Journal of Complexity 29 (2013) 4–26

Contents lists available at SciVerse ScienceDirect

Journal of Complexity

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

Optimal Monte Carlo integration with fixed relative precision

Lesław Gajek [a], Wojciech Niemiro [b,c], Piotr Pokarowski [c,*]

[a] Faculty of Mathematics and Computer Science, University of Łódź, Poland
[b] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
[c] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland

ARTICLE INFO

Article history:
Received 2 July 2012
Accepted 6 September 2012
Available online 14 September 2012

Keywords:
(ε, α)-approximation
Worst case complexity
Rare event simulation
Exponential inequalities
Mean square error
Sequential methods

ABSTRACT

We consider Monte Carlo algorithms for computing an integral $\theta = \int f d\pi$ which is positive but can be arbitrarily close to 0. It is assumed that we can generate a sequence X_n of uniformly bounded random variables with expectation θ. Estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_N)$ is called an (ε, α)-approximation if it has fixed relative precision ε at a given level of confidence $1 - \alpha$, that is it satisfies $\mathbb{P}(|\hat{\theta} - \theta| \leq \varepsilon\theta) \geq 1 - \alpha$ for all problem instances. Such an estimator exists only if we allow the sample size N to be random and adaptively chosen.

We propose an (ε, α)-approximation for which the cost, that is the expected number of samples, satisfies $\mathbb{E}N \sim 2\ln\alpha^{-1}/(\theta\varepsilon^2)$ for $\varepsilon \to 0$ and $\alpha \to 0$. The main tool in the analysis is a new exponential inequality for randomly stopped sums.

We also derive a lower bound on the worst case complexity of the (ε, α)-approximation. This bound behaves as $2\ln\alpha^{-1}/(\theta\varepsilon^2)$. Thus the worst case efficiency of our algorithm, understood as the ratio of the lower bound to the expected sample size $\mathbb{E}N$, approaches 1 if $\varepsilon \to 0$ and $\alpha \to 0$.

An L^2 analogue is to find $\hat{\theta}$ such that $\mathbb{E}(\hat{\theta} - \theta)^2 \leq \varepsilon^2\theta^2$. We derive an algorithm with the expected cost $\mathbb{E}N \sim 1/(\theta\varepsilon^2)$ for $\varepsilon \to 0$. To this end, we prove an inequality for the mean square error of randomly stopped sums. A corresponding lower bound also behaves as $1/(\theta\varepsilon^2)$. The worst case efficiency of our algorithm, in the L^2 sense, approaches 1 if $\varepsilon \to 0$.

* Corresponding author.
E-mail address: pokar@mimuw.edu.pl (P. Pokarowski).

0885-064X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2012.09.001

0. Introduction

Typical Monte Carlo (MC) algorithms aim at approximating an integral

$$\theta = \int_{\mathcal{Y}} f(y)p(y)dy, \tag{0.1}$$

where p is a probability density and f is a real function on $\mathcal{Y} \subset \mathbb{R}^d$ (alternatively, $\mathcal{Y}$ can be a large finite space and the integral should be replaced by a sum). Assume f is bounded, say $0 \le f \le 1$. The crude MC method is to generate i.i.d. samples $Y_1, \ldots, Y_n, \ldots$ from the probability density p and use the fact that $\theta = \mathbb{E}X_n$, where $X_n = f(Y_n)$. In particular, if $f = \mathbb{I}_E$, then we obtain a Bernoulli scheme with the probability of success $\mathbb{P}(X_n = 1) = \theta$. In the sequel we need not refer to the samples Y_n in the "original" space $\mathcal{Y}$ and will work only with variables X_n in $[0, 1]$. The i.i.d. assumption will later be relaxed.

If the order of magnitude of $\theta > 0$ is unknown then the absolute error $|\hat{\theta} - \theta|$ is meaningless and the *relative* error $|\hat{\theta} - \theta|/\theta$ should be considered. The main motivation comes from rare event simulation; see e.g. [2, Chapter 6] or [1,3,4,32,40,41]. We say that MC estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_N)$ is an (ε, α)-*approximation* of θ (see e.g. [25] or [19, Section 2.5]) if

$$\mathbb{P}(|\hat{\theta} - \theta| \le \varepsilon\theta) \ge 1 - \alpha \tag{0.2}$$

holds for all instances of the problem. Here ε is a bound on the relative error and $1 - \alpha$ is a given level of confidence.

If θ is not bounded away from zero, then *no* estimator which uses a number of samples N fixed in advance can satisfy (0.2) for all i.i.d. sequences X_n satisfying $\mathbb{E}X_n = \theta$ and $0 \le X_n \le 1$. This fact is not surprising and it follows from much stronger results which we prove in this paper. Therefore, to ensure (0.2), in general we have to use algorithms with *adaptive stop criteria*. This means that we allow the number N of generated samples to be random and to depend on the results of previous computations. In the statistical literature, procedures which use adaptively chosen numbers of samples are known as *sequential methods* [44]. Throughout this paper, the terms "adaptively stopped" and "sequential" are treated as synonyms.

Fixed relative precision estimation was considered by Nádas [36] in 1969. His solution follows earlier work of Chow and Robbins [12] and it is *asymptotic*. More precisely, the procedure proposed in [36] only guarantees that the probability in (0.2) approaches $1 - \alpha$ in the limit, as $\varepsilon \to 0$. The asymptotics with $\alpha \to 0$ and ε fixed is considered in [43].

An *exact* solution of the (ε, α)-approximation problem appeared in 1995 and was given by Dagum et al. [15]. They prove an elegant theorem about the complexity of the problem and optimality of the proposed procedure. Specifically, they show that the *relative efficiency* of their algorithm $\mathcal{AA}$ versus an arbitrary (ε, α)-approximation $\mathcal{BB}$ is bounded by a universal constant, in their notation equal to (c/c'): the expected number of samples used by in $\mathcal{AA}$ is at most (c'/c) times as big as in $\mathcal{BB}$.

The main aim of our paper is to show the *worst case optimality* of an (ε, α)-approximation with the stopping rule $N_r = \min\{n : S_n \ge r\}$, where $S_n = \sum_{i=1}^{n} X_i$ (this rule is also considered in [15]). An *upper bound* for the cost of our algorithm, given in Theorem 3.1, has the form

$$\theta\mathbb{E}N_r \le \mathrm{Up}(\varepsilon, \alpha),$$

where $\mathrm{Up}(\varepsilon, \alpha) \sim 2\ln\alpha^{-1}/\varepsilon^2$ for $\varepsilon \to 0$ and $\alpha \to 0$. This result plays an analogous role to Theorem $\mathcal{AA}$ in [15]. Our main tool is a sequential analogue of Hoeffding's inequality [22], given in Theorem 2.3.

We also prove a *lower bound* for the worst case complexity of the problem. We consider an *arbitrary* (ε, α)-*approximation* with the number of samples N. Theorem 3.2 shows that, when the algorithm is applied to the *Bernoulli scheme* with the probability of success θ, then

$$\liminf_{\theta \to 0} \theta\mathbb{E}N \ge \mathrm{Low}(\varepsilon, \alpha).$$

This result plays a role analogous to the "Lower Bound Theorem" concerning algorithm $\mathcal{BB}$ in [15]. The main difference is that our bounds satisfy $\mathrm{Low}(\varepsilon, \alpha)/\mathrm{Up}(\varepsilon, \alpha) \to 1$ for $\varepsilon \to 0$ and $\alpha \to 0$.

Roughly speaking, our counterpart of the constant (c'/c) of [15] approaches 1 for the most difficult problems, if great precision and high confidence level are required. In this way we improve upon the results of [15].

An alternative way of formalizing the notion of fixed relative precision is to use the *mean square error* (MSE) and require that

$$\mathbb{E}(\hat{\theta} - \theta)^2 \leq \varepsilon^2 \theta^2. \tag{0.3}$$

The mean square criterion is widely accepted in complexity theory and thus (0.3) is in our paper considered in parallel with (0.2). To our knowledge, there are no earlier nonasymptotic results concerning fixed relative precision estimation in the MSE sense. Bounds for MCMC algorithms derived in [30,31,28,29,34,35,42] are nonasymptotic but they concern absolute, not relative, error.

In Theorem 4.1 we prove an *upper bound* for the MSE of an estimator based on the stopping rule N_r. From this we deduce that the sample size sufficient to ensure (0.3) satisfies

$$\theta \mathbb{E} N_r \leq \mathrm{Up}_{L^2}(\varepsilon),$$

where $\mathrm{Up}_{L^2}(\varepsilon) \sim 1/\varepsilon^2$ for $\varepsilon \to 0$, Corollary 4.4. This can be compared with a *lower bound* given in Theorem 4.5 which shows that, when an algorithm satisfies (0.3) for every Bernoulli scheme with probability of success θ, then

$$\limsup_{\theta \to 0} \theta \mathbb{E} N \geq \mathrm{Low}_{L^2}(\varepsilon),$$

where $\mathrm{Low}_{L^2}(\varepsilon) \sim 1/\varepsilon^2$ for $\varepsilon \to 0$. Since $\mathrm{Low}_{L^2}(\varepsilon)/\mathrm{Up}_{L^2}(\varepsilon) \to 1$ with $\varepsilon \to 0$, the algorithm based on N_r samples is worst case optimal also in the L^2 sense.

Although in this paper the emphasis is on lower bounds, our *upper* bounds do not require that X_n must be i.i.d. Instead we make a much weaker martingale-type assumption, Assumption 1.1. To explain the motivation, recall (0.1). The *crude* MC method uses i.i.d. samples $X_n = f(Y_n)$, with Y_n being drawn from the probability density p. To enhance efficiency, *importance sampling* (IS) is usually applied. We generate i.i.d. samples Y_n from some other probability density q and compute $X_n = f(Y_n)p(Y_n)/q(Y_n)$ so that to have $\mathbb{E}X_n = \theta$. Ingenious ways of choosing q in specific problems are described in virtually every monograph on MC. We obtain the *adaptive* version of importance sampling (AIS) if we allow the current sampling distribution q to depend on previous samples; see e.g. [7,17,23,26,27,39,45]. The idea is to "learn what the optimal or good q should look like" from past experience. More precisely, at stage n we generate a sample Y_n from a probability density $q_n = q_n(\cdot|Y_1, \ldots, Y_{n-1})$. Random variables $X_n = f(Y_n)p(Y_n)/q_n(Y_n)$ are neither independent nor have the same distribution. However, it is still true that $\mathbb{E}X_n = \theta$ and $X_n - \theta$ are martingale differences. In this setting we state our main results.

Uniform boundedness of variables X_n is, unfortunately, essential for our optimality results. The case of *unbounded* i.i.d. samples is also considered, but in this setting we are only able to prove *upper* bounds. We consider estimators based on the median trick of Jerrum et al. [24], optimized in [38]. We discuss the relation between the efficiency of *rigorous* (ε, α)-approximation and that of its *asymptotic* counterpart, which is based on the normal approximation. We construct an adaptively stopped (ε, α)-approximation under the assumption that $\mathrm{Var}\, X_n/\theta$ is bounded. Note that this condition is much weaker than "logarithmic efficiency", i.e. boundedness of $\mathrm{Var}\, X_n/\theta^{2-\delta}$ for every $\delta > 0$.

