Annals of Operations Research manuscript No.
(will be inserted by the editor)

Minimization of a Class of Rare Event Probabilities and Buffered Probabilities of Exceedance

Amarjit Budhiraja · Shu Lu · Yang Yu · Quoc Tran-Dinh

Received: date / Accepted: date

Acknowledgements Research of SL was supported in part by the National Science Foundation (DMS-1814894). Research of AB was supported in part by the National Science Foundation (DMS-1814894 and DMS-1853968).

Abstract We consider the problem of choosing design parameters to minimize the probability of an undesired rare event that is described through the average of n i.i.d. random variables. Since the probability of interest for near optimal design parameters is very small, one needs to develop suitable accelerated Monte-Carlo methods for estimating its value. One of the challenges in the study is that simulating from exponential twists of the laws of the summands may be computationally demanding since these transformed laws may be non-standard and intractable. We consider a setting where the summands are given as a nonlinear functional of random variables, the exponential twists of whose distributions take a simpler form than those for the original summands. We use techniques from Dupuis and Wang (2004,2007) to identify the appropriate Issacs equations whose subsolutions are used to construct tractable importance sampling (IS) schemes. We also study the closely related problem of estimating buffered probability of exceedance and provide the first rigorous results that relate the asymptotics of buffered probability and that of the ordinary probability under a large deviation scaling. The analogous minimization problem for buffered probability, under conditions, can be formulated as a convex optimization problem. We show that, under conditions, changes of measures that are asymptotically efficient (under the large deviation scaling) for estimating ordinary probability are also asymptotically efficient for estimating the buffered probability of exceedance. We embed the constructed IS scheme in gradient descent algorithms to solve the optimization problems, and illustrate these schemes through computational experiments.

Keywords Importance Sampling · Stochastic Optimization · Large Deviations · Buffered Probability

Mathematics Subject Classification (2010) 90C15 · 65K10 · 65C05 · 60F10

[1] E-mail: {budhiraj, shulu, quoctd}@email.unc.edu, yy0324@live.unc.edu
Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC27599-3260, USA.

1 Introduction

When the objective function of an optimization problem is the expectation of a random variable, one can estimate that expectation by a sample average. If the standard deviation of that random variable is large relative to its mean, variance reduction techniques such as IS can be used to reduce the sample size needed for a reliable estimate of the mean, see (Chen et al. 1993; Dupuis and Wang 2004, 2007; Evans and Swartz 2000; Glasserman et al. 1997; Owen and Zhou 2000; L'Ecuyer and Tuffin 2011; Ridder 2005) and references therein for related research. Here we consider problems of the form

$$\min_{\theta \in \Theta} \mathbb{E} \exp \left\{ -nF\left(\frac{1}{n} \sum_{i=1}^{n} G(X_i, \theta) \right) \right\}, \tag{1}$$

where $X_i, i = 1, \cdots, n$ are i.i.d random variables in $\mathbb{R}^h$, $\Theta \subset \mathbb{R}^d$, $G : \mathbb{R}^h \times \Theta \to \mathbb{R}^m$ is continuous, and $F : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is measurable. For notational simplicity we write $U_i = G(X_i, \theta)$ and $Y_n = \frac{1}{n} \sum_{i=1}^{n} U_i$, so (1) can be equivalently written as $\min_{\theta \in \Theta} \mathbb{E} \exp \{-nF(Y_n)\}$.

The formulation (1) includes a special case in which $F(y) = 0$ for y in a measurable set $A \subset \mathbb{R}^m$ and ∞ otherwise. In this case (1) becomes

$$\min_{\theta \in \Theta} \mathbb{P}\left(Y_n \in A\right) = \min_{\theta \in \Theta} \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^{n} G(X_i, \theta) \in A \right). \tag{2}$$

In many applications in engineering, finance, and insurance, decisions need to be made to reduce the probability for an undesirable event (such as system breakdown) to occur. Such an event is often the result of the accumulative effects of a large number of individual events over a long period, which we model as $\{Y_n \in A\}$, with n a fixed large number. Under conditions, for values of θ such that $\mathbb{E}[U_1] \notin \operatorname{cl} A$, $\mathbb{P}(Y_n \in A)$ converges to 0 exponentially fast as $n \to \infty$ by the theory of large deviations, so its value is very small for a large n, making it hard to estimate using i.i.d samples of Y_n.

A natural approach for estimating probabilities of the form on the left side of (2) is by computing Monte-Carlo averages of the form $\frac{1}{N} \sum_{j=1}^{N} 1_{\{Y_n^{(j)} \in A\}}$ where $Y_n^{(1)}, \cdots, Y_n^{(N)}$ are iid distributed as Y_n. Note that there are two parameters: the parameter n is fixed and describes the event of interest while the parameter N represents the size of the Monte-Carlo sample to estimate the probability in (2) for a given value of n. When the event of interest $\{Y_n \in A\}$ has small probability, such Monte-Carlo estimates perform poorly. An effective way to estimate the probabilities of such rare events and expected values of more general risk sensitive functionals as on the right side of (1) is using IS techniques based on large deviations theory. Large deviation based IS techniques were introduced in (Siegmund 1976) in estimating the error probabilities of the sequential probability ratio test. Subsequent papers exhibited the good performance of specific estimators developed using this technique, see (Bucklew 1990; Collamore 2002; Sadowsky 1991). However, such estimators can perform poorly as shown in (Glasserman et al. 1997), if the necessary and sufficient conditions for effective variance reduction in (Chen et al. 1993; Sadowsky and Bucklew 1990; Sadowsky 1996) are violated. In order to address this, later papers introduced adaptive IS schemes that are more generally applicable. Among these, (Dupuis and Wang 2004, 2007) are most related to our work. (Dupuis and Wang 2004) connects the problem of constructing asymptotically efficient adaptive IS schemes with certain deterministic dynamic games. (Dupuis and Wang 2007) uses subsolutions to the Isaacs equations for such games to construct flexible and simple dynamic IS schemes that achieve asymptotic efficiency.

For a direct application of the IS techniques from (Dupuis and Wang 2004, 2007) to the situation here, one would need to use a parametric family of exponential changes of measure to generate the replacements for the U_i given each fixed θ. Such a scheme is easy to implement when the distribution of U_i is of a simple form. For example if U_i is a normal random variable then an exponential change of measure is also a normal distribution with a shifted mean. However, for more general distributions and when the dimension m is large, sampling from the exponential tilt distribution can be computationally expensive (see discussion at the end of Section 2.1). This problem gets much more severe in the optimization problem we study, in which estimates for the objective function need to be computed for many different values of θ. By writing $U_i = G(X_i, \theta)$, we aim to capture the complexity of the distribution of U_i through the function G and leave the distribution of X_i in a fixed simple form. In particular, we are interested in a setting where simulating from exponential tilts of distributions of X_i is simpler than that from exponential tilts of U_i. In this work we develop an IS technique based on a change of measure on the distribution of X_i, which is computationally much less demanding compared to a scheme that uses a change of measure directly based on U_i. The scheme is inspired by (Dupuis and Wang 2004, 2007) and, as in these papers, is guided by the Issacs equation of a certain dynamic game. The Issacs equation is given in terms of a different Hamiltonian (see (24)) than the one that arises in the formulation where the change of measure is done directly on the sequence $\{U_i\}$ (see (11)). We show that generalized subsolutions of this Issacs equation can be used to construct IS algorithms, with guaranteed lower bounds on asymptotic performance (as measured by the asymptotic exponential decay rate of the second moment), that are based on dynamic change of measure for the sequence $\{X_i\}$. Similar to (Dupuis and Wang 2004, 2007), the decay rate is governed by the initial value of the subsolution (i.e. at $(t, x) = (0, 0)$), with larger initial values implying a higher decay rate.

When we embed the above IS procedure in a gradient descent method to find the optimal θ for (1), both the objective values and the gradients need to be estimated by samples. If the function F violates certain continuity/differentiability conditions, as in the case of (2), the gradients cannot be directly estimated from sample functions. Thus, we will approximate the original F by an a.e. differentiable and Lipschitz continuous function and apply the IS methods to the expected values of the resulting risk sensitive functional. Moreover, as shown in Theorem 2, the logarithm of the objective function of (1) after scaled by $1/n$ converges to a limiting function under certain conditions. The optimal solution and value of the limiting problem, when available, can be used as approximations of those of the original problem with fixed large n.

The problem (2) or its smooth approximation will not be convex in general, so the gradient descent method will not distinguish local solutions from global solutions. For the case $m = 1$ and $A = [c, \infty)$, there is an alternative reliability measure called the buffered failure probability (Rockafellar and Royset 2010) or the buffered probability of exceedance (Mafusalov et al. 2015). Under mild conditions, minimization of the buffered failure probability over a class of probability distributions can be transformed into a convex optimization problem and is therefore more tractable. The buffered failure probability is always greater than or equal to the corresponding probability, and the two values are often close to each other when the probability of the random variable of interest taking on large values is small (see e.g. (Rockafellar and Royset 2010) for a discussion of this point). In this work we make the second statement precise in one particular setting. Specifically, we show that under conditions, probabilities of the form on the right side of (2) have the same exponential decay rate, as $n \to \infty$, as the corresponding buffered failure probabilities (see Theorem 3). To the best of our knowledge this is the first rigorous result that relates the asymptotics of a buffered failure probability and ordinary probability under a large deviation scaling. This result in particular suggests that the IS change of measure that are appropriate for estimating the probability on the right side of (2) should also be suitable for constructing estimators for the corresponding buffered failure probability. Under appropriate conditions, this is indeed the case as is shown in Theorem 4 and Theorem 5. One can view the buffered failure probability as a reliability measure that is of independent interest or, in view of its closeness to the ordinary exceedance probability, the solution to the buffered failure probability minimization problem can be used as an intermediate step for selecting the initial point in the algorithm for the probability minimization problem.

For comprehensive overviews on optimization under probabilistic (chance) constraints see (Prékopa 2013) and (Shapiro et al. 2009, Chapter 4). Various methods for solving chance-constrained optimization have been proposed, see (Bremer et al. 2015; Calafiore and Campi 2006; Dentcheva and Martinez 2013; Nemirovski and Shapiro 2006; Pagnoncelli et al. 2009; Van Ackooij and Henrion 2014). When the chance constraints involve a rare event probability, in some cases, IS can be combined with the SAA approach by exploiting problem structure to reduce the sample estimation variance uniformly with respect to the decision variables (Barrera et al. 2016).

The paper is organized as follows. Section 2 reviews IS techniques that are based on large deviation analyses and proposes a new IS scheme that is based on changes of laws of the sequence $\{X_i\}$ rather than directly transforming the probability laws of the sequence $\{U_i\}$. This section also provides an asymptotic bound on the second moment of the IS estimator. Section 3 studies the limiting behavior of the problem (1) as $n \to \infty$, as well as convergence properties of the approximation problem for (2) in which probabilities are replaced by expected values of certain risk sensitive functionals. Section 4 studies the buffered failure probability in the present setting and its estimation using IS methods. Section 5 presents the optimization algorithm and uses several numerical examples to illustrate the method. Throughout the paper, $\mathscr{P}(\mathbb{R}^h)$ denotes the space of all probability measures on $\mathbb{R}^h$.

2 Importance sampling based on large deviations analysis

In this section, we consider the estimation of the objective value of (1) for a fixed θ. Since θ is fixed, we suppress it from notation in this section and consider the estimation of

$$\mathbb{E} \exp\{-nF(Y_n)\}, \tag{3}$$

where $Y_n = \frac{1}{n}\sum_{i=1}^n U_i$ is the average of i.i.d. random variables $U_i = G(X_i)$ for $i = 1, \cdots, n$. The function $G : \mathbb{R}^h \to \mathbb{R}^m$ is continuous, and $F : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is measurable. Let η be the distribution of X_1 and ξ be the distribution of U_1, namely $\xi = \eta \circ G^{-1}$.

If the distribution of the random variable Y_n takes a simple form, then one may consider a change of measure with respect to the distribution of Y_n directly. However, by its definition, the distribution of Y_n is in general rather complicated and so one needs to construct the change of measure through the underlying distributions of U_i. Even in situations where the distribution of Y_n is of simple form, e.g. Gaussian, it may be advantageous to construct a change of measure that exploits the form of Y_n and transforms the distributions of summands U_i in a systematic manner. Subsection 2.1 below reviews the estimation methods from (Dupuis and Wang 2004, 2007) that construct a dynamic change of measure on the distributions of $\{U_i\}$ and provide results characterizing the asymptotic performance of the resulting estimator. One of the challenges in implementing these methods is that even if the distribution η of X_i were of a simple form, for a general G the distribution of U_i may be rather complicated, so sampling from the exponential twists of the distribution of U_i may become hard. In Subsection 2.2 we provide an alternative approach that constructs an estimator using a dynamic change of measure with respect to the distributions of X_i, and establish an asymptotic bound on the second moment for the resulting IS estimator.

In either approach, the replacement random variables will in general not be i.i.d., and the conditional distribution of the jth random variable given the previous $j-1$ variables is related to the original distribution by an exponential tilt, i.e., the Radon-Nikodym derivative of the replacement measure with respect to the original measure is an exponential function with a linear exponent (see e.g. (6)). Parameters for these exponents are chosen based on solutions of certain partial differentiable equations. These equations arise when one considers the problem of minimizing the second moment as a certain stochastic control problem and studies the associated dynamic programming equations. The asymptotic performance of the resulting change of measure is established using methods from the theory of large deviations.

The starting point of the analysis are the logarithms of moment generating functions of the original random variables. For $(a, \alpha) \in \mathbb{R}^{h+m}$, we define

$$H(a, \alpha) = \log \mathbb{E}\left[e^{\langle a, X_1\rangle + \langle \alpha, G(X_1)\rangle}\right]. \tag{4}$$

We also consider functions $H_1 : \mathbb{R}^h \to \mathbb{R}$ and $H_2 : \mathbb{R}^m \to \mathbb{R}$ as

$$H_1(a) = H(a, 0), \text{ and } H_2(\alpha) = H(0, \alpha), \; a \in \mathbb{R}^h, \; \alpha \in \mathbb{R}^m. \tag{5}$$

Thus, H_1 is the log-moment generating function of X_1 and H_2 is that of $U_1 = G(X_1)$.

2.1 The exponential change of measure on variables U_i

In this subsection we review results from (Dupuis and Wang 2004, 2007). Assume $H_2(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^m$. We will replace the original random variables $U_1, \cdots, U_n$ by new random variables $\bar{U}_1^n, \cdots, \bar{U}_n^n$, that have (conditional) distributions of the form

$$e^{\langle \alpha, u\rangle - H_2(\alpha)}\xi(du) \tag{6}$$

where $\alpha \in \mathbb{R}^m$ and ξ is the distribution of U_1. In general, the parameter α that defines the sampling distribution does not need to be a constant, and can depend on values of summands that precede the current variable. Formally, suppose a function $\bar{\alpha}(x,t) : \mathbb{R}^m \times [0,1] \to \mathbb{R}^m$ is given. To construct a dynamic change of measure based on $\bar{\alpha}$ one proceeds as follows. Suppose $\bar{U}_1^n, \cdots, \bar{U}_j^n$ have been simulated. Define

$$\bar{Y}_j^n = \frac{1}{n}\sum_{i=1}^{j} \bar{U}_i^n \tag{7}$$

and simulate $\bar{U}_{j+1}^n$ from the distribution

$$e^{\langle \bar{\alpha}(\bar{Y}_j^n, \frac{j}{n}), u\rangle - H_2(\bar{\alpha}(\bar{Y}_j^n, \frac{j}{n}))}\xi(du). \tag{8}$$

Thus the conditional distribution of $\bar{U}_{j+1}^n$ given $\left\{\bar{Y}_i^n, i = 1, \ldots j\right\}$ is given by (8). Iterating we obtain $\left\{\bar{U}_j^n\right\}_{1\le j\le n}$ and $\left\{\bar{Y}_j^n\right\}_{1\le j\le n}$. Using successive conditioning

$$Z^n = e^{-nF(\bar{Y}_n^n)} \prod_{j=0}^{n-1} e^{-\langle \bar{\alpha}(\bar{Y}_j^n, \frac{j}{n}), \bar{U}_{j+1}^n\rangle + H_2(\bar{\alpha}(\bar{Y}_j^n, \frac{j}{n}))} \tag{9}$$

is an unbiased estimator for (3), and the above product of exponentials is the Radon-Nikodym derivative of the distribution of $(U_1, \cdots, U_n)$ with respect to that of $(\bar{U}_1^n, \cdots, \bar{U}_n^n)$.

