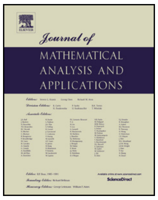
J. Math. Anal. Appl. 483 (2020) 123608

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa


Journal of
MATHEMATICAL
ANALYSIS AND
APPLICATIONS

Adaptive importance sampling and control variates[☆]

Reiichiro Kawai[1]

ARTICLE INFO

Article history:
Received 1 August 2019
Available online 21 October 2019
Submitted by A. Jentzen

Keywords:
Importance sampling
Control variates
Stochastic approximation
Sample average approximation
Variance reduction

ABSTRACT

We construct and investigate an adaptive variance reduction framework in which both importance sampling and control variates are employed. The three lines (Monte Carlo averaging and two variance reduction parameter search lines) run in parallel on a common sequence of uniform random vectors on the unit hypercube. Given that these two variance reduction techniques are effective often in a complementary way, their combined application is well expected to widen the applicability of adaptive variance reduction. We derive convergence rates of the theoretical estimator variance towards its minimum as a fixed computing budget increases, when stochastic approximation runs with optimal constant learning rates. We derive sufficient conditions for the proposed algorithm to attain the minimal estimator variance in the limit, by stochastic approximation with decreasing learning rates or by sample average approximation, when computing budget is unlimitedly available. Numerical results support our theoretical findings and illustrate the effectiveness of the proposed framework.

1. Introduction

The adaptive Monte Carlo variance reduction method aims to avoid the need for frequent recalibration of the parameters of the variance reduction techniques due to small changes in the experimental conditions governing system performance, by concurrently running the primary Monte Carlo averaging and the associated parameter search lines for the variance reduction techniques. The concept of adaptive Monte Carlo variance reduction methods and their practical use have been investigated for a long time [1,5,7,8,12,13], to mention just a few. More recently, the adaptive importance sampling framework is generalized [10], where both target and proposal laws are the uniform law on the unit hypercube. The parameter search line is further parametrized through changes of measures so as to accelerate the parameter search line, performed with either the sample average approximation [9] or the stochastic approximation [11].

[☆] The author would like to thank the anonymous referee and the editor for valuable comments and suggestions that helped improve and clarify the paper.

E-mail address: reiichiro.kawai@sydney.edu.au.

[1] Postal address: School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia.

https://doi.org/10.1016/j.jmaa.2019.123608
0022-247X/© 2019 Elsevier Inc. All rights reserved.

The aim of the present paper is twofold. We first improve the existing adaptive importance sampling framework [9–11] by incorporating control variates into the framework at little additional computing cost. We then provide a convergence analysis of the combined adaptive framework of importance sampling and control variates when stochastic approximation runs with constant learning rates and a finite computing budget, through much more involved derivations than the preceding work [11], due to the presence of control variates. We also derive sufficient conditions for the combined framework to converge to the minimal estimator variance by stochastic approximation with decreasing learning rates or sample average approximation, when computing budget can be progressively increased.

The control variates method is, on the one hand, effective in a variety of forms [4,15,16] particularly when its variates have a high correlation, positive or negative, with the estimator, which is when the importance sampling technique is often not very effective. On the other hand, importance sampling is effective, for instance, when the estimator returns zero-valued realizations with a very high probability, which is when the control variates method is of almost no use due to a nearly zero correlation between the variates and the estimator. Experientially speaking, it is rather rare that the estimator variance cannot be reduced considerably via any one of those variance reduction techniques. In this sense, the combined application of importance sampling and control variates techniques is naturally expected to be effective at least in a complementary way, which helps improve the applicability of the proposed adaptive Monte Carlo variance reduction method to a large extent.

The ultimate goal in improving the theory and implementation of Monte Carlo methods is an increase, with minimal additional computing effort, in the precision of the evaluation of the integral

$$\mu := \int_{(0,1)^d} \Psi(\mathbf{u})\, d\mathbf{u} = \mathbb{E}[\Psi(U)], \tag{1.1}$$

where Ψ is a function mapping from the unit hypercube $(0,1)^d$ to $\mathbb{R}$, and where U is a uniform random vector on $(0,1)^d$. 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 reformulated with the standard uniform random vector on the unit hypercube in the same dimension with a suitable change of variables or the principle of inverse transform sampling.

We summarize the direction and objectives of the present work in brief without completely defining some notation, so that the numerical results (Section 5) are fairly accessible without going through technical details on the problem formulation (Section 2) and the algorithm (Section 3). First, by applying the concept of bypass distribution $G(\mathbf{z};\boldsymbol{\theta})$ with density $g(\mathbf{z};\boldsymbol{\theta})$ for importance sampling [10] and then incorporating control variates parameterized by $\boldsymbol{\xi}$, we obtain the following expression (Section 2):

$$\mu = \int_{(0,1)^d} \Psi(\mathbf{u})\, d\mathbf{u} = \int_{(0,1)^d} \left[\frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta})}\Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)) + \langle \boldsymbol{\xi}, \mathbf{u} - \mathbb{1}_d/2\rangle\right] d\mathbf{u}, \tag{1.2}$$

where we may, yet do not, adopt the other way around (that is, control variates first and then importance sampling) for efficient computation (Section B.1). The estimator variance of (1.2) can be decomposed into three components as follows:

$$\int_{(0,1)^d} \left[\frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta})}\Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)) + \langle \boldsymbol{\xi}, \mathbf{u} - \mathbb{1}_d/2\rangle - \mu\right]^2 d\mathbf{u} = V(\boldsymbol{\theta}) + W(\boldsymbol{\theta},\boldsymbol{\xi}) - \mu^2, \tag{1.3}$$

where the integrals (1.2) and (1.3) are taken with respect to the Lebesgue measure on the unit hypercube, irrespective of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. We remark that the existing importance sampling framework [9–11] can be recovered in full by suppressing control variates ($\boldsymbol{\xi} = 0_d$) in (1.2) and (1.3). The first term $V(\boldsymbol{\theta})$ represents the estimator second moment when importance sampling $\boldsymbol{\theta}$ is applied alone, and the second term $W(\boldsymbol{\theta}, \boldsymbol{\xi})$ represents a possible further reduction of the estimator variance when both importance sampling $\boldsymbol{\theta}$ and control variates $\boldsymbol{\xi}$ are jointly applied. Due to the convexity of $V(\boldsymbol{\theta})$ and the lack of such strict convex structure in the sum $V(\boldsymbol{\theta}) + W(\boldsymbol{\theta}, \boldsymbol{\xi})$, it would be most natural to first find the point $\boldsymbol{\theta}^*$ to minimize the term $V(\boldsymbol{\theta})$, without interference from the parameter $\boldsymbol{\xi}$, and then look for the point $\boldsymbol{\xi}^*$ to minimize the second term $W(\boldsymbol{\theta}^*, \boldsymbol{\xi})$, rather than jointly searching two parameters $(\boldsymbol{\theta}, \boldsymbol{\xi})$.

We construct and investigate algorithms to estimate the value μ of the integral (1.1) with a smaller estimator variance $V(\boldsymbol{\theta}) + W(\boldsymbol{\theta}, \boldsymbol{\xi}) - \mu^2$ on the basis of the integral representation (1.3) by adaptively searching the parameters $(\boldsymbol{\theta}, \boldsymbol{\xi})$. First, we built the following algorithm by successively averaged stochastic approximation (Section 3.1):

$$\begin{cases} \mu(k) = k^{-1}[\sum_{l=1}^{k\wedge\tau_n} R(U_l; \boldsymbol{\theta}_0) + \sum_{l=\tau_n+1}^{k} R(U_l; \overline{\boldsymbol{\theta}}_{\tau_n}^{l-1}, \overline{\boldsymbol{\xi}}_{\tau_n}^{l-1})], \\ \boldsymbol{\theta}_k = \prod_{\mathscr{T}_k}[\boldsymbol{\theta}_{k-1} - \gamma_{k-1}\nabla_{\boldsymbol{\theta}} N(U_k; \boldsymbol{\theta}_{k-1}, \boldsymbol{\lambda}_{(k-1)\wedge\tau_n})], \\ \overline{\boldsymbol{\theta}}_{\tau_n}^{k} = \sum_{l=\tau_n}^{k} \frac{\gamma_l}{\sum_{t=\tau_n}^{k}\gamma_t}\boldsymbol{\theta}_l, \\ \boldsymbol{\xi}_k = \prod_{\mathscr{X}_k}[\boldsymbol{\xi}_{k-1} - \epsilon_{k-1}\nabla_{\boldsymbol{\xi}} S(U_k; \boldsymbol{\theta}_{k-1}, \boldsymbol{\xi}_{k-1}, \boldsymbol{\lambda}_{(k-1)\wedge\tau_n})], \\ \overline{\boldsymbol{\xi}}_{\tau_n}^{k} = \sum_{l=\tau_n}^{k} \frac{\epsilon_l}{\sum_{t=\tau_n}^{k}\epsilon_t}\boldsymbol{\xi}_l, \end{cases} \tag{1.4}$$

where $\prod_D[\mathbf{x}]$ denotes the metric projection of $\mathbf{x}$ onto the compact set D. (We will define the functions R, N and S in Section 2, and the domains $\{\mathscr{T}_k\}_{k\in\mathbb{N}_0}$ and $\{\mathscr{X}_k\}_{k\in\mathbb{N}_0}$, the parameter $\boldsymbol{\lambda}$ and the stopping time τ_n in Section 3.) We remark that only one common sequence $\{U_k\}_{k\in\mathbb{N}}$ of uniform random vectors is required throughout the implementation of the algorithm (1.4), enabling one to run all the five lines of (1.4) concurrently, rather than sequentially. In this line of research [1,5,7,8,12–14], such a parallelized run is said to be *adaptive*.

The practical difficulty when running such stochastic approximation algorithms as (1.4) is the extreme performance sensitivity to the choice of decreasing learning rates $\{\gamma_k\}_{k\in\mathbb{N}_0}$ and $\{\epsilon_k\}_{k\in\mathbb{N}_0}$, that is, the usual $\ell^2 \setminus \ell^1$-condition on the decreasing learning rates is far too loose to guide how to choose. As in the preceding work [11], we place our main focus on the situation where the learning rates are constant $(\gamma_k, \epsilon_k) \equiv (\gamma, \epsilon)$ when only a finite computing budget is available. In particular, we derive constant learning rates via minimization of (an upper bound for) the theoretical variance of the empirical mean $\mathrm{Var}(\mu(n))$ (Section 4.1). We also discuss the convergence of the algorithm (1.4) with decreasing learning rates under the $\ell^2 \setminus \ell^1$-condition to attain the minimal estimator variance when computing budget is unlimitedly available (Section 4.2).

As a possible alternative to the algorithm (1.4) by stochastic approximation, we construct (Section 3.2) and investigate (Section 4.2) the following algorithm by sample average approximation:

$$\begin{cases} \mu(k) = k^{-1}\sum_{l=1}^{k} R(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}), \\ \boldsymbol{\theta}_k = \mathrm{argmin}_{\boldsymbol{\theta}\in\Theta_2} k^{-1}\sum_{j=1}^{k} N(U_j; \boldsymbol{\theta}, \boldsymbol{\lambda}_{(k-1)\wedge\tau_n}), \\ \boldsymbol{\xi}_k = -12\frac{1}{k}\sum_{j=1}^{k} R(U_j; \boldsymbol{\theta}_k)(U_j - \mathbb{1}_d/2). \end{cases} \tag{1.5}$$

In general, sample average approximation (1.5) provides more robust performance than stochastic approximation (1.4) in return for heavier computing cost as well as the requirement of an external optimization tool. In the literature, the convergence of adaptive importance sampling scheme (without control variates $\boldsymbol{\xi}_k \equiv 0_d$), in one formulation or another, has been investigated in [5,9,10]. Note that no optimization procedure is required for the control variates parameter $\boldsymbol{\xi}_k$ here, which is an unbiased estimate for the unique optimum based on a explicit formula (Section 2.5), unlike the argmin required for the parameter $\boldsymbol{\theta}_k$. Hence,

Table 1
Existing work and the scope of the present work in adaptive variance reduction by stochastic approximation (SA) and sample average approximation (SAA).

	Finite budget	Infinite budget
SA	[11]	[1,10,13]
SAA	×	[5,9,10]

(a) Existing work in adaptive importance sampling.

	Finite budget	Infinite budget
SA	Theorem 4.1	Theorem 4.2
SAA	×	Theorem 4.3

(b) Combined importance sampling and control variates.

in comparison to the additional computing effort that we had to pay when applying importance sampling $\boldsymbol{\theta}$ by sample average approximation, this further addition of control variates $\boldsymbol{\xi}$ in (1.5) costs almost none.

Although the addition of control variates in the algorithms (1.4) and (1.5) does not cost serious additional computing effort, this addition turns nontrivial from a theoretical point of view. That is, the convergence analysis (Section 4) requires surprisingly different lines of proofs with more careful treatments, due to the lack of joint convexity of the estimator second moment $V(\boldsymbol{\theta}) + W(\boldsymbol{\theta}, \boldsymbol{\xi})$. In order to support our theoretical findings and illustrate the effectiveness of the proposed framework and algorithms, we present numerical results on a high-dimensional example (Section 5), which is the one examined in the preceding work [11] so that a direct comparison can be made in terms of the addition of control variates. To sum up, we summarize the relevant existing work in the literature (Table 1 (a)) and the contribution of the present work (Table 1 (b)).

2. Problem formulation

We begin with general notation which will be used throughout. We use the notation $\mathbb{N} := \{1, 2, \cdots\}$ and $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$, and denote by $|\cdot|$ and $\|\cdot\|$, respectively, the magnitude and the Euclidean norm. We denote by $\mathrm{Leb}(D)$, $\mathrm{int}(D)$, ∂D, $\overline{D}$, $\mathscr{B}(D)$ and $\mathrm{diam}(D)$, respectively, the Lebesgue volume, the interior, the boundary, the closure, the Borel σ-field and the diameter of a set D. All essential supremums and infimums are taken with respect to the Lebesgue measure. For a matrix A, we denote by $A^\top$ the transpose of the matrix A and write $A^{\otimes 2} := AA^\top$. 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 q-th partial derivative with respect to the univariate variable x. Moreover, $\nabla_{\mathbf{x}}$ and $\mathrm{Hess}_{\mathbf{x}}$ denote the gradient and the Hessian matrix with respect to the multivariate variable $\mathbf{x}$. We denote by $\mathbb{I}_d$, $\mathbb{1}_d$ and 0_d, respectively, the identity matrix of size d, the vector in $\mathbb{R}^d$ with all unit-valued components, and the zero vector in $\mathbb{R}^d$. We reserve ϕ, Φ and Φ^{-1} for, respectively, the standard normal density function, the standard normal cumulative distribution function and its inverse.

We define the filtration $(\mathscr{F}_k^u)_{k\in\mathbb{N}}$ generated by the sequence $\{U_k\}_{k\in\mathbb{N}}$ of iid uniform random vectors on $(0,1)^d$, that is, for each $n \in \mathbb{N}$, $\mathscr{F}_n^u = \sigma(\{U_k\}_{k\in\{1,\cdots,n\}})$ is the σ-field generated by the first n iid uniform random vectors. We then construct the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$ starting from zero by augmentation with the collection of $\mathbb{P}$-null sets, that is, $\mathscr{F}_0 := \sigma(\mathscr{N})$ and $\mathscr{F}_n := \mathscr{F}_n^u \bigvee \mathscr{F}_0$ for $n \in \mathbb{N}$, where $\mathscr{N} := \{N \subseteq \Omega : \exists\, A \in \bigvee_{k\in\mathbb{N}} \mathscr{F}_k^u,\ \mathbb{P}(A) = 0,\ N \subseteq A\}$. In practice, the σ-field $\mathscr{F}_0$ can be interpreted as the information available at the beginning of the experiment. In our framework, there will be no need to specify under what probability measure the expectation $\mathbb{E}$ is taken, since we end up taking expectations under a single probability measure $\mathbb{P}$ all the time, although we do change probability measures back and forth in the middle of derivations. Therefore, with the σ-field $\mathscr{F} := \bigvee_{k\in\mathbb{N}_0} \mathscr{F}_k$, we fix $(\Omega, \mathscr{F}, (\mathscr{F}_k)_{k\in\mathbb{N}_0}, \mathbb{P})$ as our underlying filtered probability space throughout. We denote by $\mathbb{P}_k(\cdot)$ the probability measure restricted to the σ-field $\mathscr{F}_k$, and by $\mathbb{E}_k[\cdot]$ and $\mathrm{Var}_k(\cdot)$, respectively, conditional expectation and conditional variance given the σ-field $\mathscr{F}_k$. The preceding expectation (1.1) may be considered conditional on the σ-field $\mathscr{F}_0$, that is, unconditional in effect.

We keep the integrand $\Psi(\mathbf{u})$ in the integral (1.1) general without special structure, as the primary interest of this study does not lie in a specialized problem class, such as the rare event simulation. Since we are concerned with variance reduction, it loses no essential generality to impose the existence of a finite second moment $\int_{(0,1)^d}|\Psi(\mathbf{u})|^2 d\mathbf{u} = \mathbb{E}[|\Psi(U)|^2] < +\infty$ as well as non-degeneracy $\mathbb{P}(|\Psi(U)| \neq 0) > 0$.

2.1. Bypass transform

For the rest of this section, we briefly review the problem formulation, adopting from the preceding work [9–11], to which we refer the reader for details omitted in some instances. First, we define the family of probability distributions, which we call 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 $\mathscr{D}$ of the probability density function $g(\cdot;\boldsymbol{\theta})$ is open and independent of the parameter $\boldsymbol{\theta}$;
(b) For almost every $\mathbf{z} \in \mathscr{D}$ (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 $B \in \mathscr{B}((0,1)^d)$, it holds that $\int_{\mathscr{D}} \mathbb{1}(G(\mathbf{z};\boldsymbol{\theta}) \in B) g(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} = \mathrm{Leb}(B)$;
(d) For each $\boldsymbol{\theta} \in \Theta_0$, the inverse $G^{-1}(\mathbf{u};\boldsymbol{\theta})$ (with respect to $\mathbf{u}$) is continuous in $\mathbf{u}$ on $(0,1)^d$;
(e) For each $\boldsymbol{\theta} \in \Theta_0$ and $B \in \mathscr{B}(\mathscr{D})$, 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 $\mathbf{z} \in \mathscr{D}$ (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 $\cup_{n\in\mathbb{N}} K_n = \Theta_0$ and $K_n \subsetneq \mathrm{int}(K_{n+1})$;
(g) The function $G(\mathbf{z};\boldsymbol{\theta})$ is Lipschitz continuous in $\boldsymbol{\theta}$, that is, there exists $c \geq 0$ such that $\mathrm{esssup}_{\mathbf{z}\in\mathscr{D}}|G(\mathbf{z};\boldsymbol{\theta}_a) - G(\mathbf{z};\boldsymbol{\theta}_b)| \leq c\|\boldsymbol{\theta}_a - \boldsymbol{\theta}_b\|$ for every $(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b) \in \Theta_0^2$.

