scope of this paper, although it would deserve a separate investigation.

5.2. Non-uniform laws

We next demonstrate that the proposed framework is general enough to equip adaptive importance sampling for non-uniform multivariate probability laws.

5.2.1. Gaussian law

The first example is the multivariate Gaussian law, under which adaptive importance sampling can be formulated parametrically by the exponential change of measure, applied in [1,2,13,4]. Consider the expected value $C = \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\theta}_0}}[F(Z)]$, where $F : \mathbb{R}^d \to \mathbb{R}$ and $\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\theta}}}$ denotes the expectation taken under the probability measure $\mathbb{Q}_{\boldsymbol{\theta}}$, under which $Z \sim \mathcal{N}(\boldsymbol{\theta}, \mathbb{I}_d)$. (Recall that the probability measure $\mathbb{Q}_{\boldsymbol{\theta}}$ has appeared in (3.3).) We denote $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^\top$ and $\mathbf{z} = (z_1, \dots, z_d)^\top$, and fix $\boldsymbol{\theta}_0 = (0, \dots, 0)^\top$ without loss of generality, since an appropriate matrix multiplication (such as the Cholesky decomposition) on Z can be embedded in the integrand function F so as to generate non-unit variance–covariance structure. Then, without changing the expected value C, the mean of the normal random vector can be shifted as

$$C = \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\theta}_0}}[F(Z)] = \int_{\mathbb{R}^d} F(\mathbf{z})\phi_d(\mathbf{z})d\mathbf{z} = \int_{\mathbb{R}^d} \frac{\phi_d(\mathbf{z})}{\phi_d(\mathbf{z}-\boldsymbol{\theta})} F(\mathbf{z})\phi_d(\mathbf{z}-\boldsymbol{\theta})d\mathbf{z} = \int_{\mathbb{R}^d} \frac{\phi_d(\mathbf{z}+\boldsymbol{\theta})}{\phi_d(\mathbf{z})} F(\mathbf{z}+\boldsymbol{\theta})\phi_d(\mathbf{z})d\mathbf{z},$$

where ϕ_d denotes the joint standard normal probability density function on $\mathbb{R}^d$, that is, $\phi_d(\mathbf{z}+\boldsymbol{\theta}) := \prod_{k=1}^d \phi(z_k + \theta_k)$. The practical relevance of this density transform is that the likelihood ratio $\phi_d(\mathbf{z}+\boldsymbol{\theta})/\phi_d(\mathbf{z})$ reduces to the exponential tilting (also called the Esscher transform) of the standard normal random vector Z, that is, $\phi_d(\mathbf{z}+\boldsymbol{\theta})/\phi_d(\mathbf{z}) = e^{-\langle\boldsymbol{\theta},\mathbf{z}\rangle}/\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\theta}_0}}[e^{-\langle\boldsymbol{\theta},Z\rangle}]$. To represent those identities in our framework, first rewrite the left-hand side as

$$C = \mathbb{E}_{\boldsymbol{\theta}_0}[F(Z)] = \int_{\mathbb{R}^d} F(\mathbf{z})\phi_d(\mathbf{z})d\mathbf{z} = \int_{(0,1)^d} F\left(\Phi_d^{-1}(\mathbf{u})\right) d\mathbf{u} = \mathbb{E}_{\mathbb{P}}\left[F\left(\Phi_d^{-1}(U)\right)\right],$$

where $\Phi_d^{-1}(\mathbf{u}) + \boldsymbol{\theta} := (\Phi^{-1}(u_1) + \theta_1, \dots, \Phi^{-1}(u_d) + \theta_d)^\top$. This indicates that the original expectation (2.1) recovers this by setting $\Psi(\mathbf{u}) = F(\Phi_d^{-1}(\mathbf{u}))$ and $g(\mathbf{z};\boldsymbol{\theta}) = \phi_d(\mathbf{z}-\boldsymbol{\theta})$, which is clearly in the canonical exponential family. Moreover, we have

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\phi_d(\mathbf{z}+\boldsymbol{\theta})}{\phi_d(\mathbf{z})} F(\mathbf{z}+\boldsymbol{\theta})\phi_d(\mathbf{z})d\mathbf{z} &= \int_{(0,1)^d} \frac{\phi_d(\Phi_d^{-1}(\mathbf{u})+\boldsymbol{\theta})}{\phi_d(\Phi_d^{-1}(\mathbf{u}))} F\left(\Phi_d^{-1}(\mathbf{u})+\boldsymbol{\theta}\right) d\mathbf{u} \\ &= \mathbb{E}_{\mathbb{P}}\left[\frac{\phi_d(\Phi_d^{-1}(U)+\boldsymbol{\theta})}{\phi_d(\Phi_d^{-1}(U))} F\left(\Phi_d^{-1}(U)+\boldsymbol{\theta}\right)\right],\end{aligned}$$

which is nothing but the expression (3.2). In the framework of Section 4, all those can be represented with

$$T(\mathbf{z}) = \mathbf{z}, \qquad S(\mathbf{z}) = -\frac{1}{2}\|\mathbf{z}\|^2 - d\ln\sqrt{2\pi}, \qquad \kappa(\boldsymbol{\theta}) = -\frac{1}{2}\|\boldsymbol{\theta}\|^2, \qquad D = \mathbb{R}^d,$$
$$H(\mathbf{u};\boldsymbol{\theta}) = \exp\left[-\left\langle\boldsymbol{\theta}, \Phi_d^{-1}(\mathbf{u}) + \boldsymbol{\theta}_0\right\rangle + \frac{1}{2}\|\boldsymbol{\theta}\|^2\right].$$

In a similar manner, we obtain the second moment and its derivatives:

$$\begin{cases} V(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}\left[e^{-\langle\boldsymbol{\theta},\Phi_d^{-1}(U)\rangle + \|\boldsymbol{\theta}\|^2/2}\left|F\left(\Phi_d^{-1}(U)\right)\right|^2\right], \\ \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}\left[\left(\boldsymbol{\theta} - \Phi_d^{-1}(U)\right)e^{-\langle\boldsymbol{\theta},\Phi_d^{-1}(U)\rangle + \|\boldsymbol{\theta}\|^2/2}\left|F\left(\Phi_d^{-1}(U)\right)\right|^2\right], \\ \mathrm{Hess}_{\boldsymbol{\theta}}\left(V(\boldsymbol{\theta})\right) = \mathbb{E}_{\mathbb{P}}\left[\left(\left(\boldsymbol{\theta} - \Phi_d^{-1}(U)\right)^{\otimes 2} + \mathbb{I}_d\right)e^{-\langle\boldsymbol{\theta},\Phi_d^{-1}(U)\rangle + \|\boldsymbol{\theta}\|^2/2}\left|F\left(\Phi_d^{-1}(U)\right)\right|^2\right]. \end{cases}$$