If the function $\bar{\alpha}$ is a constant, then the above scheme reduces to a static change of measure in which $(\bar{U}_1^n, \cdots, \bar{U}_n^n)$ are i.i.d.. Different choices of the function $\bar{\alpha}$ will produce different distributions for Z^n. In order to reduce the number of samples needed to the greatest extent, the idea is to choose $\bar{\alpha}$ in a way to minimize the variance of Z^n. It is hard to characterize the optimal choice of $\bar{\alpha}$ for a fixed value of n, as the distribution of Y_n is rather complicated. However, as $n \to \infty$ the (tails of the) distribution of Y_n can be described using large deviations theory, which leads to a characterization of an asymptotically optimal choice of $\bar{\alpha}$ in terms of the solution of a partial differential equation known as the Isaacs equation(Dupuis and Wang 2004). We now introduce this equation. Let L_2 be the Legendre transform of H_2 defined as

$$L_2(\beta) = \sup_{\alpha\in\mathbb{R}^m} \left(\langle \alpha, \beta\rangle - H_2(\alpha)\right),\ \beta \in \mathbb{R}^m. \tag{10}$$

It is possible that $L_2(\beta) = \infty$ for some β. Define $\mathbb{H}_2 : \mathbb{R}^{3m} \to \mathbb{R} \cup \{\infty\}$ as

$$\mathbb{H}_2(s; \alpha, \beta) = \langle s, \beta\rangle + L_2(\beta) + \langle \alpha, \beta\rangle - H_2(\alpha). \tag{11}$$

The Isaacs equation is then given as

$$W_t(y,t) + \sup_{\alpha\in\mathbb{R}^m} \inf_{\beta\in\mathbb{R}^m} \mathbb{H}_2(DW(y,t); \alpha, \beta) = 0 \tag{12}$$

where $W : \mathbb{R}^m \times [0,1] \to \mathbb{R}$ is a continuously differentiable function, $W_t(y,t)$ is its derivative w.r.t. t, and $DW(y,t)$ is its derivative w.r.t. y. If W satisfies

$$W_t(y,t) + \sup_{\alpha \in \mathbb{R}^m} \inf_{\beta \in \mathbb{R}^m} \mathbb{H}_2(DW(y,t); \alpha, \beta) \geq 0 \tag{13}$$

instead of (12) then it is a (classical) subsolution to (12). If such a subsolution W also satisfies the terminal condition $W(y,1) \leq 2F(y)$ for all $y \in \mathbb{R}^m$, then, as is shown in (Dupuis and Wang 2004, 2007), the dynamic change of measure as in (8), constructed using the supremizer $\alpha(y,t)$ for the second term in (13), produces an estimator Z^n as in (9) (with $\bar{\alpha}$ replaced by α) whose second moment decays exponentially at rate $W(0,0)$:

$$\liminf_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{E}\left[(Z^n)^2 \right] \right\} \geq W(0,0). \tag{14}$$

On the other hand, under standard conditions, the limit

$$\gamma = \lim_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{E} \exp\left\{ -nF(Y_n) \right\} \right\} \tag{15}$$

exists (Dupuis and Ellis 2011). By Jensen's inequality, if $\tilde{Z}^n$ is any unbiased estimator of (3)

$$\limsup_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{E}\left[(\tilde{Z}^n)^2 \right] \right\} \leq \limsup_{n\to\infty} \left\{ -\frac{1}{n} \log \left(\mathbb{E}\left[\tilde{Z}^n \right] \right)^2 \right\} = 2\gamma,$$

so 2γ is the largest decay rate for the second moment among all unbiased estimators. Sometimes one can find a subsolution with $W(0,0) = 2\gamma$, in which case the estimator in (9) constructed from the supermizer α in (12) is *asymptotically efficient.*

Sometimes one needs more than one subsolution in order to construct an IS estimator that achieves asymptotic efficiency. This leads to the following notion of a generalized subsolution/control(Dupuis and Wang 2007).

Definition 1 Given $K \in \mathbb{N}$, consider functions $\bar{W} : \mathbb{R}^m \times [0,1] \to \mathbb{R}$, $\rho_k : \mathbb{R}^m \times [0,1] \to \mathbb{R}$, and $\bar{\alpha}_k : \mathbb{R}^m \times [0,1] \to \mathbb{R}^m$ for $1 \leq k \leq K$. The collection $(\bar{W}, \rho_k, \bar{\alpha}_k)$ is called a generalized subsolution/control to the Isaacs equation (12), and $(\bar{\alpha}_k, \rho_k)$ the corresponding generalized control pair, if the following conditions hold:

(i) For all (y,t), $\{\rho_k(y,t)\}$ is a probability vector.

(ii) $\bar{W}$ is continuously differentiable, $\bar{W}_t(y,t) = \sum_{k=1}^K \rho_k(y,t) r_k(y,t)$, and $D\bar{W}(y,t) = \sum_{k=1}^K \rho_k(y,t) s_k(y,t)$.

(iii) For each $k = 1, \ldots, K$, it holds that

$$r_k(y,t) + \inf_{\beta \in \mathbb{R}^m} \mathbb{H}_2(s_k(y,t); \bar{\alpha}_k(y,t), \beta) \geq 0. \tag{16}$$

(iv) The functions $(r_k, s_k, \rho_k, \bar{\alpha}_k)$ are uniformly bounded and continuous.

Roughly speaking, a generalized subsolution/control as above is used as follows. At each step, one of the K functions $\bar{\alpha}_k$ is randomly selected to determine the change of measure for the summand and the likelihood of a selection is determined by a probability vector valued function $\{\rho_k\}_{k=1}^K$. More precisely, with a generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$ in hand, one can construct a dynamic change of measure as follows. Let $\bar{Y}_0^n = 0$. For $j = 0, \ldots, n-1$, having constructed $\left\{\bar{U}_i^n\right\}_{1\le i\le j}$ and $\left\{\bar{Y}_i^n\right\}_{1\le i\le j}$, we generate a multinomial random variable I with $\mathbb{P}\left(I = k\right) = \rho_k(\bar{Y}_j^n, \frac{j}{n})$ for $k \in \{1, 2, \ldots, K\}$. Next, we simulate $\bar{U}_{j+1}^n$ from the distribution

$$e^{\langle \bar{\alpha}_I(\bar{Y}_j^n, \frac{j}{n}), u\rangle - H_2(\bar{\alpha}_I(\bar{Y}_j^n, \frac{j}{n}))}\xi(du), \tag{17}$$

namely the conditional distribution of $\bar{U}_{j+1}^n$ given $\left\{\bar{U}_i^n\right\}_{i\le j}$ and I is given by (17). Define $\bar{Y}_{j+1}^n = \bar{Y}_j^n + \frac{1}{n}\bar{U}_{j+1}^n$. It follows from a simple calculation (see (Dupuis and Wang 2007)) that

$$Z^n = e^{-nF(\bar{Y}_n^n)} \prod_{j=0}^{n-1} \Big[\sum_{k=1}^{K} \rho_k(\bar{Y}_j^n, \tfrac{j}{n}) e^{\langle \bar{\alpha}_k(\bar{Y}_j^n, \frac{j}{n}), \bar{U}_{j+1}^n\rangle - H_2(\bar{\alpha}_k(\bar{Y}_j^n, \frac{j}{n}))} \Big]^{-1} \tag{18}$$

is an unbiased estimator for (3) with the n-fold product above defining the Radon-Nikodym derivative of the distribution of $(U_1, \cdots, U_n)$ w.r.t that of $(\bar{U}_1^n, \cdots, \bar{U}_n^n)$ (evaluated at $(\bar{U}_1^n, \cdots, \bar{U}_n^n)$). Once again, when the terminal condition $\bar{W}(y, 1) \le 2F(y)$ holds for all $y \in \mathbb{R}^m$, the second moment of Z^n decays exponentially at a rate no slower than $\bar{W}(0, 0)$, namely (14) is satisfied with W replaced by $\bar{W}$. Thus if one can find a $\bar{W}$ with $\bar{W}(0, 0) = 2\gamma$, one has an asymptotically efficient IS estimator. One seeks a $\bar{W}$ which has the largest possible value at $(0, 0)$.

When ξ is a simple form distribution (such as a Normal, Gamma, Poisson, exponential or a binomial), the tilted distribution (6) typically belongs to the same distribution family with a different parameter. In such cases, samples from (8) can be generated easily. However, in general the distribution of $U_i = G(X_i)$ may not take a simple form. To simulate from (8) in such a general situation, one needs to invert the conditional cumulative distributions and then evaluate them at uniform random variables. However, with a general nonlinear function G, the distribution ξ is rarely available in a tractable form, making such a procedure difficult to start with. Even when ξ is available in a closed form, inverting the conditional cumulative distributions requires iteratively carrying out numerical integrations, which is highly computationally intensive. For these reasons, the practical utility of changing measures on U_i is limited to situations in which ξ takes a simple form.

2.2 The exponential change of measure on variables X_i

The computational issue of simulating from the tilted distribution (17) is largely due to the complexity of ξ, the distribution of $U_i = G(X_i)$. This motivates us to consider the alternative approach of conducting the change of measure on variable X_i, whose distribution η is assumed to be of a simpler form. In this subsection, we assume that $H(a, \alpha) < \infty$ for all $(a, \alpha) \in \mathbb{R}^{h+m}$, and let L be the Legendre transformation of H:

$$L(b, \beta) = \sup_{(a,\alpha)\in\mathbb{R}^{h+m}} \big(\langle a, b\rangle + \langle \alpha, \beta\rangle - H(a, \alpha)\big), \quad (b, \beta) \in \mathbb{R}^{h+m}. \tag{19}$$

Then L has the following representation (Dupuis and Ellis 2011, Lemma 6.2.3):

$$L(b,\beta) = \inf_{\mu\in\mathscr{P}(\mathbb{R}^h)} \left\{ R(\mu\|\eta) \mid \int_{\mathbb{R}^h} x\mu(dx) = b, \int_{\mathbb{R}^h} G(x)\mu(dx) = \beta \right\}, \tag{20}$$

where $R(\mu\|\eta)$ is the relative entropy of μ with respect to η defined as

$$R(\mu\|\eta) = \int_{\mathbb{R}^h} \log \frac{d\mu}{d\eta} d\mu \tag{21}$$

when μ is absolutely continuous w.r.t. η, and ∞ otherwise.

Recall that H_1 is the log-moment generating function of X_1. In the change of measure scheme, we will replace random variables $X_1, \cdots, X_n$ by new variables $\bar{X}_1^n, \cdots, \bar{X}_n^n$ that have (conditional) distributions η_a of the form

$$\eta_a(dx) = e^{\langle a,x\rangle - H_1(a)}\eta(dx), \tag{22}$$

where $a \in \mathbb{R}^h$ and η as before is the distribution of X_1. The values of a will be determined dynamically by a function $\bar{a} : \mathbb{R}^m \times [0,1] \to \mathbb{R}^h$ as follows. Let $\bar{Y}_0^n = 0$. For $j = 0, \cdots, n-1$, having constructed $\{\bar{X}_i^n\}_{1\le i\le j}$, $\{\bar{U}_i^n = G(\bar{X}_i^n)\}_{1\le i\le j}$ and $\{\bar{Y}_i^n\}_{1\le i\le j}$ via (7), let $\eta_{\bar{a}(\bar{Y}_j^n,\frac{j}{n})}$ be the distribution of $\bar{X}_{j+1}^n$ conditioned on $\bar{X}_1^n, \ldots, \bar{X}_j^n$ and draw a sample $\bar{X}_{j+1}^n$ from this conditional distribution. Let $\bar{Y}_{j+1}^n = \bar{Y}_j^n + \frac{1}{n}G(\bar{X}_{j+1}^n)$. Thus recursively we obtain $\{\bar{Y}_i^n, \bar{U}_i^n, \bar{X}_i^n\}_{i=1}^n$. Using these we define the estimator

$$Z^n = e^{-nF(\bar{Y}_n^n)} \prod_{j=0}^{n-1} e^{H_1(\bar{a}(\bar{Y}_j^n,\frac{j}{n})) - \langle \bar{a}(\bar{Y}_j^n,\frac{j}{n}), \bar{X}_{j+1}^n\rangle}, \tag{23}$$

which as before is an unbiased estimator for (3). In comparison to schemes introduced in Subsection 2.1, the main advantage of the scheme proposed in the current section is the ease of implementation because, as discussed earlier, when G takes a complex form, one can simulate from $\eta_{\bar{a}(\bar{Y}_j^n,\frac{j}{n})}$ more easily than from the distribution in (6). We now introduce an Issacs equation associated with the control problem of minimizing the asymptotic second moment of Z^n. The equation is derived using similar formal dynamic programming heuristics as in (Dupuis and Wang 2004) however we omit these details. Define $\mathbb{H} : \mathbb{R}^{2m+2h} \to \mathbb{R}\cup\{\infty\}$ as

$$\mathbb{H}(s,a,b,\beta) = \langle a,b\rangle + \langle s,\beta\rangle + L(b,\beta) - H_1(a), \quad s,\beta\in\mathbb{R}^m, \quad a,b\in\mathbb{R}^h. \tag{24}$$

Then the Issacs equation is given as

$$W_t(y,t) + \sup_{a\in\mathbb{R}^h} \inf_{(b,\beta)\in\mathbb{R}^{h+m}} \mathbb{H}(DW(y,t),a,b,\beta) = 0, \tag{25}$$

along with the terminal condition $W(y,1) = 2F(y)$. We will now use this equation to construct IS schemes. As in Section 2.1, we begin with some definitions. A continuously differentiable function $\bar{W} : \mathbb{R}^m \times [0,1] \to \mathbb{R}$ is a classical subsolution to (25) if it satisfies

$$\bar{W}_t(y,t) + \sup_{a\in\mathbb{R}^h} \inf_{(b,\beta)\in\mathbb{R}^{h+m}} \mathbb{H}(D\bar{W}(y,t),a,b,\beta) \ge 0 \tag{26}$$

for each $(y,t) \in \mathbb{R}^m \times [0,1]$. If functions $\bar{W} : \mathbb{R}^m \times [0,1] \to \mathbb{R}$, $\rho_k : \mathbb{R}^m \times [0,1] \to \mathbb{R}$, $\bar{a}_k : \mathbb{R}^m \times [0,1] \to \mathbb{R}^h$, $1 \le k \le K$ satisfy all conditions in Definition 1 (with $\bar{\alpha}_k$ replaced by $\bar{a}_k$) except that (16) is replaced by

$$r_k(y,t) + \inf_{(b,\beta)\in\mathbb{R}^{h+m}} \mathbb{H}(s_k(y,t); \bar{a}_k(y,t), b, \beta) \ge 0, \tag{27}$$

then $(\bar{W}, \rho_k, \bar{a}_k)$ is said to be a generalized subsolution/control to (25). For the special case in which $K = 1$ and $\rho_1 = 1$, we abbreviate the notation $(\bar{W}, \rho_k, \bar{a}_k)$ as $(\bar{W}, \bar{a})$ and call it a subsolution/control pair.