Assumptions 2.1 **(c)**, **(d)** and **(e)** indicate that if Z is a random vector in $\mathscr{D}(\subseteq \mathbb{R}^d)$ with density $g(\mathbf{z};\boldsymbol{\theta})$ and $U \sim U(0,1)^d$, then it holds that $G(Z;\boldsymbol{\theta}) \stackrel{\mathscr{L}}{=} U$, and $Z \stackrel{\mathscr{L}}{=} G^{-1}(U;\boldsymbol{\theta})$. Assumption 2.1 **(f)** will serve as a technical condition for convexity of the estimator variance shortly in Proposition 2.2. Assumption 2.1 **(g)** is another technical condition, which was not imposed in the preceding work [9–11], but is imposed here to ensure that when the control variates method is applied, the estimator variance tends to its desired minimum (Theorems 4.1 and 4.2). Indeed, Assumption 2.1 **(g)** needs to be imposed or can be removed, depending on whether control variates is applied (2.4) or suppressed (2.3). To ease the presentation, however, we have included this condition as a standing assumption within Assumption 2.1, mainly for the reason that this assumption does not seem to be too restrictive. For instance, it is satisfied by the exponential and Gaussian bypass distributions [10, Section 4].

2.2. Combined importance sampling and control variates

Fix a point $\boldsymbol{\theta}_0 \in \Theta_0$ and pick another point $\boldsymbol{\theta} \in \Theta_0$, which can be distinct from $\boldsymbol{\theta}_0$. Under Assumption 2.1, the integral (1.1) of interest can be rewritten as follows:

$$\mu = \int\limits_{(0,1)^d} \Psi(\mathbf{u})\, d\mathbf{u} = \int\limits_{(0,1)^d} \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta})} \Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)) d\mathbf{u}, \tag{2.1}$$

where we have applied change of variables $\mathbf{u} = G(\mathbf{z};\boldsymbol{\theta}_0)$ first, then $\mathbf{z} = G^{-1}(\mathbf{u};\boldsymbol{\theta})$ and the assumption that the support $\mathscr{D}$ is independent of the parameter $\boldsymbol{\theta}$. The integral representation (2.1) is indeed the base framework of the preceding work [9–11]. The bypass transform [10] enables one to introduce the parameter $\boldsymbol{\theta}$ without affecting the underlying Lebesgue measure $d\mathbf{u}$, which is nothing but the uniform law if the integral is interpreted as a mathematical expectation. We next introduce control variates inside the integral (2.1) as follows:

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

$$= \int_{(0,1)^d} \left[\frac{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta})} \Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)) + \langle \boldsymbol{\xi}, \mathbf{u} - \mathbb{1}_d/2 \rangle \right] d\mathbf{u}, \tag{2.2}$$

where the equality holds true irrespective of the parameter $\boldsymbol{\xi} \in \mathbb{R}^d$, since $\int_{(0,1)^d} \langle \boldsymbol{\xi}, \mathbf{u} - \mathbb{1}_d/2 \rangle d\mathbf{u} = \langle \boldsymbol{\xi}, \int_{(0,1)^d} (\mathbf{u} - \mathbb{1}_d/2) d\mathbf{u} \rangle = 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$. Clearly, if $\boldsymbol{\xi} = 0_d$, then the second line (2.2) reduces to the first line (2.1). The control variates here (that is, $\mathbf{u} - \mathbb{1}_d/2$) is linear in its current form (2.2), whereas, in practice, the control variates needs to be transformed nonlinear in $\mathbf{z}$ (Section 5), for instance, $\Phi_d(\mathbf{z}) - \mathbb{1}_d/2$ when transformed to the Gaussian law. If moreover $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then all three integrals above are back identical. If $(\boldsymbol{\theta}, \boldsymbol{\xi}) \neq (\boldsymbol{\theta}_0, 0_d)$, however, three integrands may not be identical in law with respect to the uniform law $d\mathbf{u}$, without changing the value $\int_{(0,1)^d} \Psi(\mathbf{u}) d\mathbf{u}$. Hence, by wisely choosing the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, we may achieve a smaller variance with the expression on the right-hand side. (We discuss some alternatives to the formulation (2.2), such as control variates first and nonlinear variates, in Section B.1.) Hereafter, for brevity, we use the notation: for $\mathbf{u} \in (0,1)^d$ and $(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\lambda}) \in \Theta_0 \times \mathbb{R}^d \times \Theta_0$,

$$H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) := \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta})}, \quad R(\mathbf{u};\boldsymbol{\theta}) := H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}) \Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)), \tag{2.3}$$

and

$$R(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi}) := R(\mathbf{u};\boldsymbol{\theta}) + \langle \boldsymbol{\xi}, \mathbf{u} - \mathbb{1}_d/2 \rangle . \tag{2.4}$$

The argument $\boldsymbol{\lambda}$ in (2.3) will be introduced as the auxiliary parameter shortly through (2.10) and (2.12), whereas it remains fixed at $\boldsymbol{\lambda} = \boldsymbol{\theta}_0$ until then. The estimator variance of the right-hand side in (2.2) is defined as the L^2-distance of the integrand from the integral value μ with respect to the uniform distribution $d\mathbf{u}$, that is,

$$\int_{(0,1)^d} (R(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi}) - \mu)^2 d\mathbf{u} = V(\boldsymbol{\theta}) + W(\boldsymbol{\theta},\boldsymbol{\xi}) - \mu^2, \tag{2.5}$$

where

$$V(\boldsymbol{\theta}) := \int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0) |\Psi(\mathbf{u})|^2 d\mathbf{u}, \tag{2.6}$$

and

$$W(\boldsymbol{\theta},\boldsymbol{\xi}) := 2 \left\langle \boldsymbol{\xi}, \int_{(0,1)^d} \Psi(\mathbf{u}) \left(G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}) - \mathbb{1}_d/2 \right) d\mathbf{u} \right\rangle + \frac{1}{12} \|\boldsymbol{\xi}\|^2, \tag{2.7}$$

provided that the integrals exist. The progressions in both (2.6) and (2.7) follow from a change of variables $\mathbf{u} = G(\mathbf{z};\boldsymbol{\theta})$ first, then $\mathbf{z} = G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)$.

2.3. Estimator variance without control variates

In order to further investigate the estimator variance indexed by the parameter $\boldsymbol{\theta}$, we restrict our attention to the following domains:

$$\Theta_1 := \mathrm{int}\left\{\boldsymbol{\theta} \in \Theta_0 : \int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)|\Psi(\mathbf{u})|^2 d\mathbf{u} < +\infty\right\}, \tag{2.8}$$

and

$$\Theta_2 := \mathrm{int}\bigcup_{B}\left\{B \subseteq \Theta_1 : \int_{(0,1)^d} \sup_{\boldsymbol{\theta}\in\overline{B}} \max\left\{1, \left\|\frac{\nabla_{\boldsymbol{\theta}} H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)}{H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)}\right\|, \left\|\frac{\mathrm{Hess}_{\boldsymbol{\theta}}(H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0))}{H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)}\right\|\right\} \right.$$
$$\times H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)|\Psi(\mathbf{u})|^2 d\mathbf{u} < +\infty,$$
$$\left. \text{and for almost every } \mathbf{z} \in \mathscr{D},\ (g(\mathbf{z};\boldsymbol{\theta}))^{-1} \text{ is strictly convex in } \boldsymbol{\theta} \text{ on } \overline{B}\right\}. \tag{2.9}$$

In view of the expression (2.3), the convexity condition (2.9) ensures the convexity of $H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})$ 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. The convexity condition (2.9) is only concerned with the bypass distribution $g(\mathbf{z};\boldsymbol{\theta})$ and is thus verifiable irrespective of the integrand $\Psi(\mathbf{u})$. With those domains in mind, the regularity and convexity of the second moment function $V(\boldsymbol{\theta})$ is given as follows.

Proposition 2.2. **(i)** *If* $\mathrm{Leb}(\Theta_2) > 0$, *then it holds that* $V(\boldsymbol{\theta})$ *is twice continuously differentiable and strictly convex on* Θ_2, *with*

$$\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} H(U;\boldsymbol{\theta},\boldsymbol{\theta}_0)|\Psi(U)|^2\right], \quad \mathrm{Hess}_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})) = \mathbb{E}\left[\mathrm{Hess}_{\boldsymbol{\theta}}\left(H(U;\boldsymbol{\theta},\boldsymbol{\theta}_0)\right)|\Psi(U)|^2\right].$$

(ii) *If moreover* $\Theta_1 = \Theta_2$, *then* $\boldsymbol{\theta}^* := \mathrm{argmin}_{\boldsymbol{\theta}\in\Theta_2} V(\boldsymbol{\theta})$ *exists uniquely in the domain* Θ_2 *satisfying* $\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}^*) = 0$.

We remark that for implementation purposes, the twice (continuous) differentiability requirement in Proposition 2.2 **(i)** is redundant. Moreover, unless trying to attain the minimum value $V(\boldsymbol{\theta}^*)$ and/or find a minimizer $\boldsymbol{\theta}^*$, the existence and uniqueness result of Proposition 2.2 **(ii)** is not really necessary.

2.4. Auxiliary parameter

We next review the concept of the auxiliary parameter [9] into the second moment. In brief, when searching the optimal importance sampling parameter $\boldsymbol{\theta}^*$ on the basis of Proposition 2.2 by stochastic approximation (Section 3.1) or sample average approximation (Section 3.2), we approximate the gradient of the second moment $\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$ or the second moment $V(\boldsymbol{\theta})$ itself by sampling their respective integrands $\nabla_{\boldsymbol{\theta}} H(U;\boldsymbol{\theta},\boldsymbol{\theta}_0)|\Psi(U)|^2$ or $H(U;\boldsymbol{\theta},\boldsymbol{\theta}_0)|\Psi(U)|^2$. This approximation is however fatally inefficient when the term $|\Psi(U)|^2$ is zero-valued with a very high probability [9, Section 2.2]. Hence, we inject another parameter (which we call the auxiliary parameter) into the function $|\Psi(\cdot)|^2$ inside the second moment $V(\boldsymbol{\theta})$ so as to avoid zero-valued realizations.

We introduce the auxiliary parameter $\boldsymbol{\lambda}$ into the second moment function $V(\boldsymbol{\theta})$ through change of probability measure as follows: for each $(\boldsymbol{\theta},\boldsymbol{\lambda}) \in \Theta_1^2$,

$$V(\boldsymbol{\theta}) = \int_{(0,1)^d} \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta})} \frac{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0)}{g(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\lambda})} |\Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0))|^2 d\mathbf{u} = \int_{(0,1)^d} N(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) d\mathbf{u}, \tag{2.10}$$

where

$$N(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) := H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})\, H(\mathbf{u};\boldsymbol{\lambda},\boldsymbol{\lambda})\, |\Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0))|^2. \tag{2.11}$$

Above, we have applied change of variables $\mathbf{z} = G^{-1}(\mathbf{u};\boldsymbol{\theta}_0)$ and $\mathbf{u} = G(\mathbf{z};\boldsymbol{\lambda})$, each of which requires no additional integrability condition, since then the support $\mathscr{D}$ is independent of the parameter, as imposed in Assumption 2.1 **(a)**. In a similar manner, we incorporate the same auxiliary parameter $\boldsymbol{\lambda}$ into the control variates component $W(\boldsymbol{\theta},\boldsymbol{\xi})$ as follows: for each $(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda}) \in \Theta_1 \times \mathbb{R}^d \times \Theta_1$,

$$\begin{aligned} W(\boldsymbol{\theta},\boldsymbol{\xi}) &= 2\left\langle \boldsymbol{\xi}, \int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\lambda},\boldsymbol{\lambda})\Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0))\left(G(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}) - \mathbb{1}_d/2\right) d\mathbf{u} \right\rangle + \frac{1}{12}\|\boldsymbol{\xi}\|^2 \\ &= \int_{(0,1)^d} S(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda}) d\mathbf{u}, \end{aligned} \tag{2.12}$$

where

$$\begin{aligned} Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) &:= H(\mathbf{u};\boldsymbol{\lambda},\boldsymbol{\lambda})\Psi(G(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}_0))\left(G(G^{-1}(\mathbf{u};\boldsymbol{\lambda});\boldsymbol{\theta}) - \mathbb{1}_d/2\right), \\ S(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda}) &:= 2\left\langle \boldsymbol{\xi}, Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})\right\rangle + \frac{1}{12}\|\boldsymbol{\xi}\|^2. \end{aligned} \tag{2.13}$$

The auxiliary parameter $\boldsymbol{\lambda}$ inside the integrals acts as importance sampling in the estimator of the second moment functions $V(\boldsymbol{\theta})$ and $W(\boldsymbol{\theta},\boldsymbol{\xi})$. We refer the reader to [9, Section 4] for more details, such as how and when to set the auxiliary parameter effectively.

2.5. Estimator variance with importance sampling and control variates

In the expression (2.5), the estimator variance is a sum of the two terms $V(\boldsymbol{\theta})$ and $W(\boldsymbol{\theta},\boldsymbol{\xi})$ (minus the unknown constant μ^2), whereas Proposition 2.2 only addresses the smoothness of the first term $V(\boldsymbol{\theta})$ with respect to the parameter $\boldsymbol{\theta}$. We here turn to the second term $W(\boldsymbol{\theta},\boldsymbol{\xi})$. First of all, the term $W(\boldsymbol{\theta},\boldsymbol{\xi})$ is clearly finite valued as soon as the parameter $\boldsymbol{\theta}$ stays in the domain Θ_1, since the additional term $(\mathbf{u} - \mathbb{1}_d/2)$ in the integrand is bounded. Its smoothness with respect to $\boldsymbol{\xi}$ requires no additional conditions, due to its quadratic structure:

$$\nabla_{\boldsymbol{\xi}} W(\boldsymbol{\theta},\boldsymbol{\xi}) = 2\int_{(0,1)^d} Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) d\mathbf{u} + \frac{1}{6}\boldsymbol{\xi} = \int_{(0,1)^d} \nabla_{\boldsymbol{\xi}} S(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda}) d\mathbf{u}. \tag{2.14}$$

Therefore, it follows readily that for each $\boldsymbol{\theta} \in \Theta_2$,

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^d} W(\boldsymbol{\theta},\boldsymbol{\xi}) = 12\left\| \int_{(0,1)^d} R(\mathbf{u};\boldsymbol{\theta})(\mathbf{u} - \mathbb{1}_d/2) d\mathbf{u} \right\|^2 = -12\left\| \int_{(0,1)^d} Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}) d\mathbf{u} \right\|^2,$$

where the minimum is attained uniquely at

$$\operatorname*{argmin}_{\boldsymbol{\xi}\in\mathbb{R}^d} W(\boldsymbol{\theta},\boldsymbol{\xi}) = -12\int_{(0,1)^d} R(\mathbf{u};\boldsymbol{\theta})(\mathbf{u}-\mathbb{1}_d/2)d\mathbf{u} = -12\int_{(0,1)^d} Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})d\mathbf{u}. \tag{2.15}$$

Accordingly, define the optimal point $\boldsymbol{\xi}^*$, given the optimal importance sampling parameter $\boldsymbol{\theta}^*$, by

$$\boldsymbol{\xi}^* := \operatorname*{argmin}_{\boldsymbol{\xi}\in\mathbb{R}^d} W(\boldsymbol{\theta}^*,\boldsymbol{\xi}) = -12\int_{(0,1)^d} R(\mathbf{u};\boldsymbol{\theta}^*)(\mathbf{u}-\mathbb{1}_d/2)d\mathbf{u} = -12\int_{(0,1)^d} Q(\mathbf{u};\boldsymbol{\theta}^*,\boldsymbol{\lambda})d\mathbf{u}, \tag{2.16}$$

and thus the corresponding optimal value is given in the very simple form:

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^d} W(\boldsymbol{\theta}^*,\boldsymbol{\xi}) = W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*) = -\frac{1}{12}\|\boldsymbol{\xi}^*\|^2.$$

As should now be clear from the structure (2.16) as well as that the gradient (2.14), our algorithm will be constructed to first search the point $\boldsymbol{\theta}^*$ to minimize the term $V(\boldsymbol{\theta})$, without interference from the control variates parameter $\boldsymbol{\xi}$, and then look for the point $\boldsymbol{\xi}^*$ to minimize the second term $W(\boldsymbol{\theta}^*,\boldsymbol{\xi})$, rather than jointly searching two parameters $(\boldsymbol{\theta},\boldsymbol{\xi})$ so as to minimize the sum $V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})$ altogether. In fact, our intended point $(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$ could be sub-optimal, that is, there may exist distinct points, say, $(\boldsymbol{\theta}^\diamond,\boldsymbol{\xi}^\diamond)$ within the search domain, which coincides with or even outperforms our target, that is, $V(\boldsymbol{\theta}^\diamond)+W(\boldsymbol{\theta}^\diamond,\boldsymbol{\xi}^\diamond) \le V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$. Let us point out a few, both technical and practical, reasons for not pursuing this joint optimality. First of all, the convexity of $V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})$ does not hold true in general and indeed does not in practice. (We will provide simple yet illustrative examples in Appendix B.3.) Hence, there is no strong theoretical backup to pursue a global optimality. From a computational point of view, the gradient with respect to the importance sampling parameter $\boldsymbol{\theta}$ involves two terms $\nabla_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})) = \nabla_{\boldsymbol{\theta}}V(\boldsymbol{\theta})+\nabla_{\boldsymbol{\theta}}W(\boldsymbol{\theta},\boldsymbol{\xi})$, where the second term costs extra for sure. Since the ultimate interest lies in the estimation of the expectation $\mathbb{E}[\Psi(U)]$, not in the variance reduction parameter search for minimizers $\boldsymbol{\theta}^*$ and $\boldsymbol{\xi}^*$, the relative priority on the parameter search should be kept lower.

3. Algorithms

We are now in a position to construct our algorithms, which are meant to be easy-to-implement and of all-purpose type. To this end, we prepare the notation for relevant key elements.