For instance, consider one-dimensional barrier option under the standard Black–Scholes model, discussed in [2, Section 3.2.1], where the integrand is given by

$$F(\mathbf{z}) = F\left(\Phi_d^{-1}(\mathbf{u})\right) = e^{-rT}\left(S_0 e^{(r-\sigma^2/2)T + \sigma\sqrt{T/d}\sum_{k=1}^{d}\Phi^{-1}(u_k)} - K\right)_+ \times \mathbb{1}\left(S_0 e^{(r-\sigma^2/2)jT/d + \sigma\sqrt{jT/d}\sum_{k=1}^{j}\Phi^{-1}(u_k)} \geq L,\ j = 1,\ldots,d\right).$$

This integrand contains discontinuities in the variable $\mathbf{z}$ (and thus in the variable $\mathbf{u}$ as well), due to the indicator function. It is reported that with $\sigma = 0.2$, $r = 0.05$, $T = 2$, $S_0 = 100$, $K = 110$ and $d = 24$ (that is, 24-dimensional normal random vector), the exponential change of measure has the potential for reduction of variance by a factor of roughly 8 to 12 for different L's between 70 and 95.

5.2.2. Gamma distribution

The gamma distribution is another example where adaptive importance sampling can be constructed parametrically by the exponential change of measure. Here, for ease of notation, we restrict ourselves to the univariate setting but can obviously extend to the multivariate setting with independent components. Consider the expected value $\mathbb{E}[F(Z)]$, where $F : (0,+\infty) \to \mathbb{R}$ and Z here denotes the gamma random variable with density $f(z;b) := b^a/\Gamma(a)x^{a-1}e^{-bz}$ on $(0,+\infty)$, where a is a fixed positive constant. The gamma distribution possesses the so-called scaling property; for $\theta < b$,

$$\begin{aligned}\int_{(0,+\infty)} F(z)f(z;b)dz &= \int_{(0,+\infty)} \frac{f(z;b)}{f(z;b-\theta)}F(z)f(z;b-\theta)dz \\ &= \int_{(0,+\infty)} \frac{f(bz/(b-\theta);b)}{f(bz/(b-\theta);b-\theta)}F\left(\frac{b}{b-\theta}z\right)f(z;b)dz, \end{aligned} \tag{5.6}$$

provided that the integrals are all well defined. That is, the integrals at the both ends of (5.6) are taken with respect to the common probability measure $f(z;b)dz$, but the integrands are written on two different scales. The scaling property (5.6) has been employed for adaptive variance reduction methods as well as sensitivity analysis [14,19–21]. Again, the likelihood ratio reduces to an exponential tilting of the original gamma random variable Z, as

$$\frac{f(bz/(b-\theta);b)}{f(bz/(b-\theta);b-\theta)} = \left(\frac{b}{b-\theta}\right)^a e^{-\frac{b\theta}{b-\theta}z} = \frac{e^{-\frac{b\theta}{b-\theta}z}}{\mathbb{E}\left[e^{-\frac{b\theta}{b-\theta}Z}\right]}.$$

The original expectation (2.1) recovers this by setting $g(z;\theta) = f(z;b-\theta)$, which is in the canonical exponential family, and $\Psi(u) = F(\gamma^{-1}(a,\Gamma(a)u)/b)$, where $\gamma^{-1}(a,z)$ denotes the inverse of the lower incomplete gamma function $\gamma(a,x) := \int_0^x t^{a-1}e^{-t}dt$ with respect to the variable x, that is, the identity (5.6) can be rewritten as

$$\int_{(0,1)} F\left(\frac{\gamma^{-1}(a,\Gamma(a)u)}{b}\right)du = \int_{(0,1)}\left(\frac{b}{b-\theta}\right)^a e^{-\frac{\theta}{b-\theta}\gamma^{-1}(a,\Gamma(a)u)}F\left(\frac{\gamma^{-1}(a,\Gamma(a)u)}{b-\theta}\right)du.$$

For instance, adaptive importance sampling methods for the Gamma copula model (101 independent components) and the intensity Gamma model (2121 independent components) in the context of credit derivatives pricing [14, Section 5] can be rewritten in the formulation above.

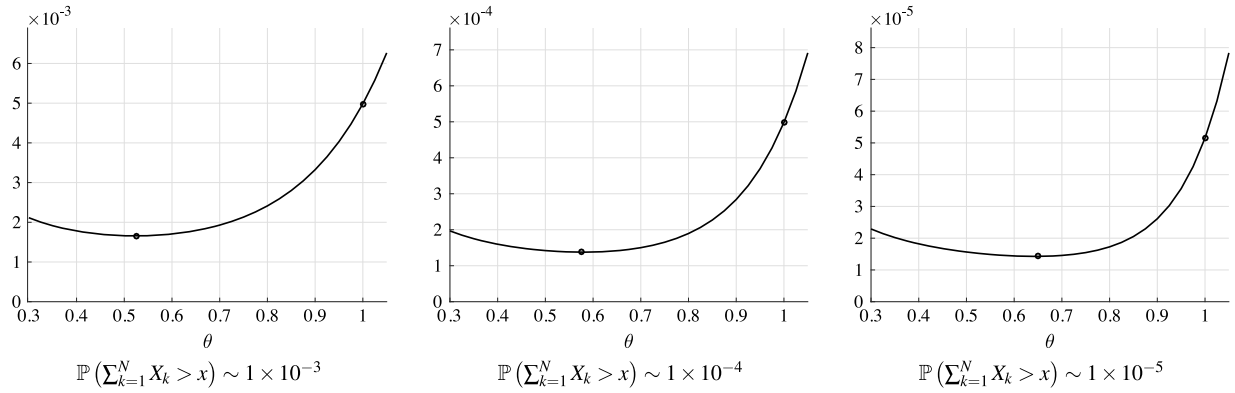


Fig. 6. The estimator second moment $V(\theta)$ against θ, with the success probability $\rho = 0.8$ and the Pareto index $\alpha = 0.5$. Two circles indicate no importance sampling $\theta = 1 (= \theta_0)$ and the unique optimum θ^*.