A dynamic change of measure, analogous to Section 2.1, based on a generalized subsolution/control $(\bar{W}, \rho_k, \bar{a}_k)$ is constructed as follows. Let $\bar{Y}_0^n = 0$. For $j = 0, \ldots, n-1$, having constructed $\{\bar{X}_i^n\}_{1\le i\le j}$ and $\{\bar{Y}_i^n\}_{1\le i\le j}$, we generate a multinomial random variable I with (conditional) probabilities $\mathbb{P}(I = k) = \rho_k(\bar{Y}_j^n, \frac{j}{n})$ for $k \in \{1, 2, \ldots, K\}$. Next, we sample $\bar{X}_{j+1}^n$ from the distribution

$$e^{\langle \bar{a}_I(\bar{Y}_j^n, \frac{j}{n}), x\rangle - H_1(\bar{a}_I(\bar{Y}_j^n, \frac{j}{n}))} \eta(dx), \tag{28}$$

and define $\bar{Y}_{j+1}^n = \bar{Y}_j^n + \frac{1}{n} G(\bar{X}_{j+1}^n)$. Finally, we define

$$Z^n = e^{-nF(\bar{Y}_n^n)} \prod_{j=0}^{n-1} \left[\sum_{k=1}^{K} \rho_k(\bar{Y}_j^n, \tfrac{j}{n}) e^{\langle \bar{a}_k(\bar{Y}_j^n, \frac{j}{n}), \bar{X}_{j+1}^n\rangle - H_1(\bar{a}_k(\bar{Y}_j^n, \frac{j}{n}))} \right]^{-1}, \tag{29}$$

which as before is an unbiased estimator for (3). The appeal of the estimator in (29) over that in (18) is that, frequently it is simpler to simulate from (28) than from (17). Theorem 1 below is an analogue of (Dupuis and Wang 2007, Theorem 8.1) and shows that the second moment of Z^n decays exponentially at a rate no slower than $\bar{W}(0,0)$.

Theorem 1 *Suppose $H(a,\alpha) < \infty$ for all $(a,\alpha) \in \mathbb{R}^{h+m}$, that $(\bar{W}, \rho_k, \bar{a}_k)$ is a generalized subsolution/control to* (25) *and satisfies the terminal condition $\bar{W}(y,1) \le 2F(y)$ for all $y \in \mathbb{R}^m$, and that Z^n is as in* (29). *Then*

$$\liminf_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{E}\left[(Z^n)^2 \right] \right\} \ge \bar{W}(0,0).$$

Proof This proof is adapted from (Dupuis and Wang 2007). For $1 \in k \in K$, $j = 0, \cdots, n-1$ and $y \in \mathbb{R}^m$, define $\rho_{k,j}^n(y) = \rho_k(y, \frac{j}{n})$ and $\bar{a}_{k,j}^n(y) = \bar{a}_k(y, \frac{j}{n})$. Using a property of Radon-Nikodym derivatives, we write the second moment of Z^n in terms of the original variables $X_1, \cdots, X_n$ as

$$\begin{aligned} V^n =& \mathbb{E}\left[(Z^n)^2 \right] \\ =& \mathbb{E}\left[e^{-2nF(Y_n^n)} \prod_{j=0}^{n-1} \left(\sum_{k=1}^{K} \rho_{k,j}^n(Y_j^n) e^{\langle \bar{a}_{k,j}^n(Y_j^n), X_{j+1}\rangle - H_1(\bar{a}_{k,j}^n(Y_j^n))} \right)^{-1} \right], \end{aligned}$$

where $Y_j^n = \frac{1}{n}\sum_{i=1}^{j} G(X_i),\ j = 1, \cdots, n,\ Y_0^n = 0$. Next, letting $B(y) = \bar{W}(y,1)$, we have by assumption that $B(y) \le 2F(y)$. Define $\tilde{V}^n$ by replacing $e^{-2nF(Y_n^n)}$ in the above display with $e^{-nB(Y_n^n)}$ and let $\tilde{W}^n = -\frac{1}{n}\log \tilde{V}^n$. The fact $B(Y_n^n) \le$

$2F(Y_n^n)$ and the convexity of the exponential function imply that $V^n \le \tilde{V}^n$. Hence it suffices to show $\liminf \tilde{W}^n \ge \bar{W}(0,0)$.

Recall from the definition of generalized solutions that ρ_k, r_k and s_k are uniformly bounded which implies the Lipschitz continuity of $\bar{W}$. By assumption, H_1 is finite everywhere. Since it is convex, it is continuous and bounded on any compact set. Using these properties one can establish the following representation (see (Dupuis and Wang 2007, Lemma A.1))

$$\begin{aligned}\tilde{W}^n = \inf_{\bar{\nu}^n \in \mathscr{P}(\mathbb{R}^{nh})} & \Big\{ \frac{1}{n} R(\bar{\nu}^n \| \eta^{\bigotimes n}) + \\ & \mathbb{E}\Big[\tfrac{1}{n} \textstyle\sum_{j=0}^{n-1} \sum_{k=1}^{K} \rho_{k,j}^n(\tilde{Y}_j^n) \Big[\langle \bar{a}_{k,j}^n(\tilde{Y}_j^n), \tilde{X}_{j+1}^n \rangle - H_1(\bar{a}_{k,j}^n(\tilde{Y}_j^n)) \Big] + B(\tilde{Y}_n^n) \Big] \Big\},\end{aligned}$$

where $\eta^{\bigotimes n}$ is the n-fold product measure of η, $(\tilde{X}_1^n, \cdots, \tilde{X}_n^n)$ follows the distribution $\bar{\nu}^n$, and $\tilde{Y}_j^n = \frac{1}{n}\sum_{i=1}^{j} G(\tilde{X}_i^n)$, $j = 1, \cdots, n$, $\tilde{Y}_0^n = 0$. Using the chain rule for the relative entropy, we can rewrite $\tilde{W}^n$ as

$$\begin{aligned}\tilde{W}^n = & \inf_{\bar{\nu}^n \in \mathscr{P}(\mathbb{R}^{nh})} \\ & \mathbb{E}\left[\frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=1}^{K} \rho_{k,j}^n(\tilde{Y}_j^n) \Big[R(\nu_j^n \| \eta) + \langle \bar{a}_{k,j}^n(\tilde{Y}_j^n), \tilde{X}_{j+1}^n \rangle - H_1(\bar{a}_{k,j}^n(\tilde{Y}_j^n)) \Big] + B(\tilde{Y}_n^n) \right],\end{aligned}$$

where ν_j^n is the conditional distribution of $\tilde{X}_{j+1}^n$ given $(\tilde{X}_1^n, \cdots, \tilde{X}_j^n)$ (a random probability measure on $\mathbb{R}^h$). By defining

$$\begin{aligned} & J^n(\bar{\nu}^n) \\ & = \mathbb{E}\left[\frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=1}^{K} \rho_{k,j}^n(\tilde{Y}_j^n) \Big[R(\nu_j^n \| \eta) - H_1(\bar{a}_{k,j}^n(\tilde{Y}_j^n)) + \langle \bar{a}_{k,j}^n(\tilde{Y}_j^n), \tilde{X}_{j+1}^n \rangle \Big] + B(\tilde{Y}_n^n) \right],\end{aligned} \tag{30}$$

we have $\tilde{W}^n = \inf_{\bar{\nu}^n \in \mathscr{P}(\mathbb{R}^{nh})} J^n(\bar{\nu}^n)$. To prove the theorem, it suffices to prove

$$\liminf J^n(\bar{\nu}^n) \ge \bar{W}(0,0), \tag{31}$$

for an arbitrary sequence $\bar{\nu}^n$ of probability measures on $\mathbb{R}^{nh}$.

To prove (31), we will use a continuous time interpolation. To this end, for $j = 0, \ldots, n-1$ and $t \in [\frac{j}{n}, \frac{j+1}{n})$, define $\tilde{Y}^n(t) = \tilde{Y}_j^n$ and $\nu^n(dx|t) = \nu_j^n(dx)$, and let $\tilde{Y}^n(1) = \tilde{Y}_n^n$. Then define a probability measure ν^n on $\mathbb{R}^h \times [0,1]$ by $\nu^n(A \times C) = \int_C \nu^n(A|t)dt$ for $A \in \mathcal{B}(\mathbb{R}^h)$ and $C \in \mathcal{B}([0,1])$. In addition, define another probability measure η' on $\mathbb{R}^h \times [0,1]$ as the product measure

$$\eta'(dx \times dt) = \eta(dx)dt. \tag{32}$$

Note that ν^n is a random probability measure on $\mathbb{R}^h \times [0,1]$. The distribution of ν^n is determined by $\bar{\nu}^n$, a non-random probability measure on $\mathbb{R}^{nh}$. Another application of the chain rule for the relative entropy gives $\mathbb{E}[R(\nu^n \| \eta')] = \mathbb{E}\left[\frac{1}{n}\sum_{j=0}^{n-1} R(\nu_j^n \| \eta)\right]$.

We can then write $J^n(\bar{\nu}^n)$ defined in (30) as

$$\begin{aligned} J^n(\bar{\nu}^n) = \mathbb{E}\Bigg[& R(\nu^n \| \eta') - \sum_{k=1}^{K} \int_0^1 \rho_k\left(\tilde{Y}^n(t), \frac{\lfloor tn \rfloor}{n}\right) H_1\left(\bar{a}_k\left(\tilde{Y}^n(t), \frac{\lfloor tn \rfloor}{n}\right)\right) dt \\ & + \sum_{k=1}^{K} \int_{\mathbb{R}^h \times [0,1]} \rho_k\left(\tilde{Y}^n(t), \frac{\lfloor tn \rfloor}{n}\right) \langle \bar{a}_k\left(\tilde{Y}^n(t), \frac{\lfloor nt \rfloor}{n}\right), x \rangle \nu^n(dx \times dt) + B(\tilde{Y}^n(1)) \Bigg].\end{aligned}$$

We define a time-continuous version of J^n as

$$\bar{J}^n(\bar{\nu}^n) = \mathbb{E}\left[R(\nu^n\|\eta') - \sum_{k=1}^{K}\int_0^1 \rho_k(\tilde{Y}^n(t),t)H_1\big(\bar{a}_k(\tilde{Y}^n(t),t)\big)dt \right.$$
$$\left. + \sum_{k=1}^{K}\int_{\mathbb{R}^h\times[0,1]} \rho_k(\tilde{Y}^n(t),t)\big\langle\bar{a}_k(\tilde{Y}^n(t),t),x\big\rangle\nu^n(dx\times dt) + B(\tilde{Y}^n(1))\right].$$

We will show

$$\liminf_{n\to\infty} J^n(\bar{\nu}^n) = \liminf_{n\to\infty} \bar{J}^n(\bar{\nu}^n) \text{ and } \liminf_{n\to\infty} \bar{J}^n(\bar{\nu}^n) \geq \bar{W}(0,0). \tag{33}$$

The theorem is an immediate consequence of the statements in (33). The proofs of these statements rely on the following lemma, the proof of which is omitted since it is analogous to (Dupuis and Wang 2007, Lemma A.2 and Lemma A.3).

Lemma 1 *Assume that $H(a,\alpha) < \infty$ for all $(a,\alpha) \in \mathbb{R}^{h+m}$, and that $(\bar{W},\rho_k,\bar{a}_k)$ is a generalized subsolution/control to (25). Consider a subsequence of $\{\bar{\nu}^n\}$ along which $J^n(\bar{\nu}^n)$ is bounded. Then, relabeling this sequence as $\{n\}$,*

$$\lim_{C\to\infty}\sup_n \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\Big(\|G(\tilde{X}_j^n)\|1_{\{\|G(\tilde{X}_j^n)\|>C\}} + \|\tilde{X}_j^n\|1_{\{\|\tilde{X}_j^n\|>C\}}\Big)\right] = 0, \tag{34}$$

the sequence $\{(\tilde{Y}^n,\nu^n)\}$ is tight, $\{\tilde{Y}^n(1)\}$ is uniformly integrable and $\{\nu^n\}$ satisfies

$$\lim_{C\to\infty}\sup_n \mathbb{E}\left[\int_{\mathbb{R}^h\times[0,1]}\big(\|G(x)\|1_{\{\|G(x)\|\geq C\}} + \|x\|1_{\{\|x\|\geq C\}}\big)\,\nu^n(dx\times dt)\right] = 0. \tag{35}$$

Further suppose that $(\tilde{Y}^n,\nu^n) \to (\tilde{Y},\nu)$ in distribution. Then $\nu(dx\times dt)$ can be factored as $\nu(dx\times dt) = \nu(dx|t)dt$, with

$$\tilde{Y}(t) = \int_{[0,t]}\int_{\mathbb{R}^h} G(x)\nu(dx|s)ds, \text{ for all } t\in[0,1], \quad a.s.. \tag{36}$$

With this lemma we can now complete the proof of (33). Without loss of generality we can assume that $J^n(\bar{\nu}^n)$ is bounded. The uniform boundedness and Lipschitz continuity of ρ_k and $\bar{a}_k$, the continuity of H_1 and the uniform integrability of ν^n in (35) imply $\lim_{n\to\infty}|J^n(\bar{\nu}^n) - \bar{J}^n(\bar{\nu}^n)| = 0$. In the remainder of the proof we show $\liminf_{n\to\infty}\bar{J}^n(\bar{\nu}^n) \geq \bar{W}(0,0)$ along any such sequence.

Since $\{(\tilde{Y}^n,\nu^n)\}$ is tight along such a subsequence (Lemma 1), by passing to a further subsequence if necessary we may assume that $(\tilde{Y}^n,\nu^n) \to (\tilde{Y},\nu)$ in distribution. Below we consider the limit of each term of $\bar{J}^n(\bar{\nu}^n)$. For its first term, note that

$$\liminf_{n\to\infty}\mathbb{E}\left[R(\nu^n\|\eta')\right] \geq \mathbb{E}\left[\liminf_{n\to\infty}R(\nu^n\|\eta')\right] \geq \mathbb{E}\left[R(\nu\|\eta')\right], \tag{37}$$

where η' is as introduced in (32) and the first inequality is by Fatou's Lemma while the second follows from the lower semi-continuity of the relative entropy. For the second term in $\bar{J}^n(\bar{\nu}^n)$, using the continuity and boundedness of ρ_k and $\bar{a}_k$, and the weak convergence of $\tilde{Y}^n$ to $\tilde{Y}$, an application of the dominated convergence theorem gives

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\left[\sum_{k=1}^{K}\int_0^1 \rho_k(\tilde{Y}^n(t),t)H_1(\bar{a}_k(\tilde{Y}^n(t),t))dt\right]\\&=\mathbb{E}\left[\sum_{k=1}^{K}\int_0^1 \rho_k(\tilde{Y}(t),t)H_1(\bar{a}_k(\tilde{Y}(t),t))dt\right].\end{aligned}\tag{38}$$

For the third term, the uniform integrability of ν^n and continuity and boundedness of ρ_k and $\bar{a}_k$ implies

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\left[\sum_{k=1}^{K}\int_{\mathbb{R}^h\times[0,1]}\rho_k(\tilde{Y}^n(t),t)\langle\bar{a}_k(\tilde{Y}^n(t),t),x\rangle\nu^n(dx\times dt)\right]\\&=\mathbb{E}\left[\sum_{k=1}^{K}\int_{\mathbb{R}^h\times[0,1]}\rho_k(\tilde{Y}(t),t)\langle\bar{a}_k(\tilde{Y}(t),t),x\rangle\nu(dx\times dt)\right].\end{aligned}\tag{39}$$

For the last term, note that the Lipschitz continuity of $\bar{W}$ implies $B(y)=\bar{W}(y,1)$ has linear growth. From the uniform integrability of $\{\tilde{Y}^n(1)\}$ in Lemma 1 we then have that

$$\lim_{n\to\infty}\mathbb{E}[B(\tilde{Y}^n(1))]=\mathbb{E}[B(\tilde{Y}(1))].\tag{40}$$