Let $\{\mathscr{T}_k\}_{k\in\mathbb{N}_0}$ and $\{\mathscr{X}_k\}_{k\in\mathbb{N}_0}$ be $\mathscr{F}_0$-measurable sequence of non-expanding compact convex subsets, respectively, of Θ_2 and $\mathbb{R}^d$, that is, $\mathscr{T}_{k+1}\subseteq\mathscr{T}_k\subset\Theta_2$ and $\mathscr{X}_{k+1}\subseteq\mathscr{X}_k\subset\mathbb{R}^d$ for all $k\in\mathbb{N}_0$. The compact sets $\mathscr{T}_k$ and $\mathscr{X}_k$ indicate the domains, respectively, where $\boldsymbol{\theta}_k$ and $\boldsymbol{\xi}_k$ are allowed to reside in. We assume that each is large enough to contain the corresponding minimizer: $\boldsymbol{\theta}^*\in\cap_{k\in\mathbb{N}_0}\mathscr{T}_k$ and $\boldsymbol{\xi}^*\in\cap_{k\in\mathbb{N}_0}\mathscr{X}_k$. The $\mathscr{F}_0$-measurable randomness of the search domains $\{\mathscr{T}_k\}_{k\in\mathbb{N}_0}$ and $\{\mathscr{X}_k\}_{k\in\mathbb{N}_0}$, as well as the initial point $\boldsymbol{\theta}_0$ correspond to the usual practice that the prior knowledge $\mathscr{F}_0$ is employed for algorithm design.

Hereafter, we reserve the notation $n\in\mathbb{N}$ for the available computing budget, that is, the maximum total number of iterations. Let τ be an $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$-stopping time taking non-negative integer values and define its truncation $\tau_n := \tau\wedge n$ by the given computing budget n. We define the σ-field $\mathscr{F}_{\tau_n}$ at the $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$-stopping time τ_n, that is, $\mathscr{F}_{\tau_n} := \{B\in\mathscr{F} : B\cap\{\tau_n\in\{0,1,\cdots,k\}\}\in\mathscr{F}_k \text{ for all } k\in\{0,1,\cdots,n\}\}$. Clearly, the tail $\{U_k\}_{k\in\{\tau_n+1,\cdots\}}$ of the sequence is independent of the σ-field $\mathscr{F}_{\tau_n}$, while the truncated sequence $\{U_k\}_{k\in\{1,\cdots,\tau_n\}}$ is $\mathscr{F}_{\tau_n}$-measurable. The stopping time τ_n here represents the time point until which one is allowed to conduct the pilot run to decide on the relevant problem parameters, such as the learning rates $\{\gamma_k\}_{k\in\mathbb{N}_0}$ and $\{\epsilon_k\}_{k\in\mathbb{N}_0}$ as well as the auxiliary parameter $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$. In particular, the pilot run may be prohibited by suppressing the stopping time ($\tau_n = 0$). Let $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ be an $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$-adapted sequence of random vectors, corresponding to the auxiliary parameter, in a compact subset Λ_0 of the domain Θ_1, satisfying $\boldsymbol{\theta}_0\in\operatorname{int}(\Lambda_0)$, with $\boldsymbol{\lambda}^\star := \boldsymbol{\lambda}_{\tau_n}$, corresponding to the argument $\boldsymbol{\lambda}$ in the expressions (2.10) and

(2.11). In what follows, we will call $\boldsymbol{\lambda}^\star$ the long-run auxiliary parameter. The superscript "$\star$" is given to the long-run auxiliary parameter $\boldsymbol{\lambda}^\star$ here, instead of "$*$", to emphasize the difference that the optimal important sampling parameter $\boldsymbol{\theta}^*$ is deterministic, while the long-run auxiliary parameter $\boldsymbol{\lambda}^\star$ is generally random.

3.1. Stochastic approximation

The algorithm we propose consists of the following five concurrent lines **(A)-(E)**, along with possibly one more line (not to be specified here) for the auxiliary parameter $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$. Recall that for a compact set B, we define $\prod_B[\mathbf{x}] := \mathrm{argmin}_{\mathbf{z}\in B}\|\mathbf{x}-\mathbf{z}\|$ the metric projection of $\mathbf{x}$ onto the compact set B.

(A) Adaptive empirical mean throughout:

$$\mu(k) := \frac{1}{k}\left[\sum_{l=1}^{k\wedge\tau_n} R(U_l;\boldsymbol{\theta}_0) + \sum_{l=\tau_n+1}^{k} R(U_l;\overline{\boldsymbol{\theta}}_{\tau_n}^{l-1},\overline{\boldsymbol{\xi}}_{\tau_n}^{l-1})\right], \quad k\in\{1,\cdots,n\}. \tag{3.1}$$

(B) Importance sampling parameter search throughout by stochastic approximation:

$$\boldsymbol{\theta}_k := \prod\nolimits_{\mathscr{T}_k}\left[\boldsymbol{\theta}_{k-1} - \gamma_{k-1}\nabla_{\boldsymbol{\theta}} N(U_k;\boldsymbol{\theta}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})\right], \quad k\in\{1,\cdots,n-1\}. \tag{3.2}$$

(C) Successive averaging of the parameter of the line **(B)**, starting from the stopping time τ_n:

$$\overline{\boldsymbol{\theta}}_{\tau_n}^{k} := \sum_{l=\tau_n}^{k}\frac{\gamma_l}{\sum_{t=\tau_n}^{k}\gamma_t}\boldsymbol{\theta}_l, \quad k\in\{\tau_n,\cdots,n-1\}.$$

(D) Control variates parameter search throughout by stochastic approximation:

$$\boldsymbol{\xi}_k := \prod\nolimits_{\mathscr{X}_k}\left[\boldsymbol{\xi}_{k-1} - \epsilon_{k-1}\nabla_{\boldsymbol{\xi}} S(U_k;\boldsymbol{\theta}_{k-1},\boldsymbol{\xi}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})\right], \quad k\in\{1,\cdots,n-1\}, \quad \boldsymbol{\xi}_0 := 0_d. \tag{3.3}$$

(E) Successive averaging of the parameter of the line **(D)**, starting from the stopping time τ_n:

$$\overline{\boldsymbol{\xi}}_{\tau_n}^{k} := \sum_{l=\tau_n}^{k}\frac{\epsilon_l}{\sum_{t=\tau_n}^{k}\epsilon_t}\boldsymbol{\xi}_l, \quad k\in\{\tau_n,\cdots,n-1\}. \tag{3.4}$$

For the sake of clarity, we summarize the algorithm in Table 2; random number generation, the integrand, the five concurrent lines, and the auxiliary parameter, all on the event $\{\tau_n = m\}$. The iterations until the stopping time τ_n provide some time for a pilot run to collect relevant knowledge for setting unspecified quantities, such as the learning rate (Section 4.1) and the long-run auxiliary parameter. In order to secure a reasonable amount of such knowledge, it is rather necessary to set the stopping time τ_n wisely, since, as discussed in Section 2.4, the recursions (3.2) and (3.3) may not make an actual update for a long time if the gradients there tend to be zero too often.

3.2. Sample average approximation

The second approach, not quite an algorithm on its own, is the so-called sample average approximation. Let $\{\ell_a(k)\}_{k\in\mathbb{N}}$ and $\{\ell_b(k)\}_{k\in\mathbb{N}}$ be sequences of $\mathscr{F}_0$-measurable random variables taking values in $\{0,1\}$, representing the update timings of respective parameters. We impose $\sum_{k\in\mathbb{N}}\ell_a(k) = +\infty$ and $\sum_{k\in\mathbb{N}}\ell_b(k) = +\infty$ to ensure that the parameters will be updated infinitely often. In light of the formulas (2.15) and (2.16), starting with $\widetilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$ and $\widetilde{\boldsymbol{\xi}}_0 = 0_d$, we iterate for $k\in\mathbb{N}$:

Table 2
Random number generation, integrand, **(A)**, **(B)**, **(C)**, the auxiliary parameter, **(D)** and **(E)**, on the event $\{\tau_n = m\}$. The symbol "×" indicates "not defined", while the symbol, say, "$\times(\boldsymbol{\lambda}^\star)$" indicates that "we apply the parameter $\boldsymbol{\lambda}^\star$, which is available without computing".

iteration (k)	0	1	$\cdots$	$m-1$	$\{\tau_n = m\}$	$m+1$	$\cdots$	$n-1$	n
$U(0,1)^d$	$\times$	U_1		U_{m-1}	U_m	U_{m+1}		U_{n-1}	U_n
integrand		$R(U_1;\boldsymbol{\theta}_0)$	$\cdots$	$R(U_{m-1};\boldsymbol{\theta}_0)$	$R(U_m;\boldsymbol{\theta}_0)$	$R(U_{m+1};\overline{\boldsymbol{\theta}}^m_m,\overline{\boldsymbol{\xi}}^m_m)$		$R(U_{n-1};\overline{\boldsymbol{\theta}}^{n-2}_m,\overline{\boldsymbol{\xi}}^{n-2}_m)$	$R(U_n;\overline{\boldsymbol{\theta}}^{n-1}_m,\overline{\boldsymbol{\xi}}^{n-1}_m)$
(A)		$\mu(1)$	$\cdots$	$\mu(m-1)$	$\mu(m)$	$\mu(m+1)$	$\cdots$	$\mu(n-1)$	$\mu(n)$
(B)	$\times(\boldsymbol{\theta}_0 \in \mathscr{F}_0)$	$\boldsymbol{\theta}_1$		$\boldsymbol{\theta}_{m-1}$	$\boldsymbol{\theta}_m$	$\boldsymbol{\theta}_{m+1}$		$\boldsymbol{\theta}_{n-1}$	
(C)	$\times$	$\times$		$\times$	$\times\ (\overline{\boldsymbol{\theta}}^m_m = \boldsymbol{\theta}_m)$	$\overline{\boldsymbol{\theta}}^{m+1}_m$		$\overline{\boldsymbol{\theta}}^{n-1}_m$	
auxiliary	$\boldsymbol{\lambda}_0$	$\boldsymbol{\lambda}_1$	$\cdots$	$\boldsymbol{\lambda}_{m-1}$	$\boldsymbol{\lambda}^\star(=\boldsymbol{\lambda}_{\tau_n})$	$\times(\boldsymbol{\lambda}^\star)$	$\cdots$	$\times(\boldsymbol{\lambda}^\star)$	$\times$
(D)	$\times\ (\boldsymbol{\xi}_0 = 0_d)$	$\boldsymbol{\xi}_1$		$\boldsymbol{\xi}_{m-1}$	$\boldsymbol{\xi}_m$	$\boldsymbol{\xi}_{m+1}$		$\boldsymbol{\xi}_{n-1}$	
(E)	$\times$	$\times$	$\cdots$	$\times$	$\times\ (\overline{\boldsymbol{\xi}}^m_m = \boldsymbol{\xi}_m)$	$\overline{\boldsymbol{\xi}}^{m+1}_m$	$\cdots$	$\overline{\boldsymbol{\xi}}^{n-1}_m$	$\times$

(F) Adaptive empirical mean throughout:

$$\mu(k) := \frac{1}{k}\sum_{l=1}^{k} R(U_l;\widetilde{\boldsymbol{\theta}}_{l-1},\widetilde{\boldsymbol{\xi}}_{l-1}). \tag{3.5}$$

(G) Importance sampling parameter search via sample average approximation:

$$\widetilde{\boldsymbol{\theta}}_k \leftarrow \begin{cases} \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta_2} \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} N(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n}), & \text{if } \ell_a(k) = 1, \\ \widetilde{\boldsymbol{\theta}}_{k-1}, & \text{if } \ell_a(k) = 0. \end{cases}$$

(H) Control variates parameter search via sample average approximation:

$$\widetilde{\boldsymbol{\xi}}_k \leftarrow \begin{cases} -12\dfrac{1}{k}\displaystyle\sum_{j=1}^{k} Q(U_j;\widetilde{\boldsymbol{\theta}}_k,\boldsymbol{\lambda}_{(k-1)\wedge\tau_n}), & \text{if } \ell_b(k) = 1, \\ \widetilde{\boldsymbol{\xi}}_{k-1}, & \text{if } \ell_b(k) = 0. \end{cases} \tag{3.6}$$

The line **(H)** can be implemented with elementary operations only, without external optimization tools, unlike the line **(G)**. Given that the implementation of the line **(G)** has been thought of as a serious bottleneck of the adaptive importance sampling framework by sample average approximation [9], the further addition of the line **(H)** can safely be considered computationally almost free of charge in comparison.

3.3. Theoretical variance of empirical mean

Recall that in the algorithm **(A)**-**(E)** of Section 3.1, there exist two phases transiting from one to the other at the stopping time τ_n (unless $\tau_n = 0$). In particular, as can be seen in the "integrand" row of Table 2, the algorithm applies no variance reduction methods until this stopping time. As will be seen shortly in Section 4.1, this stopping time is incorporated into the algorithm so as to make use of the $\mathscr{F}_{\tau_n}$-measurable information to enhance the implementation of the remaining iterations. From a theoretical point of view, this stopping time can be made implicit in the expression (3.1), by re-defining the parameters as follows:

$$\widetilde{\boldsymbol{\theta}}_k := \begin{cases} \boldsymbol{\theta}_0, & \text{if } k \in \{0,1,\cdots,\tau_n - 1\}, \\ \overline{\boldsymbol{\theta}}^k_{\tau_n}, & \text{if } k \in \{\tau_n,\tau_n+1,\cdots,n-1\}, \end{cases} \qquad \widetilde{\boldsymbol{\xi}}_k := \begin{cases} 0_d, & \text{if } k \in \{0,1,\cdots,\tau_n - 1\}, \\ \overline{\boldsymbol{\xi}}^k_{\tau_n}, & \text{if } k \in \{\tau_n,\tau_n+1,\cdots,n-1\}, \end{cases} \tag{3.7}$$

both of which clearly remain adapted to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$. Then, for the both algorithms by stochastic approximation (Section 3.1) and by sample average approximation (Section 3.2), the empirical means (3.1) and (3.5) at the computing budget n can be expressed in the following unified way, as well as we define the corresponding empirical variance $\sigma^2(n)$ by

$$\begin{aligned}\mu(n) &= \frac{1}{n}\sum_{k=1}^{n} R(U_k;\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}),\\ \sigma^2(n) &= \frac{1}{n}\sum_{k=1}^{n}\left[N(U_k;\widetilde{\boldsymbol{\theta}}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n}) + S(U_k;\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})\right] - \mu^2(n).\end{aligned} \tag{3.8}$$

As long as the random sequences $\{U_k\}_{k\in\mathbb{N}}$, $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\in\mathbb{N}_0}$ and $\{\widetilde{\boldsymbol{\xi}}_k\}_{k\in\mathbb{N}_0}$, as well as $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ are adapted to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$, the empirical mean and variance (3.8) satisfy the following ideal properties [11, Proposition 2.5].

Proposition 3.1. *Let $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\in\mathbb{N}_0}$, $\{\widetilde{\boldsymbol{\xi}}_k\}_{k\in\mathbb{N}_0}$ and $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ be $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$-adapted sequences of random vectors defined in either Section 3.1 with (3.7) or Section 3.2. It then holds $\mathbb{P}_0$-a.s. that for each $n\in\mathbb{N}$,*

$$\begin{aligned}&\mathbb{E}_0[\mu(n)] = \mu, \quad n\mathrm{Var}_0(\mu(n)) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_0\left[V(\widetilde{\boldsymbol{\theta}}_{k-1}) + W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1})\right] - \mu^2,\\ &\mathbb{E}_0\left[\sigma^2(n)\right] = (n-1)\mathrm{Var}_0(\mu(n)).\end{aligned}$$

4. Convergence analysis

We conduct convergence analysis of the algorithms developed in Sections 3.1 and 3.2. We prepare some notation for presentation of the results. First, we define a sequence $\{\alpha_k\}_{k\in\mathbb{N}_0}$ of non-negative random variables, clearly representing the strong convexity parameter, by

$$\alpha_k := \operatorname*{argmax}_{\alpha\geq 0}\left\{V(\boldsymbol{\theta}^*) \geq V(\boldsymbol{\theta}) + \langle\nabla_{\boldsymbol{\theta}}V(\boldsymbol{\theta}),\boldsymbol{\theta}^*-\boldsymbol{\theta}\rangle + \frac{\alpha}{2}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\right\|^2, \text{ for all } \boldsymbol{\theta}\in\mathscr{T}_k\right\}, \tag{4.1}$$

which is guaranteed to exist since $\mathscr{T}_k \subseteq \mathscr{T}_0 \subseteq \Theta_2$ for all $k\in\mathbb{N}_0$. Note that although the second moment $V(\boldsymbol{\theta})$ is deterministic, the convexity parameters α_k may contain randomness because the search domains $\mathscr{T}_k$ may be random, yet $\mathscr{F}_0$-measurable. Since the search domains are non-increasing with successive dominance $\mathscr{T}_{k+1}\subseteq\mathscr{T}_k$, the convexity parameters α_k are non-decreasing.

4.1. Finite computing budget: stochastic approximation with constant learning rates

Here, we focus on the algorithm by stochastic approximation (Section 3.1) when the computing budget n is fixed and cannot be increased progressively afterwards. For each computing budget $n\in\mathbb{N}$ and strictly positive constant learning rates γ and ϵ, define

$$\Upsilon(n;\gamma,\epsilon) := \Upsilon_a(n;\gamma) + \Upsilon_b(n;\epsilon) + \Upsilon_c(n) + \Upsilon_d(n), \tag{4.2}$$

where the four terms are defined by

$$\begin{aligned}\Upsilon_a(n;\gamma) := &\frac{1}{2n\gamma}\mathbb{E}_0\left[\mathrm{diam}^2(\mathscr{T}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right] - \frac{1}{2n}\mathbb{E}_0\left[\alpha_{\tau_n}\mathrm{diam}^2(\mathscr{T}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right]\\ &+\frac{\gamma}{2n}\mathbb{E}_0\left[L^2(\mathscr{T}_{\tau_n};\boldsymbol{\lambda}^\star)(n-\tau_n)\right],\end{aligned}$$

$$\begin{aligned}\Upsilon_b(n;\epsilon) &:= \frac{1}{2n\epsilon}\mathbb{E}_0\left[\operatorname{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right] - \frac{1}{12n}\mathbb{E}_0\left[\operatorname{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right]\\ &\quad + \frac{\epsilon}{2n}\mathbb{E}_0\left[J^2(\mathscr{T}_{\tau_n},\mathscr{X}_{\tau_n};\boldsymbol{\lambda}^\star)(n-\tau_n)\right],\\ \Upsilon_c(n) &:= 2\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k}\left\langle \boldsymbol{\xi}^*-\boldsymbol{\xi}_{l-1}, W_0(\boldsymbol{\theta}_{l-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right],\\ \Upsilon_d(n) &:= 2\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\left\langle \overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}, W_0(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right],\end{aligned}$$

where $W_0(\boldsymbol{\theta}) := \int_{(0,1)^d} R(\mathbf{u};\boldsymbol{\theta})(\mathbf{u}-\mathbb{1}_d/2)d\mathbf{u}(= \int_{(0,1)^d} Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})d\mathbf{u})$, which represents the optimal parameter $\boldsymbol{\xi}$ (up to a negative constant), based on (2.15) and (2.16). Here, the quantity $\Upsilon(n;\gamma,\epsilon)$ represents an upper bound for the excess of the (scaled) theoretical variance $n\operatorname{Var}_0(\mu(n))$ from the variance at the destination $(\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*)$ when the computing budget n is fixed and the learning rates are set constant $(\gamma_k,\epsilon_k)\equiv(\gamma,\epsilon)$. With computing budget n fixed, we find the constant learning rates (γ,ϵ) that minimizes this upper bound $\Upsilon(n;\gamma,\epsilon)$. Note that the numerical computation of the upper bound $\Upsilon(n;\gamma,\epsilon)$ is not required for implementation.