5.2.3. Random sums of random variables

We here demonstrate that the proposed framework is able to construct adaptive importance sampling in problem settings, where adaptive importance sampling *cannot* be constructed parametrically by the exponential change of measure in their original form. For instance, consider the tail probability of the random sum of random variables $\mathbb{P}(X_1 + \cdots + X_N > x)$, where N is a non-negative integer valued random variable, and $\{X_k\}_{k\in\mathbb{N}}$ is a sequence of i.i.d. non-negative random variables. Here, we let N be a geometric random variable with success probability ρ, where $\mathbb{P}(N = n) = (1-\rho)^n\rho$, for $n = 0, 1, 2, \ldots$ We let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of i.i.d. Lomax (Pareto type II) random variables [22] with probability density function $f(x) = \alpha(1+x)^{-\alpha-1}$ on $(0, +\infty)$. It holds by appropriate conditioning and asymptotics that

$$\mathbb{P}(X_1 + \cdots + X_N > x) = (1-\rho)\mathbb{P}(X_0 + X_1 + \cdots + X_N > x) \sim \frac{1-\rho}{\rho}\frac{1}{(1+x)^\alpha}, \quad x \uparrow +\infty, \tag{5.7}$$

where X_0 is an additional i.i.d. Lomax random variable. Such tail probabilities are of practical importance in various fields of application, such as the ruin probability in insurance, the operational risk in finance, and the steady-state waiting time in the queueing theory. Without Monte Carlo methods, the accurate evaluation would be difficult, particularly due to random and unbounded dimensionality of the problem.

Instead of the original expression of the tail probability above (without the term X_0), it is convenient to use the second expression (with the term X_0), so that the term X_0 necessarily exists, irrespective of the geometric random variable N. For the sake of simplicity and clarity, we describe the potential for reduction of estimator variance when importance sampling is applied to the term X_0 (and no other variance reduction methods are applied). Noting that i.i.d. Lomax random variables can be generated exactly by the inverse transform method as $X_k \overset{\mathscr{L}}{=} U_k^{-1/\alpha} - 1$, we parametrize the tail probability above with the parameter θ through the exponential bypass distribution with $\theta_0 = 1$, as

$$\begin{aligned}\mathbb{P}(X_0 + X_1 + \cdots + X_N > x) &= \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\left(U^{-1/\alpha} - 1 + \sum_{k=1}^N X_k > x\right)\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[\theta^{-1}U^{\frac{1-\theta}{\theta}}\mathbb{1}\left(U^{-\frac{1}{\theta\alpha}} - 1 + \sum_{k=1}^N X_k > x\right)\right],\end{aligned}$$

whose value is independent of θ, while the second moment of the rightmost expression above is given by

$$V(\theta) = \mathbb{E}_{\mathbb{P}}\left[\theta^{-1}U^{1-\theta}\mathbb{1}\left(U^{-1/\alpha} - 1 + \sum_{k=1}^N X_k > x\right)\right],$$

which now depends on θ. (Note that there exist numerous techniques proposed in the literature for improving the estimation efficiency for tail probabilities of this type. In combination of those existing methods, it is naturally expected to induce a further variance reduction. We do not go into this direction, as it is outside the scope of the present paper.) To illustrate the effectiveness of this importance sampling, we plot in Fig. 6 the estimator second moment $V(\theta)$ against the parameter θ for the tail probabilities $\mathbb{P}\left(\sum_{k=1}^N X_k > x\right) \sim 1 \times 10^{-k}$, for $k = 3, 4, 5$, with the success probability $\rho = 0.8$ and the Pareto index $\alpha = 0.5$. The threshold x is fixed on the basis of the asymptotic equivalence (5.7). Those show a reduction of variance by a factor of 3 to 4 quite easily through importance sampling on the term X_0 alone.

6. Monte Carlo simulation concurrently with optimal parameter search

In this section, we discuss the concept of importance sampling through bypass distributions in the framework of adaptive Monte simulation, that is, a Monte Carlo simulation concurrently with optimal parameter search. Recall that $(\mathscr{F}_k)_{k\in\mathbb{N}}$ denotes the filtration generated by a sequence $\{U_k\}_{k\in\mathbb{N}}$ of i.i.d. uniform random vectors on $(0,1)^d$.

The following summarizes convergence results of the *adaptive* empirical mean and variance. In short, the condition (6.3) is imposed to justify the Lindeberg condition for the martingale central limit theorem, while the condition (6.5) ensures the convergence of martingale difference related to the sequence $\{\sigma_n^2\}_{n\in\mathbb{N}}$ of empirical variances. The conditions (6.3) and (6.5) are left in their current form on purpose, rather than newly setting up (almost unverifiable) domains of $\boldsymbol{\theta}$ for those conditions to be satisfied. Recall that $\boldsymbol{\theta}^*$ denotes a deterministic minimizer of $V(\boldsymbol{\theta})$, defined in (3.12), whereas $V(\boldsymbol{\theta})$ is well defined for each $\boldsymbol{\theta} \in \Theta_1$. Also, in Assumption 3.2, we have imposed $V(\boldsymbol{\theta}^*) > C^2$, with strict inequality, that is, perfect importance sampling is impossible. (We refer the reader to, for instance, [3] for details.)

Theorem 6.1. *Let $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ be a sequence of random vectors in Θ_2, adapted to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}}$ and $\boldsymbol{\theta}_k \to \boldsymbol{\theta}^*$, $\mathbb{P}$-a.s. Define*

$$\mu_n := \frac{1}{n}\sum_{k=1}^{n} M\left(U_k;\boldsymbol{\theta}_{k-1}\right)\Psi\left(G\left(G^{-1}\left(U_k;\boldsymbol{\theta}_{k-1}\right);\boldsymbol{\theta}_0\right)\right),$$

$$\sigma_n^2 := \frac{1}{n}\sum_{k=1}^{n} H\left(U_k;\boldsymbol{\theta}_{k-1}\right)\left|\Psi\left(U_k\right)\right|^2 - \mu_n^2.$$

(i) *It holds that*

$$\mathbb{E}_{\mathbb{P}}[\mu_n] = C, \quad n \in \mathbb{N}, \tag{6.1}$$

and that as $n \uparrow +\infty$,

$$\mu_n \to \mathbb{E}_{\mathbb{P}}\left[\Psi(U)\right], \quad \mathbb{P}\text{-a.s.} \tag{6.2}$$

