Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

LOGARITHMICALLY EFFICIENT SIMULATION FOR MISCLASSIFICATION PROBABILITIES IN SEQUENTIAL MULTIPLE TESTING

Yanglei Song
Georgios Fellouris

Department of Statistics
& Coordinated Science Lab
University of Illinois, Urbana–Champaign
725 S. Wright Street
Champaign, IL 61820, USA

ABSTRACT

We consider the problem of estimating via Monte Carlo simulation the misclassification probabilities of two sequential multiple testing procedures. The first one stops when all local test statistics exceed simultaneously either a positive or a negative threshold. The second assumes knowledge of the true number of signals, say m, and stops when the gap between the top m test statistics and the remaining ones exceeds a threshold. For each multiple testing procedure, we propose an importance sampling algorithm for the estimation of its misclassification probability. These algorithms are shown to be logarithmically efficient when the data for the various statistical hypotheses are independent, and each testing problem satisfies an asymptotic stability condition and a symmetry condition. Our theoretical results are illustrated by a simulation study in the special case of testing the drifts of Gaussian random walks.

1 INTRODUCTION

One of the prototypical applications of rare-event simulation is the computation of the error probabilities of hypothesis testing procedures. In fact, the motivation for the celebrated, asymptotically efficient, importance sampling algorithm of Siegmund (1976) for the estimation of gambler's ruin probabilities was the computation of the type-I and type-II error probabilities of Wald's Sequential Probability Ratio Test (Wald 1945). More recently, Chan and Lai (2005) and Chan and Lai (2007) proposed asymptotically efficient importance sampling algorithms for the type-I error probability of truncated sequential tests based on the generalized likelihood ratio statistic.

Our work differs from the previous papers in that we consider a sequential *multiple* testing problem, where the goal is to solve simultaneously a multitude of binary testing problems. Specifically, we are interested in controlling the *misclassification rate*, that is the probability of at least one error, of two sequential multiple testing procedures that have been introduced in De and Baron (2012a) and Song and Fellouris (2016). The first procedure stops when all log-likelihood ratio test statistics exceed *simultaneously* either a positive or a negative threshold, and selects the alternative hypothesis in those streams with positive log-likelihood ratios upon stopping. The second one assumes knowledge of the true number of signals, say m, stops when the gap between the top m log-likelihood ratio statistics and the remaining ones is larger than a user-specified threshold, and selects the alternative hypotheses that correspond to the streams with the top m statistics.

Critical values for these testing procedures can be derived based on general, non-asymptotic upper bounds for the corresponding misclassification probabilities that have been obtained in De and Baron (2012a), Song and Fellouris (2016). However, due to the crudeness of these bounds, the critical values are very conservative and lead to sub-optimal performance compared to the one that could be achieved for the given tolerance to error. This can be avoided if the error probabilities can be computed efficiently via Monte Carlo simulation, in which case non-conservative critical values can be determined numerically.

In this work we propose a Monte Carlo approach, based on importance sampling, for the estimation of the misclassification probabilities of interest. Our main contribution is that we establish the asymptotic (logarithmic) efficiency of the proposed algorithms when the data streams that correspond to the various hypotheses are independent and each testing problem satisfies a symmetric and an asymptotic stability condition. To our knowledge, these are the first optimality results regarding the efficient simulation of error probabilities in sequential multiple testing, even in the special case of independent and identically distributed (i.i.d.) observations, where the corresponding test statistics become random walks. Indeed, the problem of interest is to estimate the probability that a *multi*-dimensional stochastic process (random walk in the case of i.i.d. observations) *exits a set in a "wrong" way*. This is fundamentally different from the probability that the same process *ever hits a "rare" set*, which has been considered for example by Glasserman and Wang (1997), Collamore (2002), Blanchet and Liu (2010). However, it is worth noting that the proposed importance sampling algorithms are based on (finite) mixtures of measures, which are known to be useful for estimating the probabilities of unions of rare events, as for example in Glasserman and Juneja (2008).

The rest of the paper is organized as follows: In Section 2 we formulate the problem of interest. In Section 3 we introduce and analyze the proposed importance sampling algorithm for a multiple testing procedure without any prior information on the number of signals. In Section 4 we do the same for a multiple testing procedure that knows a priori the number of signals. In Section 5 we present a simulation study that illustrates our theoretical results. In Section 6 we discuss potential generalizations of our work. Finally, we present a technical lemma and its proof in the Appendix.