We are now in a position to present the theoretical background of the algorithm by stochastic approximation with a finite computing budget and constant learning rates, which is deemed the most appropriate in practice. The constant learning rates we employ are given as follows: for each fixed computing budget n,

$$\gamma(n) := \left(\frac{\mathbb{E}_0[\operatorname{diam}^2(\mathscr{T}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}]}{\mathbb{E}_0[L^2(\mathscr{T}_{\tau_n};\boldsymbol{\lambda}^\star)(n-\tau_n)]}\right)^{1/2}, \quad \epsilon(n) := \left(\frac{\mathbb{E}_0[\operatorname{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}]}{\mathbb{E}_0[J^2(\mathscr{T}_{\tau_n},\mathscr{X}_{\tau_n};\boldsymbol{\lambda}^\star)(n-\tau_n)]}\right)^{1/2}, \tag{4.3}$$

where the functionals $L(\cdot;\boldsymbol{\lambda})$ and $J(\cdot,\cdot;\boldsymbol{\lambda})$ are defined on compact subsets $\mathscr{T}$ and $\mathscr{X}$, respectively, of Θ_2 and $\mathbb{R}^d$ by

$$L^2(\mathscr{T};\boldsymbol{\lambda}) := \sup_{\boldsymbol{\theta}\in\mathscr{T}}\int_{(0,1)^d}\left\|\nabla_{\boldsymbol{\theta}}N(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda})\right\|^2 d\mathbf{u}, \tag{4.4}$$

$$J^2(\mathscr{T},\mathscr{X};\boldsymbol{\lambda}) := \sup_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathscr{T}\times\mathscr{X}}\int_{(0,1)^d}\left\|\nabla_{\boldsymbol{\xi}}S(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda})\right\|^2 d\mathbf{u}, \tag{4.5}$$

for $\boldsymbol{\lambda}\in\Lambda_0$.

Theorem 4.1. *Let $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\in\mathbb{N}_0}$, $\{\widetilde{\boldsymbol{\xi}}_k\}_{k\in\mathbb{N}_0}$ and $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ be $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$-adapted sequences of random vectors defined in Section 3 with (3.7).*

(i) *With arbitrary constant learning rates $\gamma_k\equiv\gamma(>0)$ and $\epsilon_k\equiv\epsilon(>0)$, it holds that for each $n\in\mathbb{N}$,*

$$n\operatorname{Var}_0(\mu(n)) \le \left(V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)-\mu^2\right)+\left(V(\boldsymbol{\theta}_0)-(V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*))\right)\frac{\mathbb{E}_0[\tau_n]}{n}+\Upsilon(n;\gamma,\epsilon). \tag{4.6}$$

The constants $\gamma(n)$ and $\epsilon(n)$ of (4.3) are the unique joint minimizer of $\Upsilon(n;\gamma,\epsilon)$, that is, $\Upsilon(n;\gamma(n),\epsilon(n))\le\Upsilon(n;\gamma,\epsilon)$ for all $(\gamma,\epsilon)\in(0,+\infty)^2$.

(ii) *If no control variates is performed, then the function $\Upsilon(n;\gamma,\epsilon)$ in (4.6) is to be replaced by $\Upsilon_a(n;\gamma)$, of which the constant $\gamma(n)$ given in (4.3) remains the unique minimizer of $\Upsilon_a(n;\gamma)$. If $\tau_n=o_{\mathbb{P}_0}(n)$, then $\Upsilon_a(n;\gamma(n))=\mathscr{O}(\sqrt{\ln(n)/n})$.*

(iii) *Suppose control variates is performed. If $\liminf_{n\uparrow+\infty}\alpha_k>0$ and $\tau_n=o_{\mathbb{P}_0}(n)$, then it holds $\mathbb{P}_0$-a.s. that $\Upsilon(n;\gamma(n),\epsilon(n))=\mathscr{O}(\sqrt[4]{\ln(n)/n})$.*

In the upper bound (4.6), the first chunk $(V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)-\mu^2)$ is the minimal variance one wishes to attain in our framework. The second chunk describes how much reduction of variance we are missing out on average in exchange for waiting until the stopping time τ_n, that is, the maximum reduction of variance (the crude variance $V(\boldsymbol{\theta}_0)-\mu^2$ minus the desired optimum $V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)-\mu^2$) multiplied by the relative inactive time $\mathbb{E}_0[\tau_n]/n$. The constant learning rates (4.3) may not be optimal in reducing the estimator variance in the proposed framework. As stated in Theorem 4.1, those are the unique joint minimizer of the right hand side of the inequality (4.6), which is merely an upper bound for the (scaled) theoretical estimator variance $n\mathrm{Var}_0(\mu(n))$, or equivalently that of the (scaled) mean empirical variance $\mathbb{E}_0[\sigma^2(n)]$, in view of Proposition 3.1. Still, the constant learning rates (4.3) are a reasonable choice in the sense that, as Theorem 4.1 **(ii)** and **(iii)** assert, the minimized upper bound (4.6) decays to zero if the finite computing budget n is set large at the outset. It is worth emphasizing that Theorem 4.1 **(ii)** and **(iii)** are not asymptotic results in the usual sense [18]. In the present context, the implementation of the algorithm is not supposed to be incremental in computing budget n, that is, with a larger budget $m(>n)$, we obtain the minimized upper bound $\Upsilon(m;\gamma(m),\epsilon(m))$ only if the algorithm runs with the constant learning rates $(\gamma(m),\epsilon(m))$ *from the outset*, not as a continuation from a shorter run with a smaller budget n and its corresponding constant learning rates $(\gamma(n),\epsilon(n))$.

This result is a proper superset of the existing result [11, Theorem 4.1] (without control variates), as it is fully recovered in Theorem 4.1 **(ii)**. With control variates performed (Theorem 4.1 **(iii)**), however, the proof of convergence demands significantly more delicate treatments, mainly because the function $V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})$ is not jointly convex in two arguments $(\boldsymbol{\theta},\boldsymbol{\xi})$, and thus the intended point $(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$ is generally not be globally optimal, as we will illustrate in Appendix B.3.

4.2. Infinite computing budget

We next turn to the case where the computing budget is unlimitedly available. The availability of infinite computing budget is not very realistic, and such requirement contradicts the essential concept of variance reduction, where one wishes to terminate iterations sooner. Hence, rather than constructing algorithms exclusively for the case of infinite computing budget, we focus on deriving sufficient conditions for the proposed algorithms (Sections 3.1 and 3.2) to achieve the following convergences for the purpose of Monte Carlo simulation:

$$\mu(n)\to\mu, \quad a.s., \tag{4.7}$$

$$\sigma^2(n)\to V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)-\mu^2, \tag{4.8}$$

$$\sqrt{n}\frac{\mu(n)-\mu}{\sigma(n)}\overset{\mathscr{L}}{\to}\mathscr{N}(0,1), \tag{4.9}$$

as $n\uparrow+\infty$, under the probability measure $\mathbb{P}_0$. The mode of the convergence (4.8) is left unspecified at the moment on purpose, since we derive its $L^1(\Omega,\mathscr{F},\mathbb{P}_0)$ and almost sure convergences respectively for the algorithms by stochastic approximation (Theorem 4.2) and by sample average approximation (Theorem 4.3).

We first address the algorithm by stochastic approximation (Section 3.1). This problem has long been investigated under many different problem settings (such as [1–3,10,13,14,17]). From a practical point of view, stochastic approximation with decreasing learning rates is not necessarily the most ideal form of its implementation, since then the extreme performance sensitivity to the choice of learning rates comes back in question.

Theorem 4.2. (Stochastic Approximation with Decreasing Learning Rates) *Let $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\in\mathbb{N}_0}$, $\{\widetilde{\boldsymbol{\xi}}_k\}_{k\in\mathbb{N}_0}$ and $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ be $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$-adapted sequences of random vectors defined in Section 3.1 with (3.7).*

(i) *The almost sure convergence (4.7) holds.*

(ii) *Assume* $\{\gamma_k\}_{k\in\mathbb{N}_0}$ *and* $\{\epsilon_k\}_{k\in\mathbb{N}_0}$ *are in* $\ell^2\setminus\ell^1$*;* $\sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in\mathscr{T}_0\times\Lambda_0}\int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\lambda},\boldsymbol{\theta}_0)|H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)|^2\times|\Psi(\mathbf{u})|^4 d\mathbf{u}<+\infty$*;* $\lim_{n\uparrow+\infty}\tau_n/n=0$*; and* $\liminf_{k\uparrow+\infty}\alpha_k>0$, $\mathbb{P}_0$*-a.s. Then, the convergence* (4.8) *holds in* $L^1(\Omega,\mathscr{F},\mathbb{P}_0)$.

(iii) *If moreover* $\inf_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathscr{T}_0\times\mathscr{X}_0}(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi}))>\mu^2$ *and* $\sup_{\boldsymbol{\theta}\in\mathscr{T}_0}\int_{(0,1)^d}|H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)|^{q-1}|\Psi(\mathbf{u})|^q d\mathbf{u}<+\infty$ *for some* $q>2$*, then the weak convergence* (4.9) *holds.*

We have pointed out a possible existence of a point $(\boldsymbol{\theta}^\diamond,\boldsymbol{\xi}^\diamond)$, that outperforms the proposed framework in the sense of $V(\boldsymbol{\theta}^\diamond)+W(\boldsymbol{\theta}^\diamond,\boldsymbol{\xi}^\diamond)<V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$. It is worth adding here that the limiting value of the convergence result **(ii)** cannot be such a strictly smaller estimator variance $V(\boldsymbol{\theta}^\diamond)+W(\boldsymbol{\theta}^\diamond,\boldsymbol{\xi}^\diamond)-\mu^2$, because the sequence $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\in\mathbb{N}_0}$ converges necessarily to $\boldsymbol{\theta}^*$ since the stochastic approximation algorithm (3.2) runs on the basis of the convex function $V(\boldsymbol{\theta})$ alone, without interference from the control variates parameter $\boldsymbol{\xi}$. Then, due to the definition (2.15), the minimizer $\boldsymbol{\xi}^*$ is unique as soon as $\boldsymbol{\theta}^*$ is given.

The convergence result **(ii)** is given in the $L^1(\Omega,\mathscr{F},\mathbb{P}_0)$ mode, since its derivation follows quite naturally from the proof of Theorem 4.1, as well as the $L^1(\Omega,\mathscr{F},\mathbb{P}_0)$ convergence is enough for the subsequent result **(iii)**.

The condition $\inf_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathscr{T}_0\times\mathscr{X}_0}(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi}))>\mu^2$ in **(iii)** means that perfect variance reduction is impossible in any way. It is possible to come up with such problem settings (Appendix B.3), whereas perfect variance reduction seems possible merely artificially. In practice, verifiability of this condition however requires no serious attention for the reason that one would be happier with perfect variance reduction than with the theoretical result **(iii)**.

We next turn to the algorithm by sample average approximation (Section 3.2). The only essential difference from Theorem 4.2 is the almost sure mode of the convergence **(iii)**, which can be derived relatively easily from the preceding almost sure convergence of the parameter sequences **(i)**, whereas this difference in the mode of convergence is not directly relevant in the context of variance reduction.

Theorem 4.3. (Sample Average Approximation) *Let* $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\in\mathbb{N}_0}$, $\{\widetilde{\boldsymbol{\xi}}_k\}_{k\in\mathbb{N}_0}$ *and* $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ *be* $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$*-adapted sequences of random vectors defined in Section* 3.2.

(i) *The almost sure convergence* $(\widetilde{\boldsymbol{\theta}}_n,\widetilde{\boldsymbol{\xi}}_n)\to(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$ *holds, as* $n\uparrow+\infty$.

(ii) *The almost sure convergence* (4.7) *holds.*

(iii) *If* $\sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in\mathscr{T}_0\times\Lambda_0}\int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\lambda},\boldsymbol{\theta}_0)|H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)|^2|\Psi(\mathbf{u})|^4 d\mathbf{u}<+\infty$*, then the convergence* (4.8) *holds* $\mathbb{P}_0$*-a.s.*

(iv) *If moreover* $\inf_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathscr{T}_0\times\mathscr{X}_0}(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi}))>\mu^2$ *and* $\sup_{\boldsymbol{\theta}\in\mathscr{T}_0}\int_{(0,1)^d}|H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)|^{q-1}|\Psi(\mathbf{u})|^q d\mathbf{u}<+\infty$ *for some* $q>2$*, then the weak convergence* (4.9) *holds.*

The results above may be made more precise, such as the asymptotic Gaussianity of the parameter sequence $\sqrt{n}((\widetilde{\boldsymbol{\theta}}_n,\widetilde{\boldsymbol{\xi}}_n)-(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*))$, in a similar manner to the preceding work [11]. In the present work, we do not go into this direction, again as our primary interest is not in the parameter search but mainly in the estimator convergence (4.7) and the asymptotic Gaussianity of the estimator sequence (4.9).

5. Numerical illustrations

We have constructed adaptive Monte Carlo variance reduction algorithms in Section 3, by the stochastic approximation (Section 3.1) and the sample average approximation (Section 3.2), and have justified the relevance of the proposed framework and algorithms in Section 4, on the basis of finite computing budget (Section 4.1) and infinite computing budget (Section 4.2). We close this study with a high-dimensional example to support our theoretical findings and illustrate the effectiveness of the proposed framework and algorithms, adopting the problem setting from [11, Section 6.2] for direct comparison purposes, where control variates is absent (that is, $\boldsymbol{\xi}=0_d$).

Let $Z := (Z_1, \cdots, Z_d)$ be a standard normal random vector in $\mathbb{R}^d$ with independent components. Consider the random variable

$$F(Z) = e^{-rT} \max\left[\frac{1}{d}\sum_{n=1}^{d} S_0 \exp\left[\sum_{k=1}^{n}\left(\left(r - \frac{1}{2}\sigma^2\right)\frac{T}{d} + \sqrt{\sigma^2 \frac{T}{d}} Z_k\right)\right] - K, 0\right],$$

whose expected value is then equipped with the proposed variance reduction techniques as follows:

$$\mu = \mathbb{E}\left[F(Z)\right] = \mathbb{E}\left[e^{-\langle \boldsymbol{\theta}, Z\rangle - \|\boldsymbol{\theta}\|^2/2} F\left(Z + \boldsymbol{\theta}\right) + \langle \boldsymbol{\xi}, \Phi_d(Z) - \mathbb{1}_d/2\rangle\right],$$

where the rightmost term is derived by first applying importance sampling and then control variates as in the progression (2.2). Here, we let Φ_d denote the componentwise standard normal cumulative distribution function on $\mathbb{R}^d$, and Φ_d^{-1} is its componentwise inverse on $(0,1)^d$, so that $\Phi_d(Z) \overset{\mathscr{L}}{=} U$ and $\Phi_d^{-1}(U) \overset{\mathscr{L}}{=} Z$ for $Z \sim \mathscr{N}(0, \mathbb{I}_d)$ and $U \sim U(0,1)^d$. The same random variable $F(Z)$ was examined in [6, Section 6], but there with the raw normal random vector as simple linear control variates $\langle \boldsymbol{\xi}, \mathbf{z}\rangle$, as opposed to the nonlinear control variates $\langle \boldsymbol{\xi}, \Phi_d(\mathbf{z}) - \mathbb{1}_d/2\rangle$ here, since the underlying vector in our context is $\mathbf{u}(= \Phi_d(\mathbf{z}) - \mathbb{1}_d/2)$ on the unit hypercube. The importance sampling parameter $\boldsymbol{\theta}$ is chosen to minimize the second moment of the first term, as of (2.10):

$$V(\boldsymbol{\theta}) = \mathbb{E}\left[e^{-\langle \boldsymbol{\theta} + \boldsymbol{\lambda}, Z\rangle + \|\boldsymbol{\theta}\|^2/2 - \langle \boldsymbol{\theta}, \boldsymbol{\lambda}\rangle - \|\boldsymbol{\lambda}\|^2/2} \left|F(Z + \boldsymbol{\lambda})\right|^2\right],$$

which is certainly independent of the auxiliary parameter $\boldsymbol{\lambda}$. The control variates component as of (2.12) is then given by

$$W(\boldsymbol{\theta}, \boldsymbol{\xi}) = 2\left\langle \boldsymbol{\xi}, \mathbb{E}_0\left[e^{-\langle \boldsymbol{\lambda}, Z\rangle - \|\boldsymbol{\lambda}\|^2/2} F(Z + \boldsymbol{\lambda})\left(\Phi_d(Z + \boldsymbol{\lambda} - \boldsymbol{\theta}) - \mathbb{1}_d/2\right)\right]\right\rangle + \frac{1}{12}\|\boldsymbol{\xi}\|^2,$$

which is also independent of the auxiliary parameter $\boldsymbol{\lambda}$. All those formulas remain within the scope of the original formulation (1.1) with $\mathbf{u} = \Phi_d(\mathbf{z})$, $\mathbf{z} = \Phi_d^{-1}(\mathbf{u})$ and $F(\mathbf{z}) = \Psi(\Phi_d(\mathbf{z}))$ on $\mathscr{D} = \mathbb{R}^d$.

5.1. Finite computing budget

The finite-budget approach (Theorem 4.1) provides an objective way by stochastic approximation to address its extreme performance sensitivity to the choice of decreasing learning rates. We fix $S_0 = 50$, $r = 0.05$, $T = 0.5$, $d = 16$, $\sigma = 0.10$ and $K = 55$, and then we have the mean $\mu = 2.445 \times 10^{-2}$ and the crude estimator variance $V(0_d) - \mu^2 = 3.960 \times 10^{-2}$. With $d = 16$, the variance reduction parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are both 16-dimensional as well. Here, to focus on the effectiveness of the additional control variates, we suppress various problem parameters. For instance, we fix the parameter search domains $\mathscr{X}_k \equiv [-0.1, 0.6]^{16}$ and $\mathscr{T}_k \equiv [-0.06, 0.02]^{16}$, and disable both a pilot run and the auxiliary parameter throughout by setting $\tau_n = 0$ and $\boldsymbol{\lambda}_k \equiv \boldsymbol{\theta}_0 (= 0_d)$. The quantities $L^2(\mathscr{T}_{\tau_n}; \boldsymbol{\lambda}^\star)$ and $J^2(\mathscr{T}_{\tau_n}, \mathscr{X}_{\tau_n}; \boldsymbol{\lambda}^\star)$ in (4.4) and (4.5) and, consequently, the constant learning rates $\gamma(n)$ and $\epsilon(n)$ in (4.3) are now all deterministic. We conduct the supremums within the quantities $L^2(\mathscr{T}_0; \boldsymbol{\theta}_0)$ and $J^2(\mathscr{T}_0, \mathscr{X}_0; \boldsymbol{\theta}_0)$ numerically using MATLAB's 'fmincon' function, as the integrals (4.4) and (4.5) do not seem to have particular geometric structures, such as convexity, concavity and monotonicity. We refer the reader to the preceding work [9–11] for some strategic setting of those problem parameters. Let us recall that the convexity parameter α_k defined in (4.1) improves the upper bound (4.6) for the estimator variance if it is strictly positive, whereas it is not required for implementation at all.