(ii) *If there exists $q > 2$ such that*

$$\limsup_{n\uparrow+\infty} \frac{1}{n}\sum_{k=1}^{n}\left(\int_{(0,1)^d}\left(H(\mathbf{u};\boldsymbol{\theta}_{k-1})\right)^{q-1}\left|\Psi(\mathbf{u})\right|^q d\mathbf{u}\right)^{2/q} < +\infty, \quad \mathbb{P}\text{-a.s.}, \tag{6.3}$$

then it holds under the probability measure $\mathbb{P}$ that

$$\sqrt{n}\left(\mu_n - \mathbb{E}_{\mathbb{P}}\left[\Psi(U)\right]\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, V\left(\boldsymbol{\theta}^*\right) - C^2\right), \quad n \uparrow +\infty. \tag{6.4}$$

(iii) *If moreover*

$$\limsup_{n\uparrow+\infty} \frac{1}{n}\sum_{k=1}^{n}\int_{(0,1)^d}\left(H(\mathbf{u};\boldsymbol{\theta}_{k-1})\right)^2\left|\Psi(\mathbf{u})\right|^4 d\mathbf{u} < +\infty, \quad \mathbb{P}\text{-a.s.}, \tag{6.5}$$

then it holds under the probability measure $\mathbb{P}$ that $\lim_{n\uparrow+\infty}\sigma_n^2 = V(\boldsymbol{\theta}^*) - C^2$, a.s., and

$$\sqrt{n}\frac{\mu_n - \mathbb{E}_{\mathbb{P}}\left[\Psi(U)\right]}{\sigma_n} \xrightarrow{\mathscr{L}} \mathscr{N}(0,1), \quad n \uparrow +\infty. \tag{6.6}$$

Remark 6.2. In the context of Theorem 6.1, specification of the sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ is not necessary, as long as the sequence resides almost surely in Θ_2 and $(\mathscr{F}_k)_{k\in\mathbb{N}}$-adapted. In particular, the strong law of large numbers (6.2) requires only the existence of $V(\boldsymbol{\theta})$ and the almost sure convergence $\boldsymbol{\theta}_k \to \boldsymbol{\theta}^*$. Moreover, the sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ does not have to be convergent, but the condition $\limsup_{n\uparrow+\infty} n^{-1}\sum_{k=1}^{n} V(\boldsymbol{\theta}_k) < +\infty$ is sufficient for (6.2). We however do not go into this direction as the key results are the central limit theorems (6.4) and (6.6), in the context of the Monte Carlo simulation. □

For the rest of this section, we demonstrate some ways for constructing the parameter sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ convergent to $\boldsymbol{\theta}^*$, using the stochastic approximation and the sample average approximation. In the present paper, in order to achieve our demonstration purpose in a concise manner, we intend not to go beyond the most standard versions of the stochastic approximation and the sample average approximation and do not go into too much technicality. We present the performance in the simple problem setting of Section 5.1.1, with the integrand $\Psi(\mathbf{u}) = \mathbb{1}(u_2 \le c_2 - (c_2/c_1)u_1)$ and the exponential–normal bypass distributions, so that the convergence of the parameter sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ towards the minimizer $\boldsymbol{\theta}^*$ can be illustrated effectively in two-dimensional plots.

6.1. Parameter search by stochastic approximation

Here, we examine the sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ constructed recursively by

$$\boldsymbol{\theta}_k = \prod_K \left(\boldsymbol{\theta}_{k-1} - \varepsilon_k \nabla_{\boldsymbol{\theta}} H\left(U_k; \boldsymbol{\theta}_{k-1}\right) \left|\Psi\left(U_k\right)\right|^2\right), \quad k \in \mathbb{N},$$

where K denotes a nonempty compact convex subset of Θ_2 and $\prod_K$ denotes the metric projection onto the set K, that is, $\prod_K(\boldsymbol{\theta}) := \operatorname{argmin}_{\boldsymbol{\theta}'\in K} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, and where $\{\varepsilon_k\}_{k\in\mathbb{N}}$ is a sequence of (deterministic) positive non-increasing constants satisfying $\varepsilon_{k+1} \le \varepsilon_k$, $\sum_{k\in\mathbb{N}} \varepsilon_k = +\infty$ and $\sum_{k\in\mathbb{N}} \varepsilon_k^2 < +\infty$. Clearly, the sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ is adapted to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}}$, whereas the initial point $\boldsymbol{\theta}_0$ is a deterministic vector chosen in Section 3. When the parameter search is conducted by the stochastic approximation algorithm above, it is known [6] that the martingale central limit theorem (6.4) can be made a little more specific with theoretical variance of the adaptive empirical mean at each iteration. We present Theorem 6.3, which is a consequence of the existing results [6,23] adopted in our framework. Note that the order of the upper bound of the difference in variances below in Theorem 6.3 is at most $n^{-3/2}$ by the Cauchy–Schwarz inequality $\left(\sum_{k=1}^n \varepsilon_k\right)^2 \le n\sum_{k=1}^n \varepsilon_k^2$ and the condition $\sum_{k\in\mathbb{N}} \varepsilon_k^2 < +\infty$.

Theorem 6.3. *Let K be a compact and convex subset of Θ_2 with $\mathrm{Leb}(K) > 0$, and define $L_K := \sup_{\boldsymbol{\theta}\in K} \int_{(0,1)^d} \|\nabla_{\boldsymbol{\theta}} H(\mathbf{u}; \boldsymbol{\theta})\|^2 |\Psi(\mathbf{u})|^4 d\mathbf{u}$ and $D_K := \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in K} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2$. If $\boldsymbol{\theta}^* \in K$ and $L_K < +\infty$, then it holds $\mathbb{P}$-a.s. that $\boldsymbol{\theta}_k \to \boldsymbol{\theta}^*$, and that for each $n \in \mathbb{N}$,*

$$0 \le \mathrm{Var}_{\mathbb{P}}\left(\mu_n\right) - \frac{1}{n}\left(V\left(\boldsymbol{\theta}^*\right) - C^2\right) \le \frac{D_K}{2n^2\varepsilon_n} + \frac{L_K}{2n^2}\sum_{k=1}^n \varepsilon_k.$$