Combining (37), (38), (39), (40), we get a lower bound for $\liminf_{n\to\infty}\bar{J}^n(\bar{\nu}^n)$ as given below

$$\begin{aligned}&\mathbb{E}\left[R(\nu\|\eta')-\sum_{k=1}^{K}\int_0^1\rho_k(\tilde{Y}(t),t)H_1(\bar{a}_k(\tilde{Y}(t),t))dt\right.\\&\left.+\sum_{k=1}^{K}\int_{\mathbb{R}^h\times[0,1]}\rho_k(\tilde{Y}(t),t)\langle\bar{a}_k(\tilde{Y}(t),t),x\rangle\nu(dx\times dt)+B(\tilde{Y}(1))\right].\end{aligned}\tag{41}$$

Next, using the chain rule of the relative entropy and the representation (20), we have

$$R(\nu\|\eta')=\int_0^1 R(\nu(\cdot|t)\|\eta)dt\geq\int_0^1 L(b(t),\beta(t))dt,$$

where $b(t)=\int_{\mathbb{R}^h}x\nu(dx|t)$ and $\beta(t)=\int_{\mathbb{R}^h}G(x)\nu(dx|t)$. From the definition of $b(t)$,

$$\int_{\mathbb{R}^h\times[0,1]}\langle\bar{a}_k(\tilde{Y}(t),t),x\rangle\nu(dx\times dt)=\int_{[0,1]}\langle\bar{a}_k(\tilde{Y}(t),t),b(t)\rangle dt.$$

This gives the following lower bound for (41):

$$\mathbb{E}\left[\int_0^1\sum_{k=1}^{K}\rho_k(\tilde{Y}(t),t)\Big[L(b(t),\beta(t))-H_1(\bar{a}_k(\tilde{Y}(t),t))+\langle\bar{a}_k(\tilde{Y}(t),t),b(t)\rangle\Big]dt+B(\tilde{Y}(1))\right].\tag{42}$$

By the definition of generalized solutions (see (27)),

$$\begin{aligned}
&\bar{W}_t(\tilde{Y}(t),t) + \langle D\bar{W}(\tilde{Y}(t),t), \beta(t)\rangle \\
&= \sum_{k=1}^{K} \rho_k(\tilde{Y}(t),t) \left[r_k(\tilde{Y}(t),t) + \left\langle s_k(\tilde{Y}(t),t), \beta(t)\right\rangle \right] \\
&\geq - \sum_{k=1}^{K} \rho_k(\tilde{Y}(t),t) \left[L(b(t),\beta(t)) - H_1(\bar{a}_k(\tilde{Y}(t),t)) + \left\langle \bar{a}_k(\tilde{Y}(t),t), b(t)\right\rangle \right].
\end{aligned}$$

From (36) we have $\beta(t) = d\tilde{Y}(t)/dt$ for almost every t. Integrating over $[0,1]$ and taking expectations, we get

$$\begin{aligned}
&\bar{W}(0,0) - \mathbb{E}\left[\bar{W}(\tilde{Y}(1),1)\right] \\
\leq &\mathbb{E}\left[\int_0^1 \sum_{k=1}^{K} \rho_k(\tilde{Y}(t),t) \left[L(b(t),\beta(t)) - H_1(\bar{a}_k(\tilde{Y}(t),t)) + \langle \bar{a}_k(\tilde{Y}(t),t), b(t)\rangle \right] dt \right].
\end{aligned}$$

Since $B(\tilde{Y}(1) = \bar{W}(\tilde{Y}(1),1)$, we have shown that $\bar{W}(0,0)$ is a lower bound of (42) and thereby completed the proof of Theorem 1. □

In practice, one wants to construct a subsolution/control $(\bar{W}, \rho_k, \bar{a}_k)$ that has a simple form and for which the value of $\bar{W}(0,0)$ is as large as possible. For this, we first consider subsolution/control pairs $(\bar{W}, \bar{a})$, as defined below (27), for which $\bar{W}$ is an affine function of (y,t) and $\bar{a}$ is in fact a constant. If we write $\bar{W}$ in the form

$$\bar{W}(y,t) = \bar{c} + \langle u, y\rangle - (1-t)v \quad \text{for some} \quad \bar{c} \in \mathbb{R}, u \in \mathbb{R}^m, v \in \mathbb{R}, \tag{43}$$

then $(\bar{W}, \bar{a})$ is a subsolution/control pair if the following holds for all $(y,t) \in \mathbb{R}^{m+1}$:

$$\bar{W}_t(y,t) + \inf_{(b,\beta)\in\mathbb{R}^{h+m}} \mathbb{H}(D\bar{W}(y,t), \bar{a}, b, \beta) \geq 0, \tag{44}$$

namely

$$v + \inf_{(b,\beta)\in\mathbb{R}^{h+m}} \mathbb{H}(u, \bar{a}, b, \beta) \geq 0. \tag{45}$$

Next, we select a finite collection of pairs $\left\{(\bar{W}_k, \bar{a}_k), k = 1, \ldots, K\right\}$ from this family of subsolution/control pairs, such that the point-wise minimum $\bar{W} \doteq \wedge_{k=1}^{K} \bar{W}_k$ defined as $\bar{W}(y,t) = \wedge_{k=1}^{K} \bar{W}_k(y,t) = \min_{k=1,\cdots,K} \bar{W}_k(y,t)$ satisfies

$$\wedge_{k=1}^{K} \bar{W}_k(y,1) \leq 2F(y) \quad \text{for all} \quad y \in \mathbb{R}^m. \tag{46}$$

In the process of choosing $\{(\bar{W}_k, \bar{a}_k), k = 1, \cdots, K\}$ we also maximize $\wedge_{k=1}^{K} \bar{W}_k(0,0)$ among all qualified choices. Finally, we choose a small positive number δ, and define

$$\bar{W}^{\delta}(y,t) \doteq -\delta \log\left(\sum_{k=1}^{K} e^{-(1/\delta)\bar{W}_k(y,t)} \right), \tag{47}$$

and

$$\rho_k^{\delta}(y,t) \doteq \frac{e^{-(1/\delta)\bar{W}_k(y,t)}}{\sum_{i=1}^{K} e^{-(1/\delta)\bar{W}_i(y,t)}} \quad \text{for} \quad 1 \leq k \leq K. \tag{48}$$

Then, following (Dupuis and Wang 2007), we see that $(\bar{W}^{\delta}, \rho_k^{\delta}, \bar{a}_k)$ is a subsolution/control with

$$\wedge_{k=1}^{K} \bar{W}_k(y,t) \geq \bar{W}^{\delta}(y,t) \geq \wedge_{k=1}^{K} \bar{W}_k(y,t) - \delta \log K \quad \text{for all } (y,t).$$

In particular, the difference between $\bar{W}^{\delta}(0,0)$ and $\wedge_{k=1}^{K} \bar{W}_k(0,0)$ is not larger than $\delta \log K$. Thus the estimator Z^n based on this generalized subsolution/control satisfies

$$\liminf_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{E}\left[(Z^n)^2 \right] \right\} \geq \bar{W}^{\delta}(0,0) \geq \bar{W}(0,0) - \delta \log K. \tag{49}$$

In Section 5 we illustrate the implementation of such a construction for some examples.

3 Analysis of some approximate problems

It is possible for the objective function of (1) to be differentiable even if F is not differentiable everywhere. However, the gradient of the objective function is not given by the expectation of the gradient of the function inside the expectation w.r.t. θ, unless additional conditions hold (see, e.g., (Shapiro et al. 2009, Theorem 7.49)). Those conditions are not satisfied with $F(y) = \infty 1_{A^c}(y)$, the main problem we are interested in. To use a gradient based optimization algorithm to solve (1), we approximate F by a continuous function $\varphi : \mathbb{R}^m \to \mathbb{R}$, and use a solution to the problem

$$\min_{\theta\in\Theta} p(\theta) = \mathbb{E} \exp \left\{ -n\varphi\left(\frac{1}{n} \sum_{i=1}^{n} G(X_i, \theta) \right) \right\} \tag{50}$$

as an estimate for the solution of (1).

Next, we consider the problem (50) with a fixed continuous function φ, and study its convergence as $n \to \infty$. While our main interest is in solving (1) or its approximation (50) for a fixed n, this convergence ensures stability of the solution of (50) as n increases, and can be used in computation to find an initial point for solving (50). For this purpose, we define functions $g^n : \Theta \to \mathbb{R}$ and $g : \Theta \to \mathbb{R}$ as

$$g^n(\theta) = -\frac{1}{n} \log \mathbb{E} \exp \left\{ -n\varphi \left(\frac{1}{n} \sum_{i=1}^{n} G(X_i, \theta) \right) \right\}, \tag{51}$$

and

$$g(\theta) = \inf_{\nu \in \mathscr{P}(\mathbb{R}^h)} \left\{ \varphi \left(\int_{\mathbb{R}^h} G(x,\theta)\nu(dx) \right) + R(\nu||\eta) \right\}. \tag{52}$$

Clearly, (50) is equivalent to

$$\max_{\theta\in\Theta} g^n(\theta). \tag{53}$$

Theorem 2 below shows that g^n converges to g uniformly under suitable conditions, which implies the convergence of solutions of (53) to that of the limiting problem:

$$\max_{\theta\in\Theta} g(\theta). \tag{54}$$

Let H_2^θ denote the log moment generating function of $G(X_1, \theta)$, namely,

$$H_2^\theta(\alpha) = \log \mathbb{E}\, e^{\langle \alpha, G(X_1,\theta)\rangle}, \quad \alpha \in \mathbb{R}^m. \tag{55}$$

Also, let L_2^θ denote the Legendre transform of H_2^θ, i.e.,

$$L_2^\theta(\beta) = \sup_{\alpha\in\mathbb{R}^m} \Big(\langle \alpha, \beta\rangle - H_2^\theta(\alpha)\Big), \quad \beta \in \mathbb{R}^m. \tag{56}$$

Theorem 2 *Let Θ be a compact subset of $\mathbb{R}^d$. Assume that $\sup_{\theta\in\Theta} H_2^\theta(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^m$. If φ is continuous and bounded, then $g^n \to g$ uniformly on Θ.*

Proof Let $\{X_i\}_{i\in\mathbb{N}}$ be i.i.d. random variables with distribution η, and let $\mathcal{L}^n$ be the empirical measure in $\mathbb{R}^h$ that puts mass $1/n$ at each of the first n points $X_1, \cdots, X_n$, namely $\mathcal{L}^n(dx) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(dx)$. From the representation established in (Dupuis and Ellis 2011, Section 2.3), for $\theta \in \Theta$, we have

$$g^n(\theta) = \inf_{\bar\nu^n} \mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x,\theta)\bar{\mathcal{L}}^n(dx)\right) + \frac{1}{n}\sum_{i=1}^n R(\bar\nu_i^n \| \eta)\right], \tag{57}$$

where the infimum is over all probability distributions $\bar\nu^n \in \mathscr{P}(\mathbb{R}^{nh})$, with $(\bar X_1^n, \cdots, \bar X_n^n)$ being a random variable with distribution $\bar\nu^n$, $\bar{\mathcal{L}}^n$ being the empirical measure in $\mathbb{R}^h$ of the n points $\bar X_1^n, \cdots, \bar X_n^n$, and $\bar\nu_i^n$ being the conditional distribution of $\bar X_i^n$ given $\bar X_1^n, \cdots, \bar X_{i-1}^n$. Since φ is bounded, the infimum in (57) is bounded above by $\|\varphi\|_\infty = \sup_{y\in\mathbb{R}^m} \mid \varphi(y)| < \infty$. It follows that for any fixed value of $n \in \mathbb{N}$, in taking the infimum in (57) we can restrict to distributions $\bar\nu^n$ for which

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n R(\bar\nu_i^n\|\eta)\right] \le 2\|\varphi\|_\infty + 1. \tag{58}$$

Under our assumption $\sup_\theta H_2^\theta(\alpha) < \infty$, by a standard argument (see, e.g. the proof of Lemma 1), for any sequence $\{\bar\nu^n\}_{n\in\mathbb{N}}$ that satisfies (58) for all n we see that

$$\lim_{C\to\infty} \sup_{n\in\mathbb{N}} \sup_{\theta\in\Theta} \mathbb{E}\left[\int_{\mathbb{R}^h} \|G(x,\theta)\| 1_{\{\|G(x,\theta)\|\ge C\}} \bar{\mathcal{L}}^n(dx)\right] = 0. \tag{59}$$

Now let $\{\theta^n\} \subset \Theta$ such that $\theta^n \to \theta$ as $n \to \infty$. Fix $\varepsilon > 0$ and let $\{\bar\nu^n\}$ satisfy

$$\begin{aligned} &-\frac{1}{n}\log \mathbb{E}\left[e^{-n\varphi\left(\int_{\mathbb{R}^h} G(x,\theta^n)\mathcal{L}^n(dx)\right)}\right] + \varepsilon \\ \ge &\mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x,\theta^n)\bar{\mathcal{L}}^n(dx)\right) + \frac{1}{n}\sum_{i=1}^n R(\bar\nu_i^n\|\eta)\right], \end{aligned}$$

as well as (58) for each n, and define $\hat\nu^n \doteq \frac{1}{n}\sum_{i=1}^n \bar\nu_i^n$. Using arguments similar to (Dupuis and Ellis 2011, Proposition 8.2.5 and Lemma 8.2.7), $\{(\bar{\mathcal{L}}^n, \hat\nu^n)\}_{n\in\mathbb{N}}$ is tight. Consider a subsequence along which $(\bar{\mathcal{L}}^n, \hat\nu^n)$ converges weakly to $(\bar{\mathcal{L}}, \hat\nu)$. Then it is easy to check that

$$\lim_{n\to\infty} \mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x,\theta^n)\bar{\mathcal{L}}^n(dx)\right)\right] = \mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x,\theta)\bar{\mathcal{L}}(dx)\right)\right]. \tag{60}$$

Consequently, we have

$$\begin{aligned}\liminf_{n\to\infty} g^n(\theta^n)+\varepsilon &= \liminf_{n\to\infty}\left\{-\frac{1}{n}\log\mathbb{E}\exp\left\{-n\varphi\left(\int_{\mathbb{R}^h}G(x,\theta^n)\mathcal{L}^n(dx)\right)\right\}\right\}+\varepsilon\\ &\geq \liminf_{n\to\infty}\mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h}G(x,\theta^n)\bar{\mathcal{L}}^n(dx)\right)+\frac{1}{n}\sum_{i=1}^n R(\bar{\nu}_i^n\|\eta)\right]\\ &\geq \liminf_{n\to\infty}\mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h}G(x,\theta^n)\bar{\mathcal{L}}^n(dx)\right)+R(\hat{\nu}^n\|\eta)\right]\\ &\geq \mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h}G(x,\theta)\bar{\mathcal{L}}(dx)\right)+R(\hat{\nu}\|\eta)\right]\\ &\geq \inf_{\nu\in\mathscr{P}(\mathbb{R}^h)}\left[\varphi\left(\int_{\mathbb{R}^h}G(x,\theta)\nu(dx)\right)+R(\nu\|\eta)\right]=g(\theta),\end{aligned}$$

where the second inequality holds by Jensen's inequality and convexity of relative entropy, the third inequality follows from the convergence in distribution, Fatou's Lemma and lower semicontinuity of relative entropy, and the fourth inequality follows from the fact that $\bar{\mathcal{L}}=\hat{\nu}$ a.s., see (Dupuis and Ellis 2011, Theorem 8.2.8). Since $\varepsilon>0$ is arbitrary, we have $\liminf g^n(\theta^n)\geq g(\theta)$.