Unlike in the existing study of stochastic approximation algorithms, our primary focus is more on reduction of the estimator variance $n\mathrm{Var}_0(\mu(n))$, than on improvements on the convergence to the (sub)optimal

Table 3
Estimator variances $n\mathrm{Var}_0(\mu(n))$ $(\times 10^{-3})$ with a variety of finite computing budgets and constant learning rates, estimated using 5000 iid replications of the adaptive empirical mean $\mu(n)$. The numbers outside parentheses represent estimator variances with both importance sampling and control variates, while the numbers inside parentheses are estimator variances with importance sampling alone (with $\boldsymbol{\xi}_k \equiv 0_d$ fixed). The crude estimator variance is $V(0_d) - \mu^2 = 39.60 \times 10^{-3}$.

n \ c	0.1	0.2	0.5	1	2	5	10
5000	6.962 (8.667)	4.156 (5.393)	2.974 (3.898)	2.268 (3.014)	1.961 (2.643)	2.056 (2.645)	2.292 (2.775)
10000	5.307 (7.075)	3.039 (4.253)	2.112 (3.053)	1.606 (2.323)	1.405 (1.994)	1.468 (2.054)	1.741 (2.278)
15000	4.516 (6.062)	2.613 (3.708)	1.794 (2.660)	1.415 (2.101)	1.166 (1.684)	1.211 (1.765)	1.485 (2.033)
20000	4.356 (5.982)	2.267 (3.383)	1.566 (2.296)	1.249 (1.857)	1.021 (1.563)	1.100 (1.634)	1.349 (1.898)
25000	3.866 (5.425)	2.114 (3.117)	1.435 (2.150)	1.105 (1.718)	0.9535 (1.501)	0.9822 (1.529)	1.223 (1.775)
30000	3.228 (4.991)	1.917 (2.925)	1.349 (2.060)	1.046 (1.687)	0.9374 (1.456)	0.9566 (1.495)	1.168 (1.718)

point $(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$. With this in mind, we present in Table 3 estimator variances $n\mathrm{Var}_0(\mu(n))$, estimated using 5000 iid replications of the adaptive empirical mean $\mu(n)$, with a variety of constant learning rates, up to the constant multiple c in common on both constant learning rates $c\gamma(n)$ and $c\epsilon(n)$. It is certainly possible and more realistic that such miscomputation occurs differently to two (for instance, $2\gamma(n)$ and $0.1\epsilon(n)$), but we do not go into such exhaustive presentation. The 42 $(= 7 \times 6)$ experiments are conducted separately for each finite computing budget n and miscomputation multiple c.

The proposed algorithm is built to seek a better application of variance reduction techniques adaptively, rather than to run Monte Carlo simulation with optimized variance reduction techniques applied from the outset. The set of the results here suggests that the best, or nearly best, possible variance ratios in this problem setting with both importance sampling and control variates applied is as large as 42.24 $(= (V(0_d) - \mu^2)/(n\mathrm{Var}_0(\mu(n))))$ gained with $n = 30000$ and $c = 2$. Remarkably, even starting from no information $\boldsymbol{\theta}_0 = \boldsymbol{\xi}_0 = 0_d$, a right choice of constant learning rates leads us to well reduced estimator variance at an early stage. For instance, with $n = 5000$ and $c = 1$, the proposed algorithm reaches the following high variance ratios:

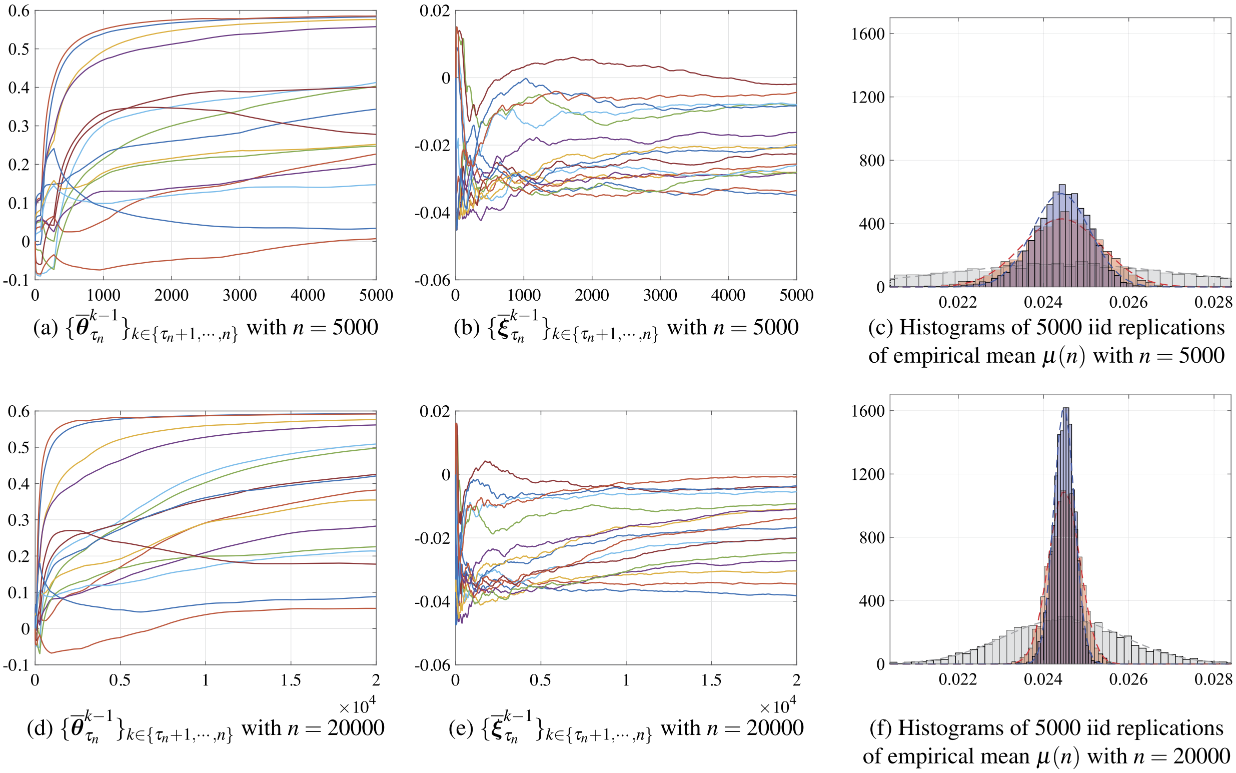
$$\frac{39.60 \times 10^{-3}}{2.268 \times 10^{-3}} = 17.46, \quad \frac{39.60 \times 10^{-3}}{3.014 \times 10^{-3}} = 13.14,$$

respectively, if both importance sampling and control variates are applied or if control variates is suppressed (with $\boldsymbol{\xi}_k \equiv 0_d$ fixed), relative to the crude Monte Carlo simulation (with $\boldsymbol{\theta}_k = \boldsymbol{\xi}_k \equiv 0_d$ fixed). It is encouraging that the addition of control variates contributes to reduce a lot more estimator variance with a little additional computing effort, without an exception.

Observe also that the experiments with $c = 2$ (that is, constant learning rates are doubly miscomputed) produce the smallest estimator variance across all computing budgets. This is not very surprising in the sense that the constant learning rates (4.3) are derived by minimizing an upper bound for the estimator variance (4.6), not the estimator variance itself. It is more essential that the performance is not very sensitive to the learning rates as long as those are not too far from the formulas (4.3).

To better illustrate how the proposed algorithm achieves a well reduced estimator variance even at an early stage, we plot in Fig. 1 typical trajectories of the successive averaging of the variance reduction parameters $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k \in \{\tau_n+1,\cdots,n\}}$ and $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k \in \{\tau_n+1,\cdots,n\}}$, as well as histograms (normalized as probability density functions) of 5000 iid replications of the empirical mean $\mu(n)$ with both importance sampling and control variates (blue), importance sampling alone with $\boldsymbol{\xi}_k \equiv 0_d$ fixed (red), and no variance reduction techniques with $\boldsymbol{\theta}_k = \boldsymbol{\xi}_k \equiv 0_d$ fixed (grey), where the computing budget is either $n = 5000$ or $n = 20000$. In view of the formulas (4.3) with $\tau_n = 0$ and $\boldsymbol{\lambda}^\star = \boldsymbol{\theta}_0$ fixed, we have the ratio of two sizes with different computing budgets:

$$\frac{\gamma(5000)}{\gamma(20000)} = \frac{\epsilon(5000)}{\epsilon(20000)} = \left(\frac{\frac{1}{5000}\sum_{k=1}^{5000} k^{-1}}{\frac{1}{20000}\sum_{k=1}^{20000} k^{-1}} \right)^{1/2} = 1.863,$$



(a) $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ with $n=5000$

(b) $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ with $n=5000$

(c) Histograms of 5000 iid replications of empirical mean $\mu(n)$ with $n=5000$

(d) $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ with $n=20000$

(e) $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ with $n=20000$

(f) Histograms of 5000 iid replications of empirical mean $\mu(n)$ with $n=20000$

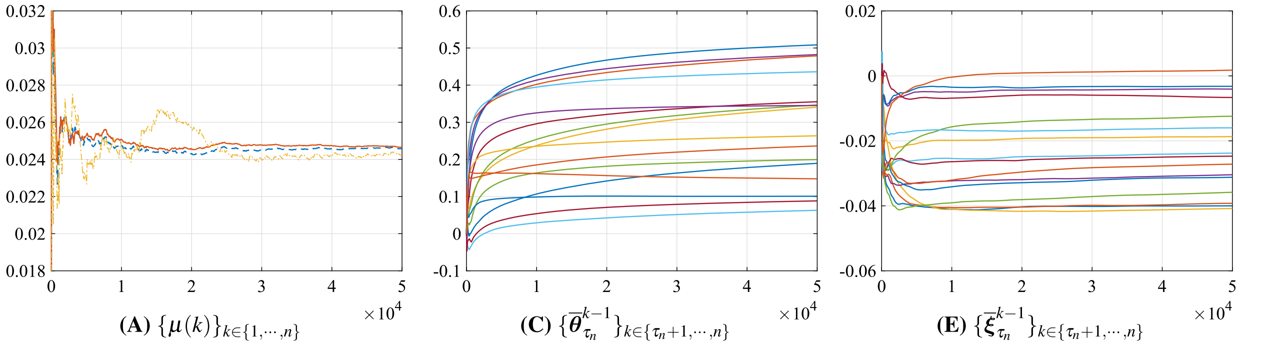
Fig. 1. Numerical results with constant learning rates $\gamma(n)$ and $\epsilon(n)$. Figures (a), (b) and (c) correspond to the computing budget $n = 5000$, while figures (d), (e) and (f) the computing budget $n = 20000$. Figures (a), (b), (d) and (e) plot typical trajectories of the 16 components of $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ and $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$. Figures (c) and (f) are histograms of 5000 iid replications of the empirical mean $\mu(n)$ with both important sampling and control variates (blue), importance sampling alone (red), and no variance reduction techniques applied (grey). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

that is, the learning rates of the stochastic approximation algorithms (3.2) and (3.3) are both roughly twice as fast with $n = 5000$ as with $n = 20000$. As can be seen in Fig. 1 (a) and (b), even with such a small computing budget $n = 5000$, the 16 components seem to get stable at an early stage (as early as around 1000 steps), and then the remaining 4000 steps ($= 5000 - 1000$) can run Monte Carlo simulation with well-tuned variance reduction techniques. For illustration purpose, we attach the Gaussian density functions (dashed lines) based on the corresponding empirical mean and variances. Although we have no such limiting Gaussianity of the empirical mean $\mu(n)$ at any finite computing budget, the rightmost histograms (c) and (f) indicate that the law is fairly close to Gaussian (even at a small finite budget $n = 5000$), which is an encouraging outcome for constructing confidence intervals.

5.2. *Infinite computing budget*

We next turn to the case where the computing budget is progressively and unlimitedly available. We carry over the problem setting of $S_0 = 50$, $r = 0.05$, $T = 0.5$, $d = 16$, $\sigma = 0.10$ and $K = 55$, and set $n = 5 \times 10^4$ iterations, which can be well considered infinite. As earlier, we disable a pilot run and the auxiliary parameter by setting $\tau_n = 0$ and $\boldsymbol{\lambda}_k \equiv \boldsymbol{\theta}_0 (= 0_d)$, and thus $\boldsymbol{\lambda}^\star = \boldsymbol{\theta}_0$ as well. We again fix the parameter search domains $\mathscr{X}_k \equiv \mathscr{X}_0 = [-0.1, 0.6]^{16}$ and $\mathscr{T}_k \equiv \mathscr{T}_0 = [-0.06, 0.02]^{16}$.

In accordance with Theorem 4.2, we let the learning rates be decreasing as $\gamma_k = \epsilon_k = (k+1)^{-0.6}$. We plot in Fig. 2 typical trajectories of the adaptive empirical mean $\{\mu(k)\}_{k\in\{1,\cdots,n\}}$ of the line **(A)**, the successive averaging of importance sampling parameters $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ of the line **(C)**, and the



(A) $\{\mu(k)\}_{k\in\{1,\cdots,n\}}$ **(C)** $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ **(E)** $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$

Fig. 2. Numerical results by stochastic approximation with decreasing learning rates $\gamma_k = \epsilon_k = (k+1)^{-0.6}$ for $k \in \mathbb{N}_0$. In the leftmost figure, the dash-dot line (yellow) is a typical trajectory of the crude Monte Carlo simulation (with $\boldsymbol{\theta}_k = \boldsymbol{\xi}_k \equiv 0_d$ fixed), the dashed line (blue) corresponds to the Monte Carlo simulation with importance sampling alone (that is, with $\boldsymbol{\xi}_k \equiv 0_d$ fixed), whereas the sold line (red) indicates the Monte Carlo simulation with both importance sampling and control variates employed.

successive averaging of importance sampling parameters $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ of the line **(E)**. In each of the middle and rightmost figures, the 16 trajectories correspond to the 16 components of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, respectively. The 16 components of the parameter vectors $\{\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ and $\{\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}\}_{k\in\{\tau_n+1,\cdots,n\}}$ of Fig. 2 **(C)** and **(E)** seem to move towards quite distinct values from each other, distinct enough not to project the 16-dimensional vector onto one degree of freedom (that is, $\boldsymbol{\theta} = \theta\mathbb{1}_d$ [9, Section 5]). With all these 16 components kept free, a frequent search for the minimizer $\boldsymbol{\theta}^*$ on a 16-dimensional space (or the joint sub-optimum $(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$ without convexity on a $32(= 2 \times 16)$-dimensional space) by sample average approximation (Theorem 4.3) is often computationally prohibitive, which warrants the choice of stochastic approximation (Section 3.1) over sample average approximation (Section 3.2). As has long been widely known, however, the performance depends on the choice of the decreasing learning rates $\{\gamma_k\}_{k\in\mathbb{N}_0}$ and $\{\epsilon_k\}_{k\in\mathbb{N}_0}$, which is essentially arbitrary as long as the $\ell^2 \setminus \ell^1$-condition is satisfied. Let us add that the performance depends largely on the random seed chosen for experiments as well.

6. Concluding remarks

We have advanced an adaptive variance reduction framework in such a way that both importance sampling and control variates can be applied in parallel. We have derived convergence rates of an upper bound for the theoretical estimator variance towards its minimum as a fixed computing budget increases, when stochastic approximation runs with optimal constant learning rates. We have also proved that the proposed algorithm attains the minimal estimator variance in the limit by stochastic approximation with decreasing learning rates or by sample average approximation, when computing budget is unlimitedly available. We close this study by highlighting additional practical considerations and future directions of research stemming from this work.

First, we have kept the problem setting as general as possible, rather than developing a tailor-made methodology for a specialized problem class, such as the rare event simulation. A natural question is how far the proposed framework can be specialized for performance improvements if the problem setting is more focused. One such approach is a non-linearization of the variates (Section B.2), which however would cost us significantly more intricate proofs for convergence results. Specialization of the problem setting may also guide us as to the choice of the bypass distribution, which is currently exponential or normal distribution for easy and light computation.

Next, despite that the proposed algorithm runs and the convergence results hold true in their current form no matter how large the problem dimension is in theory with almost no additional coding effort, a serious increase of computational complexity in high-dimensional problems is a crucial matter from a practical point of view. As is clear, parametrizing every component of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ (just like Section 5) would not reduce the variance relative to the increased computing cost required for implementation, such as (say, if $d = 10^6$) computing 10^6-dimensional gradients, projecting and averaging 10^6-dimensional vectors (**(B)**-**(E)** of Section 3.1) and optimizing over a 10^6-dimensional parameter domain (**(G)** of Section 3.2) at every step, whereas such explosive complexity would never pay off. A possible evasion is to adopt the simple averaging **(H)** (rather than **(D)**-**(E)**) for the control variates component so as to skip many computation steps, such as metric projection and optimization. Another and more straightforward one is a direct projection of the parameter vectors (the most extreme one is onto scaler as mentioned in Section 5.2, that is, $\boldsymbol{\theta} = \theta \mathbb{1}_d$ and $\boldsymbol{\xi} = \xi \mathbb{1}_d$), which can directly lower the vector dimensions in the lines **(B)**-**(E)** of Section 3.1 and **(G)**-**(H)** of Section 3.2.

Last but not least, we are still left to fix criteria for designing the auxiliary parameter, the stopping time, and the shrinking parameter domains. In particular, the supremums (4.4) and (4.5) require an additional optimization procedure, which could potentially stand as a fatal bottleneck, especially in high-dimensional problems. Relevant difficulties here may be mitigated somehow by forcing the algorithm to make a decision on such components at predetermined timings, for instance, through batching of the run, which would be an interesting direction of research towards improvements in both performance and implementation.

Appendix A. Proofs

We collect all proofs here with main focus on additional intricate derivations due to the lack of joint convexity of the sum $V(\boldsymbol{\theta}) + W(\boldsymbol{\theta}, \boldsymbol{\xi})$ with respect to the parameter $(\boldsymbol{\theta}, \boldsymbol{\xi})$. To avoid overloading the paper, we skip nonessential details of somewhat routine nature in some instance, particularly where the existing results on the stochastic programming [18] and on the importance sampling term $V(\boldsymbol{\theta})$ alone [11] can be applied.