2 PROBLEM FORMULATION

Consider K *independent* streams of observations, $X^k = \{X_n^k : n \in \mathbb{N}\}$, where $k \in [K] := \{1, \ldots, K\}$ and $\mathbb{N} := \{1, 2, \ldots\}$. We denote by $\mathcal{F}_n$ the σ-field generated by all streams up to time n, i.e., $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$, where $X_n := (X_n^1, \ldots, X_n^K)$ is the observed vector at time n. For each $k \in [K]$, we denote by P^k the distribution of X^k and consider two simple hypotheses for it:

$$H_0^k \ : \mathsf{P}^k = \mathsf{P}_0^k \quad \text{versus} \quad H_1^k \ : \mathsf{P}^k = \mathsf{P}_1^k. \tag{1}$$

We will say that there is "noise" in the k^{th} stream under P_0^k and "signal" under P_1^k. For each $k \in [K]$ and $n \in \mathbb{N}$, the probability measures P_0^k and P_1^k are assumed to be equivalent, i.e., mutually absolutely continuous, when they are both restricted to the σ-algebra $\mathcal{F}_n^k := \sigma(X_1^k, \ldots, X_n^k)$ and we denote by $\lambda^k(n)$ the corresponding log-likelihood ratio, i.e.,

$$\lambda^k(n) := \log \frac{d\mathsf{P}_1^k}{d\mathsf{P}_0^k}(\mathcal{F}_n^k). \tag{2}$$

Moreover, we assume that the two hypotheses in each stream are well-separated, in the sense that

$$\mathsf{P}_0^k \left(\lim_{n\to\infty} \lambda^k(n) = -\infty \right) = \mathsf{P}_1^k \left(\lim_{n\to\infty} \lambda^k(n) = \infty \right) = 1, \quad \text{for each } \ k \in [K]. \tag{3}$$

We denote by P_A the underlying probability measure when $A \subset [K]$ is the true subset of signals, i.e.,

$$\mathsf{P}_A := \bigotimes_{k=1}^{K} \mathsf{P}^k; \quad \mathsf{P}^k = \begin{cases} \mathsf{P}_0^k, & \text{if } k \notin A \\ \mathsf{P}_1^k, & \text{if } k \in A \end{cases}. \tag{4}$$

For any $C \subset [K]$ and $n \in \mathbb{N}$ we denote by $\lambda^{A,C}(n)$ the log-likelihood ratio of P_A versus P_C when both measures are restricted to $\mathcal{F}_n$; from (4) it is clear that

$$\lambda^{A,C}(n) := \log \frac{d\mathsf{P}_A}{d\mathsf{P}_C}(\mathcal{F}_n) = \sum_{k \in A \setminus C} \lambda^k(n) - \sum_{j \in C \setminus A} \lambda^j(n). \tag{5}$$

We assume that data are acquired sequentially in all streams and that the goal is to stop sampling (simultaneously in all streams) as soon as there is sufficient evidence in order to identify the correct hypothesis in all K testing problems of interest. Formally, a *sequential multiple testing procedure* is a family of pairs $\{(T_b, D_b) : b > 0\}$, where b is a user-specified parameter, T_b is an $\{\mathcal{F}_n\}$-stopping time at which we stop sampling in all streams, and $D_b := (D_b^1, \ldots, D_b^K)$ is an $\mathcal{F}_{T_b}$-measurable, K-dimensional random vector with values in $\{0,1\}^K$, which represents the decision upon stopping. Specifically, for each $k \in [K]$ hypothesis H_i^k is selected on the event $\{D_b^k = i, T_b < \infty\}$, $i = 0, 1$. With an abuse of notation, we will also identify D_b with the subset $\{k \in [K] : \; D_b^k = 1\}$ of streams in which the alternative hypothesis is selected upon stopping.

If A is the true subset of signals, then $\alpha(b) := \mathsf{P}_A(D_b \neq A)$ is the misclassification probability of (T_b, D_b), i.e., its probability of making at least one mistake. We will assume that $\alpha(b) \to 0$ as $b \to \infty$; thus, the user can control $\alpha(b)$ at arbitrarily small levels with an appropriate selection of the parameter b. In the absence of closed-form expressions, sharp bounds, or good approximations for $\alpha(b)$, Monte Carlo simulation provides an attractive method for its computation. The plain Monte Carlo approach suggests averaging independent realizations of the indicator of the event $\{D_b \neq A\}$, generated under P_A. However, it is well understood that this approach is inefficient when $\alpha(b)$ is very small as the number of simulations required to guarantee a certain relative error is inversely proportional to $\alpha(b)$.