Proof. For ease, we use the notation $Q_1(\mathbf{u}; \boldsymbol{\theta}) := M(\mathbf{u}; \boldsymbol{\theta})\Psi(G(G^{-1}(\mathbf{u}; \boldsymbol{\theta}); \boldsymbol{\theta}_0)) - C$, and $Q_2(\mathbf{u}; \boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} H(\mathbf{u}; \boldsymbol{\theta})|\Psi(\mathbf{u})|^2$, so that we have $\mu_n - C = n^{-1}\sum_{k=1}^n Q_1(U_k; \boldsymbol{\theta}_{k-1})$, $\boldsymbol{\theta}_k = \prod_K(\boldsymbol{\theta}_{k-1} - \varepsilon_k Q_2(U_k; \boldsymbol{\theta}_{k-1}))$, and $\mathbb{E}_{\mathbb{P}}[Q_2(U; \boldsymbol{\theta})] = \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$. We omit the proof of the first claim, which is quite well known, and refer the reader to, for instance, [23, Chapter 5]. The second claim is due to [6]. For the sake of completeness and since our framework is slightly more general, we provide a proof in a concise manner. Due to (6.1), it holds that

$$\begin{aligned}
\mathrm{Var}_{\mathbb{P}}\left(\mu_n\right) &= \mathbb{E}_{\mathbb{P}}\left[(\mu_n - C)^2\right] = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}_{\mathbb{P}}\left[Q_1\left(U_k; \boldsymbol{\theta}_{k-1}\right)^2\right] \\
&\quad + \frac{2}{n^2}\sum_{1\le k_1<k_2\le n} \mathbb{E}_{\mathbb{P}}\left[Q_1\left(U_{k_1}; \boldsymbol{\theta}_{k_1-1}\right) Q_1\left(U_{k_2}; \boldsymbol{\theta}_{k_2-1}\right)\right] = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[Q_1\left(U_k; \boldsymbol{\theta}_{k-1}\right)^2 \,\middle|\, \mathscr{F}_{k-1}\right]\right] \\
&\quad + \frac{2}{n^2}\sum_{1\le k_1<k_2\le n} \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[Q_1\left(U_{k_1}; \boldsymbol{\theta}_{k_1-1}\right) Q_1\left(U_{k_2}; \boldsymbol{\theta}_{k_2-1}\right) \,\middle|\, \mathscr{F}_{k_2-1}\right]\right] \\
&= \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}_{\mathbb{P}}\left[V\left(\boldsymbol{\theta}_{k-1}\right) - C^2\right] + \frac{2}{n^2}\sum_{1\le k_1<k_2\le n} \mathbb{E}_{\mathbb{P}}\left[Q_1\left(U_{k_1}; \boldsymbol{\theta}_{k_1-1}\right)\mathbb{E}_{\mathbb{P}}\left[Q_1\left(U_{k_2}; \boldsymbol{\theta}_{k_2-1}\right) \,\middle|\, \mathscr{F}_{k_2-1}\right]\right] \\
&= \frac{1}{n^2}\sum_{k=1}^n \left(\mathbb{E}_{\mathbb{P}}\left[V\left(\boldsymbol{\theta}_{k-1}\right)\right] - C^2\right),
\end{aligned}$$

where we have used the fact that U_{k_1}, $\boldsymbol{\theta}_{k_1-1}$ and $\boldsymbol{\theta}_{k_2-1}$ are $\mathscr{F}_{k_2-1}$-measurable for the last equality. We obtain the lower bound immediately by the definition of the minimizer $\boldsymbol{\theta}^*$. For the upper bound, by noting that the projection $\|\prod_K(\boldsymbol{\theta}_1) - \prod_K(\boldsymbol{\theta}_2)\|^2 \le \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2$ and let U be a uniform random vector on $(0, 1)^d$ independent of $\{U_k\}_{k\in\mathbb{N}}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*\right\|^2 \middle|\, \mathscr{F}_k\right] &= \mathbb{E}_{\mathbb{P}}\left[\left\|\prod_K\left(\boldsymbol{\theta}_k - \varepsilon_{k+1}Q_2\left(U_{k+1}; \boldsymbol{\theta}_k\right)\right) - \prod_K\left(\boldsymbol{\theta}^*\right)\right\|^2 \middle|\, \mathscr{F}_k\right] \\
&\le \mathbb{E}_{\mathbb{P}}\left[\left\|\boldsymbol{\theta}_k - \varepsilon_{k+1}Q_2\left(U; \boldsymbol{\theta}_k\right) - \boldsymbol{\theta}^*\right\|^2 \middle|\, \mathscr{F}_k\right] \\
&= \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|^2 + \varepsilon_{k+1}^2\mathbb{E}_{\mathbb{P}}\left[\left\|Q_2\left(U; \boldsymbol{\theta}_k\right)\right\|^2 \middle|\, \mathscr{F}_k\right] - 2\varepsilon_{k+1}\mathbb{E}_{\mathbb{P}}\left[\left\langle Q_2\left(U; \boldsymbol{\theta}_k\right), \boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\rangle \middle|\, \mathscr{F}_k\right] \\
&= \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|^2 + \varepsilon_{k+1}^2\mathbb{E}_{\mathbb{P}}\left[\left\|Q_2\left(U; \boldsymbol{\theta}_k\right)\right\|^2 \middle|\, \mathscr{F}_k\right] - 2\varepsilon_{k+1}\left\langle \nabla_{\boldsymbol{\theta}} V\left(\boldsymbol{\theta}_k\right), \boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\rangle \\
&\le \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|^2 + \varepsilon_{k+1}^2\mathbb{E}_{\mathbb{P}}\left[\left\|Q_2\left(U; \boldsymbol{\theta}_k\right)\right\|^2 \middle|\, \mathscr{F}_k\right] - 2\varepsilon_{k+1}\left(V\left(\boldsymbol{\theta}_k\right) - V\left(\boldsymbol{\theta}^*\right)\right),
\end{aligned}$$