Proof of Theorem 4.1. Throughout, we fix $n \in \mathbb{N}$ and let $m \in \{0, 1, \cdots, n\}$ and $l \in \{m+1, \cdots, n\}$. On the event $A_m := \{\tau_n = m\}$, we have $\boldsymbol{\lambda}_{(l-1)\wedge\tau_n} = \boldsymbol{\lambda}_{(l-1)\wedge m} = \boldsymbol{\lambda}_m = \boldsymbol{\lambda}^\star$ for all $l \in \{m+1, \cdots, n\}$.

(i) First, as for the second moment function $V(\boldsymbol{\theta})$ on the importance sampling component, we employ the upper bounds derived in the preceding work [11]:

$$\begin{aligned}\mathbb{E}_0\left[V(\overline{\boldsymbol{\theta}}_m^{k-1})\mathbb{1}(A_m)\right] &\leq V\left(\boldsymbol{\theta}^*\right)\mathbb{P}_0(A_m)\\ &\quad + \frac{(1-\alpha_m\gamma_m)\mathrm{diam}^2(\mathscr{T}_m)\mathbb{P}_0(A_m) + \mathbb{E}_0[L^2(\mathscr{T}_m;\boldsymbol{\lambda}^\star)\mathbb{1}(A_m)]\sum_{l=m+1}^k \gamma_{l-1}^2}{2\sum_{t=m+1}^k \gamma_{t-1}},\end{aligned} \tag{A.1}$$

and

$$\begin{aligned}\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^n V(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1})\right] &\leq V\left(\boldsymbol{\theta}^*\right)\left(1-\frac{\mathbb{E}_0[\tau_n]}{n}\right) + \mathbb{E}_0\left[\mathrm{diam}^2(\mathscr{T}_{\tau_n})\frac{1}{2n}\sum_{k=\tau_n+1}^n \frac{1-\alpha_{\tau_n}\gamma_{\tau_n}}{\sum_{t=\tau_n+1}^k \gamma_{t-1}}\right]\\ &\quad + \mathbb{E}_0\left[L^2(\mathscr{T}_{\tau_n};\boldsymbol{\lambda}^\star)\frac{1}{2n}\sum_{k=\tau_n+1}^n \frac{\sum_{l=\tau_n+1}^k \gamma_{l-1}^2}{\sum_{t=\tau_n+1}^k \gamma_{t-1}}\right].\end{aligned} \tag{A.2}$$

To avoid a simple repetition, we refer the reader to [11, Appendix A] for their derivation.

Next, we derive an upper bound for $n^{-1}\mathbb{E}_0[\sum_{k=\tau_n+1}^n W(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}, \overline{\boldsymbol{\xi}}_{\tau_n}^{k-1})]$. Observe that the function $W(\boldsymbol{\theta}, \boldsymbol{\xi})$ is strongly convex in the second argument with parameter $1/6$:

$$W(\boldsymbol{\theta}, \boldsymbol{\xi}^*) = W(\boldsymbol{\theta}, \boldsymbol{\xi}) + \langle \boldsymbol{\xi}^* - \boldsymbol{\xi}, \nabla_{\boldsymbol{\xi}} W(\boldsymbol{\theta}, \boldsymbol{\xi})\rangle + \frac{(1/6)}{2}\left\|\boldsymbol{\xi}^* - \boldsymbol{\xi}\right\|^2, \quad (\boldsymbol{\xi}, \boldsymbol{\xi}^*) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{A.3}$$

Fix $l \in \{m+1, \cdots, n\}$ and $k \in \{l, \cdots, n\}$. It holds that, on the event A_m,

$$\begin{aligned}
\|\boldsymbol{\xi}_l - \boldsymbol{\xi}^*\|^2 &= \left\|\prod\nolimits_{\mathscr{X}_l} \left(\boldsymbol{\xi}_{l-1} - \epsilon_{l-1}\nabla_{\boldsymbol{\xi}} S(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right) - \prod\nolimits_{\mathscr{X}_l}(\boldsymbol{\xi}^*)\right\|^2 \\
&\le \left\|\boldsymbol{\xi}_{l-1} - \epsilon_{l-1}\nabla_{\boldsymbol{\xi}} S(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star) - \boldsymbol{\xi}^*\right\|^2 \\
&= \left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2 + \epsilon_{l-1}^2 \left\|\nabla_{\boldsymbol{\xi}} S(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right\|^2 - 2\epsilon_{l-1}\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, \nabla_{\boldsymbol{\xi}} S(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right\rangle \\
&= \left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2 + \epsilon_{l-1}^2 \left\|\nabla_{\boldsymbol{\xi}} S(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right\|^2 - 2\epsilon_{l-1}\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, \nabla_{\boldsymbol{\xi}} S(U_l; \boldsymbol{\theta}^*, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right\rangle \\
&\quad - 4\epsilon_{l-1}\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, Q(U_l; \boldsymbol{\theta}_{l-1}, \boldsymbol{\lambda}^\star) - Q(U_l; \boldsymbol{\theta}^*, \boldsymbol{\lambda}^\star)\right\rangle.
\end{aligned}$$

Taking conditional expectation $\mathbb{E}_{l-1}$ yields

$$\begin{aligned}
\mathbb{E}_{l-1}\left[\|\boldsymbol{\xi}_l - \boldsymbol{\xi}^*\|^2\right] &\le \left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2 + \epsilon_{l-1}^2 \int\limits_{(0,1)^d} \left\|\nabla_{\boldsymbol{\xi}} S(\mathbf{u}; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right\|^2 d\mathbf{u} \\
&\quad - 2\epsilon_{l-1}\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, \nabla_{\boldsymbol{\xi}} W(\boldsymbol{\theta}^*, \boldsymbol{\xi}_{l-1})\right\rangle - 4\epsilon_{l-1}\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, W_0(\boldsymbol{\theta}_{l-1}) - W_0(\boldsymbol{\theta}^*)\right\rangle.
\end{aligned}$$

Combining this with the strong convexity (A.3) yields

$$\begin{aligned}
&2\epsilon_{l-1}\left(W(\boldsymbol{\theta}^*, \boldsymbol{\xi}_{l-1}) - W(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)\right) \\
&\le -\frac{\epsilon_{l-1}}{6}\left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2 - \left(\mathbb{E}_{l-1}\left[\|\boldsymbol{\xi}_l - \boldsymbol{\xi}^*\|^2\right] - \left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2\right) \\
&\quad + \epsilon_{l-1}^2 \int\limits_{(0,1)^d} \left\|\nabla_{\boldsymbol{\xi}} S(\mathbf{u}; \boldsymbol{\theta}_{l-1}, \boldsymbol{\xi}_{l-1}, \boldsymbol{\lambda}^\star)\right\|^2 d\mathbf{u} - 4\epsilon_{l-1}\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, W_0(\boldsymbol{\theta}_{l-1}) - W_0(\boldsymbol{\theta}^*)\right\rangle,
\end{aligned}$$

where both $\boldsymbol{\theta}_{l-1}$ and $\boldsymbol{\xi}_{l-1}$ are $\mathscr{F}_{l-1}$-measurable, while the random vector U_l is independent of the σ-field $\mathscr{F}_{l-1}$. Recall the convex combination (3.4), and again the strong convexity (A.3). We multiply the inequality above by the $\mathscr{F}_{l-1}$-measurable indicator $\mathbb{1}(A_m)$, divide by the strictly positive constant $\sum_{t=m+1}^k \epsilon_{t-1}$, and take conditional expectation $\mathbb{E}_0$, which overall yields

$$\begin{aligned}
\mathbb{E}_0\left[W(\boldsymbol{\theta}^*, \overline{\boldsymbol{\xi}}_m^{k-1})\mathbb{1}(A_m)\right] &\le \mathbb{E}_0\left[\sum_{l=m+1}^{k} \frac{\epsilon_{l-1}}{\sum_{t=m+1}^k \epsilon_{t-1}} W(\boldsymbol{\theta}^*, \boldsymbol{\xi}_{l-1})\mathbb{1}(A_m)\right] \\
&\le W(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)\mathbb{P}_0(A_m) + \frac{1 - \epsilon_m/6}{2\sum_{t=m+1}^k \epsilon_{t-1}}\mathbb{P}_0(A_m)\mathrm{diam}^2(\mathscr{X}_m) \\
&\quad + \frac{1}{2}\sum_{l=m+1}^{k} \frac{\epsilon_{l-1}^2}{\sum_{t=m+1}^k \epsilon_{t-1}} J^2(\mathscr{T}_m, \mathscr{X}_m; \boldsymbol{\lambda}^\star)\mathbb{P}_0(A_m) \\
&\quad - 2\sum_{l=m+1}^{k} \frac{\epsilon_{l-1}}{\sum_{t=m+1}^k \epsilon_{t-1}} \mathbb{E}_0\left[\left\langle \boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*, W_0(\boldsymbol{\theta}_{l-1}) - W_0(\boldsymbol{\theta}^*)\right\rangle \mathbb{1}(A_m)\right], \quad \text{(A.4)}
\end{aligned}$$

due to the assumptions **(b)** and **(f)**. Moreover, applying the identity

$$\mathbb{E}_0\left[W(\overline{\boldsymbol{\theta}}_m^{k-1}, \overline{\boldsymbol{\xi}}_m^{k-1})\mathbb{1}(A_m)\right] = \mathbb{E}_0\left[W(\boldsymbol{\theta}^*, \overline{\boldsymbol{\xi}}_m^{k-1})\mathbb{1}(A_m)\right] + 2\mathbb{E}_0\left[\left\langle \overline{\boldsymbol{\xi}}_m^{k-1}, W_0(\overline{\boldsymbol{\theta}}_m^{k-1}) - W_0(\boldsymbol{\theta}^*)\right\rangle \mathbb{1}(A_m)\right], \tag{A.5}$$

and taking double summation $n^{-1}\sum_{m=0}^n \sum_{k=m+1}^n$, we get

$$\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} W(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}, \overline{\boldsymbol{\xi}}_{\tau_n}^{k-1})\right] \leq W(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)\left(1-\frac{\mathbb{E}_0[\tau_n]}{n}\right) + \frac{1}{2n}\mathbb{E}_0\left[\operatorname{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=\tau_n+1}^{n}\frac{1-\epsilon_{\tau_n}/6}{\sum_{t=\tau_n+1}^{k}\epsilon_{t-1}}\right]$$
$$+\frac{1}{2n}\mathbb{E}_0\left[J^2(\mathscr{T}_{\tau_n}, \mathscr{X}_{\tau_n}; \boldsymbol{\lambda}^\star)\sum_{k=\tau_n+1}^{n}\frac{\sum_{l=\tau_n+1}^{k}\epsilon_{l-1}^2}{\sum_{t=\tau_n+1}^{k}\epsilon_{t-1}}\right]$$
$$-2\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\sum_{l=\tau_n+1}^{k}\frac{\epsilon_{l-1}}{\sum_{t=\tau_n+1}^{k}\epsilon_{t-1}}\left\langle \boldsymbol{\xi}_{l-1}-\boldsymbol{\xi}^*, W_0(\boldsymbol{\theta}_{l-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right]$$
$$+2\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\left\langle \overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}, W_0(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right]. \tag{A.6}$$

Next, by combining two inequalities (A.2) and (A.6) and setting arbitrary (yet strictly positive) constant learning rates $\gamma_k \equiv \gamma(>0)$ and $\epsilon_k \equiv \epsilon(>0)$, we obtain the inequality (4.6). Note that all random vectors $\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}$ and $\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}$ in the function $\Upsilon_d(n)$ above are independent of the constant learning rates γ and ϵ, since then $\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1} = \sum_{l=\tau_n}^{k-1}\frac{\gamma}{\sum_{t=\tau_n}^{k-1}\gamma}\boldsymbol{\theta}_l = \frac{1}{k-\tau_n}\sum_{l=\tau_n}^{k-1}\boldsymbol{\theta}_l$ and $\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1} = \sum_{l=\tau_n}^{k-1}\frac{\epsilon}{\sum_{t=\tau_n}^{k-1}\epsilon}\boldsymbol{\xi}_l = \frac{1}{k-\tau_n}\sum_{l=\tau_n}^{k-1}\boldsymbol{\xi}_l$. Also, note however that the expectation in the line (A.6) cannot be replaced with $\Upsilon_d(n)$ as of yet, since the learning rates are generally not constant at the stage of (A.6). Hence, regarding the minimization, it suffices to examine the first two terms $\Upsilon_a(n;\gamma)$ and $\Upsilon_b(n;\epsilon)$, and moreover separately. First, it holds that for every $n \in \mathbb{N}$ and $\gamma > 0$,

$$\Upsilon_a(n;\gamma) \geq \frac{1}{n}\left(\mathbb{E}_0\left[\operatorname{diam}^2(\mathscr{T}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right]\mathbb{E}_0\left[L^2(\mathscr{T}_{\tau_n};\boldsymbol{\lambda}^\star)(n-\tau_n)\right]\right)^{1/2}$$
$$-\frac{1}{2n}\mathbb{E}_0\left[\alpha_{\tau_n}\operatorname{diam}^2(\mathscr{T}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right] = \Upsilon_a(n;\gamma(n)), \tag{A.7}$$

where the inequality holds true uniformly in $\gamma(>0)$ with equality uniquely with the constant $\gamma(n)$ given by (4.3). In a similar manner, it holds that for every $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\Upsilon_b(n;\epsilon) \geq \frac{1}{n}\left(\mathbb{E}_0\left[\operatorname{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right]\mathbb{E}_0\left[J^2(\mathscr{T}_{\tau_n},\mathscr{X}_{\tau_n};\boldsymbol{\lambda}^\star)(n-\tau_n)\right]\right)^{1/2}$$
$$-\frac{1}{12n}\mathbb{E}_0\left[\operatorname{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=1}^{n-\tau_n}k^{-1}\right] = \Upsilon_b(n;\epsilon(n)), \tag{A.8}$$

where the inequality holds true uniformly in $\epsilon > 0$ with equality uniquely with the constant $\epsilon(n)$ given in (4.3). It is now straightforward from the expressions (A.7) and (A.8) that $\Upsilon_a(n;\gamma(n)) = \mathscr{O}(\sqrt{\ln(n)/n})$ and $\Upsilon_b(n;\epsilon(n)) = \mathscr{O}(\sqrt{\ln(n)/n})$, as $n \uparrow +\infty$.

(ii) It suffices to set $\epsilon_k \equiv 0$, which yields $W(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}, \overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}) = W(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1}, 0_d) = 0$, $\mathbb{P}_0$-*a.s.*

(iii) It remains to derive the convergence of the two terms $\Upsilon_c(n)$ and $\Upsilon_d(n)$ to zero as $n \uparrow +\infty$. To this end, we prepare some auxiliary results. Recall (Assumption 2.1 **(g)**) that the function $G(\mathbf{z};\boldsymbol{\theta})$ is Lipschitz (in $\boldsymbol{\theta}$) on every compact subset $\mathscr{T}_k$ of Θ_2, uniformly in $\mathbf{z}$ (on $\mathscr{D}$). Hence, there exists $c > 0$ such that for every $k \in \{0, 1, \cdots, n\}$ and $\boldsymbol{\theta} \in \mathscr{T}_k$,

$$\|W_0(\boldsymbol{\theta}) - W_0(\boldsymbol{\theta}^*)\|^2 = \left\|\int_{(0,1)^d}\Psi(\mathbf{u})\left(G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}) - G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}^*)\right)d\mathbf{u}\right\|^2$$

$$\leq \int_{(0,1)^d} |\Psi(\mathbf{u})|^2 \left\|G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}) - G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta}^*)\right\|^2 d\mathbf{u} \leq c^2 V(\boldsymbol{\theta}_0) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2,$$

where the first inequality holds by the Cauchy-Schwarz inequality. Moreover, due to the assumption $\liminf_{k\uparrow+\infty} \alpha_k > 0$, there exist $\widetilde{\alpha} > 0$ and $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, $(\widetilde{\alpha}/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \leq V(\boldsymbol{\theta}) - V(\boldsymbol{\theta}^*)$ on the compact set $\mathscr{T}_k$. It thus holds that for all $\boldsymbol{\theta} \in \mathscr{T}_0$,

$$\begin{aligned}\|W_0(\boldsymbol{\theta}) - W_0(\boldsymbol{\theta}^*)\|^2 &\leq c^2 V(\boldsymbol{\theta}_0) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \\ &\leq \frac{2c^2 V(\boldsymbol{\theta}_0)}{\widetilde{\alpha}} (V(\boldsymbol{\theta}) - V(\boldsymbol{\theta}^*)) \mathbb{1}(\boldsymbol{\theta} \in \mathscr{T}_{k_0}) + c^2 V(\boldsymbol{\theta}_0) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \mathbb{1}(\boldsymbol{\theta} \notin \mathscr{T}_{k_0}), \end{aligned} \quad (A.9)$$

which further yields

$$\begin{aligned}&\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} \|W_0(\boldsymbol{\theta}_{l-1}) - W_0(\boldsymbol{\theta}^*)\|^2\right] \\ &\leq \frac{2c^2 V(\boldsymbol{\theta}_0)}{\widetilde{\alpha}}\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} (V(\boldsymbol{\theta}_{l-1}) - V(\boldsymbol{\theta}^*)) \mathbb{1}(\boldsymbol{\theta}_{l-1} \in \mathscr{T}_{k_0})\right] \\ &\quad + c^2 V(\boldsymbol{\theta}_0)\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} \|\boldsymbol{\theta}_{l-1} - \boldsymbol{\theta}^*\|^2 \mathbb{1}(\boldsymbol{\theta}_{l-1} \notin \mathscr{T}_{k_0})\right] \\ &\lesssim \frac{2c^2 V(\boldsymbol{\theta}_0)}{\widetilde{\alpha}}\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} (V(\boldsymbol{\theta}_{l-1}) - V(\boldsymbol{\theta}^*))\right] \\ &\leq \frac{2c^2 V(\boldsymbol{\theta}_0)}{\widetilde{\alpha}}\Upsilon_a(n;\gamma(n)),\end{aligned}$$

where we have applied the inequalities (A.1), (A.2) and (A.9) with the constant learning rate $\gamma(n)$ and the second inequality holds asymptotically as $n \uparrow +\infty$, up to a constant due to the fact $\boldsymbol{\theta}_{l-1} \in \mathscr{T}_{l-1}$. Moreover, it holds that for each $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} \left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2\right] &\leq 12\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} \left(W(\boldsymbol{\theta}^*,\boldsymbol{\xi}_{l-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\right)\right] \\ &\leq 12\left(\Upsilon_b(n;\epsilon(n)) + 2\Upsilon_c(n)\right),\end{aligned}$$