An alternative way to estimate $\alpha(b)$ via Monte Carlo simulation is based on Wald's likelihood ratio identity. For any probability measure $\bar{\mathsf{P}}$ under which the stopping time T_b is almost surely finite we have

$$\alpha(b) = \bar{\mathsf{E}}\left[\bar{\Lambda}_b^{-1}; \bar{D}_b \neq A\right],$$

where $\bar{\mathsf{E}}$ is expectation under $\bar{\mathsf{P}}$, and $\bar{\Lambda}_b$ is the likelihood ratio of $\bar{\mathsf{P}}$ versus P_A when both measures are restricted to the σ-algebra generated by all observations up to time T_b. This identity suggests that an alternative way to estimate $\alpha(b)$ is to average independent realizations of $\bar{\alpha}(b)$, generated under $\bar{\mathsf{P}}$, where

$$\bar{\alpha}(b) := \bar{\Lambda}_b^{-1} \, \mathbb{I}\{\bar{D}_b \neq A\}.$$

We would like to select $\bar{\mathsf{P}}$ such that the second moment, and consequently the variance, of $\bar{\alpha}(b)$ under $\bar{\mathsf{P}}$ goes to 0 at the fastest possible rate. From the non-negativity of the variance it is clear that for every $b > 0$ we have $\bar{\mathsf{E}}\left[\bar{\alpha}^2(b)\right] \geq \alpha^2(b)$, and consequently

$$\limsup_{b \to \infty} \frac{|\log \bar{\mathsf{E}}\left[\bar{\alpha}^2(b)\right]|}{|\log \alpha^2(b)|} \leq 1.$$

We will say that $\bar{\mathsf{P}}$ is *asymptotically efficient* or *asymptotically optimal* (Asmussen and Glynn 2007) if

$$\liminf_{b \to \infty} \frac{|\log \bar{\mathsf{E}}\left[\bar{\alpha}^2(b)\right]|}{|\log \alpha^2(b)|} \geq 1,$$

in which case the equality holds and $\liminf$ can be replaced by $\lim$. This notion of asymptotic optimality is also known as *logarithmic efficiency*, and it is well known to be equivalent to

$$\limsup_{b \to \infty} \frac{\bar{\mathsf{E}}\left[\bar{\alpha}^2(b)\right]}{\alpha^{2-\epsilon}(b)} = 0, \quad \forall \epsilon > 0.$$

Our main goal in this paper is to propose logarithmically efficient importance sampling estimators for the misclassification probabilities of two specific multiple testing procedures, which have been proposed by De and Baron (2012a) and Song and Fellouris (2016). In order to do so, we need to make certain assumptions on the distributions of the observed processes and the structure of the testing problems. In particular, we will assume that

1. there are positive numbers $\{\mathcal{I}_0^k, \mathcal{I}_1^k : k \in [K]\}$ such that for every $k \in [K]$ and $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathsf{P}_0^k \left(\left| \frac{1}{n} \lambda^k(n) + \mathcal{I}_0^k \right| > \epsilon \right) < \infty, \quad \sum_{n=1}^{\infty} \mathsf{P}_1^k \left(\left| \frac{1}{n} \lambda^k(n) - \mathcal{I}_1^k \right| > \epsilon \right) < \infty, \tag{6}$$

2. the null and the alternative hypothesis in each stream are symmetric, in the sense that

$$\mathcal{I}_0^k = \mathcal{I}_1^k = \mathcal{I}^k \quad \text{for every} \quad k \in [K]. \tag{7}$$

Condition (6) guarantees the asymptotic stability of the log-likelihood ratio in each testing problem, in the sense that it satisfies a strengthened version of the Strong Law of Large Numbers. Indeed, from the Borel–Cantelli Lemma it follows that (6) implies

$$\mathsf{P}_0^k \left(\lim_{n \to \infty} \frac{1}{n} \lambda^k(n) = -\mathcal{I}_0^k \right) = 1, \quad \mathsf{P}_1^k \left(\lim_{n \to \infty} \frac{1}{n} \lambda^k(n) = \mathcal{I}_1^k \right) = 1.$$

The stronger notion of convergence described in (6) was introduced by Hsu and Robbins (1947) and is known as *complete convergence*. It is satisfied by a large class of stochastic models with possibly dependent observations, such as hidden Markov models and autoregressive models (see, e.g. Sections 3.4.6 and 3.4.7 in Tartakovsky, Nikiforov, and Basseville (2014)). In the special case that each λ^k is a random walk, i.e., the increments $\{\lambda^k(n) - \lambda^k(n-1); n \in \mathbb{N}\}$ are i.i.d. under P_i^k, then it is well known that (6) is equivalent to the finiteness of the *second* moment of $\lambda^k(1)$ under P_i^k, where $i = 0, 1$ (Hsu and Robbins 1947, Erdos 1949).