Let us point out some possible applications of our results. Much work in theoretical computer science has been devoted to finding (ε, α)-approximations for problems where exact computation is NP-hard. Examples include approximating the permanent, solving the Ising model or other problems of statistical mechanics, computing the volume of a convex body and solving for network reliability; see [15] and the references therein. Another example is approximating posterior probabilities in Bayesian networks [9–11,16]. Several families of MCMC algorithms aimed at specific problems do not use adaptive stopping but instead require lower bounds, known in advance, on the target quantity θ. Formulas for the cost essentially depend on these lower bounds; see e.g. [38]. Adaptively stopped procedures, e.g. those based on our results, can substantially reduce overestimation of the cost.

The paper is organized as follows. In Section 1 we gather definitions and assumptions. Section 2 contains our main probabilistic tools, i.e. sequential inequalities. The main results are given in

Sections 3 and 4. In the former we derive an optimal (ε, α)-approximation, in the latter, an optimal relative MSE approximation. Theoretical results are complemented by a numerical study in Section 5. In Section 6 we consider extensions to the case of unbounded samples. For the convenience of the reader, necessary probabilistic results which play a crucial role in our paper are given in the Appendix.

1. Preliminaries and definitions

We assume that the MC algorithms considered in this paper take as an input a (potentially) infinite sequence of samples $X_1, X_2, \ldots$. They are regarded as random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration, that is a sequence of σ-fields $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$. Let us make the following basic assumption.

Assumption 1.1. For some $\theta > 0$ the following holds: X_n is $\mathcal{F}_n$-measurable, $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = \theta$ a.s. and $0 \le X_n \le 1$ a.s. for $n = 1, 2, \ldots$.

Note that if the X_n are independent (not necessarily identically distributed), $0 \le X_n \le 1$ a.s. and $\mathbb{E}X_n = \theta$ then Assumption 1.1 holds with $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.

An adaptive stop criterion is usually specified by a sequence of Borel measurable mappings $s_n : [0, 1]^n \to \{0, 1\}$. The number of generated samples is determined via $N = \inf\{n \ge 1 : s_n(X_1, \ldots, X_n) = 1\}$ and it must satisfy $\mathbb{P}(N < \infty) = 1$ for all input sequences X_n. In the terminology of statistical sequential analysis, N is a stopping rule [13].

Definition 1.2. A random variable $N : \Omega \to \{0, 1, \ldots\}$ is called a *stopping rule* if $\{N \le n\} \in \mathcal{F}_n$ for $n = 1, 2, \ldots$ and $\mathbb{P}(N < \infty) = 1$.

Consider another sequence of Borel measurable mappings $\hat{\theta}_n : [0, 1]^n \to \mathbb{R}$, where $\hat{\theta}_n$ is interpreted as an estimator of θ depending on $X_1, \ldots, X_n$. If N is a random variable, we can define $\hat{\theta} = \hat{\theta}_N(X_1, \ldots, X_N)$.

Definition 1.3. An adaptively stopped MC algorithm is a pair $(N, \hat{\theta})$, where N is a stopping rule and $\hat{\theta} = \hat{\theta}_N(X_1, \ldots, X_N)$.

In statistical parlance, $\hat{\theta}$ is a sequential estimator. Let us now formally define two concepts which play central roles in our paper. Consider a class $\Im$ of sequences $X_1, X_2, \ldots$ regarded as inputs of the algorithm.

Definition 1.4. An adaptively stopped MC algorithm $(N, \hat{\theta})$ is an (ε, α)-*approximation* on $\Im$ if (0.2) holds for every input sequence belonging to $\Im$.

Definition 1.5. An adaptively stopped MC algorithm $(N, \hat{\theta})$ is an ε-bounded relative root mean square error (ε-RRMSE) approximation on $\Im$ if (0.3) holds for every input sequence belonging to $\Im$.

Of course, at each reference to Definition 1.4 or 1.5 we have to specify the class $\Im$, i.e. make clear what conditions are imposed on the sequence $X_1, X_2, \ldots$.

The definition of ε-RRMSE approximation resembles that of the "bounded relative variance estimator" (BRV-estimator); see e.g. [46,47]. However, in Definition 1.5 we consider possibly sequential estimators and we do not require unbiasedness. Note that in recent literature the term "bounded relative error estimator" is often treated as a synonym of "BRV-estimator"; see e.g. [2]. In our context this terminology might be misleading, because in older papers the term "bounded relative error estimator" was used instead of (ε, α)-approximation [18,21].

The algorithms which will be later shown to be worst case optimal use the following stopping rule. Let $S_n = \sum_{i=1}^{n} X_i$. Fix $r > 0$ and define N_r by

$$N_r = \min\{n : S_n \ge r\}. \tag{1.1}$$

We have $N_r < \infty$ a.s. because $S_n \to \infty$ a.s. This latter fact follows e.g. from Lemma 2.1 in [37] or from standard results on bounded martingale differences. We will consider the following two sequential estimators, both based on the stopping rule N_r:

$$\bar{\theta}_{(r)} = \frac{S_{N_r}}{N_r}, \tag{1.2}$$

$$\tilde{\theta}_{(r)} = \frac{r}{N_r}. \tag{1.3}$$

The behavior of these estimators is very similar and analogous results could be proved for both of them. However, in Section 3 we focus on $\tilde{\theta}_{(r)}$ while in Section 4 we focus on $\bar{\theta}_{(r)}$, just because it makes formulas simpler.

2. Sequential probability inequalities

The following theorem is an analogue of the Chebyshev inequality for sequential estimators.

Theorem 2.1. *Assume that X_n are nonnegative i.i.d. random variables with $\mathbb{E}X_n = \theta > 0$ and $\operatorname{Var} X_n = \sigma^2$. If N_r and $\bar{\theta}_{(r)}$ are given by (1.1) and (1.2), respectively, then*

$$\mathbb{P}\left(|\bar{\theta}_{(r)} - \theta| \geq \varepsilon\theta\right) \leq \left(\frac{1+\varepsilon}{\varepsilon}\right)^2 \frac{\sigma^2}{\theta r}\left[1 + \frac{\sigma^2}{\theta r} + \frac{\theta}{r}\right]. \tag{2.1}$$

Proof. Observe that

$$\begin{aligned}
&\mathbb{P}\left(\bar{\theta}_{(r)}/\theta - 1 \leq -\varepsilon\right) + \mathbb{P}\left(\bar{\theta}_{(r)}/\theta - 1 \geq \varepsilon\right)\\
&= \mathbb{P}\left(\frac{1}{1-\varepsilon} \leq \frac{N_r\theta}{S_{N_r}}\right) + \mathbb{P}\left(\frac{1}{1+\varepsilon} \geq \frac{N_r\theta}{S_{N_r}}\right)\\
&= \mathbb{P}\left(\frac{\varepsilon}{1-\varepsilon} \leq \frac{N_r\theta}{S_{N_r}} - 1\right) + \mathbb{P}\left(\frac{-\varepsilon}{1+\varepsilon} \geq \frac{N_r\theta}{S_{N_r}} - 1\right)\\
&\leq \mathbb{P}\left(\frac{\varepsilon}{1+\varepsilon} \leq \frac{N_r\theta}{S_{N_r}} - 1\right) + \mathbb{P}\left(\frac{-\varepsilon}{1+\varepsilon} \geq \frac{N_r\theta}{S_{N_r}} - 1\right)\\
&= \mathbb{P}\left(\left|\frac{N_r\theta}{S_{N_r}} - 1\right| \geq \frac{\varepsilon}{1+\varepsilon}\right) = \mathbb{P}\left(|S_{N_r} - N_r\theta| \geq \frac{\varepsilon S_{N_r}}{1+\varepsilon}\right)\\
&\leq \mathbb{P}\left(|S_{N_r} - N_r\theta| \geq \frac{\varepsilon r}{1+\varepsilon}\right)\\
&\leq \left(\frac{1+\varepsilon}{\varepsilon}\right)^2 \frac{1}{r^2}\mathbb{E}\left(S_{N_r} - N_r\theta\right)^2,
\end{aligned}$$

where the last inequality follows from Chebyshev. Now we can apply the second Wald identity (A.9) and then the first Wald identity (A.8), obtaining

$$\mathbb{E}\left(S_{N_r} - N_r\theta\right)^2 = \sigma^2\mathbb{E}N_r = \frac{\sigma^2}{\theta}\mathbb{E}S_{N_r}.$$

In view of (1.1), random variable $S_{N_r} - r$ is the "overshoot". An elegant result of Lorden [33], recalled in Appendix A.3, allows us to bound its expectation. We have $\mathbb{E}S_{N_r} - r \leq \theta + \sigma^2/\theta$. Consequently,

$$\mathbb{E}\left(S_{N_r} - N_r\theta\right)^2 = \sigma^2\left[\frac{r}{\theta} + 1 + \frac{\sigma^2}{\theta^2}\right]$$

and (2.1) follows. □

Remark 2.2. If we additionally assume that $X_n \leq 1$ then the conclusion of Theorem 2.1 can be strengthened to

$$\mathbb{P}\left(|\bar{\theta}_{(r)} - \theta| \geq \varepsilon\theta\right) \leq \left(\frac{1+\varepsilon}{\varepsilon}\right)^2 \frac{r+1}{r^2}.$$

Indeed, it is enough to use inequality (A.10) in Lemma A.2. However, the main advantage of (2.1) is that it holds for unbounded variables.

Let us now turn to exponential inequalities. The main result of this section is the following sequential analogue of Lemma A.1.

Theorem 2.3. *Let N_r and $\tilde{\theta}_{(r)}$ be given by (1.1) and (1.3), respectively. If Assumption 1.1 holds then for every $\varepsilon > 0$,*

$$\mathbb{P}(\tilde{\theta}_{(r)} - \theta \geq \varepsilon\theta) \leq \exp\left\{-r\left[\ln(1+\varepsilon) - \frac{\varepsilon}{1+\varepsilon}\right]\right\} \tag{2.2}$$

$$\leq \exp\left\{-\frac{r}{2}\left(\frac{\varepsilon}{1+\varepsilon}\right)^2\right\}. \tag{2.3}$$

If moreover $r > \max(2/\varepsilon, 2)$ then

$$\mathbb{P}(\theta - \tilde{\theta}_{(r)} > \varepsilon\theta) \leq \exp\left\{-r\left[\ln(1-\varepsilon) + \frac{\varepsilon}{1-\varepsilon}\right]\right\} \tag{2.4}$$

$$\leq \exp\left\{-\frac{r}{2}\cdot\frac{\varepsilon^2}{1-\varepsilon}\right\}. \tag{2.5}$$

Proof. To prove (2.2), observe that

$$\begin{aligned}\mathbb{P}(\tilde{\theta}_{(r)} - \theta \geq \varepsilon\theta) &= \mathbb{P}\left(\frac{r}{N_r} \geq (1+\varepsilon)\theta\right)\\ &= \mathbb{P}\left(N_r \leq \frac{r}{(1+\varepsilon)\theta}\right)\\ &= \mathbb{P}\left(N_r \leq \left\lfloor\frac{r}{(1+\varepsilon)\theta}\right\rfloor\right)\\ &= \mathbb{P}(S_n \geq r) \quad \text{for } n = \left\lfloor\frac{r}{(1+\varepsilon)\theta}\right\rfloor.\end{aligned} \tag{2.6}$$