where we have applied the inequalities (A.4), (A.5) and (A.6) with the constant learning rate $\epsilon(n)$, and the strong convexity $(1/12)\|\boldsymbol{\xi} - \boldsymbol{\xi}^*\|^2 \leq W(\boldsymbol{\theta}^*,\boldsymbol{\xi}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$. On the whole, with $\boldsymbol{\theta}_{l-1} \in \mathscr{T}_{k_0}$ for all $l \in \{k_0,\cdots,n\}$, it holds that

$$\begin{aligned}|\Upsilon_c(n)|^2 &\leq \mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} \|W_0(\boldsymbol{\theta}_{l-1}) - W_0(\boldsymbol{\theta}^*)\|^2\right]\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} \frac{1}{k-\tau_n}\sum_{l=\tau_n+1}^{k} \left\|\boldsymbol{\xi}_{l-1} - \boldsymbol{\xi}^*\right\|^2\right] \\ &\lesssim \frac{24c^2 V(\boldsymbol{\theta}_0)}{\widetilde{\alpha}}\Upsilon_a(n;\gamma(n))\left(\Upsilon_b(n;\epsilon(n)) + \Upsilon_c(n)\right),\end{aligned}$$

where we have applied the Cauchy-Schwarz inequality for the first inequality. With the aid of $\Upsilon_a(n;\gamma(n)) = \mathscr{O}(\sqrt{\ln(n)/n})$ and $\Upsilon_b(n;\epsilon(n)) = \mathscr{O}(\sqrt{\ln(n)/n})$, this yields $|\Upsilon_c(n)| = o(\sqrt{\ln(n)/n})$. Similarly, using the inequality (A.9) with an additional convexity argument, it holds $\mathbb{P}_0$-$a.s.$ that

$$\begin{aligned}
&\left\|W_0(\overline{\boldsymbol{\theta}}^{k-1}_{\tau_n}) - W_0(\boldsymbol{\theta}^*)\right\|^2 \\
&\le c^2 V(\boldsymbol{\theta}_0)\left\|\overline{\boldsymbol{\theta}}^{k-1}_{\tau_n} - \boldsymbol{\theta}^*\right\|^2 \le \frac{c^2 V(\boldsymbol{\theta}_0)}{k-\tau_n}\sum_{l=\tau_n+1}^{k}\|\boldsymbol{\theta}_{l-1}-\boldsymbol{\theta}^*\|^2 \\
&\le \frac{c^2 V(\boldsymbol{\theta}_0)}{k-\tau_n}\sum_{l=\tau_n+1}^{k}\left[\frac{2}{\widetilde{\alpha}}\left(V(\boldsymbol{\theta}_{l-1})-V(\boldsymbol{\theta}^*)\right)\mathbb{1}(\boldsymbol{\theta}_{l-1}\in\mathscr{T}_{k_0}) + \|\boldsymbol{\theta}_{l-1}-\boldsymbol{\theta}^*\|^2\,\mathbb{1}(\boldsymbol{\theta}_{l-1}\notin\mathscr{T}_{k_0})\right],
\end{aligned}$$

which yields

$$\begin{aligned}
|\Upsilon_d(n)|^2 &\le \mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\left\|W_0(\overline{\boldsymbol{\theta}}^{k-1}_{\tau_n}) - W_0(\boldsymbol{\theta}^*)\right\|^2\right]\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\left\|\overline{\boldsymbol{\xi}}^{k-1}_{\tau_n}\right\|^2\right] \\
&\lesssim \frac{2c^2V(\boldsymbol{\theta}_0)\mathrm{diam}^2(\mathscr{X}_0)}{\widetilde{\alpha}}\left(1-\frac{\mathbb{E}_0[\tau_n]}{n}\right)\Upsilon_a(n;\gamma(n)).
\end{aligned}$$

Hence, $|\Upsilon_d(n)| = \mathscr{O}(\sqrt[4]{\ln(n)/n})$, as $n \uparrow +\infty$. □

As is clear from the presence of the convexity parameter α_{τ_n} in the second term of the minimized upper bound (A.7), on the one hand, a strong convexity of the expected value function $V(\boldsymbol{\theta})$ contributes to lower the upper bound. The proposed algorithm allows one to enjoy this contribution free of charge, in the sense that the prior knowledge of the convexity parameter α is not required for implementation. On the other hand, the second term of the minimized upper bound (A.8) is due to the strong convexity of the function $W(\boldsymbol{\theta},\boldsymbol{\xi})$ in $\boldsymbol{\xi}$ with parameter $1/6$. The strong convexity benefit in the both terms (A.7) and (A.8) decay in the order $\mathscr{O}(\ln(n)/n)$, that is, quadratically as fast as the respective first leading terms of order $\mathscr{O}(\sqrt{\ln(n)/n})$. Hence, the more computing budget, the more prominent the first $\mathscr{O}(\sqrt{\ln(n)/n})$-order terms will be.

It is worth mentioning that the inequality (4.6) holds true even when the second moment function $V(\boldsymbol{\theta})$ is not strongly convex. Moreover, as desired, the two minimized terms $\Upsilon_a(n;\gamma(n))$ and $\Upsilon_b(n;\epsilon(n))$ of the upper bound (4.6) tend to zero, again without strong convexity of $V(\boldsymbol{\theta})$. The strong convexity of $V(\boldsymbol{\theta})$ is only employed for the convergences of the remaining two residual terms $\Upsilon_c(n)$ and $\Upsilon_d(n)$, more precisely, for the inequality (A.9). In the absence of a particular structure of the function $W_0(\boldsymbol{\theta})$ (in particular, this function is far from convex/concave), we have not been able to place appropriate upper bounds for the remaining two residual terms $\Upsilon_c(n)$ and $\Upsilon_d(n)$ without the strong convexity requirement as of yet. Nevertheless, as we have observed through numerical results (Section 5), the constant learning rates $\gamma(n)$ and $\epsilon(n)$ seem to realize a very low estimator variance, in fact, fairly close to the intended minimum $V(\boldsymbol{\theta}^*)+W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)-\mu^2$, given a relatively large yet fixed computing budget. The strong convexity requirement is indeed sufficient to lead the upper bound to attain the intended minimum (Theorem 4.1 **(iii)**), whereas we conjecture that, in practice, this requirement is not necessary for the actual estimator variance (not its upper bound) to attain the intended minimum.

Proof of Theorem 4.2. The derivation of those results entails somewhat repetitive algebraic work similar to the proof of Theorem 4.1 and, for instance, [9, Theorem 3.4]. To avoid overloading the paper, we omit nonessential details from place to place.

(i) Define $\phi_k := R(U_k;\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1})-\mu$, which forms a martingale difference sequence with respect to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$. The stochastic process $\{\sum_{k=1}^n \phi_k : n\in\mathbb{N}\}$ is a square integrable martingale with respect to the filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$, where the second moment of each summand is bounded by $\sup_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathscr{T}_0\times\mathscr{X}_0}(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})) < +\infty$. Moreover, we have $n^{-1}\sum_{k=1}^n \mathbb{E}_{k-1}[|\phi_k|^2] = n^{-1}\sum_{k=1}^n(V(\widetilde{\boldsymbol{\theta}}_{k-1})+W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}))-\mu^2$, whose limit superior in n is almost surely finite by the standing assumption. Therefore, we obtain the desired convergence $n^{-1}\sum_{k=1}^n\phi_k = \mu(n)-\mu \to 0$, $\mathbb{P}_0$-*a.s.*

(ii) Define $\eta_{a,k} := N(U_k;\widetilde{\boldsymbol{\theta}}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n}) - V(\widetilde{\boldsymbol{\theta}}_{k-1})$ and $\eta_{b,k} := S(U_k;\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n}) - W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1})$, each of which forms a martingale difference sequence with respect to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$, so that

$$\sigma^2(n) = \frac{1}{n}\sum_{k=1}^{n}\eta_{a,k} + \frac{1}{n}\sum_{k=1}^{n}\eta_{b,k} + \frac{1}{n}\sum_{k=1}^{n}\left[V(\widetilde{\boldsymbol{\theta}}_{k-1}) + W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1})\right] - \mu^2(n). \tag{A.10}$$

We first show the first two terms in (A.10) converge $\mathbb{P}_0$-*a.s.* to zero. Using the notation

$$S_0(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda}) := \int_{(0,1)^d} (S(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda}))^2\, d\mathbf{u} = 4\left\langle \boldsymbol{\xi}, \int_{(0,1)^d} (Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}))^{\otimes 2} d\mathbf{u}\,\boldsymbol{\xi}\right\rangle + \frac{\|\boldsymbol{\xi}\|^2}{6}W(\boldsymbol{\theta},\boldsymbol{\xi}) - \frac{\|\boldsymbol{\xi}\|^4}{144},$$

observe that for each $n\in\mathbb{N}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{k-1}\left[|\eta_k^a|^2\right] + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{k-1}\left[|\eta_k^b|^2\right] \\
&= \frac{1}{n}\sum_{k=1}^{n}\Bigg[\int_{(0,1)^d} H(\mathbf{u};\boldsymbol{\lambda}_{(k-1)\wedge\tau_n},\boldsymbol{\theta}_0)\left|H(\mathbf{u};\widetilde{\boldsymbol{\theta}}_{k-1},\boldsymbol{\theta}_0)\right|^2|\Psi(\mathbf{u})|^4 d\mathbf{u} + S_0(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n}) - (V(\widetilde{\boldsymbol{\theta}}_{k-1}))^2 \\
&\qquad\qquad - (W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}))^2\Bigg],
\end{aligned}$$

where the (averaged) first term here is almost surely finite in the limit by the standing assumption, the (averaged) second term inside is finite since $\sup_{(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda})\in\mathscr{T}_0\times\mathscr{X}_0\times\Lambda_0}|S_0(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\lambda})| < +\infty$ due to $|\langle\boldsymbol{\xi},\int_{(0,1)^d}(Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}))^{\otimes 2}d\mathbf{u}\boldsymbol{\xi}\rangle| \le c_d\|\boldsymbol{\xi}\|^2\int_{(0,1)^d}H(\mathbf{u};\boldsymbol{\lambda},\boldsymbol{\theta}_0)|\Psi(\mathbf{u})|^2 d\mathbf{u}$ and the assumption (2.8) with $\Lambda_0\subseteq\Theta_1$, for a suitable positive constant c_d depending on the dimension d. The (averaged) last two terms are finite in the limit as well by the standing assumption. Hence, it suffices to show that $n^{-1}\sum_{k=1}^{n}(V(\widetilde{\boldsymbol{\theta}}_{k-1}) + W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1})) \to V(\boldsymbol{\theta}^*) + W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$ in $L^1(\Omega,\mathscr{F},\mathbb{P}_0)$.

The importance sampling component is rather straightforward, with the aid of the proof of Theorem 4.1. Using the definition of the minimizer $\boldsymbol{\theta}^*$, that is, $V(\boldsymbol{\theta})\ge V(\boldsymbol{\theta}^*)$ on the domain Θ_2, it holds that

$$\begin{aligned}
\mathbb{E}_0\left[\left|\frac{1}{n}\sum_{k=1}^{n}V(\widetilde{\boldsymbol{\theta}}_{k-1}) - V(\boldsymbol{\theta}^*)\right|\right] &= \mathbb{E}_0\left[\frac{1}{n}\sum_{k=1}^{n}V(\widetilde{\boldsymbol{\theta}}_{k-1})\right] - V(\boldsymbol{\theta}^*) \\
&\le (V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}^*))\frac{\mathbb{E}_0[\tau_n]}{n} + \mathbb{E}_0\left[\mathrm{diam}^2(\mathscr{T}_{\tau_n})\frac{1}{2n}\sum_{k=\tau_n+1}^{n}\frac{1-2\alpha_{\tau_n}\gamma_{\tau_n}}{\sum_{t=\tau_n+1}^{k}\gamma_{t-1}}\right] \\
&\qquad + \mathbb{E}_0\left[L^2(\mathscr{T}_{\tau_n};\boldsymbol{\lambda}^\star)\frac{1}{2n}\sum_{k=\tau_n+1}^{n}\frac{\sum_{l=\tau_n+1}^{k}\gamma_{l-1}^2}{\sum_{t=\tau_n+1}^{k}\gamma_{t-1}}\right] \to 0,
\end{aligned}$$

as $n\uparrow+\infty$, due to the inequality (A.2). Note that the second and third terms tend to zero faster than the case with constant learning rates of Theorem 4.1.

The control variates component needs to be dealt with separately for its positive and negative parts, as the value $W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$ may not be the joint global minimum of the function $W(\boldsymbol{\theta},\boldsymbol{\xi})$ on the domain $\Theta_2\times\mathbb{R}^d$. For the positive part, observe that

$$\begin{aligned}
&\mathbb{E}_0\left[\frac{1}{n}\sum_{k=1}^{n} W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\right]\\
&\quad= \mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n} W(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1},\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\right]\\
&\quad\leq -W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\frac{\mathbb{E}_0[\tau_n]}{n} + \frac{1}{2n}\mathbb{E}_0\left[\mathrm{diam}^2(\mathscr{X}_{\tau_n})\sum_{k=\tau_n+1}^{n}\frac{1-\epsilon_{\tau_n}/6}{\sum_{t=\tau_n+1}^{k}\epsilon_{t-1}}\right]\\
&\qquad+\frac{1}{2n}\mathbb{E}_0\left[J^2(\mathscr{T}_{\tau_n},\mathscr{X}_{\tau_n};\boldsymbol{\lambda}^{\star})\sum_{k=\tau_n+1}^{n}\frac{\sum_{l=\tau_n+1}^{k}\epsilon_{l-1}^2}{\sum_{t=\tau_n+1}^{k}\epsilon_{t-1}}\right]\\
&\qquad-2\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\sum_{l=\tau_n+1}^{k}\frac{\epsilon_{l-1}}{\sum_{t=\tau_n+1}^{k}\epsilon_{t-1}}\left\langle\boldsymbol{\xi}_{l-1}-\boldsymbol{\xi}^*, W_0(\boldsymbol{\theta}_{l-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right]\\
&\qquad+2\mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\left\langle\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}, W_0(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right]\to 0,
\end{aligned}$$

due to the inequality (A.6), while for the negative part,

$$\begin{aligned}
\mathbb{E}_0\left[\frac{1}{n}\sum_{k=1}^{n} W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\right] &= \mathbb{E}_0\left[\frac{1}{n}\sum_{k=1}^{n}\Big(W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\xi}}_{k-1})\Big)\right]\\
&\quad+\mathbb{E}_0\left[\frac{1}{n}\sum_{k=1}^{n}\Big(W(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\Big)\right]\\
&\geq \mathbb{E}_0\left[\frac{1}{n}\sum_{k=1}^{n}\Big(W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\xi}}_{k-1})\Big)\right]\\
&= \mathbb{E}_0\left[\frac{1}{n}\sum_{k=\tau_n+1}^{n}\left\langle\overline{\boldsymbol{\xi}}_{\tau_n}^{k-1}, W_0(\overline{\boldsymbol{\theta}}_{\tau_n}^{k-1})-W_0(\boldsymbol{\theta}^*)\right\rangle\right]\to 0,
\end{aligned}$$

due to the definition (2.16) of the minimizer $\boldsymbol{\xi}^*$, where both convergences hold true in a similar manner to, and again faster than, the case with constant learning rates of Theorem 4.1.

(iii) It suffices to verify the Lindeberg condition for the martingale central limit theorem, due to the rewriting

$$\sqrt{n}\frac{\mu(n)-\mu}{\sigma(n)} = \sqrt{n}\frac{n^{-1}\sum_{k=1}^{n}\phi_k}{(n^{-1}\sum_{k=1}^{n}\mathbb{E}_{k-1}[|\phi_k|^2])^{1/2}}\left(\frac{n^{-1}\sum_{k=1}^{n}\mathbb{E}_{k-1}[|\phi_k|^2]}{\sigma^2(n)}\right)^{1/2}.$$

First, by applying the Minkowski inequality twice, we obtain that for each $k\in\mathbb{N}$,

$$\mathbb{E}_{k-1}\left[|\phi_k|^q\right]^{1/q} \leq \left(\int_{(0,1)^d}\left|R(\mathbf{u};\widetilde{\boldsymbol{\theta}}_{k-1})\right|^q d\mathbf{u}\right)^{1/q} + \frac{\sqrt{d}\|\widetilde{\boldsymbol{\xi}}_{k-1}\|}{2} + |\mu|.$$

By the Hölder inequality and the Markov inequality, it holds that for each $\epsilon>0$,

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{k-1}\left[|\phi_k|^2\mathbb{1}\left(|\phi_k|>\epsilon\sqrt{n}\right)\right]$$

$$\leq \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{k-1}\left[|\phi_k|^q\right]^{2/q}\left(\mathbb{P}_{k-1}\left(|\phi_k|>\epsilon\sqrt{n}\right)\right)^{1-2/q}$$

$$\leq \frac{1}{n}\sum_{k=1}^{n}\left[\left(\int_{(0,1)^d}\left|R(\mathbf{u};\widetilde{\boldsymbol{\theta}}_{k-1})\right|^q d\mathbf{u}\right)^{1/q}+\frac{\sqrt{d}\|\widetilde{\boldsymbol{\xi}}_{k-1}\|}{2}+|\mu|\right]^2\left(\frac{\mathbb{E}_{k-1}[|\phi_k|^2]}{\epsilon^2 n}\right)^{1-2/q}$$

$$\leq \frac{1}{(\epsilon^2 n)^{1-2/q}}\sup_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathscr{T}\times\mathscr{X}}(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})-\mu^2)^{1-2/q}$$

$$\times\frac{1}{n}\sum_{k=1}^{n}\left[\left(\int_{(0,1)^d}\left|H(\mathbf{u};\widetilde{\boldsymbol{\theta}}_{k-1},\boldsymbol{\theta}_0)\right|^{q-1}|\Psi(\mathbf{u})|^q d\mathbf{u}\right)^{1/q}+\frac{\sqrt{d}\|\widetilde{\boldsymbol{\xi}}_{k-1}\|}{2}+|\mu|\right]^2,$$

which converges $\mathbb{P}_0$-*a.s.* to zero as $n\uparrow+\infty$, due to $1-2/q>0$. □

Proof of Theorem 4.3. We focus on **(i)** and **(iii)**, as the proof of Theorem 4.2 **(i)** and **(iii)** directly applies, respectively, to **(ii)** and **(iv)** here.