Condition (7) is clearly satisfied when the distribution of λ^k under P_1^k is the same as the distribution of $-\lambda^k$ under P_0^k. This is in particular the case in the fundamental problem of testing the means of Gaussian i.i.d. observations, which we consider in more detail in Section 5.

3 INTERSECTION RULE

In this section we focus on the so-called "*intersection rule*", $(\widetilde{T}_b, \widetilde{D}_b)$, according to which we stop when all local log-likelihood ratio statistics are either above b or below $-b$. Formally,

$$\widetilde{T}_b := \inf \left\{ n \in \mathbb{N} : |\lambda^k(n)| \geq b \;\; \text{for every } k \in [K] \right\},$$

$$\widetilde{D}_b^k := \begin{cases} 1, & \text{if} \quad \lambda^k(\widetilde{T}_b) \geq b \\ 0, & \text{if} \quad \lambda^k(\widetilde{T}_b) \leq -b \end{cases} \qquad k \in [K].$$

In order to compute the misclassification probability of the intersection rule,

$$\alpha_{int}(b) := \mathsf{P}_A(\widetilde{D}_b \neq A),$$

we suggest an importance sampling approach that is based on a change of measure from P_A to a uniform mixture over measures of the form $\{\mathsf{P}_C, \; |C \triangle A| = 1\}$, i.e.,

$$\widetilde{\mathsf{P}} := \frac{1}{K} \left[\sum_{j \notin A} \mathsf{P}_{A \cup \{j\}} + \sum_{k \in A} \mathsf{P}_{A \setminus \{k\}} \right]. \tag{8}$$

That is, we suggest estimating $\alpha_{int}(b)$ by averaging independent realizations of

$$\widetilde{\alpha}(b) := \widetilde{\Lambda}_b^{-1}\, \mathbb{I}\{\widetilde{D}_b \neq A\},$$

generated under $\widetilde{\mathsf{P}}$, where $\widetilde{\Lambda}_b$ is the likelihood ratio of $\widetilde{\mathsf{P}}$ versus P_A when both measures are restricted to the σ-algebra generated by all observations up to time $\widetilde{T}_b$, which takes the form

$$\widetilde{\Lambda}_b := \frac{1}{K}\left[\sum_{j\notin A} \exp\{\lambda^j(\widetilde{T}_b)\} + \sum_{k\in A} \exp\{-\lambda^k(\widetilde{T}_b)\}\right].$$

Our goal in this section is to show that, under certain assumptions on the testing problem, this is a logarithmically efficient importance sampling estimator for $\alpha_{int}(b)$.

Lemma 1 For any $b > 0$ we have $\widetilde{\mathsf{P}}(\widetilde{T}_b < \infty) = 1$. Moreover,

$$\limsup_{b\to\infty} \frac{\log \alpha_{int}(b)}{b} \leq -1, \quad \liminf_{b\to\infty} \frac{|\log \widetilde{\mathsf{E}}\left[\widetilde{\alpha}^2(b)\right]|}{b} \geq 2. \tag{9}$$

Proof. Fix $b > 0$. In order to prove that $\widetilde{T}_b$ terminates almost surely under $\widetilde{\mathsf{P}}$, it suffices to show that this is the case under any measure of the form P_C. Fix $C \subset [K]$. Then,

$$\widetilde{T}_b \leq \widetilde{T}'_b := \inf\{n \geq 1 : \lambda^k(n) \geq b \text{ and } \lambda^j(n) \leq -b \quad \text{for any } k \in C, j \notin C\}$$

and condition (3) guarantees that $\widetilde{T}'_b$, and consequently $\widetilde{T}_b$, is almost surely finite under P_C. Now, we turn to the proof of the two asymptotic bounds in (9). On the event $\{\widetilde{D}_b \neq A\}$, either there is a $j \notin A$ such that $\lambda^j(\widetilde{T}_b) \geq b$ or there is a $k \in A$ such that $\lambda^k(\widetilde{T}_b) \leq -b$, and consequently $\widetilde{\Lambda}_b \geq e^b/K$. Therefore, $\widetilde{\alpha}(b) \leq Ke^{-b}$, which implies that

$$\alpha_{int}(b) = \widetilde{\mathsf{E}}\left[\widetilde{\alpha}(b)\right] \leq Ke^{-b}, \quad \widetilde{\mathsf{E}}\left[\widetilde{\alpha}^2(b)\right] \leq K^2 e^{-2b}. \tag{10}$$