If $n < r$ then the RHS of (2.6) is 0. Let us consider separately the cases when $n > r$ and $n = r$. If $n > r$ then by Hoeffding inequality (A.1), with $a = r/n - \theta$ and $\bar{X}_n = S_n/n$, we get

$$\begin{aligned}\mathbb{P}(S_n \geq r) &= \mathbb{P}\left(\bar{X}_n - \theta \geq \frac{r}{n} - \theta\right)\\ &\leq \exp\left\{-n\left[\frac{r}{n}\ln\frac{r}{n\theta} + \left(1-\frac{r}{n}\right)\ln\frac{1-r/n}{1-\theta}\right]\right\}\\ &= \exp\left\{-r\left[\ln\frac{r}{n\theta} + \left(\frac{n}{r}-1\right)\ln\frac{1-r/n}{1-\theta}\right]\right\}\\ &= \exp\left\{-rf\left(\frac{n}{r},\theta\right)\right\},\end{aligned} \tag{2.7}$$

where

$$f(y,\theta) := -\ln(\theta y) + (y-1)\ln\frac{1-1/y}{1-\theta}. \tag{2.8}$$

Put

$$n_1 = \frac{r}{(1+\varepsilon)\theta}$$

and observe that

$$1 < \frac{n}{r} \le \frac{n_1}{r} < \frac{1}{\theta}.$$

Function $f(y,\theta)$ is decreasing in y for $1 < y \le 1/\theta$, because

$$\frac{\partial f(y,\theta)}{\partial y} = \ln\left(1-\frac{1}{y}\right) - \ln(1-\theta) < 0.$$

Consequently

$$\begin{aligned} f\left(\frac{n}{r},\theta\right) &\ge f\left(\frac{n_1}{r},\theta\right) \\ &= \ln(1+\varepsilon) - \frac{1-(1+\varepsilon)\theta}{(1+\varepsilon)\theta}\ln\frac{1-\theta}{1-(1+\varepsilon)\theta} \\ &=: f_1(\theta). \end{aligned} \tag{2.9}$$

Function $f_1(\theta)$ defined by (2.9) is increasing, because

$$f_1'(\theta) = \frac{1}{(1+\varepsilon)\theta^2}\left[-\ln\left(1-\frac{\varepsilon\theta}{1-\theta}\right) - \frac{\varepsilon\theta}{1-\theta}\right] > 0.$$

Thus, by the de L'Hôpital rule,

$$f_1(\theta) \ge \lim_{\theta\to 0} f_1(\theta) = \ln(1+\varepsilon) - \frac{\varepsilon}{1+\varepsilon} =: k(\varepsilon). \tag{2.10}$$

Combining (2.6), (2.7), (2.9) and (2.10) we get

$$\mathbb{P}(\tilde{\theta}_{(r)} - \theta \ge \varepsilon\theta) \le \exp\{-rk(\varepsilon)\},$$

and hence we obtain inequality (2.2).

To complete the proof of (2.2), it remains to consider the case $n = r$, which is much easier. Since we can assume that $(1+\varepsilon)\theta \le 1$, it follows that $(1+\varepsilon)\theta = 1$. In view of (A.7), we have

$$\mathbb{P}(S_n \ge r) \le \theta^r = \exp\{-r\ln(1+\varepsilon)\} \le \exp\{-rk(\varepsilon)\},$$

so (2.2) holds for this case, too.

Notice that the bound (2.2) could also be obtained using Bennett inequality (A.2) instead of Hoeffding inequality (A.1).

Using Bernstein inequality (A.3) in a similar way, we obtain (2.3). We omit the details.

The proof of (2.4) is similar but technically more complicated. Observe that

$$\begin{aligned} \mathbb{P}(\theta - \tilde{\theta}_{(r)} > \varepsilon\theta) &= \mathbb{P}\left(\frac{r}{N_r} < (1-\varepsilon)\theta\right) \\ &= \mathbb{P}\left(N_r > \frac{r}{(1-\varepsilon)\theta}\right) \\ &= \mathbb{P}\left(N_r \ge \left\lfloor\frac{r}{(1-\varepsilon)\theta}\right\rfloor + 1\right) \\ &= \mathbb{P}(S_n < r) \quad \text{for } n = \left\lfloor\frac{r}{(1-\varepsilon)\theta}\right\rfloor. \end{aligned} \tag{2.11}$$

To bound the RHS of (2.11) we use Hoeffding inequality (A.4), with $a = \theta - r/n$. We obtain

$$\mathbb{P}(S_n < r) \le \exp\left\{-rf\left(\frac{n}{r}, \theta\right)\right\}, \tag{2.12}$$

where $f(y, \theta)$ is defined by (2.8), exactly as in the first part of the proof. Function $f(y, \theta)$ is increasing in y for $y > 1/\theta$. Put

$$n_2 = \frac{r}{(1-\varepsilon)\theta} - 1$$

and observe that

$$\frac{1}{\theta} < \frac{n_2}{r} \le \frac{n}{r},$$

where the first inequality is implied by $r\varepsilon > 1$ and the second one is obvious. Therefore we have

$$\begin{aligned} f\left(\frac{n}{r}, \theta\right) &\ge f\left(\frac{n_2}{r}, \theta\right) \\ &=: f_2(\theta). \end{aligned} \tag{2.13}$$

If we denote $r/(1-\varepsilon)$ by R then we can express the function defined by (2.13) as follows:

$$f_2(\theta) = \ln(1-\varepsilon) + \ln\frac{R}{R-\theta} + \frac{R-(r+1)\theta}{r\theta}\ln\frac{R-(r+1)\theta}{(1-\theta)(R-\theta)}.$$

It is easy to see, using the de L'Hôpital rule, that

$$\lim_{\theta\to 0} f_2(\theta) = \ln(1-\varepsilon) + \frac{\varepsilon}{1-\varepsilon} =: k(-\varepsilon), \tag{2.14}$$

where the notation $k(-\varepsilon)$ is consistent with (2.10). Taking into account (2.11)–(2.14), we can see that to complete the proof of (2.4) it is sufficient to show that

$$f_2(\theta) \ge \lim_{\theta\to 0} f_2(\theta).$$

We will verify that $f_2(\theta)$ is increasing for $0 < \theta < 1$. It will be more convenient to examine the function

$$f_3(\theta) := rf_2(\theta) - r\ln r = \left(\frac{R}{\theta} - r - 1\right)\ln\frac{R-(r+1)\theta}{(1-\theta)(R-\theta)} - r\ln(R-\theta)$$

and show that $f_3(\theta)$ is increasing, which is clearly equivalent. Observe that

$$f_3'(\theta) = -\frac{R}{\theta^2}\ln\frac{R-(r+1)\theta}{(1-\theta)(R-\theta)} + \frac{R-r-\theta}{\theta(1-\theta)} \ge 0$$

if and only if

$$f_4(\theta) := \frac{\theta^2}{R} f_3'(\theta) \ge 0.$$

The last inequality will follow from the fact that $f_4'(\theta) \ge 0$, because $f_4(0) = 0$. We have

$$f_4'(\theta) = \frac{\theta f_5(\theta)}{R(1-\theta)^2(R-\theta)(R-(r+1)\theta)},$$

where

$$f_5(\theta) := (r+1)\theta^3 - [R+2(r+1)]\theta^2 - [(R-r)^2 - 4R + r]\theta + (R-r-2)R(R-r).$$

Since $r\varepsilon > 2$, we have $R - r - 2 > 0$. It follows that $f_5(0) > 0$ and

$$f_5(1) = (R-1)[(R-r)(R-r-2)+1] > 0.$$

Moreover, $f_5(\theta)$ is concave for $0 < \theta < 1$, because

$$\begin{aligned} f_5''(\theta) &= 6(r+1)\theta - 2[R + 2(r+1)] \\ &< 6(r+1) - 2[R + 2(r+1)] < 0, \end{aligned}$$

the last inequality being implied by $r + 1 < R$.

Therefore $f_5(\theta) \geq 0$ for $0 < \theta < 1$ and consequently $f_4'(\theta) \geq 0$, too. This implies that $f_4(\theta) \geq 0$, so $f_3(\theta)$ increases. Equivalently, $f_2(\theta)$ increases and the proof of (2.4) is complete.

Using Bennet inequality (A.5) or Bernstein inequality (A.6) in a similar way, we obtain (2.5). The proof is quite analogous but easier, so we omit the details. □

Analogous inequalities for $\bar{\theta}_{(r)}$ can be easily deduced from those for $\tilde{\theta}_{(r)}$.

Corollary 2.4. *Inequalities (2.4) and (2.5) are still valid if we replace $\tilde{\theta}_{(r)}$ by $\bar{\theta}_{(r)}$. If $\varepsilon > 1/r$ then*

$$\mathbb{P}(\bar{\theta}_{(r)} - \theta \geq \varepsilon\theta) \leq \exp\left\{-r \ln \frac{r(1+\varepsilon)}{r+1} + \frac{r\varepsilon - 1}{1+\varepsilon}\right\} \tag{2.15}$$

$$\leq \exp\left\{-\frac{3}{2}\frac{(r\varepsilon - 1)^2}{(1+\varepsilon)[r(3+\varepsilon)+2]}\right\}. \tag{2.16}$$

Proof. The first claim follows from the fact that $\bar{\theta}_{(r)} \geq \tilde{\theta}_{(r)}$. To obtain (2.15) and (2.16), note that $\bar{\theta}_{(r)} \leq (r+1)/N_r = \tilde{\theta}_{(r)}(1 + 1/r)$, so $\mathbb{P}(\bar{\theta}_{(r)} - \theta \geq \varepsilon\theta) \leq \mathbb{P}(\tilde{\theta}_{(r)} - \theta \geq [(r\varepsilon - 1)/(r+1)]\theta)$. Consequently, we can apply inequalities (2.2) and (2.3) with $\varepsilon_1 := (r\varepsilon - 1)/(r+1)$. Condition $\varepsilon > 1/r$ is needed for $\varepsilon_1 > 0$. □

3. (ε, α)-approximation

3.1. The upper bound

Theorem 2.3 allows us to construct an (ε, α)-approximation and to determine the expected number of samples. Let us rewrite inequalities (2.2) and (2.4) as follows:

$$\mathbb{P}(\tilde{\theta}_{(r)} \geq (1+\varepsilon)\theta) \leq \exp\{-rk(\varepsilon)\}, \quad (\varepsilon > 0),$$
$$\mathbb{P}(\tilde{\theta}_{(r)} < (1-\varepsilon)\theta) \leq \exp\{-rk(-\varepsilon)\}, \quad (0 < \varepsilon < 1),$$

where

$$k(\varepsilon) = \ln(1+\varepsilon) - \frac{\varepsilon}{1+\varepsilon}, \quad (\varepsilon > -1). \tag{3.1}$$

It is easy to see that $k(\varepsilon) < k(-\varepsilon)$ for $0 < \varepsilon < 1$ and $k(\varepsilon) \sim \varepsilon^2/2$ for $\varepsilon \to 0$. The graph of $k(\varepsilon)$ is displayed in Fig. 1.