We now consider the reverse inequality. Once more, let $\theta^n\to\theta$. We first argue that $g(\theta^n)\to g(\theta)$. Note that, for $\theta\in\Theta$

$$\begin{aligned}g(\theta) &= \inf_{\nu\in\mathscr{P}(\mathbb{R}^h)}\left[\varphi\left(\int_{\mathbb{R}^h}G(x,\theta)\nu(dx)\right)+R(\nu\|\eta)\right]\\ &= \inf_{\nu\in\mathscr{P}(\mathbb{R}^h):R(\nu\|\eta)\leq\|\varphi\|_\infty}\left[\varphi\left(\int_{\mathbb{R}^h}G(x,\theta)\nu(dx)\right)+R(\nu\|\eta)\right].\end{aligned}$$

Fix $\varepsilon>0$ and let ν^n, ν^0 be ε-optimal for $g(\theta^n)$ and $g(\theta)$, respectively, and such that $R(\nu^n\|\eta)\leq\|\varphi\|_\infty$, $R(\nu^0\|\eta)\leq\|\varphi\|_\infty$. Then the sequence $\{\nu^n\}$ is tight and in a similar manner as for the proof of (59), we have

$$\lim_{C\to\infty}\sup_{n\geq 0}\sup_{\theta\in\Theta}\int_{\mathbb{R}^h}\|G(x,\theta)\|1_{\{\|G(x,\theta)\|\geq C\}}\nu^n(dx)=0. \tag{61}$$

In particular, as $n\to\infty$, it holds that

$$\left|\varphi\left(\int_{\mathbb{R}^h}G(x,\theta^n)\nu^n(dx)\right)-\varphi\left(\int_{\mathbb{R}^h}G(x,\theta)\nu^n(dx)\right)\right|\to 0, \tag{62}$$

and

$$\left|\varphi\left(\int_{\mathbb{R}^h}G(x,\theta^n)\nu^0(dx)\right)-\varphi\left(\int_{\mathbb{R}^h}G(x,\theta)\nu^0(dx)\right)\right|\to 0. \tag{63}$$

From the ε-optimality of ν^n, we have

$$\begin{aligned}\limsup_{n\to\infty}(g(\theta)-g(\theta^n)) &\leq \limsup_{n\to\infty}\left[\varphi\left(\textstyle\int_{\mathbb{R}^h}G(x,\theta)\nu^n(dx)\right)-\varphi\left(\textstyle\int_{\mathbb{R}^h}G(x,\theta^n)\nu^n(dx)\right)\right]+\varepsilon\\ &\leq\varepsilon,\end{aligned}$$

where the second inequality follows from (62). Similarly, using (63), we can see that $\limsup_{n\to\infty}(g(\theta^n)-g(\theta))\leq\varepsilon$. Since $\varepsilon>0$ is arbitrary, we have shown that

$$g(\theta^n)\to g(\theta)\quad\text{as}\quad n\to\infty. \tag{64}$$

Next, with ε, ν^n as above, define $\bar{\mathcal{L}}^n$ as the empirical measure of $\{\bar{X}_i^n\}_{i=1}^n$ which are i.i.d. ν^n. Using (61) it can be seen that the sequence $\{\bar{\mathcal{L}}^n\}$ satisfies (59). Also, for every bounded $\tilde{G} : \Theta \times \mathbb{R}^h \to \mathbb{R}$, as $n \to \infty$, $\int_{\mathbb{R}^h} \tilde{G}(x, \theta^n)\bar{\mathcal{L}}^n(dx) - \int_{\mathbb{R}^h} \tilde{G}(x, \theta^n)\nu^n(dx) \to 0$, in probability. Combining these two observations with the fact that φ is continuous and bounded, we have that, as $n \to \infty$,

$$\delta^n \doteq \left| \mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x, \theta^n)\bar{\mathcal{L}}^n(dx)\right)\right] - \varphi\left(\int_{\mathbb{R}^h} G(x, \theta^n)\nu^n(dx)\right)\right| \to 0. \tag{65}$$

Finally, from the representation in (57),

$$\begin{aligned}
\limsup_{n\to\infty} g^n(\theta^n) &\le \limsup_{n\to\infty} \mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x, \theta^n)\bar{\mathcal{L}}^n(dx)\right) + \frac{1}{n}\sum_{i=1}^n R(\bar{\nu}_i^n \| \eta)\right] \\
&\le \limsup_{n\to\infty} \left(\mathbb{E}\left[\varphi\left(\int_{\mathbb{R}^h} G(x, \theta^n)\nu^n(dx)\right) + R(\nu^n \| \eta)\right] + \delta^n\right) \\
&\le \limsup_{n\to\infty} g(\theta^n) + \varepsilon = g(\theta) + \varepsilon,
\end{aligned}$$

where the second inequality uses the fact that $\bar{\nu}_i^n = \nu^n$ for each i, and the third inequality uses (65) and the ε-optimality of ν^n. Since ε is arbitrary, we have proved $\limsup_{n\to\infty} g^n(\theta^n) \le g(\theta)$. This completes the proof. □

As an immediate consequence of the above theorem we have the following corollary. For a function $\psi : \Theta \to \mathbb{R}$ and $\delta \in (0, \infty)$ we say a $\theta^* \in \Theta$ satisfies $\theta^* \in \delta - \operatorname{argmax}_{\theta\in\Theta} \psi$ if $\psi(\theta^*) \ge \sup_{\theta\in\Theta} \psi(\theta) - \delta$.

Corollary 1 *Suppose the assumptions in Theorem 2 hold. Then,* $\max_{\theta\in\Theta} g^n(\theta)$ *converges to* $\max_{\theta\in\Theta} g(\theta)$*, and for any choice of* $\delta^n \downarrow 0$ *and* $\theta^n \in \delta^n - \operatorname{argmax}_{\theta\in\Theta} g^n$*, all cluster points of the sequence* $\{\theta^n\}_{n\in\mathbb{N}}$ *belong to* $\operatorname{argmax}_{\theta\in\Theta} g$*. If* $\operatorname{argmax}_{\theta\in\Theta} g$ *consists of a unique point* θ^**, one must actually have* $\theta^n \to \theta^*$*.*

4 Minimization of the buffered failure probability

In this section, we consider the special case in which $F(y) = \delta_A(y)$, $m = 1$ and $A = [0, \infty)$. In such a setting, an alternative reliability measure known as the buffered failure probability or the buffered probability of exceedance (abbreviated as the *buffered probability* in the rest of the paper) can be used in place of the standard probability. The buffered probability was introduced in (Rockafellar and Royset 2010), which also showed how to convert optimization problems with buffered probability constraints into convex programs using a result in (Rockafellar and Uryasev 2000). An extension and more properties of the buffered probability were provided in (Mafusalov and Uryasev 2014). In general, for a continuous 1-dimensional random variable X, and a scalar $c \in (\mathbb{E}[X], \text{ess sup}X)$ (ess sup$X$ is the essential supremum of X), the buffered probability is defined as $\bar{p}_c(X) = \mathbb{P}(X > q)$, where q is the unique solution to the equation $\mathbb{E}[X \mid X > q] = c$; in addition, we define $\bar{p}_c(X) = 0$ for $c \ge \text{ess sup}X$ and $\bar{p}_c(X) = 1$ for $c \le \mathbb{E}[X]$. For a detailed discussion and the definition that applies to a general distribution, see (Mafusalov and Uryasev 2014). A direct consequence of the above definition is that $q \le c$ and $\mathbb{P}(X > c) \le \bar{p}_c(X)$. It was shown in (Mafusalov and

Uryasev 2014) that the buffered probability can be equivalently represented as $\bar{p}_c(X) = \min_{\lambda \geq 0} \mathbb{E}\left[\lambda(X-c)+1\right]^+ 1_{\{\text{ess sup} X > c\}}$.

The following theorem gives an important connection between buffered probabilities and the large deviations rate function. Specifically, it shows that, under conditions, when X is replaced by the the sample mean of i.i.d. random variables, the buffered probability and the corresponding ordinary probability have the same asymptotic decay rate.

Theorem 3 *Let U_i, $i \geq 1$ be an i.i.d. sequence of $\mathbb{R}$-valued random variables, and suppose that $M(\lambda) \doteq \mathbb{E}\left[e^{\lambda U_1}\right] < \infty$ for every $\lambda \in \mathbb{R}$. Let $H(\lambda) \doteq \log M(\lambda)$ for $\lambda \in \mathbb{R}$ and L be the Legendre transform of H, and suppose that L is finite on $(0, \infty)$. Write $Y_n \doteq \frac{1}{n}\sum_{i=1}^n U_i$ for $n \geq 1$. Then for every $c > \mathbb{E}\left[U_1\right]$ and $\gamma \geq 0$, one has*

$$\lim_{n\to\infty} \frac{1}{n} \log \min_{\lambda \geq \gamma} \mathbb{E}[\lambda(Y_n - c)+1]^+ = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Y_n > c] = -L(c).$$

Proof Without loss of generality we assume that $\mathbb{E}(U_1) = 0$. Fix $c > 0$. Since for $\lambda = 0$, $\log \mathbb{E}\left[\lambda(X-c)+1\right]^+ = 0$ and $L(c) \geq 0$, it suffices to prove the result with the minimization over $\{\lambda : \lambda > \gamma\}$ for every $\gamma \geq 0$. Note that under the assumptions of the theorem, for every $\kappa > 0$, we have $\liminf_{n\to\infty} \dfrac{1}{n} \log \mathbb{P}\left(Y_n > \kappa\right) = \limsup_{n\to\infty} \dfrac{1}{n} \log \mathbb{P}\left(Y_n \geq \kappa\right) = -L(\kappa)$. For $\lambda > 0$

$$\mathbb{E}\left[\lambda(Y_n - c)+1\right]^+ \geq \mathbb{E}\left[\left[\lambda(Y_n-c)+1\right]1_{\{Y_n > c\}}\right] \geq \mathbb{P}(Y_n > c).$$

Thus, for any $\gamma \geq 0$, $\frac{1}{n}\log\min_{\lambda > \gamma} \mathbb{E}\left[\lambda(Y_n-c)+1\right]^+ \geq \frac{1}{n}\log\mathbb{P}\left(Y_n > c\right)$. Taking limit as $n \to \infty$, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \min_{\lambda > \gamma} \mathbb{E}\left[\lambda(Y_n - c)+1\right]^+ \geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(Y_n > c\right) = -L(c).$$

Now we prove the complementary inequality. Choose $m \geq 1$ such that $L(c+m) > L(c)+1$. Note that for $\lambda > 0$

$$\begin{aligned}&\mathbb{E}\left[\lambda(Y_n - c)+1\right]^+ \\ =&\mathbb{E}\left[\left[\lambda(Y_n - c)+\ 1\right]1_{\left\{c-\frac{1}{\lambda}\leq Y_n \leq c+m\right\}}\right] + \mathbb{E}\left[\left[\lambda(Y_n-c)+1\right]1_{\{Y_n > c+m\}}\right].\end{aligned}$$

Let $\alpha_0^* \in \mathbb{R}$ be the dual point to $(c+m)$, namely

$$L(c+m) = \sup_{\alpha\in\mathbb{R}}[\alpha(c+m) - H(\alpha)] = \alpha_0^*(c+m) - H(\alpha_0^*). \tag{66}$$

Existence of the dual point is guaranteed under the assumptions here, by (Ellis 2007, Theorems VIII.4.3 and VIII.4.4). Note that $\alpha_0^* > 0$, by Jensen's inequality, $H(\alpha_0^*) \geq \log(e^{\alpha_0^* \mathbb{E}(U_1)}) = 0$. Given $\lambda > 0$, choose $n(\lambda)$ such that for all $n \geq n(\lambda)$, $\gamma_n \doteq \alpha_0^* - \frac{\lambda}{n} > 0$. Then, for such n,

$$\begin{aligned}\mathbb{E}\left[\left[\lambda(Y_n-c)+1\right]1_{\{Y_n > c+m\}}\right] &\leq e^{-\lambda c - n\gamma_n(c+m)}\mathbb{E}\left[e^{(\lambda+n\gamma_n)Y_n}\right] \\ &= e^{nH(\gamma_n+\lambda/n)}e^{-\lambda c - n\gamma_n(c+m)} \\ &= e^{nH(\alpha_0^*)}e^{-n\alpha_0^*(c+m)}e^{\lambda m} = e^{-nL(c+m)+\lambda m}.\end{aligned} \tag{67}$$

Thus $\frac{1}{n}\log\mathbb{E}\left[[\lambda(Y_n-c)+1]1_{\{Y_n>c+m\}}\right] \le -L(c+m)+\frac{\lambda m}{n} \le -L(c)-1+\frac{\lambda m}{n}$. Also, for $\lambda>0$, it holds that

$$\frac{1}{n}\log\mathbb{E}\left[[\lambda(Y_n-c)+1]1_{\{c-1/\lambda\le Y_n\le c+m\}}\right] \le \frac{1}{n}\log\mathbb{P}\left(Y_n\ge c-1/\lambda\right)+\frac{\log(m\lambda+1)}{n}.$$

We now have that for all $n\ge n(\lambda)$,

$$\begin{aligned}&\tfrac{1}{n}\log\mathbb{E}\left[\lambda(Y_n-c)+1\right]^+\\ &\le \tfrac{\log 2}{n}+\max\left\{-L(c)-1+\tfrac{\lambda m}{n},\tfrac{\log(m\lambda+1)}{n}+\tfrac{1}{n}\log\mathbb{P}\left(Y_n\ge c-\tfrac{1}{\lambda}\right)\right\}.\end{aligned}$$

Fix $\varepsilon>0$ and let $\delta_0\in(0,\gamma^{-1}\wedge c)$, $n_0\in\mathbb{N}$, be such that for all $n\ge n_0$, $\frac{1}{n}\log\mathbb{P}\left(Y_n\ge c-\delta_0\right)\le -L(c)+\varepsilon$. Let $\lambda_0=\delta_0^{-1}$ and $n_1=n_0\vee n(\lambda_0)$. Then, for $n\ge n_1$

$$\begin{aligned}&\min_{\lambda>\gamma}\frac{1}{n}\log\mathbb{E}[\lambda(Y_n-c)+1]^+\\ \le&\frac{1}{n}\log\mathbb{E}[\lambda_0(Y_n-c)+1]^+\\ \le&\frac{\log 2}{n}+\max\left\{-L(c)-1+\frac{\lambda_0 m}{n},\frac{\log(m\lambda_0+1)}{n}+\varepsilon-L(c)\right\}.\end{aligned}$$

Now, choose $n_2\ge n_1$ s.t. for all $n\ge n_2$, $\lambda_0 m/n<1$. Then, for all $n\ge n_2$, we have

$$\max\left\{-L(c)-1+\frac{\lambda_0 m}{n},\frac{\log(m\lambda_0+1)}{n}+\varepsilon-L(c)\right\}=\frac{\log(m\lambda_0+1)}{n}+\varepsilon-L(c).$$

Thus for all $n\ge n_2$,

$$\min_{\lambda>\gamma}\frac{1}{n}\log\mathbb{E}[\lambda(Y_n-c)+1]^+\le\frac{\log(m\lambda_0+1)}{n}+\varepsilon-L(c)+\frac{\log 2}{n}.$$

Since $\varepsilon>0$ is arbitrary, we have the desired inequality on sending $n\to\infty$ and $\varepsilon\to 0$. □

The above theorem suggests that the change of measure that is asymptotically optimal for IS Monte-Carlo for estimating $\mathbb{P}\left(Y_n>c\right)$ may be useful for Monte-Carlo estimation of $\min_{\lambda>\alpha}\frac{1}{n}\log\mathbb{E}\left[\lambda(Y_n-c)+1\right]^+$ as well. Recall that the asymptotically optimal probability measure for IS for estimating $\mathbb{P}\left(Y_n>c\right)$ with $\{Y_n\}$ as in Theorem 3 is given as $\nu_{\alpha^*}(dz)\doteq e^{\alpha^*z-H(\alpha^*)}\xi(dz)$, where ξ is the probability distribution of U_1 and α^* is the conjugate dual of c, namely

$$L(c)=\sup_{\alpha\in\mathbb{R}}[\alpha c-H(\alpha)]=\alpha^*c-H(\alpha^*). \tag{68}$$