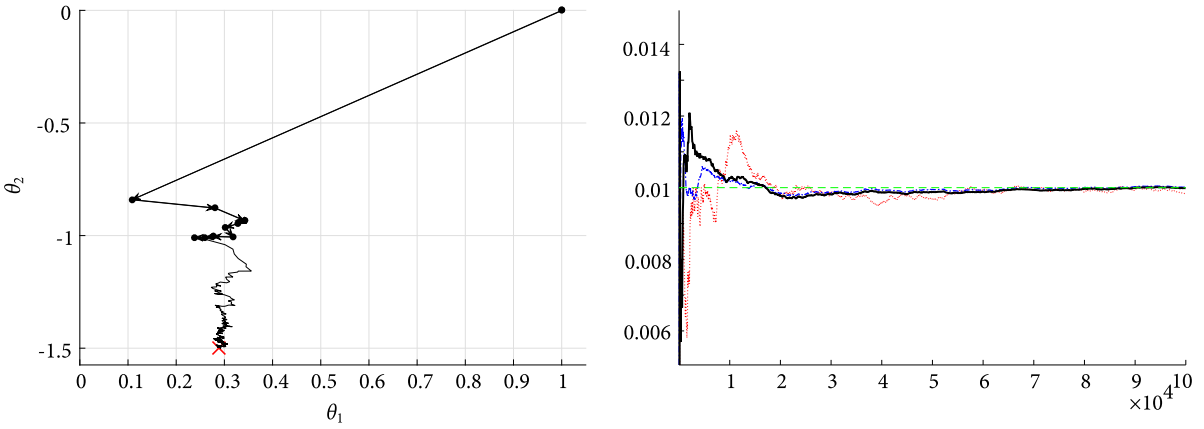


Fig. 7. A typical convergence of the sequence $\{\boldsymbol{\theta}_k\}_{k=1,\dots,10^5}$, towards the minimizer $\boldsymbol{\theta}^*$ (cross). The first filled circle is the deterministic initial point $\boldsymbol{\theta}_0$, whereas the following filled circles indicate the first 10 actual updates.

due to the smoothness and convexity of $V(\boldsymbol{\theta})$ in Θ_2 by Theorem 3.3(ii), that is, $V(\boldsymbol{\theta}^*) \geq V(\boldsymbol{\theta}) + \langle \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}), \boldsymbol{\theta}^* - \boldsymbol{\theta} \rangle$ for $\boldsymbol{\theta} \in \Theta_2$. Note that in the above progression, we have replaced U_{k+1} with U under the condition expectation on $\mathscr{F}_k$, since $\boldsymbol{\theta}_k$ is $\mathscr{F}_k$-measurable, while U_{k+1} is independent of $\mathscr{F}_k$. Taking expectation, rearranging and summing yields

$$2\sum_{k=0}^{n-1}\left(\mathbb{E}_{\mathbb{P}}\left[V\left(\boldsymbol{\theta}_k\right)\right] - V\left(\boldsymbol{\theta}^*\right)\right) \leq \sum_{k=0}^{n-1}\frac{1}{\varepsilon_{k+1}}\left(\mathbb{E}_{\mathbb{P}}\left[\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|^2\right] - \mathbb{E}_{\mathbb{P}}\left[\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*\right\|^2\right]\right) + \sum_{k=0}^{n-1}\varepsilon_{k+1}\mathbb{E}_{\mathbb{P}}\left[\left\|Q_2\left(U;\boldsymbol{\theta}_k\right)\right\|^2\right]$$
$$\leq \frac{1}{\varepsilon_1}\left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\right\|^2 + \sum_{k=1}^{n-1}\left(\frac{1}{\varepsilon_{k+1}} - \frac{1}{\varepsilon_k}\right)\mathbb{E}_{\mathbb{P}}\left[\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|^2\right] + \sum_{k=0}^{n-1}\varepsilon_{k+1}\mathbb{E}_{\mathbb{P}}\left[\left\|Q_2\left(U;\boldsymbol{\theta}_k\right)\right\|^2\right],$$

which yields the upper bound with the aid of $\mathbb{E}_{\mathbb{P}}[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2] \leq D_K$ and $\mathbb{E}_{\mathbb{P}}[\|Q_2(U;\boldsymbol{\theta}_k)\|^2] \leq L_K$ for every $k = 0, \dots, n-1$. □

Example 6.4. We have known from Section 5.1.1 that $\Theta_2 = (0, 2) \times \mathbb{R}$ and $\boldsymbol{\theta}^* \approx (0.288600, -1.499630)^\top$. For the central limit theorem (6.6), observe that $|\Psi(\mathbf{u})|^4 = |\Psi(\mathbf{u})|$, and that for each $k \in \mathbb{N}$,

$$\int_{(0,\varepsilon)} \sup_{\theta \in B_k}\left(\theta^{-1}u^{1-\theta}\right)^2 du < +\infty, \qquad \int_{(0,\varepsilon)} \sup_{\theta \in [-k,+k]}\left(e^{-\theta\Phi^{-1}(u)+\theta^2/2}\right)^2 du < +\infty,$$

where $B_k = [c/(k+1),\, 3/2 - c/(k+1)]$, for some small $c > 0$ such that B_k is not empty. Since the minimizer $\boldsymbol{\theta}^*$ lies in the set, both central limit theorems (6.4) and (6.6) can be applied.

Setting the tuning sequence $\varepsilon_k = 70k^{-3/4}$ and writing $\boldsymbol{\theta}_k = (\theta_{k,1}, \theta_{k,2})^\top$ and $U_k = (U_{k,1}, U_{k,2})^\top$, we recursively compute

$$\boldsymbol{\theta}_k = \prod_K\left(\boldsymbol{\theta}_{k-1} - \frac{70}{k^{3/4}}\left|\Psi\left(U_k\right)\right|^2 \exp\left[-\ln\left(\theta_{k-1,1}\right) + \left(1-\theta_{k-1,1}\right)\ln\left(U_{k,1}\right) + \frac{1}{2}\theta_{k-1,2}^2 - \theta_{k-1,2}\Phi^{-1}\left(U_{k,2}\right)\right]\right.$$
$$\left.\times \begin{bmatrix} -1/\theta_{k-1,1} - \ln(U_{k,1}) \\ \theta_{k-1,2} - \Phi^{-1}(U_{k,2}) \end{bmatrix}\right),$$

where we have set $K = [10^{-3}, 1] \times [-2, 0]$, using the prior knowledge of the minimizer $\boldsymbol{\theta}^*$. Note that this algorithm is known to be very sensitive to a choice of the tuning sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$. Those issues are of great practical importance but outside the scope of the present paper, so we do not go into those issues. In Fig. 7, we plot a typical convergence of the sequence $\{\boldsymbol{\theta}_k\}_{k=1,\dots,n}$ for $n = 1 \times 10^5$ iterations, towards the minimizer $\boldsymbol{\theta}^*$, indicated by the cross mark. In this particular experiment, we observe overshooting from the compact subdomain K only 7 times. Out of 1×10^5 iterations, the algorithm has made *actual* updating (that is, $\boldsymbol{\theta}_k \neq \boldsymbol{\theta}_{k-1}$) only 1065 times. The very first filled circle indicates the initial point $\boldsymbol{\theta}_0 = (1, 0)^\top$, whereas the following ones are the first 10 of such actual updating, which took place at $k = 765, 768, 770, 994, 1005$, etc.