Assume $0 < \varepsilon < 1$ and $0 < \alpha < 1$. Choose $r = r(\varepsilon, \alpha)$ such that

$$\exp\{-rk(\varepsilon)\} + \exp\{-rk(-\varepsilon)\} = \alpha. \tag{3.2}$$

It is easy to see that r defined by (3.2) satisfies the inequality

$$\frac{\ln \alpha^{-1}}{k(\varepsilon)} < r < \frac{\ln(\alpha/2)^{-1}}{k(\varepsilon)}. \tag{3.3}$$

Indeed, if $r_1 = (\ln \alpha^{-1})/k(\varepsilon)$ then $\exp[-r_1 k(\varepsilon)] + \exp[-r_1 k(-\varepsilon)] > \alpha$, so $r_1 < r$. If $r_2 = (\ln(\alpha/2)^{-1})/k(\varepsilon)$ then $\exp[-r_2 k(\varepsilon)] + \exp[-r_2 k(-\varepsilon)] < \alpha/2 + \alpha/2 = \alpha$, so $r < r_2$.

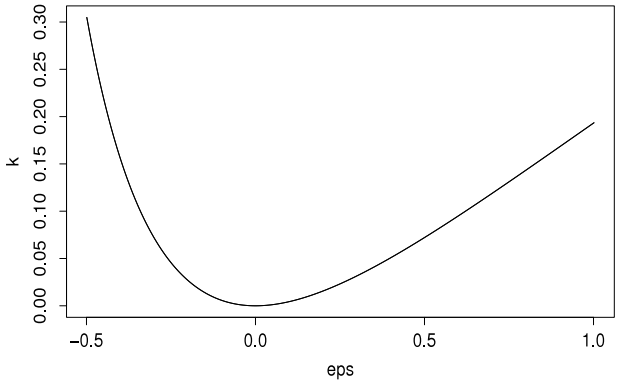

Fig. 1. Graph of $k(\varepsilon) = \ln(1+\varepsilon) - \varepsilon/(1+\varepsilon)$.

Theorem 3.1. *Let N_r and $\tilde{\theta}_{(r)}$ be given by (1.1) and (1.3), respectively, with r defined by (3.2). Then $(N_r, \tilde{\theta}_{(r)})$ is an (ε, α)-approximation on the class of input sequences $X_1, X_2, \ldots$ satisfying Assumption 1.1. Moreover,*

$$\frac{\ln \alpha^{-1}}{k(\varepsilon)} < \theta \mathbb{E} N_r < \frac{\ln(\alpha/2)^{-1}}{k(\varepsilon)} + 1 =: \mathrm{Up}(\varepsilon, \alpha).$$

Proof. The result immediately follows from Theorem 2.3 and (3.3) via a version of the first Wald lemma; see (A.8) in the Appendix. □

Note that $\mathbb{E}N_r$ depends on the probability distribution of $X_1, X_2, \ldots$, although this is not explicit in our notation. Of course, the important part of Theorem 3.1 is the upper bound. It is immediately seen that $\mathrm{Up}(\varepsilon, \alpha) \sim 2 \ln \alpha^{-1}/\varepsilon^2$ as $\varepsilon \to 0$ and $\alpha \to 0$.

3.2. The lower bound

In this section we consider an arbitrary algorithm which is an (ε, α)-approximation on a class of inputs which contains (at least) every Bernoulli scheme, i.e. the sequence of i.i.d. random variables with

$$\mathbb{P}(X_i = 1) = \theta, \qquad \mathbb{P}(X_i = 0) = 1 - \theta.$$

Theorem 3.2. *If $(N, \hat{\theta})$ is an (ε, α)-approximation with $\alpha \le 1/2$ and $X_1, X_2, \ldots$ is the Bernoulli scheme with the probability of success θ, then*

$$\liminf_{\theta \to 0} \theta \mathbb{E} N \ge \frac{\ln(2\alpha)^{-1}}{k(-\varepsilon)} \left(1 + \frac{1 - \sqrt{2 \ln(2\alpha)^{-1} + 1}}{\ln(2\alpha)^{-1}} \right) =: \mathrm{Low}(\varepsilon, \alpha).$$

Here $\mathbb{E}N$ depends on θ through the probability distribution of $X_1, X_2, \ldots$, just as in Theorem 3.1. Note also that $\mathrm{Low}(\varepsilon, \alpha) \sim 2 \ln \alpha^{-1}/\varepsilon^2$ as $\varepsilon \to 0$ and $\alpha \to 0$.

Before proceeding to the proof of Theorem 3.2, let us examine its consequences. Begin with the following simple observation.

Corollary 3.3. *There exists no (ε, α)-approximation with nonadaptive stop that is based on a number of samples $N \equiv n$ fixed in advance, which works on the class of all Bernoulli schemes.*

This is an immediate consequence of Theorem 3.2, because a constant $N \equiv n$ is a special case of stopping rule and certainly $\liminf_{\theta \to 0} \theta n = 0$. Of course Corollary 3.3 is very easy and can be proved in

many different ways. However, we think that the argument given above is enlightening and reveals the essence of the phenomenon. The necessary number of samples N must go to infinity if θ approaches 0, so it cannot be fixed without a prior information on θ.

The following corollary summarizes the main results of this section. It follows directly from Theorems 3.1 and 3.2.

Corollary 3.4. *Assume $(N_r, \tilde{\theta}_{(r)})$ is defined by (1.1), (1.3) and (3.2), whilst $(N, \hat{\theta})$ is an arbitrary other (ε, α)-approximation. If both the algorithms are applied to the Bernoulli scheme with the probability of success θ, then*

$$\liminf_{\theta\to 0} \frac{\mathbb{E}N}{\mathbb{E}N_r} \geq \frac{\mathrm{Low}(\varepsilon,\alpha)}{\mathrm{Up}(\varepsilon,\alpha)},$$

where $\mathrm{Low}(\varepsilon,\alpha)/\mathrm{Up}(\varepsilon,\alpha) \to 1$ *as* $\varepsilon \to 0$ *and* $\alpha \to 0$.

The ratio $\mathbb{E}N/\mathbb{E}N_r$ is interpreted as the relative efficiency of the two algorithms. In this sense we can say that $(N_r, \tilde{\theta}_{(r)})$ is worst case optimal, the worst case being the family of Bernoulli sequences with $\theta \to 0$.

The rest of this subsection is devoted to the proof of Theorem 3.2. We assume that $X_1, \ldots, X_n, \ldots$ is a Bernoulli sequence, so θ is the probability of success and $S_n = \sum_{i=1}^n X_i$ is the number of successes in n trials. First we need some new notation. Since our argument is based on the change of measure, it will be helpful to explicitly indicate the dependence on θ. Let us therefore write $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ instead of $\mathbb{P}$ and $\mathbb{E}$. We begin with the following lemma.

Lemma 3.5. *Assume N is a stopping rule and $\hat{\theta}_N = \hat{\theta}_N(X_1, \ldots, X_N)$ is an estimator based on N samples. Fix $\theta < 1/2$ and $0 < \varepsilon < 1 - \theta$. If we have*

$$\mathbb{P}_{\theta_1}\left[\hat{\theta}_N \geq \theta\right] = \alpha_1 \quad \text{for } \theta_1 = \frac{\theta}{1+\varepsilon},$$
$$\mathbb{P}_{\theta_2}\left[\hat{\theta}_N < \theta\right] = \alpha_2 \quad \text{for } \theta_2 = \frac{\theta}{1-\varepsilon},$$

then the expectation of N fulfills the following inequality:

$$-\ln(\alpha_1+\alpha_2) \leq \sqrt{u^2\theta\mathbb{E}_\theta N} + k\theta\mathbb{E}_\theta N, \tag{3.4}$$

where

$$u = -\ln\left(1 - \frac{\varepsilon}{1-\theta}\right),$$
$$k = \ln(1-\varepsilon) - \frac{1-\theta}{\theta}\ln\left(1 - \frac{\theta\varepsilon}{(1-\theta)(1-\varepsilon)}\right).$$

Proof. Let $p(x,\theta) = \mathbb{P}_\theta(X_1 = x)$ be the likelihood of a single observation, i.e.

$$p(x,\theta) = \theta^x(1-\theta)^{1-x}.$$

Now, for $i = 1, 2$, consider the likelihood ratios $p(x,\theta_i)/p(x,\theta)$. The ratios can be expressed as follows:

$$\frac{p(x,\theta_i)}{p(x,\theta)} = \left(\frac{\theta_i}{\theta}\right)^x \left(\frac{1-\theta_i}{1-\theta}\right)^{1-x} = \exp[-u_i(x-\theta) - k_i\theta], \tag{3.5}$$

where

$$u_1 = \ln\left(1 + \frac{\varepsilon}{1-\theta}\right),$$
$$k_1 = \ln(1+\varepsilon) - \frac{1-\theta}{\theta}\ln\left(1 + \frac{\theta\varepsilon}{(1-\theta)(1+\varepsilon)}\right)$$

and

$$u_2 = \ln\left(1 - \frac{\varepsilon}{1-\theta}\right),$$
$$k_2 = \ln(1-\varepsilon) - \frac{1-\theta}{\theta}\ln\left(1 - \frac{\theta\varepsilon}{(1-\theta)(1-\varepsilon)}\right).$$

Indeed, for $i = 1$, in view of the relation $\theta_1 = \theta/(1+\varepsilon)$, formula (3.5) is equivalent to the following system of equations:

$$u_1(1-\theta) + k_1\theta = \ln(1+\varepsilon)$$
$$u_1\theta - k_1\theta = \ln\left(1 + \frac{\varepsilon}{1-\theta}\right) - \ln(1+\varepsilon).$$

Solving these two equations with respect to u_1 and k_1 yields the first pair of expressions. The derivation of the formulas for u_2 and k_2 is analogous.

Let S be the number of successes that occur up to the random time N:

$$S := S_N = \sum_{i=1}^{N} X_i.$$

For every event $A \in \sigma(X_1, \ldots, X_N)$, we have the following "change of measure" formula:

$$\begin{aligned}\mathbb{P}_{\theta_i}(A) &= \mathbb{E}_\theta \frac{p(X_1,\theta_i)}{p(X_1,\theta)} \cdots \frac{p(X_N,\theta_i)}{p(X_N,\theta)} \mathbb{I}(A)\\ &= \mathbb{E}_\theta \exp\left[-u_i(S-\theta N) - k_i\theta N\right] \mathbb{I}(A).\end{aligned}$$

We are going to apply this formula twice: first for $i = 1$ and the event $A = \{\hat\theta_N \geq \theta\}$, then for $i = 2$ and $A = \{\hat\theta_N < \theta\}$. Elementary computations show that $u_1 < -u_2$ and $k_1 < k_2$. Therefore, if we put $u = \max(u_1, |u_2|) = -u_2$ and $k = \max(k_1, k_2) = k_2$, we can write

$$\begin{aligned}\alpha_1 = \mathbb{P}_{\theta_1}(\hat\theta_N \geq \theta) &= \mathbb{E}_\theta \exp\left[-u_1(S-\theta N) - k_1\theta N\right] \mathbb{I}(\hat\theta_N \geq \theta)\\ &\geq \mathbb{E}_\theta \exp\left[-u|S-\theta N| - k\theta N\right] \mathbb{I}(\hat\theta_N \geq \theta) \qquad (3.6)\end{aligned}$$

and

$$\begin{aligned}\alpha_2 = \mathbb{P}_{\theta_2}(\hat\theta_N < \theta) &= \mathbb{E}_\theta \exp\left[-u_2(S-\theta N) - k_2\theta N\right] \mathbb{I}(\hat\theta_N < \theta)\\ &\geq \mathbb{E}_\theta \exp\left[-u|S-\theta N| - k\theta N\right] \mathbb{I}(\hat\theta_N < \theta). \qquad (3.7)\end{aligned}$$