We will now show that this change of measure is nearly asymptotically optimal for IS estimation of $\frac{1}{n}\log\mathbb{E}\left[\lambda(Y_n-c)+1\right]^+$ for large values of λ. Note that by an elementary application of Jensen's inequality, if $T_n(\lambda)$ is any unbiased estimate of $\mathbb{E}\left[\lambda(Y_n-c)+1\right]^+$, then for any $\lambda>0$,

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[T_n^2(\lambda)\right]&\ge 2\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\lambda(Y_n-c)+1\right]^+\\ &\ge 2\liminf_{n\to\infty}\min_{\lambda'>0}\frac{1}{n}\log\mathbb{E}\left[\lambda'(Y_n-c)+1\right]^+=-2L(c).\end{aligned}$$

The following result shows that this lower asymptotic bound is nearly achieved when the estimator $T_n(\lambda)$ is constructed using the change of measure ν_{α^*} and λ is large. The second moment of this estimator is given as

$$R_n(\lambda) = \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n + nH(\alpha^*)}\right].$$

Theorem 4 *Suppose that the conditions of Theorem 3 are satisfied. Then for every $\varepsilon > 0$, there exists a $\gamma > 0$ such that*

$$\sup_{\lambda \geq \gamma} \limsup_{n\to\infty} \frac{1}{n} \log R_n(\lambda) \leq -2L(c) + \varepsilon.$$

Proof Without loss of generality, assume that $E(U_1) = 0$ and fix $c > 0$. For any $\lambda > 0$

$$\begin{aligned} \tfrac{1}{n}\log R_n(\lambda) &= \tfrac{1}{n}\log \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n + nH(\alpha^*)}\right] \\ &= H(\alpha^*) + \tfrac{1}{n}\log \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n}\right] \qquad (69) \\ &= -L(c) + \alpha^* c + \tfrac{1}{n}\log \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n}\right]. \end{aligned}$$

Choose $m \geq 1$ such that $L(c+m) \geq L(c) + \alpha^* c + 1$. Then, for $\lambda > 0$, we have

$$\begin{aligned} &\mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n}\right] \\ =&\mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n} 1_{\{c-1/\lambda \leq Y_n \leq c+m\}}\right] \\ &+ \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n} 1_{\{Y_n > c+m\}}\right]. \end{aligned}$$

For the second term on the right side we have with γ_n as in Theorem 3,

$$\begin{aligned} \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n} 1_{\{Y_n > c+m\}}\right] &\leq 4\mathbb{E}\left[\left(1 + \tfrac{(\lambda(Y_n - c))^2}{2}\right) 1_{\{Y_n > c+m\}}\right] \\ &\leq 4\mathbb{E}\left[e^{\lambda(Y_n - c)} e^{n\gamma_n (Y_n - c - m)}\right], \end{aligned}$$

where the first inequality is a consequence of $(1+x)^2 \leq 4(1 + \frac{x^2}{2})$ and $\alpha^* \geq 0$. Therefore, from (67), for all $n \geq n(\lambda)$, where $n(\lambda)$ is as in Theorem 3,

$$\begin{aligned} &\frac{1}{n}\log \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n} 1_{\{Y_n > c\}}\right] \\ \leq& - L(c+m) + \frac{\lambda m}{n} + \frac{\log 4}{n} \\ \leq& - L(c) - \alpha^* c - 1 + \frac{\lambda m}{n} + \frac{\log 4}{n}. \end{aligned}$$

Next,

$$\begin{aligned} &\frac{1}{n}\log \mathbb{E}\left[\left([\lambda(Y_n - c) + 1]^+\right)^2 e^{-n\alpha^* Y_n} 1_{\left\{c - \frac{1}{\lambda} \leq Y_n \leq c+m\right\}}\right] \\ \leq& - \alpha^*(c - \frac{1}{\lambda}) + \frac{2\log(1 + m\lambda)}{n} + \frac{1}{n}\log \mathbb{P}\left(Y_n > c - \frac{1}{\lambda}\right). \end{aligned}$$

Therefore, for all $n \geq n(\lambda)$,

$$\begin{aligned}&\frac{1}{n}\log\mathbb{E}\left[\left([\lambda(Y_n-c)+1]^+\right)^2 e^{-n\alpha^* Y_n}\right]\\ &\leq \frac{\log 2}{n} + \max\Big\{-L(c)-\alpha^*c-1+\frac{\lambda m+\log 4}{n},\\ &\quad -\alpha^*(c-\frac{1}{\lambda})+\frac{1}{n}\log\mathbb{P}\left(Y_n>c-\frac{1}{\lambda}\right)+\frac{2\log(1+m\lambda)}{n}\Big\}.\end{aligned}$$

Fix $\varepsilon > 0$ and let $0 < \delta_0 \leq c$ and $n_0 \in \mathbb{N}$ be such that for all $n \geq n_0$, we have $\frac{1}{n}\log\mathbb{P}(Y_n \geq c-\delta_0) \leq -L(c)+\frac{\varepsilon}{2}$. Then, for all $n \geq n_0$ and $\delta < \delta_0$,

$$\frac{1}{n}\log\mathbb{P}(Y_n \geq c-\delta) \leq \frac{1}{n}\log\mathbb{P}(Y_n \geq c-\delta_0) \leq -L(c)+\frac{\varepsilon}{2}.$$

Let $\gamma \doteq \max\left\{\frac{1}{\delta_0}, \frac{2\alpha^*}{\varepsilon}\right\}$. Then for every $\lambda \geq \gamma$ and $n \geq \max\{n_0, n(\lambda)\}$, we have

$$\begin{aligned}&\tfrac{1}{n}\log\mathbb{E}\left[\left([\lambda(Y_n-c)+1]^+\right)^2 e^{-n\alpha^* Y_n}\right] \leq \tfrac{\log 2}{n}\\ &\quad+\max\left\{-L(c)-\alpha^*c-1+\tfrac{\lambda m+\log 4}{n}, -L(c)-\alpha^*c+\varepsilon+\tfrac{2\log(1+m\lambda)}{n}\right\}.\end{aligned}$$

Choose $n_1 \geq n_0$ such that $\frac{\lambda m+\log 4}{n_1} < 1$. Then for $n \geq \max\{n_1, n(\lambda)\}$ the maximum on the right side equals $-L(c)-\alpha^*c+\varepsilon+\frac{2}{n}\log(1+m\lambda)$. Combining the above with (69), for every $\lambda \geq \gamma$, one has $\limsup_{n\to\infty}\frac{1}{n}\log R_n(\lambda) \leq -L(c)+\alpha^*c-L(c)-\alpha^*c+\varepsilon = -2L(c)+\varepsilon$. The result follows. □

We now return to our main optimization problem. Replacing the probability in (2) with the corresponding buffered probability for the random variable $Y_n = \frac{1}{n}\sum_{i=1}^n G(X_i,\theta)$, and assuming $c = 0 <$ ess supY_n, we obtain the following problem:

$$\inf_{\lambda\geq 0,\theta\in\Theta}\mathbb{E}\left[\lambda\left(\frac{1}{n}\sum_{i=1}^n G(X_i,\theta)-c\right)+1\right]^+. \tag{70}$$

As discussed below Theorem 5, the above optimization problem has some appealing features. We now present a result that makes connections between a change of measure used for solving the minimization problem in (2) and the minimization problem for the corresponding buffered probability, namely the problem in (70). For this result we recall the definition of a subsolution of (25) and the associated generalized subsolution/control, given in Subsection 2.2. We will use the notation and setting of Subsection 2.2 but here $m = 1$ and $F(y) = \infty 1_{(-\infty,c]}(y)$. The following is the main theorem which gives the same lower bound on the exponential decay rate of the second moment of the estimator for $\mathbb{E}[\lambda(Y_n-c)+1]^+$ as was obtained in Theorem 1.

Theorem 5 *Let $c > 0$. Assume that $H(a,\alpha) < \infty$ for all $(a,\alpha) \in \mathbb{R}^{n+1}$, and that $(\bar{W}, \{\rho_k, \bar{a}_k\}_{k=1}^K)$ is a generalized subsolution/control to (25) with $\bar{W}(y,1) < 0$ for all $y \geq c$. Let $\{\bar{X}_j^n\}_{1\leq j\leq n}$ and $\{\bar{Y}_j^n\}_{0\leq j\leq n}$ be as defined above Theorem 1. For $\lambda > 0$, define $Z^n(\lambda) \doteq [\lambda(\bar{Y}_n^n - c)+1]^+\bar{\Upsilon}^n$, where*

$$\bar{\Upsilon}^n \doteq \prod_{j=0}^{n-1}\left[\sum_{k=1}^K \rho_k(\bar{Y}_j^n, \tfrac{j}{n})e^{\langle \bar{a}_k(\bar{Y}_j^n,\frac{j}{n}),\bar{X}_{j+1}^n\rangle - H_1(\bar{a}_k(\bar{Y}_j^n,\frac{j}{n}))}\right]^{-1}.$$

Then, $Z^n(\lambda)$ is unbiased for $\mathbb{E}\left[\lambda(Y_n - c) + 1\right]^+$ and there exists $\gamma > 0$ s.t.

$$\sup_{\lambda \geq \gamma} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[(Z^n(\lambda))^2\right] \leq -\bar{W}(0,0).$$

Proof The unbiasedness of $Z^n(\lambda)$ is easy to check. Consider now $V^n(\lambda) \doteq \mathbb{E}[(Z^n(\lambda))^2]$. Let $m \geq 1$. Then with

$$\Upsilon^n \doteq \prod_{j=0}^{n-1} \left[\sum_{k=1}^{K} \rho_k(Y_j^n, \tfrac{j}{n}) e^{\langle \bar{a}_k(Y_j^n, \frac{j}{n}), X_{j+1}^n\rangle - H_1(\bar{a}_k(Y_j^n, \frac{j}{n}))}\right]^{-1},$$

we have

$$\begin{aligned} V^n(\lambda) &= \mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{Y_n \geq c - 1/\lambda\}}\right) \\ &= \mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{c-1/\lambda \leq Y_n \leq c+m\}}\right) \\ &\quad + \mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{Y_n > c+m\}}\right). \end{aligned} \tag{71}$$

For the second term on the last line, we have by the Cauchy-Schwarz inequality

$$\begin{aligned} &\left[\mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{Y_n > c+m\}}\right)\right]^2 \\ \leq &\mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^4 1_{\{Y_n > c+m\}}\right) \mathbb{E}\left(\Upsilon^n\right)^2. \end{aligned}$$

By Jensen's inequality

$$0 \leq \Upsilon^n \leq \tilde{\Upsilon}^n \doteq \prod_{j=0}^{n-1} \exp\left\{\sum_{k=1}^{K} \rho_k(Y_j^n, \tfrac{j}{n})(\langle \bar{a}_k(Y_j^n, \tfrac{j}{n}), X_{j+1}^n\rangle - H_1(\bar{a}_k(Y_j^n, \tfrac{j}{n})))\right\}.$$

From this, the boundedness of ρ_k and $\bar{a}_k$, and our assumption on the finiteness of H, we have for some $c_1 < \infty$ that $\left[\mathbb{E}\left(\Upsilon^n\right)^2\right]^{1/2} \leq e^{nc_1}$ for all $n \geq 1$. Also, for some $c_2 < \infty$

$$\mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^4 1_{\{Y_n > c+m\}}\right) \leq c_2 \mathbb{E}\left(e^{\lambda(Y_n - c)} e^{n\gamma_n(Y_n - c - m)}\right),$$

where γ_n is as introduced above (67). The same calculation as in (67) now shows that

$$\frac{1}{n} \log\left[\mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^4 1_{\{Y_n > c+m\}}\right)\right]^{1/2} \leq -\frac{L(c+m)}{2} + \frac{\lambda m + \log c_2}{2n}.$$

Thus

$$\frac{1}{n} \log \mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{Y_n > c+m\}}\right) \leq -\frac{L(c+m)}{2} + c_1 + \frac{\lambda m + \log c_2}{2n}.$$

Now fix $m \geq 1$ such that $L(c+m)/2 \geq \bar{W}(0,0) + 1 + c_1$. Consider the first term on the right side of (71). We have

$$\mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{c-1/\lambda \leq Y_n \leq c+m\}}\right) \leq (\lambda m + 1)^2 \mathbb{E}\left(\tilde{\Upsilon}^n 1_{\{Y_n \geq c - 1/\lambda\}}\right).$$

Choose γ large enough so that $\bar{W}(y,1) \leq 0$ for $y \geq c - 1/\gamma$. Then with B as in the proof of Theorem 1 we have $1_{\{Y_n \geq c-1/\lambda\}} \leq e^{-nB(Y_n)}$ for $\lambda \geq \gamma$. Thus we have

$$\frac{1}{n} \log \mathbb{E}\left(([\lambda(Y_n - c) + 1]^+)^2 \Upsilon^n 1_{\{c-1/\lambda \leq Y_n \leq c+m\}}\right) \leq \frac{2\log(\lambda m + 1)}{n} + \frac{1}{n} \log \tilde{V}^n,$$

where $\tilde{V}^n$ is as in the proof of Theorem 1. Choose $n_1 \in \mathbb{N}$ such that $\frac{(\lambda m+\log c_2)}{2n_1} < 1$. Thus for all $\lambda \geq \gamma$ and $n \geq n_1$

$$\frac{1}{n}\log V^n(\lambda) \leq \frac{\log 2}{n} + \max\left\{\frac{2\log(\lambda m+1)}{n} + \frac{1}{n}\log\tilde{V}^n, -\bar{W}(0,0)\right\}.$$

Taking limit as $n \to \infty$, we now have from the proof of Theorem 1 that for all $\lambda \geq \gamma$, $\limsup_{n\to\infty} \frac{1}{n}\log V^n(\lambda) \leq -\bar{W}(0,0)$. The result follows. □

Suppose that $c = 0 <$ ess supY_n and suppose further that $G(x,\theta)$ can be decomposed as $G(x,\theta) = G_1(x,\theta) + G_2(x)$, where G_1 is positively homogeneous, i.e., $G_1(\lambda x, \lambda\theta) = \lambda G_1(x,\theta)$ for $\lambda \geq 0$. Then (70) can be rewritten as

$$\begin{aligned}&\inf_{\lambda\geq 0,\theta\in\Theta} \mathbb{E}\left[\frac{\lambda}{n}\sum_{i=1}^n G_1(X_i,\theta) + \frac{\lambda}{n}\sum_{i=1}^n G_2(X_i) + 1\right]^+ \\ &= \inf_{\lambda\geq 0,\bar{\theta}\in\lambda\Theta} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n G_1(\lambda X_i,\bar{\theta}) + \frac{\lambda}{n}\sum_{i=1}^n G_2(X_i) + 1\right]^+ .\end{aligned} \tag{72}$$

If Θ is a convex set and G_1 is convex in (x,θ), the above minimization is a convex problem with variables λ and $\bar{\theta}$, and thus can be solved with well studied methods such as the gradient (sub)descent method, as shown in Example 3 in Section 5.

5 Computational experiments

We consider problems of the form (2) with $A = \mathbb{R}^m_+$ and Θ a compact, convex set, and approximate those problems by (50) in which $\varphi : \mathbb{R}^m \to \mathbb{R}$ is defined as

$$\varphi(y) = \Lambda \min(\|\min(y,0)\|_2^2, \varepsilon^2), \quad y \in \mathbb{R}^m, \tag{73}$$

with $\varepsilon > 0$ and $\Lambda > 0$ being fixed parameters. The function φ is bounded and Lipschitz continuous, and can be written as the point-wise minimum $\varphi_1 \wedge \varphi_2$ of the constant function $\varphi_1(y) \equiv \Lambda\varepsilon^2$ and $\varphi_2(y) = \Lambda\,\|\min(y,0)\|_2^2$.