The right figure plots a typical convergence of the resulting adaptive Monte Carlo simulation, on top of Fig. 1. Until the first actual updating (at the 765th iteration), the adaptive version remains identical to the crude version. Although the adaptive version is less stable than optimal version (with the minimum variance from the beginning), it shows a faster convergence than the crude version. This is so because an adaptive Monte Carlo simulation with improving importance sampling parameters $\boldsymbol{\theta}_k$ (even before reaching the minimizer $\boldsymbol{\theta}^*$) is much more efficient than running the crude version with $\boldsymbol{\theta}_0$. □

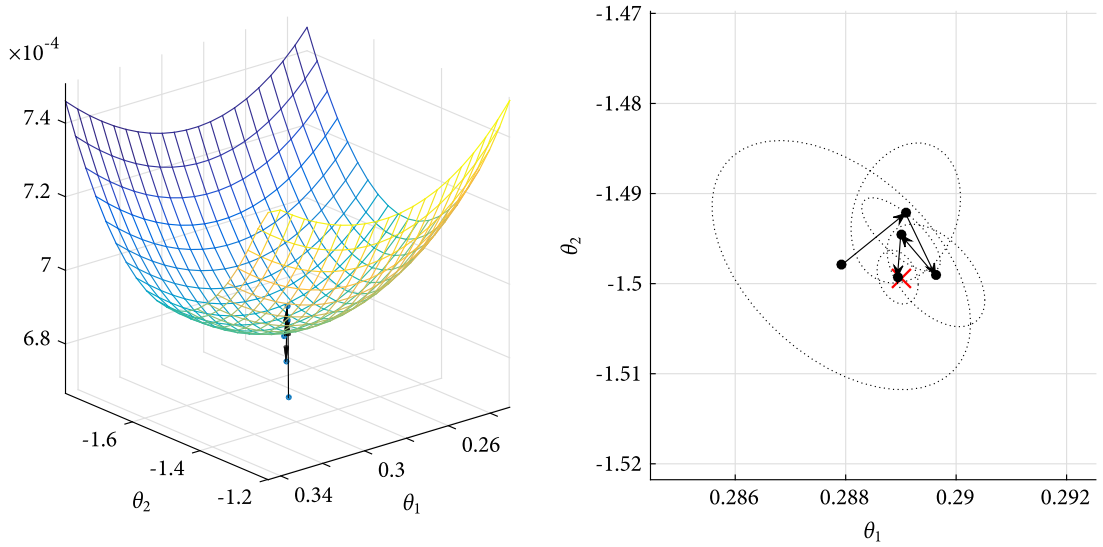


Fig. 8. The left figure plots typical results of $\boldsymbol{\theta}_n$ and the corresponding estimates of the reduced estimator variance at $n = 5 \times 10^4$, 1×10^5, 3×10^5, 5×10^5, and 1×10^6. The right figure illustrates a typical convergence of $\boldsymbol{\theta}_n$ towards the minimizer $\boldsymbol{\theta}^*$ (cross) along with a one-sigma contour ellipse.

6.2. Parameter search by sample average approximation

We next construct the sequence $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ by the sample average approximation. For $\boldsymbol{\theta} \in \Theta_1$ and $n \in \mathbb{N}$, define

$$V_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{k=1}^{n} H(U_k;\boldsymbol{\theta})\,|\Psi(U_k)|^2,$$

which is an unbiased estimator of $V(\boldsymbol{\theta})$, that is, $V(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}[V_n(\boldsymbol{\theta})]$ for every $\boldsymbol{\theta} \in \Theta_1$ and $n \in \mathbb{N}$, and define

$$\boldsymbol{\theta}_n := \operatorname*{argmin}_{\boldsymbol{\theta}\in K} V_n(\boldsymbol{\theta}), \tag{6.7}$$

where K denotes a nonempty bounded compact subset of the domain Θ_2, as in Section 6.1. Clearly, the random vector $\boldsymbol{\theta}_n$ is measurable with respect to the σ-field $\sigma(U_k : k = 1, \dots, n)$, whereas the initial point $\boldsymbol{\theta}_0$ is again a deterministic vector chosen in Section 3. We further define the square matrix

$$\Sigma := \left[\mathrm{Hess}_{\boldsymbol{\theta}}\left(V(\boldsymbol{\theta}^*)\right)\right]^{-1} \mathbb{E}_{\mathbb{P}}\left[\left(\nabla_{\boldsymbol{\theta}} H(U;\boldsymbol{\theta}^*)\right)^{\otimes 2} |\Psi(U)|^4\right]\left[\mathrm{Hess}_{\boldsymbol{\theta}}\left(V(\boldsymbol{\theta}^*)\right)\right]^{-1} \in \mathbb{R}^{d\times d}, \tag{6.8}$$

where the expectation in the middle is the variance–covariance matrix of the random vector $\nabla_{\boldsymbol{\theta}} H(U;\boldsymbol{\theta}^*)|\Psi(U)|^2$, since its first moment vanishes due to the first-order necessary optimality condition. It then holds $\mathbb{P}$-*a.s.* that $\boldsymbol{\theta}_n \to \boldsymbol{\theta}^*$ and $V_n(\boldsymbol{\theta}_n) \to V(\boldsymbol{\theta}^*)$, as $n \uparrow +\infty$. Also, if the matrix Σ exists, then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \Sigma)$, as $n \uparrow +\infty$. (We refer the reader to, for instance, [23, Chapter 5] for details.)

Example 6.5. We continue the problem setting of Example 6.4. In this experiment, we examine the performance at some fixed n's on the compact subdomain $K = [10^{-3}, 1] \times [-2, 0]$, as before. Fig. 8 plots typical numerical results of $\boldsymbol{\theta}_n$ and the corresponding estimates of the reduced estimator variance, at $n = 5\times 10^4, 1\times 10^5, 3\times 10^5, 5\times 10^5$, and 1×10^6. The sample average approximation (6.7) cannot be considered to be an algorithm in the sense that it needs to be solved by numerical approximation. We use the MATLAB fmincon for the operation (6.7). We feed empirical gradient and Hessian based on the expressions (3.10) and (3.11) along with (5.3) and (5.4).