Now, let us add the two sides of (3.6) and (3.7) and apply the Jensen inequality:

$$\begin{aligned}\alpha_1 + \alpha_2 &\geq \mathbb{E}_\theta \exp\left[-u|S-\theta N| - k\theta N\right]\\ &\geq \exp \mathbb{E}_\theta \left[-u|S-\theta N| - k\theta N\right].\end{aligned}$$

Hence we obtain

$$\begin{aligned}-\ln(\alpha_1+\alpha_2) &\leq u\mathbb{E}_\theta|S-\theta N| + k\theta\mathbb{E}_\theta N\\ &\leq u\sqrt{\mathbb{E}_\theta(S-\theta N)^2} + k\theta\mathbb{E}_\theta N.\end{aligned}$$

To complete the proof, notice that the second Wald identity gives

$$\mathbb{E}_\theta(S-\theta N)^2 = \theta(1-\theta)\mathbb{E}_\theta N \leq \theta\mathbb{E}_\theta N. \quad \square$$

Remark 3.6. It is perhaps worth mentioning that $k_i\theta$ in the preceding proof is nothing but the Kullback–Leibler information:

$$-\mathbb{E}_\theta \ln\frac{p(X_1,\theta_i)}{p(X_1,\theta)} = \mathbb{E}_\theta[u_i(X_1-\theta) + k_i\theta] = k_i\theta.$$

Proof of Theorem 3.2. Consider a sequential estimator $\hat{\theta}_N = \hat{\theta}_N(X_1, \ldots, X_N)$ which satisfies, for all θ,

$$\mathbb{P}_\theta\left[(1-\varepsilon)\theta \le \hat{\theta}_N < (1+\varepsilon)\theta\right] \ge 1-\alpha. \tag{3.8}$$

(The strict inequality in (3.8) simplifies the argument below and can be assumed without loss of generality.) Clearly, (3.8) implies $\mathbb{P}_\theta\big[\hat{\theta}_N \ge \theta(1+\varepsilon)\big] \le \alpha$ and $\mathbb{P}_\theta\big[\hat{\theta}_N < \theta(1-\varepsilon)\big] \le \alpha$. These inequalities hold for every θ, so we can apply them to θ_1 and θ_2 defined as in Lemma 3.5, i.e. $\theta_1 = \theta/(1+\varepsilon)$ and $\theta_2 = \theta/(1-\varepsilon)$. It follows that

$$\mathbb{P}_{\theta_1}\big[\hat{\theta}_N \ge \theta_1(1+\varepsilon)\big] = \mathbb{P}_{\theta_1}\big[\hat{\theta}_N \ge \theta\big] \le \alpha,$$
$$\mathbb{P}_{\theta_2}\big[\hat{\theta}_N < \theta_2(1-\varepsilon)\big] = \mathbb{P}_{\theta_2}\big[\hat{\theta}_N < \theta\big] \le \alpha.$$

Therefore the assumptions of Lemma 3.5 are fulfilled for every $\theta < 1/2$ with $\alpha_1 + \alpha_2 \le 2\alpha$. Let us apply this lemma and pass to the limit with $\theta \to 0$. Set

$$\nu = \liminf_{\theta\to 0} \theta \mathbb{E}_\theta N$$

and examine the limiting behavior of u and k which appear in (3.4). If $\theta \to 0$, then

$$\begin{aligned} u &= -\ln\left(1 - \frac{\varepsilon}{1-\theta}\right) \to -\ln(1-\varepsilon), \\ k &= \ln(1-\varepsilon) - \frac{\varepsilon}{1-\varepsilon}\cdot\frac{(1-\theta)(1-\varepsilon)}{\theta\varepsilon}\ln\left(1 - \frac{\theta\varepsilon}{(1-\theta)(1-\varepsilon)}\right) \\ &\to \ln(1-\varepsilon) + \frac{\varepsilon}{1-\varepsilon} \\ &= k(-\varepsilon), \end{aligned}$$

where the function k is given by (3.1).

Therefore, Lemma 3.5 implies the inequality

$$-\ln(2\alpha) \le \sqrt{[\ln(1-\varepsilon)]^2\nu} + k(-\varepsilon)\nu.$$

It is easily seen that $[\ln(1-\varepsilon)]^2 \le 2k(-\varepsilon)$. Consequently,

$$-\ln(2\alpha) \le \sqrt{2k(-\varepsilon)\nu} + k(-\varepsilon)\nu.$$

This is in fact a quadratic inequality with respect to $y = \sqrt{2k(-\varepsilon)\nu}$. If we solve this inequality, we get $y \ge \sqrt{1+2l} - 1$, where $l = -\ln(2\alpha)$. Since $k(-\varepsilon)\nu = y^2/2 \ge l + 1 - \sqrt{2l+1}$, we obtain

$$k(-\varepsilon)\nu \ge l\left(1 + \frac{1-\sqrt{2l+1}}{l}\right),$$

which is equivalent to the conclusion of the theorem. □

4. The relative mean square error

4.1. The upper bound

We examine the relative MSE of the sequential estimator $\bar{\theta}_{(r)}$ under Assumption 1.1. The notation and assumptions will be the same as in Sections 2 and 3.1.

Theorem 4.1. *Let N_r and $\bar{\theta}_{(r)}$ be given by (1.1) and (1.2), respectively. If the input sequence $X_1, X_2, \ldots$ satisfies Assumption 1.1 and $r > 3$ then*

$$\mathbb{E}\left(\frac{\bar{\theta}_{(r)} - \theta}{\theta}\right)^2 \le \frac{1}{r} + \frac{w(r)}{r^{3/2}}, \tag{4.1}$$

where $\lim_{r\to\infty} w(r) = 6\sqrt{2\pi}$.

The proof will be preceded by the following elementary lemma.

Lemma 4.2. *We have*

$$\left(\frac{S_{N_r}}{\theta N_r}-1\right)^2 \le \begin{cases} \left(\dfrac{\theta N_r}{S_{N_r}}-1\right)^2 & \text{for } \dfrac{\theta N_r}{S_{N_r}} \ge 1; \\ \left(\dfrac{\theta N_r}{S_{N_r}}-1\right)^2 + 2\left(\dfrac{S_{N_r}}{\theta N_r}-1\right)^3 & \text{for } \dfrac{\theta N_r}{S_{N_r}} < 1. \end{cases}$$

Proof. Put $z = S_{N_r}/(\theta N_r) - 1$ and notice that $1 - \theta N_r/S_{N_r} = z/(z+1)$. If $z \le 0$ then

$$z^2 \le \left(\frac{z}{z+1}\right)^2.$$

If $z > 0$ then

$$z^2 - \left(\frac{z}{z+1}\right)^2 = z^3\frac{z+2}{(z+1)^2} \le 2z^3. \quad \square$$

Proof of Theorem 4.1. From Lemma 4.2 it follows that

$$\begin{aligned} \mathbb{E}\left(\frac{\bar\theta_{(r)}-\theta}{\theta}\right)^2 &= \mathbb{E}\left(\frac{S_{N_r}}{\theta N_r}-1\right)^2 \\ &\le \mathbb{E}\left(\frac{\theta N_r}{S_{N_r}}-1\right)^2 \qquad (4.2) \\ &\quad + 2\mathbb{E}\left(\frac{S_{N_r}}{\theta N_r}-1\right)^3 \mathbb{I}\left\{\frac{\theta N_r}{S_{N_r}} < 1\right\}. \qquad (4.3) \end{aligned}$$

We bound the first term, i.e. (4.2), using the generalized Wald identities. Since $r \le S_{N_r}$, by inequality (A.10) in Lemma A.2 we obtain

$$\mathbb{E}\left(\frac{\theta N_r}{S_{N_r}}-1\right)^2 \le \frac{\mathbb{E}(\theta N_r - S_{N_r})^2}{r^2} \le \frac{r+1}{r^2} = \frac{1}{r} + \frac{1}{r^2}. \qquad (4.4)$$

To bound (4.3), we use the exponential inequality (2.15) and integration by parts:

$$\begin{aligned} &2\mathbb{E}\left(\frac{S_{N_r}}{\theta N_r}-1\right)^3 \mathbb{I}\left\{\frac{\theta N_r}{S_{N_r}} < 1\right\} \\ &= 6\int_0^\infty \varepsilon^2 \mathbb{P}\left(\frac{S_{N_r}}{\theta N_r} - 1 \ge \varepsilon\right) d\varepsilon \\ &= 6\int_0^\infty \varepsilon^2 \mathbb{P}\left(\bar\theta_{(r)} - \theta \ge \varepsilon\theta\right) d\varepsilon \\ &\le 6\int_{1/r}^\infty \varepsilon^2 \exp\left\{-r\ln\frac{r(1+\varepsilon)}{r+1} + \frac{r\varepsilon - 1}{1+\varepsilon}\right\} d\varepsilon + 6\int_0^{1/r} \varepsilon^2 d\varepsilon \\ &\le 6\int_0^\infty \varepsilon^2 \left(\frac{r+1}{1+\varepsilon}\right)^r r^{-r} \exp\left\{-\frac{r+1}{1+\varepsilon} + r\right\} d\varepsilon + \frac{2}{r^3} \\ &= 6r^{-r}e^r \int_0^\infty \varepsilon^2 \left(\frac{r+1}{1+\varepsilon}\right)^r \exp\left\{-\frac{r+1}{1+\varepsilon}\right\} d\varepsilon + \frac{2}{r^3} \\ &=: I + \frac{2}{r^3}. \qquad (4.5) \end{aligned}$$

Putting $x = (r+1)/(1+\varepsilon)$ we can express the integral in (4.5) in the following form:

$$\begin{aligned} I &= 6r^{-r}e^{r}\int_0^{r+1}\left(\frac{r+1}{x}-1\right)^2 x^{r-2}e^{-x}(r+1)dx \\ &\le 6r^{-r}e^{r}\int_0^{\infty}\left(\frac{r+1}{x}-1\right)^2 x^{r-2}e^{-x}(r+1)dx \\ &= 6r^{-r}e^{r}\left[(r+1)^3\int_0^{\infty}x^{r-4}e^{-x}dx - 2(r+1)^2\int_0^{\infty}x^{r-3}e^{-x}dx + (r+1)\int_0^{\infty}x^{r-2}e^{-x}dx\right]. \end{aligned}$$

Therefore

$$\begin{aligned} I &\le 6r^{-r}e^{r}\left[(r+1)^3\Gamma(r-3) - 2(r+1)^2\Gamma(r-2) + (r+1)\Gamma(r-1)\right] \\ &= 6r^{-r}e^{r}\Gamma(r+1)\left[\frac{(r+1)^3}{(r-3)(r-2)(r-1)r} - \frac{2(r+1)^2}{(r-2)(r-1)r} + \frac{r+1}{(r-1)r}\right] \\ &= 6r^{-r}e^{r}\Gamma(r+1)\frac{(r+1)(r+13)}{(r-3)(r-2)(r-1)r}. \end{aligned}$$

By the Stirling formula

$$I \le 6\sqrt{2\pi r}e^{1/(12r)}\frac{(r+1)(r+13)}{(r-3)(r-2)(r-1)r}.$$

Taking into account (4.2)–(4.5), we can see that (4.1) holds with $w(r)$ defined by

$$w(r) = 6\sqrt{2\pi}e^{1/(12r)}\frac{r(r+1)(r+13)}{(r-3)(r-2)(r-1)} + \frac{1}{\sqrt{r}} + \frac{2}{r\sqrt{r}} \tag{4.6}$$

and the proof is complete. □

Remark 4.3. Theorem 4.1 remains true if we replace estimator $\bar{\theta}_{(r)}$ by $\tilde{\theta}_{(r)}$. However, the function $w(r)$ is then different and strictly greater. We omit the details and focus attention on $\bar{\theta}_{(r)}$ in the rest of this section.