As noted below (52), the problem (50) is equivalent to (53), which converges to the limiting problem (54) as $n \to \infty$ as shown in Corollary 1. In view of this convergence, we first solve (54) for which no use of IS is needed, and then use its solution as the initial point to solve (53) with a gradient based method in which the function and gradient values are computed using IS. For problems with $m = 1$ we also solve the buffered probability problem (72) and compare its solution with that of (53). Subsection 5.1 below discusses a reformulation of (54) and its properties, Subsection 5.2 gives details on implementing IS in solving (53), and Subsection 5.3 summarizes results from numerical examples.

5.1 Reformulation of the limiting problem

To solve (54), we reformulate it as a constrained optimization problem. As before we assume that $H_2^\theta(\alpha) < \infty$ for all $\theta \in \Theta$ and $\alpha \in \mathbb{R}^m$. Note that $L_2^\theta(\beta) \geq 0$ for all $\beta \in \mathbb{R}^m$ and $\theta \in \Theta$. Suppose also

$$\sup_{\theta\in\Theta}\inf_{\beta\geq 0} L_2^\theta(\beta) < \infty. \tag{74}$$

Then, by choosing the parameters Λ and ε in the definition of φ in (73) to satisfy $\Lambda\varepsilon^2 \geq \sup_{\theta\in\Theta} \inf_{\beta\geq 0} L_2^\theta(\beta)$, for each $\theta \in \Theta$ and $\beta \in \mathbb{R}^m$, we have

$$\varphi_1(\beta) + L_2^\theta(\beta) \geq \Lambda\varepsilon^2 \geq \inf_{\beta'\geq 0} L_2^\theta(\beta') = \inf_{\beta'\geq 0} (L_2^\theta(\beta') + \varphi_2(\beta')), \tag{75}$$

where the first inequality holds because $\varphi_1 \equiv \Lambda\varepsilon^2$ and $L_2^\theta(\beta) \geq 0$, and the last equality holds because $\varphi_2(\beta) = 0$ for $\beta \geq 0$. Consequently, for any $\theta \in \Theta$ we have

$$\begin{aligned} g(\theta) &= \inf\nolimits_{\beta\in\mathbb{R}^m} (\varphi(\beta) + L_2^\theta(\beta)) = \inf\nolimits_{\beta\in\mathbb{R}^m} (\varphi_2(\beta) + L_2^\theta(\beta)) \\ &= \inf\nolimits_{\beta\in\mathbb{R}^m} \sup\nolimits_{\alpha\in\mathbb{R}^m} \left[\varphi_2(\beta) + \langle\alpha,\beta\rangle - \log \mathbb{E}e^{\langle\alpha, G(X_1,\theta)\rangle}\right], \end{aligned} \tag{76}$$

where the first equality follows from Cramér's Theorem, and the second is from (75).

For each $\theta \in \Theta$ define a function $\Phi^\theta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ as

$$\Phi^\theta(\alpha,\beta) = \varphi_2(\beta) + \langle\alpha,\beta\rangle - \log \mathbb{E}e^{\langle\alpha, G(X_1,\theta)\rangle}. \tag{77}$$

Clearly, Φ^θ is continuous, convex with respect to β and concave with respect to α. The following proposition gives the existence of saddle-points of Φ^θ. We use $S_\theta \subset \mathbb{R}^m$ to denote the support of the random variable $G(X_1,\theta)$, i.e. the smallest closed set in $\mathbb{R}^m$ such that $\mathbb{P}(G(X_1,\theta) \in S_\theta) = 1$, and denote the closed convex hull of S_θ as ccS_θ.

Proposition 1 *Suppose that ccS_θ has a nonempty interior. Then, for each $\theta \in \Theta$, the set of saddle points of Φ^θ is nonempty and compact.*

Proof Fix $\theta \in \Theta$. By (Bertsekas 2009, Proposition 5.5.7), it suffices to show that for some $\bar\alpha \in \mathbb{R}^m$, $\bar\beta \in \mathbb{R}^m$ and $\bar\gamma \in \mathbb{R}$, the following sets

$$\{\alpha \in \mathbb{R}^m \mid \Phi^\theta(\alpha,\bar\beta) \geq \bar\gamma\} \text{ and } \{\beta \in \mathbb{R}^m \mid \Phi^\theta(\bar\alpha,\beta) \leq \bar\gamma\} \tag{78}$$

are nonempty and compact.

First, choose $\bar\alpha > 0$, and we show that the level sets of $\Phi^\theta(\bar\alpha,\cdot)$ (namely sets of the form $\{\beta \in \mathbb{R}^m \mid \Phi^\theta(\bar\alpha,\beta) \leq \bar\gamma\}$ for $\bar\gamma \in \mathbb{R}$) are compact. It is not hard to check that the recession function of $\Phi^\theta(\bar\alpha,\cdot)$ evaluated at a direction $d \in \mathbb{R}^m$ takes the value of $\langle\bar\alpha, d\rangle$ for $d \geq 0$ and ∞ for all other d. The recession function is nonpositive only at $d = 0$. By (Bertsekas 2009, Propositions 1.4.5-1.4.6), all level sets of $\Phi^\theta(\bar\alpha,\cdot)$ are compact.

Second, choose $\bar\beta$ from the interior of ccS_θ; then 0 belongs to the interior of cc$(S_\theta - \bar\beta)$, where $S_\theta - \bar\beta$ is the support of the random variable $G(X_1,\theta) - \bar\beta$. As shown in the proof of (Ellis 2007, Theorem VIII.4.3), the level sets of the log-moment generating function of $G(X_1,\theta) - \bar\beta$ are all compact, which are exactly sets of the form $\{\alpha \in \mathbb{R}^m \mid \Phi^\theta(\alpha,\bar\beta) \geq \bar\gamma\}$. We have thus shown that the sets (78) are compact for all $\bar\gamma \in \mathbb{R}$. By choosing $\bar\gamma$ to be sufficiently large, these sets are also nonempty. □

When saddle points of Φ^θ exist, they provide solutions to the outer minimization and inner maximization problems of $\inf_\beta \sup_\alpha \Phi^\theta(\alpha,\beta)$. When Φ^θ is differentiable, saddle points of Φ^θ can be further characterized by points where the partial derivatives vanish, so (54) can be written as

$$\begin{cases} \max\limits_{\theta\in\Theta,\alpha\in\mathbb{R}^m,\beta\in\mathbb{R}^m} & \Phi^\theta(\alpha,\beta) = \varphi_2(\beta) + \langle\alpha,\beta\rangle - \log \mathbb{E} e^{\langle\alpha, G(X_1,\theta)\rangle} \\ \text{s.t.} & \mathbb{E}[e^{\langle\alpha,G(X_1,\theta)\rangle}]\beta = \mathbb{E}[G(X_1,\theta)e^{\langle\alpha,G(X_1,\theta)\rangle}], \\ & 2\Lambda \min(\beta,0) + \alpha = 0. \end{cases}$$

With the equality constraints the above problem is nonconvex, but it has a favorable feature that evaluating the expected values in the objective function and the constraints does not necessitate the use of IS. In our numerical examples, we solve the problem by replacing the expected values by a numerical quadrature, or a sample average approximation when the latter is not available.

5.2 Implementing IS in the gradient method

In the numerical examples, X_i follows a normal distribution and the function $G(x,\theta)$ is piecewise linear in (x,θ). Using (Shapiro et al. 2009, Theorem 7.49), the gradient of g^n is

$$\triangledown g^n(\theta) = \frac{\mathbb{E}\left[\exp\left\{-n\varphi\left(\frac{1}{n}\sum_{i=1}^n G(X_i,\theta)\right)\right\} \triangledown_\theta \left[\varphi(\frac{1}{n}\sum_{i=1}^n G(X_i,\theta))\right]\right]}{\mathbb{E}\left[\exp\left\{-n\varphi\left(\frac{1}{n}\sum_{i=1}^n G(X_i,\theta)\right)\right\}\right]}. \quad (79)$$

For given $\theta\in\Theta$, let $\hat{\triangledown} g^n(\theta)$ be an SAA estimator for $\triangledown g^n(\theta)$. The gradient ascent update at the lth step is given by $\theta^{l+1} = \Pi_\Theta(\theta^l + o_l \hat{\triangledown} g^n(\theta^l))$, where o_l is the step size and Π_Θ is the projection operator onto the set Θ. The algorithm stops when the distance from $-\hat{\triangledown} g^n(\theta^l)$ to the normal cone to Θ at θ^l, is less than a threshold Δ.

Because the denominator of (79) is in the form of (3), with φ and $G(\cdot,\theta)$ playing roles of F and $G(\cdot)$ respectively, we can follow the procedures in Section 2 to estimate it using IS. Although the IS methods give guaranteed asymptotic performance bounds only for estimators of the denominator in (79), for our numerical studies we use the same change of measure to estimate the numerator as well. As discussed in Section 2, there are two approaches depending on whether X_i or $U_i = G(X_i,\theta)$ is used for the change of measure. Below we outline the implementation for both approaches.

Change of measure on X_i To implement IS based on a change of measure on X_i, we follow the procedure outlined below Theorem 1 to construct a generalized subsolution/control. We select $\{(\bar{W}_k,\bar{a}_k)\}_{k=1,2}$ from the family of affine subsolution/control pairs $(\bar{W},\bar{a})$, where $\bar{W}$ is of the form (43) and $\bar{a}$ satisfies (45). We impose the requirements $\bar{W}_1(y,1) \le 2\phi_1(y)$ and $\bar{W}_2(y,1)\le 2\phi_2(y)$ for all $y\in\mathbb{R}^m$, to guarantee (46) holds with φ in place of F. Since $\varphi_1(y)\equiv \Lambda\varepsilon^2$, we simply let $\bar{W}_1(y,t)\equiv 2\Lambda\varepsilon^2$; it can be verified that $\bar{a}_1 = 0$ satisfies (45). The coefficients for $\bar{W}_2$ and the corresponding $\bar{a}_2$ are determined by the following optimization problem:

$$\max_{\bar{a}_2,\bar{c},u} \left\{ \bar{c} - H(-\bar{a}_2,-u) - H_1(\bar{a}_2) \quad \text{s.t.} \;\; u\le 0, \quad \bar{c}\le 0, \;\; \bar{c} + \frac{u^T u}{8\Lambda} \le 0 \right\}.$$

The constraints arise from the requirement $\bar{W}_2(y,1) \leq 2\phi_2(y)$ for all y and the objective function reflects our aim to maximize $\bar{W}_2(0,0)$ while satisfying (44) $\bar{W}$ replaced with $\bar{W}_2$. After solving the above optimization problem, we define $\bar{W}_2$ as

$$\bar{W}_2(y,t) = \bar{c} + \langle u, y\rangle - (1-t)\big(H(-\bar{a}_2, -u) + H_1(\bar{a}_2)\big).$$

It is easy to check $(\bar{W}_2, \bar{a}_2)$ is a subsolution/control pair. With $\big\{(\bar{W}_k, \bar{a}_k)\big\}_{k=1,2}$ obtained, we next construct a generalized subsolution/control by defining $\bar{W}^\delta$ and ρ_k^δ as in (47) and (48), and then follow the procedure below (27) to obtain an unbiased sample average estimator for the denominator of (79), in the form (29) with F replaced by φ and $(\bar{W}, \rho_k)$ by $(\bar{W}^\delta, \rho_k^\delta)$. For the numerator of (79), we use the same generalized subsolution/control to construct the change of measure on X_i, so the unbiased estimator for the numerator is similar to (29) except that $e^{-nF(\bar{Y}_n^n)}$ is replaced by $e^{-n\varphi(\bar{Y}_n^n)}\nabla_\theta\varphi(\bar{Y}_n^n)$.

Change of measure on U_i To conduct IS based on a change of measure on U_i we follow (Dupuis and Wang 2007). For $k=1,2$ we let

$$\beta_k \in \operatorname*{argmin}_{\beta\in\mathbb{R}^m} \Big[L_2^\theta(\beta) + \varphi_k(\beta)\Big] \qquad \text{and} \qquad \alpha_k \in \operatorname*{argmax}_{\alpha\in\mathbb{R}^m} \Big[\langle\alpha, \beta_k\rangle - H_2^\theta(\alpha)\Big],$$

where H_2^θ and L_2^θ are defined in (55) and (56), and define $\bar{W}_k : \mathbb{R}^m \times [0,1] \to \mathbb{R}$ as

$$\bar{W}_k(y,t) = -2\langle\alpha_k, y\rangle + 2[\varphi_k(\beta_k) + \langle\alpha_k, \beta_k\rangle] - 2(1-t)H_2^\theta(\alpha_k).$$

Since φ_1 is a constant, $\alpha_1 = 0$. Next we define $\bar{W}^\delta$ and compute ρ_k^δ similarly as above to obtain a generalized subsolution/control as in Definition 1 (see (Dupuis and Wang 2007)), and then follow the procedure below (16) to obtain an unbiased estimator for the denominator of (79). Again the numerator of (79) is estimated using the same change of measure. The above definitions of $\bar{W}_k$, α_k and β_k imply that

$$\bar{W}_1(0,0) \wedge \bar{W}_2(0,0) = \min_{k=1,2}\Big(2(\varphi_k(\beta_k) + L_2^\theta(\beta_k)\Big) = 2\inf_{\beta\in\mathbb{R}^m}[\varphi(\beta) + L_2^\theta(\beta)] = 2\gamma,$$

where γ is as defined in (15) with φ in place of F. It follows that the estimator for the denominator constructed using $(\bar{W}^\delta, \rho_k^\delta)$ is $\delta\log 2$ - asymptotically optimal.

5.3 Numerical results

5.3.1 Example 1

We use this example with $h = m = d = 1$ to first compare the two IS schemes discussed in Sections 2.1 and 2.2 with ordinary Monte-Carlo simulation, and then solve problems (54) and (53). The parameters of φ in (73) are $(\Lambda, \varepsilon) = (10^5, 0.01)$. The function G is defined as $G(x,\theta) = (x-\theta)^+ - 0.4(1.5-\theta)$. We let $\Theta = [0, 1.5]$, $n = 100$ and η be the standard normal distribution.

Tables 1, 2 and 3 report the performance of estimators for $p(\theta)$ as defined in (50), using ordinary Monte-Carlo simulation, the scheme based on a change of measure on $U_i = G(X_i, \theta)$ (i.e, the method in (Dupuis and Wang 2004, 2007)), and our proposed scheme based on a change of measure on X_i respectively. In each of the three schemes, we generate N independent realizations of the unbiased estimator for $p(\theta)$, which is $e^{-n\varphi\left(\frac{1}{n}\sum_{i=1}^{n} G(X_i,\theta)\right)}$ in ordinary Monte-Carlo simulation, and is in the form of (18) or (29) in the two IS schemes. We report the natural logarithms of the sample average and the sample standard deviation (scaled by $1/\sqrt{N}$), denoted as "log sample mean" and "log sample std" respectively, for different values of θ and N. We also report the CPU time, which includes time spent on sampling and calculating the reported values. In Table 1, some values under $N = 5 \times 10^3$ are $-\infty$, because the event $\frac{1}{n}\sum_{i=1}^{n} G(X_i, \theta) > 0$ does not occur in any realization. Moreover, the log sample std values are close to log sample means in many cases, which means that the sample mean estimates are not stable. With $N = 5 \times 10^5$ we get better estimates for $p(\theta)$, though the number of realizations in which the rare event occurs is still very small.