Although $\boldsymbol{\theta}_n$ tends to the minimizer $\boldsymbol{\theta}^*$ as n increases, the estimation quality of the estimator variance is very poor (way below the true variance in this particular experiment). We conjecture that the reason is that the majority of iterations gave $|\Psi(U_k)|^2 = 0$, which makes no good contribution to the estimation. This issue is of great practical importance and will be investigated in a subsequent paper [18].

Next, for sufficiently large n, we expect $\boldsymbol{\theta}_n \overset{\mathscr{L}}{\approx} \mathcal{N}(\boldsymbol{\theta}^*, n^{-1}\Sigma)$. In the right figure of Fig. 8, we replot the progression of $\boldsymbol{\theta}_n$ on a plane (that is, a 2D version of the 3D plot in the left figure), along with a one-sigma contour ellipse $\boldsymbol{\theta}_n + \sqrt{\Sigma/n}(\cos(\xi), \sin(\xi))^\top$, $\xi \in [0, 2\pi)$, for illustration purposes. We estimate the variance–covariance matrix $\Sigma(\boldsymbol{\theta}_n)$ using the expressions (3.11), (5.1), (5.3), (5.4) and (6.8). □

7. Concluding remarks

We have constructed adaptive parametrized importance sampling variance reduction methods for general multivariate probability laws on the basis of the principle of bypass distribution, without relying no particular properties of the target and proposal distributions. We establish the asymptotic normality of the estimator of the mean and of the importance sampling parameter when running Monte Carlo estimation and optimal parameter search concurrently. Throughout the paper, we have provided numerical results to illustrate the applicability and effectiveness of the principle of bypass distribution as well as the proposed simulation algorithm with a mixture of exponential and normal bypass distribution.

An important future research direction is the trade-off issue between a reduction in estimator variance and its required additional computing time, to which we did not pay much attention in the present study. In particular, we have observed that the convergence of importance sampling parameters could be extremely slow, unfortunately together with poor estimation quality. An acceleration of the parameter search phase is certainly beneficial and would deserve a separate research.

The proposed adaptive Monte Carlo framework can perhaps be considered as one of the most general parametric forms, relative to the existing adaptive Monte Carlo methods under rather restrictive distributional assumptions. In particular, considering all random elements involved is the standard uniform law, the most interesting field of application may be, for instance, natural and computer sciences, rather than the social science in which the existing adaptive methods are often effective due to particular assumptions on the underlying distribution such as multivariate Gaussian and gamma distributions.

Acknowledgments

The author would like to thank the anonymous reviewers for their careful reading and valuable suggestions to improve the quality of the paper.

References

[1] B. Arouna, Adaptative Monte Carlo method, a variance reduction technique, Monte Carlo Methods Appl. 10 (1) (2004) 1–24.
[2] B. Jourdain, J. Lelong, Robust adaptive importance sampling for normal random vectors, Ann. Appl. Probab. 19 (5) (2009) 1687–1718.
[3] B. Lapeyre, J. Lelong, A framework for adaptive Monte Carlo procedures, Monte Carlo Methods Appl. 17 (1) (2011) 77–98.
[4] V. Lemaire, G. Pagès, Unconstrained recursive importance sampling, Ann. Appl. Probab. 20 (3) (2010) 1029–1067.
[5] M.-S. Oh, J.O. Berger, Adaptive importance sampling in Monte Carlo integration, J. Stat. Comput. Simul. 41 (3–4) (1992) 143–168.
[6] E.K. Ryu, S.P. Boyd, Adaptive importance sampling via stochastic convex programming. ArXiv e-prints, Dec. 2014.
[7] G.H. Givens, A.E. Raftery, Local adaptive importance sampling for multivariate densities with strong nonlinear relationships, J. Amer. Statist. Assoc. 91 (433) (1996) 132–141.
[8] G.P. Lepage, A new algorithm for adaptive multidimensional integration, J. Comput. Phys. 27 (2) (1978) 192–203.
[9] T. Ohl, Vegas revisited: Adaptive Monte Carlo integration beyond factorization, Comput. Phys. Comm. 120 (1) (1999) 13–19.
[10] A. Owen, Y. Zhou, Adaptive importance sampling by mixtures of products of beta distributions. Technical report 1999-25, Stanford University Department of Statistics, 1999.
[11] P. Zhang, Nonparametric importance sampling, J. Amer. Statist. Assoc. 91 (435) (1996) 1245–1253.
[12] M.H. Alrefaei, H.M. Abdul-Rahman, An adaptive Monte Carlo integration algorithm with general division approach, Math. Comput. Simulation 79 (1) (2008) 49–59.
[13] R. Kawai, Adaptive Monte Carlo variance reduction with two-time-scale stochastic approximation, Monte Carlo Methods Appl. 13 (3) (2007) 197–217.
[14] R. Kawai, Adaptive Monte Carlo variance reduction for Lévy processes with two-time-scale stochastic approximation, Methodol. Comput. Appl. Probab. 10 (2) (2008) 199–223.
[15] R. Kawai, Asymptotically optimal allocation of stratified sampling with adaptive variance reduction by strata, ACM Trans. Model. Comput. Simul. 20 (2) (2010) 9:1–9:17.
[16] S. Kim, S.G. Henderson, Adaptive control variates for finite-horizon simulation, Math. Oper. Res. 32 (3) (2007) 508–527.
[17] T. Pennanen, M. Koivu, An adaptive importance sampling technique, in: H. Niederreiter, D. Talay (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2004, Springer, Berlin, Heidelberg, 2006, pp. 443–455.
[18] R. Kawai, Acceleration of adaptive importance sampling with sample average approximation, preprint.
[19] R. Kawai, A. Takeuchi, Sensitivity analysis for averaged asset price dynamics with gamma processes, Statist. Probab. Lett. 80 (1) (2010) 42–49.
[20] R. Kawai, A. Takeuchi, Greeks formulas for an asset price model with gamma processes, Math. Finance 21 (4) (2011) 723–742.
[21] R. Kawai, A. Takeuchi, Computation of Greeks for asset price dynamics driven by stable and tempered stable processes, Quant. Finance 13 (8) (2013) 1303–1316.
[22] K.S. Lomax, Business failures: Another example of the analysis of failure data, J. Amer. Statist. Assoc. 49 (268) (1954) 847–852.
[23] A. Shapiro, D. Dentcheva, A. Ruszczyński, Lectures on Stochastic Programming, Society for Industrial and Applied Mathematics, 2009.