Theorem 4.1 allows us to construct an ε-RRMSE approximation (Definition 1.5) with a tight bound for the expected number of required samples. Choose $r = r(\varepsilon)$ such that

$$\frac{1}{r} + \frac{w(r)}{r^{3/2}} = \varepsilon^2, \tag{4.7}$$

where $w(r)$ is the function defined by (4.6). Note that $r(\varepsilon) \sim 1/\varepsilon^2$ for $\varepsilon \to 0$. From Theorem 4.1 we immediately obtain the following result.

Corollary 4.4. *Let N_r and $\bar{\theta}_{(r)}$ be given by (1.1) and (1.2), with $r = r(\varepsilon)$ defined by (4.7). Then $(N_r, \bar{\theta}_{(r)})$ is an ε-RRMSE approximation on the class of input sequences satisfying Assumption 1.1 and*

$$r(\varepsilon) \le \theta\mathbb{E}N_r < r(\varepsilon) + 1 =: \mathrm{Up}_{L^2}(\varepsilon).$$

4.2. The lower bound

The following theorem is an L^2 analogue of Theorem 3.2. We consider an arbitrary algorithm which is an ε-RRMSE approximation in the sense of Definition 1.5 on a class of inputs which contains (at least) every Bernoulli scheme.

Theorem 4.5. *If $(N, \hat{\theta})$ is an ε-RRMSE approximation and $X_1, X_2, \ldots$ is the Bernoulli sequence with the probability of success θ, then*

$$\limsup_{\theta\to 0} \theta\mathbb{E}N \geq \frac{1-\varepsilon^2}{\varepsilon^2} =: \mathrm{Low}_{L^2}(\varepsilon).$$

Although Theorem 4.5 gives a bound only for the upper limit, in contrast with the bound for the lower limit in Theorem 3.2, it is sufficient to establish the worst case optimality of $(N_r, \bar{\theta}_{(r)})$.

Corollary 4.6. *Assume $(N_r, \bar{\theta}_{(r)})$ is defined by (1.1), (1.2) and (4.7), whilst $(N, \hat{\theta})$ is an arbitrary other ε-RRMSE approximation. If both the algorithms are applied to the Bernoulli scheme with the probability of success θ, then*

$$\limsup_{\theta\to 0} \frac{\mathbb{E}N}{\mathbb{E}N_r} \geq \frac{\mathrm{Low}_{L^2}(\varepsilon)}{\mathrm{Up}_{L^2}(\varepsilon)},$$

where $\mathrm{Low}_{L^2}(\varepsilon)/\mathrm{Up}_{L^2}(\varepsilon) \to 1$ as $\varepsilon \to 0$.

In this sense $(N_r, \bar{\theta}_{(r)})$ is worst case optimal. In the present L^2 setting we can only say that the worst case is some (unspecified) subfamily of Bernoulli schemes with parameters $\theta \to 0$.

In the following proof, just as in Section 3.2, we will explicitly write $\mathbb{E}_\theta$ instead of $\mathbb{E}$.

Proof of Theorem 4.5. Fix $\varepsilon > 0$ and assume that $\hat{\theta}_N$ is such that (0.3) holds. Using the Cramér–Rao–Wolfowitz inequality (see [44]), we get

$$\varepsilon^2 \geq \mathbb{E}_\theta\left(\frac{\hat{\theta}_N - \theta}{\theta}\right)^2 \geq \frac{b^2(\theta)}{\theta^2} + \frac{[1+b'(\theta)]^2}{\theta^2 I(\theta)\mathbb{E}_\theta N},$$

where $b(\theta) = \mathbb{E}_\theta\hat{\theta}_N - \theta$ denotes the bias of $\hat{\theta}_N$ and $I(\theta) = [\theta(1-\theta)]^{-1}$ denotes the Fisher information for a single Bernoulli observation (for a discussion of regularity conditions, see [6]). Thus

$$\begin{aligned}
\varepsilon^2 &\geq \frac{b^2(\theta)}{\theta^2} + \frac{1-\theta}{\theta}\frac{[1+b'(\theta)]^2}{\mathbb{E}_\theta N} \\
&= \left[b'(\theta)\right]^2 + \frac{1-\theta}{\theta}\frac{[1-b'(\theta)]^2}{\mathbb{E}_\theta N} + \frac{b^2(\theta)}{\theta^2} - [b'(\theta)]^2 \\
&\geq \frac{1-\theta}{1-\theta+\theta\mathbb{E}_\theta N} + \frac{b^2(\theta)}{\theta^2} - [b'(\theta)]^2.
\end{aligned}$$

Using argumentation analogous to that in the proof of formula (2.6) in [20], we can find a sequence $\theta_m \to 0$ such that

$$\lim_{m\to\infty}\left[\left(\frac{b(\theta_m)}{\theta_m}\right)^2 - \left(b'(\theta_m)\right)^2\right] \geq 0.$$

Hence for arbitrary $\delta > 0$ and sufficiently large m,

$$\varepsilon^2 + \delta \geq \frac{1-\theta_m}{1-\theta_m+\theta_m\mathbb{E}_{\theta_m}N},$$

which is equivalent to

$$\theta_m\mathbb{E}_{\theta_m}N \geq (1-\theta_m)\frac{1-\varepsilon^2-\delta}{\varepsilon^2+\delta}.$$

Consequently,

$$\limsup_{\theta\to 0} \theta\mathbb{E}_\theta N \geq \frac{1-\varepsilon^2-\delta}{\varepsilon^2+\delta}.$$

Since δ is arbitrary, the conclusion of the theorem follows. □

Table 1
Values of $r(\varepsilon,\alpha)$ and $r_{L^2}(\varepsilon)$.

ε	α				
	0.0001	0.001	0.01	0.05	r_{L^2}
0.1	2.106e+03	1.595e+03	1.093e+03	7.510e+02	
	2.125e+03	1.613e+03	1.109e+03	7.652e+02	2.134e+02
0.01	1.982e+05	1.521e+05	1.060e+05	7.379e+04	
	1.983e+05	1.522e+05	1.061e+05	7.389e+04	1.141e+04
0.001	1.981e+07	1.520e+07	1.060e+07	7.378e+06	
	1.981e+07	1.520e+07	1.060e+07	7.379e+06	1.015e+06
0.0001	1.981e+09	1.520e+09	1.060e+09	7.378e+08	
	1.981e+09	1.520e+09	1.060e+09	7.378e+08	1.002e+08

5. Numerical results

The optimality results in Section 3 are of asymptotic type. To evaluate numerically the efficiency of our estimators, we will adopt a slightly different approach. We compare the expected number of samples needed by an adaptively stopped procedure with that needed by an "oracle" nonadaptive algorithm, which is based on *a priori* knowledge which is in reality unavailable. This approach was introduced in [12,36].

Let us begin with the computation of a parameter $r = r(\varepsilon,\alpha)$ of the estimators $\tilde{\theta}_{(r)}$ and $\bar{\theta}_{(r)}$. For $\tilde{\theta}_{(r)}$, we solve Eq. (3.2). For $\bar{\theta}_{(r)}$, by Corollary 2.4, r can be obtained by solving the following equation:

$$\exp\left\{-r\ln\frac{r(1+\varepsilon)}{r+1}+\frac{r\varepsilon-1}{1+\varepsilon}\right\}+\exp\{-rk(-\varepsilon)\}=\alpha. \tag{5.1}$$

Table 1 gives $r(\varepsilon,\alpha)$ computed from (3.2) and (5.1) (these are upper and lower entries, respectively) for some pairs (ε,α) which seem to be realistic in practice. In the last column there are also included the $r_{L^2}(\varepsilon)$ computed from (4.7).

The values $r = r(\varepsilon,\alpha)$ obtained from (3.2) and (5.1) are very close. To study the efficiency of (ε,α)-approximation, we focus on the estimator $\tilde{\theta}_{(r)}$ defined by (1.1) and (1.3), with $r = r(\varepsilon,\alpha)$ given by (3.2). The mean number of observations needed for this sequential procedure is $\mathbb{E}N_r < (r+1)/\theta$. If we knew θ, we could determine $n = n(\varepsilon,\alpha,\theta)$ such that $\mathbb{P}(|\bar{X}_n-\theta|\le\varepsilon\theta)\ge 1-\alpha$, by using the non-sequential Hoeffding inequalities (A.1) and (A.4):

$$\exp\{-nh(\varepsilon,\theta)\}+\exp\{-nh(-\varepsilon,\theta)\}=\alpha,$$

where

$$h(\varepsilon,\theta)=\theta(1+\varepsilon)\ln(1+\varepsilon)+(1-\theta(1+\varepsilon))\ln\frac{1-\theta(1+\varepsilon)}{1-\theta}.$$

Then $\bar{X}_n$ would be a non-sequential estimator satisfying (0.2), if it were not for the vicious circle in its construction. Nonetheless, n obtained in this way is a good point of reference for our *bona fide* sequential estimator. For the purposes of the numerical study below, define the efficiency as

$$\mathrm{eff}(\alpha,\varepsilon,\theta)=\frac{n(\varepsilon,\alpha,\theta)}{[r(\varepsilon,\alpha)+1]/\theta}.$$

To study the efficiency of ε-RRMSE approximation, it will be more convenient to consider the estimator $\bar{\theta}_{(r)}$ defined by (1.2), with r given by (4.7). The corresponding oracle-estimator is constructed very simply. Clearly, $\mathbb{E}(\bar{X}_n-\theta)^2/\theta^2\le \mathrm{Var}\,X_1/(n\theta^2)\le(1-\theta)/(n\theta)$. Therefore, if $n=n(\varepsilon,\theta)=(1-\theta)/(\theta\varepsilon^2)$ then $\bar{X}_n$ estimates θ with the relative mean square error bounded by ε^2. It is natural to define the mean square efficiency as

$$\mathrm{eff}_{L^2}(\varepsilon,\theta)=\frac{n(\varepsilon,\theta)}{[r(\varepsilon)+1]/\theta}=\frac{(1-\theta)/\varepsilon^2}{r(\varepsilon)+1}.$$

Table 2
Efficiency of (ε, α)-approximation and ε-RRMSE approximation.