Table 1 Estimation of $p(\theta)$ using ordinary Monte-Carlo simulation in Example 1

	θ	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
$N = 5 \times 10^3$	log sample mean	-7.1308	-35.4495	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-5.2430	-0.8957
	log sample std	-7.8243	-35.4495	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-6.8907	-4.9723
	CPU time (sec)	0.0500	0.0200	0.0200	0.0600	0.0600	0.0400	0.0500	0.0299
$N = 5 \times 10^5$	log sample mean	-7.2532	-8.9291	-10.8197	-11.3306	-10.9251	-8.9533	-5.3179	-0.9424
	log sample std	-10.1896	-11.0293	-11.9710	-12.2264	-12.0237	-11.0423	-9.2264	-7.2831
	CPU time (sec)	2.3699	2.8100	2.6100	2.4000	2.5100	2.3999	2.3899	2.2400

The log sample std values in Table 2 are considerably smaller than the log sample mean values, which means that the sample mean estimates in the IS scheme that changes measure on U_i are stable. Here, we add a row labeled "prop" to report the proportions of rare events among all realizations. For each fixed θ, we observe the event $\frac{1}{n}\sum_{i=1}^{n} \bar{U}_i > 0$ ($\bar{U}_i$ is the random variable replacing U_i in the change of measure) to occur for about 50% of the realizations, a dramatic improvement from ordinary Monte-Carlo simulation. On the other hand, computation time needed in this scheme is significantly larger as expected, because constructing any realization of $\bar{U}_i$ requires numerically solving an equation to invert the cumulative distribution function of $\bar{U}_i$, and we need to solve Nn such equations to estimate $p(\theta)$ for a fixed θ using N samples.

Table 2 Estimation of $p(\theta)$ with the change of measure on U_i in Example 1

	θ	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
$N = 5 \times 10^3$	log sample mean	-7.2107	-9.2225	-10.7867	-11.4086	-10.9536	-8.8700	-5.1780	-0.8774
	log sample std	-10.8261	-12.7341	-14.2155	-14.8412	-14.3785	-12.3555	-8.8396	-5.1316
	CPU time (sec)	24.9300	27.4700	22.0700	20.5300	20.4700	19.4100	17.6200	17.0400
	prop	0.4904	0.4908	0.4830	0.4798	0.4792	0.4850	0.4768	0.4650

In Table 3 the log sample std values are also smaller than the log sample mean values by a clear margin, so the sample mean estimates are more stable than those in ordinary Monte-Carlo simulation. The prop values are smaller than 50% but significantly larger than the sample mean values, which means that our proposed scheme is not as efficient as the scheme based on the change of measure on U_i, but still significantly better than ordinary Monte-Carlo simulation. The reported CPU times are significantly lower than those in Table 2 even with the larger sample size $N = 5 \times 10^5$, because to simulate the replacement variable $\bar{X}_i$ we only need to draw samples from the standard normal distribution and then suitably translate and scale these values, a main advantage of the proposed IS scheme over the existing scheme.

Table 3 Estimation of $p(\theta)$ with the change of measure on X_i in Example 1

	θ	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
$N = 5 \times 10^3$	log sample mean	-7.2653	-9.1440	-10.7649	-11.5153	-10.9374	-8.3427	-5.1589	-0.8984
	log sample std	-9.9762	-11.2694	-12.0563	-12.7407	-11.8451	-9.5270	-7.3090	-4.9920
	CPU time	0.2700	0.0900	0.1299	0.1100	0.1499	0.1000	0.1199	0.0999
$N = 5 \times 10^5$	log sample mean	-7.2719	-9.2986	-10.8466	-11.6375	-11.0927	-9.0575	-5.2923	-0.9423
	log sample std	-12.0156	-13.4715	-14.4857	-14.7782	-14.0001	-12.2827	-9.6416	-7.3022
	CPU time	3.7100	5.0000	4.9199	4.3499	3.8299	3.5699	3.3500	3.3699
	prop	0.1216	0.0530	0.0195	0.0068	0.0034	0.0040	0.0198	0.3937

Finally, to find the optimal θ that maximizes $p(\theta)$ or equivalently $g^n(\theta)$, we first solve an SAA problem of the limiting problem (54) to find $\theta^* = 0.6229$ with an optimal value 0.0898, by directly using the Matlab nonlinear programming solver `fmincon`. We then implement the gradient ascent method to (53) with an initial point $\theta^0 = \theta^*$ and a diminishing step size $o_l = \frac{0.1}{\sqrt{l+1}}$ for 50 iterations. Figure 1 shows nine trajectories of objective values where the objective values of (53) and its gradients are estimated by ordinary Monte-Carlo simulation with $N = 2.5 \times 10^6$. Figure 2 shows nine trajectories where the objective values and gradients are estimated using the scheme that changes measure on X_i. The more concentrated trajectories in Figure 2 shows the effect of variance reduction.

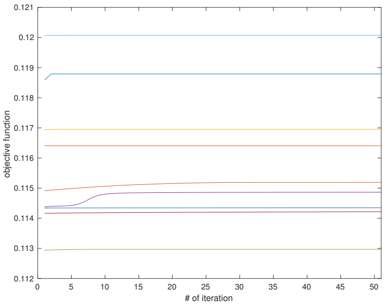

Fig. 1 Trajectories of objective values of (53) in the gradient method for Example 1 with ordinary Monte-Carlo simulation

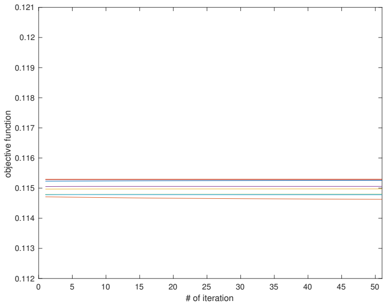

Fig. 2 Trajectories of objective values of (53) in the gradient method for Example 1 with the IS scheme from Section 2.2

5.3.2 Example 2

In this section, we consider an example with $h = m = d = 5$. The ith component of the function G is defined as

$$G_i(x_i, \theta_i) \doteq (x_i - \theta_i)^+ - b_i(c_i - \theta_i),$$

with $b = [0.3, 0.2, 0.3, 0.3, 0.2]^T$ and $c = [1, 2, 2, 1, 2]^T$. The feasible set Θ is $[0, c]$, and the random variables $\{X_i\}_{i=1}^n$ are i.i.d. multivariate normal with mean 0 and a randomly generated covariance matrix. Other parameters used to define the problem and the algorithm are summarized in Table 4.

Table 4 Parameters in Example 2

Λ	ε	n	N	o_l	Δ
10^5	0.01	50	2.5×10^6	$o_l = \frac{0.5}{\sqrt{l+1}}$	10^{-4}

Note that, in such a multidimensional setting, the scheme that changes measure on U_i becomes impractical due to its computation complexity. The scheme that changes measure on X_i remains feasible and is used in the gradient method to estimate the gradients and objective values.

We start by solving the limiting problem (54) by the Matlab function $fmincon$. Using different initial points, we find a SAA solution $\theta^* = [0.6270, 1.6872, 0.0000, 0.4105, 1.2983]^T$ with the optimal value 0.1143. We then start from θ^* and apply the gradient method to problem (53), using IS to evaluate the function and gradient values. The method stops after 337 iterations with the solution

$$\theta^{337} = [0.6007, 1.5190, 0.4807, 0.4088, 1.2165]^T,$$

the optimal value 0.3208, and $p(\theta^{337}) = 10^{-7}$. The proportion of rare event among all realization is about 0.005%, which is much larger than the probability $p(\theta^{337})$, showing the effect of IS. Figure 3 displays objective values at the iterations.

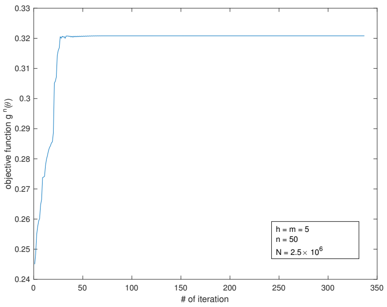

Fig. 3 Objective values of (53) for the gradient method in Example 2

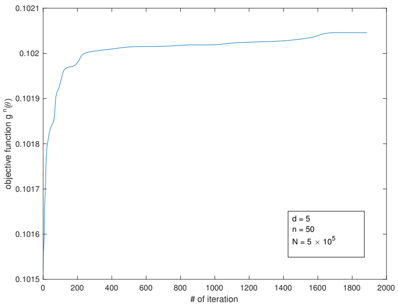

Fig. 4 Objective values of (53) for the gradient method in Example 3 ($n = 50$)

5.3.3 Example 3

In this example, we let $h = d = 5$ and $m = 1$. The function G is defined as

$$G(x, \theta) = f^T(x - \theta)^+ - b^T(c - \theta), \ \theta \in \Theta \subset \mathbb{R}^d, \ x \in \mathbb{R}^d,$$

with $b = [0.3, 0.2, 0.3, 0.3, 0.2]^T$, $c = [1, 2, 2, 1, 2]^T$ and $f = [1, 1, 1, 1, 1]^T$. The distribution of X_i is the same as in Example 2, and Θ is $[0, c]$. We use this example to compare the solutions to (53) and the buffered probability problem (72) with $n = 50$. SAA solution to the limiting problem is $\theta^* = [0.7863, 1.2361, 0.7860, 0.7647, 0.8842]^T$ with the optimal value 0.0894. For the problem (53), we let $N = 5 \times 10^5$, $o_l = \frac{0.5}{\sqrt{l+1}}$ and $\Delta = 10^{-4}$. After 1886 iterations, the stopping criterion is satisfied. The optimal solution is $\theta^{1886} = [0.7359, 1.1708, 0.7526, 0.7656, 0.8524]^T$, with the optimal value 0.1020 and $p(\theta^{1886}) = 6.1 \times 10^{-3}$. Approximately 2.78% of the realizations are rare events. Figure 4 shows objective values at the iterations, where oscillations in the paths of objective values are largely due to variations in estimating objective values and gradients (oscillations in the Figure 3 in Example 2 are due to the same reason).

For this example the problem (72) becomes

$$\min_{\lambda\geq 0,\bar{\theta}\in\lambda\Theta} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f^T(\lambda X_i-\bar{\theta})^+ - b^T(\lambda c-\bar{\theta})+1\right]^+ .$$

In the application of the gradient method, we use the IS scheme provided in Section 2.2 with $F=\infty 1_{A^c}$ to find the change of measure on X_i, to estimate the objective values and gradients. We arbitrarily select an initial point $(\theta^0,\lambda^0)=(f/2,1)$, which corresponds to $(\bar{\theta}^0,\lambda^0)=(\lambda^0\theta^0,\lambda^0)=(f/2,1)$. We then use a fixed length stepsize $o_l=0.1\|\hat{\nabla}h(\bar{\theta}^l,\lambda^l)\|_2^{-1}$ (i.e., $\|\bar{\theta}^{l+1}-\bar{\theta}^l\|_2=0.1$ for all l) to achieve a relatively large progress at each step. We also compute the value $\theta^l=\bar{\theta}^l/\lambda^l$ at each iteration. The solution we find is $\theta^{292}=[0.7314, 1.1534, 0.7312, 0.7631, 0.8369]^T$ with the optimal value 0.0159. Figure 5 shows the objective values at each iteration. Before the algorithm terminates, the objective value at iteration l is not necessarily close to the buffered probability evaluated at θ^l, because λ^l may be far from the optimal λ that defines the buffered probability. After (72) is solved, we calculate the probabilities and buffered probabilities at the θ_l value of each iteration and plot them in Figure 6, where the solid and dashed lines represent buffered probabilities and probabilities respectively.

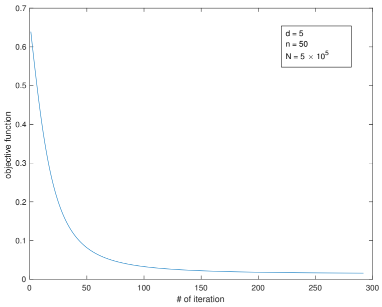

Fig. 5 Objective values of (72) for the gradient method in Example 3 ($n = 50$)

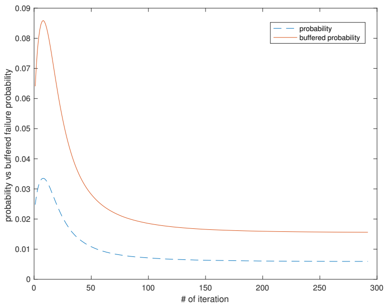

Fig. 6 Probabilities and buffered probabilities corresponding to Figure 5

References

Barrera J, Homem-de Mello T, Moreno E, Pagnoncelli BK, Canessa G (2016) Chance-constrained problems and rare events: An importance sampling approach. Mathematical Programming 157(1):153–189

Bertsekas DP (2009) Convex Optimization Theory. Athena Scientific Belmont

Bremer I, Henrion R, Möller A (2015) Probabilistic constraints via SQP solver: Application to a renewable energy management problem. Computational Management Science 12(3):435–459

Bucklew JA (1990) Large Deviation Techniques in Decision, Simulation, and Estimation. Wiley, New York

Calafiore GC, Campi MC (2006) The scenario approach to robust control design. IEEE Transactions on Automatic Control 51(5):742–753

Chen JC, Lu D, Sadowsky JS, Yao K (1993) On importance sampling in digital communications. I. Fundamentals. IEEE Journal on Selected Areas in Communications 11(3):289–299

Collamore JF (2002) Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. Annals of Applied Probability 12(1):382–421

Dentcheva D, Martinez G (2013) Regularization methods for optimization problems with probabilistic constraints. Mathematical Programming 138:223–251

Dupuis P, Ellis RS (2011) A Weak Convergence Approach to the Theory of Large Deviations. John Wiley & Sons

Dupuis P, Wang H (2004) Importance sampling, large deviations, and differential games. Stochastics: An International Journal of Probability and Stochastic Processes 76(6):481–508

Dupuis P, Wang H (2007) Subsolutions of an Isaacs equation and efficient schemes for importance sampling. Mathematics of Operations Research 32(3):723–757

Ellis RS (2007) Entropy, Large Deviations, and Statistical Mechanics. Springer

Evans M, Swartz T (2000) Approximating integrals via Monte Carlo and deterministic methods. OUP Oxford

Glasserman P, Wang Y, et al. (1997) Counterexamples in importance sampling for large deviations probabilities. The Annals of Applied Probability 7(3):731–746

L'Ecuyer P, Tuffin B (2011) Approximating zero-variance importance sampling in a reliability setting. Annals of Operations Research 189(1):277–297

Mafusalov A, Uryasev S (2014) Buffered probability of exceedance: Mathematical properties and optimization algorithms. Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, Research Report 1

Mafusalov A, Shapiro A, Uryasev S (2015) Estimation and asymptotics for buffered probability of exceedance. Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, Research Report 5

Nemirovski A, Shapiro A (2006) Convex approximations of chance constrained programs. SIAM Journal on Optimization 17(4):969–996

Owen A, Zhou Y (2000) Safe and effective importance sampling. Journal of the American Statistical Association 95(449):135–143

Pagnoncelli B, Ahmed S, Shapiro A (2009) Sample average approximation method for chance constrained programming: Theory and applications. Journal of Optimization Theory and Applications 142(2):399–416

Prékopa A (2013) Stochastic Programming. Springer Science & Business Media

Ridder A (2005) Importance sampling simulations of Markovian reliability systems using cross-entropy. Annals of Operations Research 134(1):119–136

Rockafellar RT, Royset JO (2010) On buffered failure probability in design and optimization of structures. Reliability Engineering & System Safety 95(5):499–510

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. Journal of Risk 2:21–42

Sadowsky JS (1991) Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue. IEEE Transactions on Automatic Control 36(12):1383–1394

Sadowsky JS (1996) On Monte Carlo estimation of large deviations probabilities. The Annals of Applied Probability 6(2):399–422

Sadowsky JS, Bucklew JA (1990) On large deviations theory and asymptotically efficient Monte Carlo estimation. IEEE Transactions on Information Theory 36(3):579–588

Shapiro A, Dentcheva D, Ruszczyński A (2009) Lectures on Stochastic Programming: Modeling and Theory. SIAM

Siegmund D (1976) Importance sampling in the Monte Carlo study of sequential tests. The Annals of Statistics 4(4):673–684

Van Ackooij W, Henrion R (2014) Gradient formulae for nonlinear probabilistic constraints with Gaussian and Gaussian-like distributions. SIAM Journal on Optimization 24(4):1864–1889