(i) The almost sure convergence $\widetilde{\boldsymbol{\theta}}_k\to\boldsymbol{\theta}^*$ holds true along the subsequence $\{k\in\mathbb{N}:\ell_a(k)=1\}$, just as proved in [9, Theorem 3.3], irrespective of the presence of the control variates component, as $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}_0}$ remains adapted to the filtration $(\mathscr{F}_k)_{k\in\mathbb{N}_0}$. Therefore, in light of the iteration (3.6), it suffices to show that $k^{-1}\sum_{j=1}^{k}Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})$ converges $\mathbb{P}_0$-*a.s.*, along the subsequence $\{k\in\mathbb{N}:\ell_b(k)=1\}$, to $\int_{(0,1)^d}Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}^\star)d\mathbf{u}$ uniformly in $\boldsymbol{\theta}$ on a neighborhood of the (deterministic) limiting point $\boldsymbol{\theta}^*$. Let A be a compact subset of Θ_2 with $\boldsymbol{\theta}^*\in A$ and fix $\boldsymbol{\theta}\in A$. It holds $\mathbb{P}_0$-*a.s.* that for every $k>\tau_n$,

$$\frac{1}{k}\sum_{j=1}^{k}Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})=\frac{1}{k}\sum_{j=1}^{k}Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}^\star)=\frac{1}{k}\sum_{j=1}^{\tau_n}Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}^\star)+\frac{k-\tau_n}{k}\frac{1}{k-\tau_n}\sum_{j=\tau_n+1}^{k}Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}^\star)$$
$$\to\int_{(0,1)^d}Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}^\star)d\mathbf{u}=\int_{(0,1)^d}\Psi(\mathbf{u})(G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta})-\mathbb{1}_d/2)d\mathbf{u}=:\varsigma(\boldsymbol{\theta}),\tag{A.11}$$

where the convergence holds true by the random sum strong law of large numbers since τ_n is bounded by n and the tail $\{U_j\}_{j>\tau_n}$ is independent of the $\mathscr{F}_{\tau_n}$-measurable stopped point $\boldsymbol{\lambda}^\star$. Next, it holds $\mathbb{P}_0$-*a.s.* that for every $k>\tau_n$,

$$\sup_{\boldsymbol{\zeta}\in A\cap B_\epsilon[\boldsymbol{\theta}]}\left\|\frac{1}{k}\sum_{j=1}^{k}Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})-\frac{1}{k}\sum_{j=1}^{k}Q(U_j;\boldsymbol{\zeta},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})\right\|$$
$$\leq\frac{1}{k}\sum_{j=1}^{k}\sup_{\boldsymbol{\zeta}\in A\cap B_\epsilon[\boldsymbol{\theta}]}\|Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}^\star)-Q(U_j;\boldsymbol{\zeta},\boldsymbol{\lambda}^\star)\|$$
$$\to\int_{(0,1)^d}\sup_{\boldsymbol{\zeta}\in A\cap B_\epsilon[\boldsymbol{\theta}]}\|Q(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\lambda}^\star)-Q(\mathbf{u};\boldsymbol{\zeta},\boldsymbol{\lambda}^\star)\|\,d\mathbf{u}$$
$$=\int_{(0,1)^d}|\Psi(\mathbf{u})|\sup_{\boldsymbol{\zeta}\in A\cap B_\epsilon[\boldsymbol{\theta}]}\left\|G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\theta})-G(G^{-1}(\mathbf{u};\boldsymbol{\theta}_0);\boldsymbol{\zeta})\right\|d\mathbf{u}$$

$$\leq \epsilon c_\epsilon \int\limits_{(0,1)^d} |\Psi(\mathbf{u})| d\mathbf{u},$$

where the convergence holds true as $k \uparrow +\infty$ by the random sum strong law of large numbers, the equality holds true by a change of variables, and the last inequality holds by Assumption 2.1 **(g)** with a suitable positive constant c_ϵ, which is non-increasing as $\epsilon \downarrow 0$.

Hence, there exist a finite number of points $\{\boldsymbol{\theta}_{(l)}\}_{l=1,\cdots,m}$ in the compact set A and respective neighborhoods $\{A_l\}_{l=1,\cdots,m}$, with $A \subseteq \cup_{l=1,\cdots,m} A_l$, such that for sufficiently large $k(> \tau_n)$, $\sup_{\boldsymbol{\zeta}\in A_l} \|k^{-1}\sum_{j=1}^k Q(U_j;\boldsymbol{\theta}_{(l)},\boldsymbol{\lambda}^\star) - k^{-1}\sum_{j=1}^k Q(U_j;\boldsymbol{\zeta},\boldsymbol{\lambda}^\star)\| < \epsilon$, $\mathbb{P}_0$-*a.s.*, as well as $\sup_{\boldsymbol{\zeta}\in A_l} \|\varsigma(\boldsymbol{\theta}_{(l)}) - \varsigma(\boldsymbol{\zeta})\| < \epsilon$, for every $l = 1,\cdots,m$ and $q = 0,1,2$, thanks to the continuity of the deterministic function $\varsigma(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$. Moreover, the result (A.11) asserts that for sufficiently large $k(> \tau_n)$, $\|k^{-1}\sum_{j=1}^k Q(U_j;\boldsymbol{\theta}_{(l)},\boldsymbol{\lambda}^\star) - \varsigma(\boldsymbol{\theta}_{(l)})\| < \epsilon$, $\mathbb{P}_0$-*a.s.*, and thus overall, for sufficiently large $k(> \tau_n)$, $\sup_{\boldsymbol{\theta}\in A} \|k^{-1}\sum_{j=1}^k Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}^\star) - \varsigma(\boldsymbol{\theta})\| < 3\epsilon$, $\mathbb{P}_0$-*a.s.* This asserts the desired almost sure uniform convergence of $k^{-1}\sum_{j=1}^k Q(U_j;\boldsymbol{\theta},\boldsymbol{\lambda}_{(k-1)\wedge\tau_n})$ to the deterministic function $\varsigma(\boldsymbol{\theta})$ on a neighborhood of the optimal point $\boldsymbol{\theta}^*$. Hence, the continuous mapping theorem yields the joint almost sure convergence $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\xi}}_k) \to (\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$.

(iii) The convergence to μ follows directly from **(ii)**. For the remainder, as we have shown in the proof of Theorem 4.2 (ii), it suffices to show the almost sure convergence $n^{-1}\sum_{k=1}^n (V(\widetilde{\boldsymbol{\theta}}_{k-1}) + W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1})) \to V(\boldsymbol{\theta}^*) + W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$. The convergence to $V(\boldsymbol{\theta}^*)$ is rather straightforward since $V_n(\boldsymbol{\theta}) := n^{-1}\sum_{k=1}^n V(\boldsymbol{\theta})$ (which is, in fact, $V(\boldsymbol{\theta})$ itself irrespective of n) is uniformly convergent to V on a neighborhood of $\boldsymbol{\theta}^*$ by Proposition 2.2, as well as due to the convergence result **(i)**. To show the convergence to $W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$, observe first that

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^n W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*) &= \frac{1}{n}\sum_{k=1}^n \left(W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\xi}}_{k-1})\right) \\ &\quad + \left[\frac{1}{n}\sum_{k=1}^n W(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)\right],\end{aligned}$$

where the second term on the righthand side clearly tends to zero $\mathbb{P}_0$-*a.s.* due to its simple quadratic structure (2.7) in $\boldsymbol{\xi}$, as well as due to the convergence result **(i)**. The first term tends to zero as well, since $|n^{-1}\sum_{k=1}^n (W(\widetilde{\boldsymbol{\theta}}_{k-1},\widetilde{\boldsymbol{\xi}}_{k-1}) - W(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\xi}}_{k-1}))| \leq cn^{-1}\sum_{k=1}^n \|\widetilde{\boldsymbol{\xi}}_{k-1}\|\|\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*\| \to 0$, for a suitable positive constant c, due to Assumption 2.1 **(g)**. □

Appendix B. Technical notes

We collect some relevant technical details, which are important complements to the development and analysis in the main body.

B.1. Importance sampling first, or control variates first?

As is clear from the progression (2.2), we apply change of measure first, and then introduce the control variates term without being influenced by the first change of measure. A natural question here is what if the control variates term was introduced first and then change of underlying measure, that is,

$$\mu = \int\limits_{(0,1)^d} H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta})\left(\Psi\left(G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0)\right) + \left\langle \boldsymbol{\xi}, G(G^{-1}(\mathbf{u};\boldsymbol{\theta});\boldsymbol{\theta}_0) - \mathbb{1}_d/2\right\rangle\right) d\mathbf{u}, \tag{B.1}$$

whose estimator variance is then given by

$$\left(V(\boldsymbol{\theta})+W(\boldsymbol{\theta},\boldsymbol{\xi})-\mu^2\right)+\left\langle \boldsymbol{\xi}, \int\limits_{(0,1)^d} \left(H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)-1\right)\left(\mathbf{u}-\mathbb{1}_d/2\right)^{\otimes 2} d\mathbf{u}\boldsymbol{\xi}\right\rangle, \tag{B.2}$$

where the rightmost term (quadratic in $\boldsymbol{\xi}$) indicates the difference from the estimator variance (2.5). It seems unclear, without numerical experiments, which expression, (2.2) or (B.1), reduces more variance, or more precisely, whether the rightmost term of (B.2) is positive or negative. The estimator variance (B.2) is much more demanding from both computational and theoretical standpoints, due to the additional integral, involving the likelihood ratio $H(\mathbf{u};\boldsymbol{\theta},\boldsymbol{\theta}_0)$, in the quadratic term. As our primary interest is not only the resulting performance (how much variance can be reduced), but also the ease of implementation, we adopt the original approach (2.2) and do not treat the alternative approach (B.1) in the present work.

B.2. Nonlinear variates

The control variates we have adopted is one of the simplest forms and can be generalized in various ways. For instance, the linear variates $\langle \boldsymbol{\xi}, \mathbf{u}-\mathbb{1}_d/2\rangle$ in the integrand (2.2) may be replaced with, for instance, a more general form $\langle \boldsymbol{\xi}, C(\mathbf{u})\rangle$, where $C:(0,1)^d\to\mathbb{R}^{d_c}$, $\int_{(0,1)^d} C(\mathbf{u})d\mathbf{u}=0_{d_c}$ and $\boldsymbol{\xi}\in\mathbb{R}^{d_c}$ for a suitable dimension d_c, provided that the variance-covariance matrix $\int_{(0,1)^d}(C(\mathbf{u}))^{\otimes 2}d\mathbf{u}\in\mathbb{R}^{d_c\times d_c}$ is known (or, at least readily computable with high accuracy). However, given that our focus is the integrand $\Psi(\mathbf{u})$ in the most general form (1.1), we adopted the simple linear variates $\langle \boldsymbol{\xi}, \mathbf{u}-\mathbb{1}_d/2\rangle$ with a view towards a general integrand $\Psi(\mathbf{u})$, rather than a general nonlinear variates $\langle \boldsymbol{\xi}, C(\mathbf{u})\rangle$ for a specialized integrand.

B.3. Joint suboptimality and perfect variance reduction

As a simple one-dimensional example for illustration purpose, consider $\Psi(u)=u^q$ for some $q>-1$, with the expectation $\mu=\int_0^1\Psi(u)du=(1+q)^{-1}$ and the crude second moment $V(\theta_0)=\int_0^1(\Psi(u))^2du=(1+2q)^{-1}$. (At the moment, we let θ_0 only represent no importance sampling.) To emphasize that the problem setting is one-dimensional, we make the letters $\mathbf{u}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ unbold throughout this example.

First, we employ the exponential bypass distribution $g(z;\theta)=\theta e^{-\theta z}$, $G(z;\theta)=e^{-\theta z}$, and $G^{-1}(u;\theta)=-\theta^{-1}\ln(u)$, with $\theta_0=1$. Based on the representations (2.6) and (2.7), we have

$$V(\theta)=\int\limits_{(0,1)}(R(u;\theta))^2du=\int\limits_{(0,1)}\frac{1}{\theta}u^{1+2q-\theta}du=\frac{1}{\theta(2+2q-\theta)}\geq\frac{1}{(1+q)^2}=V(\theta^*),$$

where the minimum is attained uniquely at $\theta^*=1+q$. We have $V(\theta^*)=\mu^2$, which indicates perfect importance sampling is achieved, that is, one condition in Theorem 4.2 **(iii)** is violated. In fact, we have

$$\begin{aligned}W(\theta,\xi)&=2\xi\int\limits_{(0,1)}R(u;\theta)\left(u-1/2\right)du+\frac{1}{12}\xi^2=2\xi\int\limits_{(0,1)}u^q\left(u^\theta-1/2\right)du+\frac{1}{12}\xi^2\\&=2\xi\left(\frac{1}{\theta+1+q}-\frac{1}{2(1+q)}\right)+\frac{1}{12}\xi^2,\end{aligned}$$

which yields $W(\theta^*,\xi)=W(1+q,\xi)=\xi^2/12$ and thus $\xi^*=0$. Indeed, no control variates is needed. It is however not necessarily true that the optimal value (in this case, $V(\theta)+W(\theta,\xi)=\mu^2$) is attainable uniquely at (θ^*,ξ^*). For illustration, fix $q=1$, that is, $\mu=1/2$, $V(\theta)=(\theta(4-\theta))^{-1}$, and $W(\theta,\xi)=2\xi((\theta+2)^{-1}-1/4)+\xi^2/12$, where perfect importance sampling is achieved at the point $(\theta^*,\xi^*)=(2,0)$, as we have just derived. In this case, since the estimator is linear $\Phi(u)=u$, perfect control variates is possible as well, $V(1)+W(1,-1)(=V(\theta_0)+W(\theta_0,-1))=\mu^2$. This perfect control variates is independent of the choice of bypass distribution, that is, no importance sampling is needed. Hence, there exist two distinct points $(\theta^*, \xi^*) = (2, 0)$ and $(\theta, \xi) = (1, -1)$, both of which achieve perfect variance reduction $V(2) + W(2, 0) = V(1) + W(1, -1) = \mu^2$. Since the function $V(\theta) + W(\theta, \xi)$ is not flat on the line segment joining those two points $(2, 0)$ and $(1, -1)$, the function $V(\theta) + W(\theta, \xi)$ is not convex jointly in the two variables (θ, ξ).

Next, we demonstrate that there can exist a distinct point $(\theta^\diamond, \xi^\diamond)$, which strictly outperforms the point (θ^*, ξ^*), where the latter is considered optimal in the proposed framework. Fix $q = 1$ again for the sake of simplicity, and employ a different bypass distribution $g(z; \theta) = \phi(z - \theta)$, $G(z; \theta) = \Phi(z - \theta)$, and $G^{-1}(u; \theta) = \theta + \Phi^{-1}(u)$, with $\theta_0 = 0$. Based on the representation (2.6) and (2.7), we have

$$V(\theta) = \int_{(0,1)} (R(u;\theta))^2 \, du = \int_{\mathbb{R}} e^{-\theta z + \theta^2/2} |\Phi(z)|^2 \phi(z) dz,$$

$$W(\theta, \xi) = 2\xi \int_{(0,1)} R(u;\theta)\,(u - 1/2)\, du + \frac{1}{12}\xi^2 = 2\xi \int_{\mathbb{R}} \Phi(z) \left(\Phi(z-\theta) - \frac{1}{2}\right) \phi(z) dz + \frac{1}{12}\xi^2,$$

with $\xi^* = -12 \int_{\mathbb{R}} \Phi(z)(\Phi(z - \theta^*) - 1/2)\phi(z)dz$. Note that $V(\theta_0) = 1/3$ and $W(\theta_0, \xi) = (\xi^2 + 2\xi)/12$, so $V(\theta_0) + W(\theta_0, -1) = \mu^2$, that is, perfect control variates is possible without importance sampling, just as described before. By numerical approximation, we obtain $\theta^* = 0.55072$, $\xi^* = -0.010723$, $V(\theta^*) = 0.26434$, and $W(\theta^*, \xi^*) = -9.5818 \times 10^{-6}$. That is, we have $V(\theta^*) + W(\theta^*, \xi^*) - \mu^2 = 0.014326$, compared to the crude variance $V(\theta_0) - \mu^2 = 0.083333$. The proposed framework at the point (θ^*, ξ^*) reduces a large portion (around 83%) of the estimator variance, whereas it cannot achieve perfect variance reduction, which is indeed possible at the point $(\theta^\diamond, \xi^\diamond) = (\theta_0, -1)$. Let us stress again that although we have just seen a few ways of perfect variance reduction, the examples examined here are highly artificial to demonstrate the possibility of perfect variance reduction. In reality, there is effectively no point in worrying about such peculiar perfect variance reduction.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmaa.2019.123608.

References

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

[2] V.S. Borkar, Stochastic approximation with two time scales, Systems Control Lett. 29 (1997) 291–294.

[3] R. Buche, H.J. Kushner, Rate of convergence for constrained stochastic approximation algorithms, SIAM J. Control Optim. 40 (2001) 1011–1041.

[4] P.W. Glynn, R. Szechtman, Some new perspectives on the method of control variates, in: K.-T. Fang, H. Niederreiter, F.J. Hickernell (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 27–49.

[5] B. Jourdain, J. Lelong, Robust adaptive importance sampling for normal random vectors, Ann. Appl. Probab. 19 (5) (2009) 1687–1718.

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

[7] R. Kawai, Adaptive Monte Carlo variance reduction for Lévy processes with two-time-scale stochastic approximation, Methodol. Comput. Appl. Probab. 10 (2) (Jun 2008) 199–223.

[8] 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.

[9] R. Kawai, Acceleration on adaptive importance sampling with sample average approximation, SIAM J. Sci. Comput. 39 (4) (2017) A1586–A1615.

[10] R. Kawai, Adaptive importance sampling Monte Carlo simulation for general multivariate probability laws, J. Comput. Appl. Math. 319 (2017) 440–459.

[11] R. Kawai, Optimizing adaptive importance sampling by stochastic approximation, SIAM J. Sci. Comput. 40 (4) (2018) A2774–A2800.

[12] S. Kim, S.G. Henderson, Adaptive control variates for finite-horizon simulation, Math. Oper. Res. 32 (3) (2007) 508–527.

[13] B. Lapeyre, J. Lelong, A framework for adaptive Monte Carlo procedures, Monte Carlo Methods Appl. 17 (1) (2011) 77–98.

[14] V. Lemaire, G. Pagès, Unconstrained recursive importance sampling, Ann. Appl. Probab. 20 (3) (2010) 1029–1067.

[15] C.J. Oates, M. Girolami, N. Chopin, Control functionals for Monte Carlo integration, J. R. Stat. Soc. Ser. B. Stat. Methodol. 79 (3) (2017) 695–718.

[16] A. Owen, Y. Zhou, Safe and effective importance sampling, J. Amer. Statist. Assoc. 95 (449) (2000) 135–143.

[17] B.T. Polyak, A.B. Juditsky, Acceleration of stochastic approximation by averaging, SIAM J. Control Optim. 30 (4) (1992) 838–855.

[18] A. Shapiro, D. Dentcheva, A. Ruszczyński, Lectures on Stochastic Programming, Society for Industrial and Applied Mathematics, 2009.