ε	θ	α				
		0.0001	0.001	0.01	0.05	eff_{L^2}
0.1	0.1	0.8486	0.8591	0.8723	0.8830	0.4198
	0.01	0.9343	0.9456	0.9600	0.9716	0.4618
	0.001	0.9428	0.9542	0.9687	0.9804	0.4659
	0.0001	0.9437	0.9551	0.9696	0.9813	0.4664
	0	0.9438	0.9552	0.9697	0.9814	0.4664
0.01	0.1	0.8994	0.8995	0.8997	0.8998	0.7886
	0.01	0.9893	0.9895	0.9897	0.9898	0.8675
	0.001	0.9983	0.9985	0.9987	0.9988	0.8754
	0.0001	0.9992	0.9994	0.9996	0.9997	0.8762
	0	0.9993	0.9995	0.9997	0.9998	0.8763
0.001	0.1	0.9000	0.9000	0.9000	0.9000	0.8868
	0.01	0.9900	0.9900	0.9900	0.9900	0.9754
	0.001	0.9990	0.9990	0.9990	0.9990	0.9843
	0.0001	0.9999	0.9999	0.9999	0.9999	0.9852
	0	1.0000	1.0000	1.0000	1.0000	0.9853
0.0001	0.1	0.9000	0.9000	0.9000	0.9000	0.8986
	0.01	0.9900	0.9900	0.9900	0.9900	0.9885
	0.001	0.9990	0.9990	0.9990	0.9990	0.9975
	0.0001	1.0000	1.0000	1.0000	1.0000	0.9984
	0	1.0000	1.0000	1.0000	1.0000	0.9985

Table 2 gives $\mathrm{eff}(\alpha, \varepsilon, \theta)$ and $\mathrm{eff}_{L^2}(\varepsilon, \theta)$ for some chosen parameters α, ε and θ. We additionally include entries corresponding to $\theta = 0$, which are defined by the obvious conventions $\mathrm{eff}(\alpha, \varepsilon, 0) := \lim_{\theta\to 0} \mathrm{eff}(\alpha, \varepsilon, \theta)$ and $\mathrm{eff}_{L^2}(\varepsilon, 0) := \lim_{\theta\to 0} \mathrm{eff}_{L^2}(\varepsilon, \theta)$.

6. Unbounded samples

In many applications it is desirable to construct an (ε, α)-approximation for unbounded input sequences X_n. Let us begin with a few remarks about a seemingly obvious case of bounded relative variance (BRV).

Assumption 6.1. X_n is $\mathcal{F}_n$-measurable, nonnegative, $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = \theta$ a.s. and $\mathrm{Var}(X_n|\mathcal{F}_{n-1})/\theta^2 \le V < \infty$ a.s. for $n = 1, 2, \ldots$. The constant V is independent of θ and known.

Note that $\mathrm{Var}(X_n|\mathcal{F}_{n-1}) := \mathbb{E}\left((X_n - \theta)^2|\mathcal{F}_{n-1}\right)$. In the i.i.d. case Assumption 6.1 reduces to $\sigma^2/\theta^2 \le V$, where $\sigma^2 = \mathrm{Var}\, X_n$.

Proposition 6.2. *There exists an (ε, α)-approximation $(t, \hat{\theta}_t)$ with nonadaptive stop criterion, which works on the class of input sequences $X_1, X_2, \ldots$ satisfying Assumption 6.1. The number of samples is*

$$t := 2\left\lceil 8.35\tfrac{V}{\varepsilon^2}\right\rceil \cdot \left\lceil 1.16\ln(2\alpha)^{-1}\right\rceil.$$

Proof. First observe that for $\bar{X}_n := \sum_{i=1}^n X_i/n$ we have $\mathbb{P}(|\bar{X}_n - \theta| > \varepsilon\theta) \le V/(n\varepsilon^2)$, because the Chebyshev inequality holds also for martingale differences, i.e. under Assumption 6.1. Therefore if $n \ge V/(a\varepsilon^2)$ then

$$\mathbb{P}(|\bar{X}_n - \theta| > \varepsilon\theta) \le a.$$

Now choose $t = nm$ and consider m batches of samples, each of length n, and observe that each batch satisfies Assumption 6.1. Let

$$\hat{\theta}_t := \mathrm{Med}\,(\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(m)}),$$

where $X_n^{(j)} := \sum_{i=n(j-1)+1}^{nj} X_i/n$ for $j = 1, \ldots, m$. If we choose $a = 0.12$ and let m be the least odd integer which satisfies $m \geq 2.31 \ln(2\alpha)^{-1}$ then $\mathbb{P}(|\hat{\theta}_t - \theta| > \varepsilon\theta) \leq \alpha$. For details we refer the reader to [38, Section 1], in particular Proposition 1.3 therein. □

Note that t defined in Proposition 6.2 satisfies

$$t \sim 19.34 \frac{V}{\varepsilon^2} \ln \alpha^{-1} \tag{6.1}$$

for $\varepsilon \to 0$ and $\alpha \to 0$. For a comparison let us recall the standard error approximation based on the Central Limit Theorem (CLT). If we take the sample average $\bar{X}_t$ as the estimator, then it follows from the CLT that $\mathbb{P}(|\bar{X}_t - \theta| > \varepsilon) \approx \alpha$ for

$$t \approx \frac{V}{\varepsilon^2} \left[\Phi^{-1}(1 - \alpha/2)\right]^2, \tag{6.2}$$

where Φ^{-1} is the quantile function of the standard normal distribution. Since $[\Phi^{-1}(1 - \alpha/2)]^2 \sim 2 \ln \alpha^{-1}$ for $\alpha \to 0$, the right hand side of (6.1) is bigger than (6.2) roughly by a constant factor of about 10 (for small α). The important difference is that (6.1) is sufficient for an *exact* bound while (6.2) is only for an *asymptotic* one. A similar discussion in [28, Section 10] is about bounds on the absolute error.

To our knowledge, the best *known* upper bound on the cost of an (ε, α)-approximation under Assumption 6.1 holds for the median of averages and is given by Proposition 6.2. However, we claim that it is an *open problem* to find an (ε, α)-approximation which is worst case optimal for the class of all BRV input sequences, with given V (even in the i.i.d. case).

Without a bound on relative variance, we cannot expect that an (ε, α)-approximation with nonadaptive stop criterion exists. In the rest of this section we restrict attention to i.i.d. input sequences but relax the assumption concerning the variance.

Assumption 6.3. X_n are i.i.d., nonnegative, $\mathbb{E}X_n = \theta \leq A < \infty$ a.s. and $\operatorname{Var} X_n/\theta \leq B < \infty$ a.s. for $n = 1, 2, \ldots$. The constants A and B are known.

The restriction $\theta \leq 1$ is natural in the context of rare event simulation. Assumption 6.3 is much weaker than the logarithmic efficiency condition [1,2], which stipulates that $\operatorname{Var} X_n/\theta^{2-\delta}$ is bounded for some $\delta > 0$. Under Assumption 6.3 we can combine Theorem 2.1 with the "median trick" to construct an (ε, α)-approximation.

Proposition 6.4. *There exists an (ε, α)-approximation $(T, \hat{\theta}_T)$ with adaptive stop criterion, which works on the class of input sequences $X_1, X_2, \ldots$ satisfying Assumption 6.3. For the expected number of samples we have the following inequality:*

$$rm \leq \theta \mathbb{E}T \leq (r + A + B)m,$$

where

$$r = r(\varepsilon) := \left\lceil 8.35B \left(\frac{\varepsilon + 1}{\varepsilon}\right)^2 + A + B \right\rceil, \tag{6.3}$$

$$m = m(\alpha) := 2\left\lceil 1.16 \ln(2\alpha)^{-1} \right\rceil. \tag{6.4}$$

Proof. Let $a = 0.12$, so that $1/a = 8.35$. If r is given by (6.3) then elementary algebra, similar to that in [28, Section 4], shows that

$$\frac{B}{r}\left[1+\frac{B+A}{r}\right] \le a\left(\frac{\varepsilon}{1+\varepsilon}\right)^2. \tag{6.5}$$

Theorem 2.1 now implies

$$\mathbb{P}\left(|\bar{\theta}_{(r)}-\theta| \ge \varepsilon\theta\right) \le a. \tag{6.6}$$

The number of samples used to compute $\bar{\theta}_{(r)}$ is equal to N_r and we have $r \le \theta\mathbb{E}N_r \le r+\theta+\sigma^2/\theta \le r+A+B$, as in the proof of Theorem 2.1. To obtain $\hat{\theta}$, we repeat m times the computation of $\bar{\theta}_{(r)}$ (using segments of the input sequence of random length) and take the median. If m is given by (6.4) then we have $\mathbb{P}\left(|\bar{\theta}_{(r)}-\theta| \ge \varepsilon\theta\right) \le \alpha$, as in the proof of Proposition 6.2. The total number T of samples is the sum of m independent copies of N_r. □

Note that for $\varepsilon \to 0$ and $\alpha \to 0$,

$$\theta\mathbb{E}T \sim r(\varepsilon)m(\alpha) \sim 19.34\frac{B}{\varepsilon^2}\ln\alpha^{-1}.$$

If X_n are i.i.d., $0 \le X_n \le 1$ a.s. and $\mathbb{E}X_n = \theta$ then Assumption 6.3 holds with $A = B = 1$. In particular this is true for Bernoulli schemes. Therefore the lower bound in Theorem 3.2 is applicable also under Assumption 6.3. We can easily see that $r(\varepsilon)m(\alpha)/\mathrm{Low}(\varepsilon,\alpha) \to 9.67$ as $\varepsilon \to 0$ and $\alpha \to 0$. The algorithm described in the proof of Proposition 6.4 requires roughly ten times as many samples as the optimal one if both are applied to Bernoulli inputs with small θ. In contrast with the case for bounded inputs considered in Section 3, we do not know how to construct an (ε,α)-approximation which is worst case optimal in the sense of Corollary 3.4. We even do not know whether Bernoulli sequences are still the worst case under Assumption 6.3. We think that these are *open questions*. A more important *open problem*, in our opinion, is that of constructing an optimal (ε,α)-approximation which works under the assumption $\mathrm{Var}\,X_n/\theta^{2-\delta} \le C$ (with known $\delta > 0$ and $C < \infty$). We conjecture that there exists an algorithm for which the expected number of samples is bounded by

$$(\text{absolute constant}) \cdot \frac{C}{\theta^\delta\varepsilon^2}\ln\alpha^{-1}.$$

Acknowledgment

This work was partially supported by the Polish National Science Center Grant No. N N201 608740.

Appendix. Auxiliary results

A.1. The Hoeffding inequality

We recall a classical result of Hoeffding [22] in a form generalized to martingale differences. For a new simpler proof we refer the reader to Bentkus [5].

Lemma A.1. *If Assumption 1.1 is fulfilled and* $0 < a < 1-\theta$ *then*

$$\mathbb{P}\left(\bar{X}_n-\theta \ge a\right) \le \exp\left\{-n\left[(\theta+a)\ln\frac{\theta+a}{\theta}+(1-\theta-a)\ln\frac{1-\theta-a}{1-\theta}\right]\right\} \tag{A.1}$$

$$\le \exp\left\{-n\frac{a}{1-\theta}\left[\left(1+\frac{\theta}{a}\right)\ln\left(1+\frac{\theta}{a}\right)-1\right]\right\} \tag{A.2}$$

$$\le \exp\left\{-n\frac{a^2}{1-\theta}\cdot\frac{1}{2(\theta+a/3)}\right\}. \tag{A.3}$$

If $0 < a < \theta$ *then*

$$\mathbb{P}\left(\theta - \bar{X}_n \geq a\right) \leq \exp\left\{-n\left[(\theta - a)\ln\frac{\theta - a}{\theta} + (1 - \theta + a)\ln\frac{1 - \theta + a}{1 - \theta}\right]\right\} \tag{A.4}$$

$$\leq \exp\left\{-n\frac{a}{\theta}\left[\left(1 + \frac{1 - \theta}{a}\right)\ln\left(1 + \frac{a}{1 - \theta}\right) - 1\right]\right\} \tag{A.5}$$

$$\leq \exp\left\{-n\frac{a^2}{\theta}\cdot\frac{1}{2(1 - \theta + a/3)}\right\}. \tag{A.6}$$

Proof. From the discussion on page 20 in the Hoeffding's paper [22] it follows that inequality (2.8) in his Theorem 3 implies (2.1) in his Theorem 1. Therefore our (A.1) is an immediate consequence of inequality (2.5) in Bentkus' Theorem 2.1 in [5]. Upper bounds (A.2) and (A.3) for the RHS of (A.1) are shown in the cited Hoeffding paper. Inequalities (A.4)–(A.6) are obvious, symmetric counterparts. □

In the classical case of i.i.d. random variables, (A.1)–(A.3) are known as Hoeffding, Bennett and Bernstein inequalities, respectively.

Note that if we pass to the limit with $a \to 1 - \theta$ in (A.1), we obtain

$$\mathbb{P}\left(\bar{X}_n - \theta \geq 1 - \theta\right) \leq \theta^n. \tag{A.7}$$

A.2. Wald identities

The following lemma gives martingale analogues of the two classical Wald identities. It is needed in Sections 3 and 4.

Lemma A.2. *If Assumption 1.1 or Assumption 6.3 holds then*

$$\mathbb{E}S_{N_r} = \mathbb{E}N_r\theta, \tag{A.8}$$

$$\mathbb{E}\left(S_{N_r} - N_r\theta\right)^2 = \mathbb{E}\sum_{i=1}^{N_r}\mathrm{Var}\,(X_i|\mathcal{F}_{i-1}). \tag{A.9}$$

Moreover, under Assumption 1.1,

$$\mathbb{E}\left(S_{N_r} - N_r\theta\right)^2 < r + 1. \tag{A.10}$$

Proof. To show (A.8), notice that $S_n - n\theta$ is a martingale and thus $\mathbb{E}S_{N_r\wedge n} = \mathbb{E}(N_r \wedge n)\theta$ by the optional sampling theorem. It is now enough to let $n \to \infty$ and invoke the monotone convergence theorem.

If $0 \leq X_n \leq 1$ then $\mathbb{E}\sum_{i=1}^{N_r}(X_i - \theta)^2 < \infty$, because the summands are bounded and $\mathbb{E}N_r < \infty$ in view of (A.8). Thus we can use Theorem 7.4.7 in [14] to get

$$\mathbb{E}\left(S_{N_r} - N_r\theta\right)^2 = \mathbb{E}\sum_{i=1}^{N_r}(X_i - \theta)^2 = \mathbb{E}\sum_{i=1}^{N_r}\mathrm{Var}\,(X_i|\mathcal{F}_{i-1}),$$

i.e. (A.9). If X_n are i.i.d. then (A.9) is well-known.

Under Assumption 1.1, $\mathrm{Var}\,(X_i|\mathcal{F}_{i-1}) = \mathbb{E}(X_i^2|\mathcal{F}_{i-1}) - \mathbb{E}(X_i|\mathcal{F}_{i-1})^2 \leq \mathbb{E}(X_i|\mathcal{F}_{i-1}) - \mathbb{E}(X_i|\mathcal{F}_{i-1})^2 = \theta(1-\theta) \leq \theta$, because $X_i \leq 1$. Using first (A.9), then (A.8) and finally the fact that $S_{N_r} < r + 1$ we infer that $\mathbb{E}(\theta N_r - S_{N_r})^2 \leq N_r\theta = \mathbb{E}S_{N_r} < r + 1$. □

A.3. The Lorden inequality

This useful result belongs to the renewal theory and was proved in [33]. For a newer proof see [8].

Lemma A.3. *If Assumption 6.3 holds and N_r is given by (1.1) then*

$$\mathbb{E}S_{N_r} - r \le \theta + \frac{\sigma^2}{\theta}.$$

Proof. Combine formulas (1.6) and (1.3) in [8]. □

References

[1] S. Asmussen, Ruin Probabilities, World Scientific, 2000.
[2] S. Asmussen, P.W. Glynn, Stochastic Simulation, Algorithms and Analysis, Springer-Verlag, 2007.
[3] S. Asmussen, D.P. Kroese, Improved algorithms for rare event simulation with heavy tails, Adv. Appl. Probab. 38 (2006) 545–558.
[4] S. Asmussen, R.Y. Rubinstein, Steady state rare event simulation in queuing models and its complexity properties, in: J.H. Dshalalow (Ed.), Advances in Queueing, 1995, pp. 429–461.
[5] V. Bentkus, On Hoeffding's inequalities, Ann. Probab. 32 (2004) 1650–1673.
[6] L.D. Brown, L. Gajek, Information inequalities for the Bayes risk, Ann. Statist. 18 (1990) 1578–1594.
[7] C.G. Bucher, Adaptive importance sampling—an iterative fast Monte Carlo procedure, Struct. Saf. 5 (1988) 119–126.
[8] J.T. Chang, Inequalities for the overshoot, Ann. Appl. Probab. 4 (1994) 1223–1233.
[9] J. Cheng, Sampling algorithms for estimating the mean of bounded random variables, Comput. Statist. 16 (2001) 1–23.
[10] J. Cheng, M.J. Druzdzel, AIS-BN: an adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks, J. Artificial Intelligence Res. 13 (2000) 155–188.
[11] J. Cheng, M.J. Druzdzel, Confidence inference in Bayesian networks, in: Breese, J.S. and Koller, D. (Eds.), Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, 2001, pp. 75–82.
[12] Y.S. Chow, H. Robbins, On the asymptotic theory of fixed-width sequential confidence intervals for mean, Ann. Math. Statist. 36 (1965) 457–462.
[13] Y.S. Chow, H. Robbins, D. Siegmund, Great Expectations: The Theory of Optimal Stopping, Houghton Mifflin, Boston, 1971.
[14] Y.S. Chow, H. Teicher, Probability Theory. Independence, Interchangeability, Martingales, Springer-Verlag, 1978.
[15] P. Dagum, R. Karp, M. Luby, S. Ross, An optimal algorithm for MC estimation, SIAM J. Comput. 29 (2000) 1484–1496. FOCS'95.
[16] P. Dagum, M. Luby, An optimal approximation algorithm for Bayesian inference, Artificial Intelligence 93 (1997) 1–27.
[17] D. Egloff, M. Leippold, Quantile estimation with adaptive importance sampling, Ann. Statist. 38 (2010) 1244–1278.
[18] B. Epstein, Estimates of bounded relative error for the mean life of an exponential distribution, Technometrics 3 (1961) 107–109.
[19] G.S. Fishman, Monte Carlo Concepts, Algorithms, and Applications, Springer-Verlag, New York, 1996.
[20] L. Gajek, On the minimax value in the scale model with truncated data, Ann. Statist. 16 (1988) 669–677.
[21] M.A. Girshick, H. Rubin, R. Sitgraves, Estimates of bounded relative error in particle counting, Ann. Math. Statist. 26 (1955) 276–285.
[22] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.
[23] Y. Ibrahim, An adaptive importance sampling strategy, Mech. Comput. (1990) 1288–1292. in 1990's.
[24] M.R. Jerrum, L.G. Valiant, V.V. Vazirani, Random generation of combinatorial structures from a uniform distribution, Theoret. Comput. Sci. 43 (1986) 169–188.
[25] R.M. Karp, M. Luby, Monte-Carlo algorithms for the planar multiterminal network reliability problem, J. Complexity 1 (1985) 45–64.
[26] A. Karamchandani, Adaptive importance sampling, in: Proc. Int. Conf. Structural Safety Reliability, ICOSSAR'89, 1989, pp. 855–862.
[27] Y.B. Kim, M.Y. Lee, Nonparametric adaptive importance sampling for rare event simulation, in: Joines, J.A. Barton, R.R. Kang, K. and Fishwick, P.A. (Eds.), Proc. 2000 Winter Simulation Conference, 2000, pp. 767–772.
[28] K. Łatuszyński, B. Miasojedow, W. Niemiro, Nonasymptotic bounds on the mean square error for MCMC estimates via renewal techniques, in: L. Plaskota and H. Woźniakowski (Eds.), Proceedings of the 9th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, MCQMC 2010, 2010, pp. 541–557.
[29] K. Łatuszyński, B. Miasojedow, W. Niemiro, Nonasymptotic bounds on the estimation error of MCMC algorithms, Bernoulli, (2012) (in press).
[30] K. Łatuszyński, W. Niemiro, $(\epsilon - \alpha)$-MCMC approximation under drift condition, in: Proceedings of the 6th International Workshop on Rare Event Simulation, RESIM 2006, 2006.
[31] K. Łatuszyński, W. Niemiro, Rigorous confidence bounds for MCMC under a geometric drift condition, J. Complexity 27 (2011) 23–38.
[32] P. L'Ecuyer, J. Blanchet, B. Tuffin, P.W. Glynn, Asymptotic robustness properties of estimators in rare-event simulation, ACM Trans. Model. Comput. Simul. 20 (2010) 6:1–6:41.
[33] G. Lorden, On excess over the boundary, Ann. Math. Statist. 41 (1970) 520–527.
[34] P. Mathé, Numerical integration using V-uniformly Markov chains, J. Appl. Probab. 41 (2004) 1104–1112.
[35] P. Mathé, E. Novak, Simple Monte Carlo and the Metropolis algorithm, J. Complexity 23 (2007) 673–696.
[36] A. Nádas, An extension of a theorem of Chow and Robbins of sequential confidence intervals for the mean, Ann. Math. Statist. 40 (1969) 667–671.
[37] W. Niemiro, P. Pokarowski, Tail events of some non-homogeneous Markov chains, Ann. Appl. Probab. 5 (1995) 261–293.
[38] W. Niemiro, P. Pokarowski, Fixed precision MCMC estimation by median of products of averages, J. Appl. Probab. 46 (2009) 309–329.
[39] M.S. Oh, J.O. Berger, Adaptive importance sampling in Monte Carlo integration, J. Stat. Comput. Simul. 41 (1992) 143–168.

[40] RESIM 2006, in: W. Sandman (Ed.), Proceedings of the 6th International Workshop on Rare Event Simulation, Bamberg.
[41] G. Rubino, B. Tuffin, Rare Event Simulation using Monte Carlo Methods, John Wiley & Sons, 2009.
[42] D. Rudolf, Explicit error bounds for lazy reversible Markov chain Monte Carlo, J. Complexity 25 (2009) 11–24.
[43] R.J. Serfling, D.D. Wackerly, Asymptotic theory of sequential fixed-width confidence interval procedures, J. Amer. Statist. Assoc. 71 (1976) 949–955.
[44] D. Siegmund, Sequential Analysis, in: Springer Series in Statistics, Springer-Verlag, 1985.
[45] J.S. Stadler, S. Roy, Adaptive importance sampling, IEEE J. Sel. Areas Commun. 11 (1993) 309–316.
[46] B. Tuffin, P. L'Ecuyer, W. Sandman, Robustness properties of highly reliable systems, in: Proceedings of the 6th International Workshop on Rare Event Simulation, RESIM 2006, 2006.
[47] X. Zhang, J. Blanchet, P.W. Glynn, Efficient suboptimal rare-event simulation, in: S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J.D. Tew, and R.R. Barton (Eds.), Proceedings of the 2007 Winter Simulation Conference, 2007.