Taking logarithms, dividing by b and letting $b \to \infty$ in these two inequalities completes the proof. □

Remark 1 From (9) it follows that to establish the logarithmic efficiency of $\widetilde{\mathsf{P}}$, it suffices to show that

$$\liminf_{b\to\infty} \frac{1}{b} \log \alpha_{int}(b) \geq -1. \tag{11}$$

Indeed, this asymptotic lower bound together with the first inequality in (9) implies that

$$\lim_{b\to\infty} \frac{1}{b} |\log \alpha_{int}(b)| = 1,$$

which, together with the second inequality in (9), guarantees the logarithmic efficiency property, i.e.,

$$\lim_{b\to\infty} \frac{|\log \widetilde{\mathsf{E}}\left[\widetilde{\alpha}^2(b)\right]|}{|\log \alpha^2_{int}(b)|} = 1. \tag{12}$$

Theorem 1 Suppose that assumptions (6)–(7) hold. Then, the mixture distribution $\widetilde{\mathsf{P}}$, defined in (8), is logarithmically efficient for the estimation of the misclassification probability of the intersection rule, i.e., (12) holds.

Proof. Based on Remark 1, it suffices to show that the asymptotic lower bound in (11) holds under the conditions of the theorem. In order to do so, we set

$$L_0(A) := \min_{j \notin A} \mathcal{I}^j, \quad L_1(A) := \min_{k \in A} \mathcal{I}^k, \quad L(A) := \min\{L_0(A),\, L_1(A)\}, \tag{13}$$

and we adopt the convention that $L_0([K]) = L_1(\emptyset) := \infty$. Then, either there is some $j_0 \notin A$ such that $\mathcal{I}^{j_0} = L(A)$, or some $k_0 \in A$ such that $\mathcal{I}^{k_0} = L(A)$. Without loss of generality, we assume the existence of j_0, and set $C = A \cup \{j_0\}$. Clearly,

$$\alpha_{int}(b) = \mathsf{P}_A(\widetilde{D}_b \neq A) \geq \mathsf{P}_A(\widetilde{D}_b = C).$$

From representation (5) it follows that for every $n \in \mathbb{N}$ we have $\lambda^{A,C}(n) = -\lambda^{j_0}(n)$. Therefore, from Wald's likelihood ratio identity we have

$$\mathsf{P}_A(\widetilde{D}_b = C) = \mathsf{E}_C\left[\exp\{-\lambda^{j_0}(\widetilde{T}_b)\};\ \widetilde{D}_b = C\right].$$

Then, for every $\eta > 0$ we have

$$\begin{aligned}\alpha_{int}(b) &\geq \mathsf{E}_C\left[\exp\{-\lambda^{j_0}(\widetilde{T}_b)\}; \widetilde{D}_b = C,\ \lambda^{j_0}(\widetilde{T}_b) < (1+\eta)b\right] \\ &\geq e^{-(1+\eta)b}\,\mathsf{P}_C\left(\widetilde{D}_b = C, \lambda^{j_0}(\widetilde{T}_b) < (1+\eta)b\right),\end{aligned}$$

and consequently,

$$\frac{1}{b}\log \alpha_{int}(b) \geq -(1+\eta) + \frac{1}{b}\log\left[\mathsf{P}_C\left(\widetilde{D}_b = C, \lambda^{j_0}(\widetilde{T}_b) < (1+\eta)b\right)\right].$$

Since η is an arbitrary positive number, it suffices to show that the probability in the right-hand side goes to 1 as $b \to \infty$. Since this is the probability of an intersection, it suffices to show that as $b \to \infty$ we have

$$\mathsf{P}_C(\widetilde{D}_b \neq C) \to 0 \quad \text{and} \quad \mathsf{P}_C\left(\lambda^{j_0}(\widetilde{T}_b) \geq (1+\eta)b\right) \to 0.$$

From (10) it is clear that the first convergence holds; therefore it suffices to show that the second holds as well. From the definition of the stopping rule $\widetilde{T}_b$, it follows that

$$\{\lambda^{j_0}(\widetilde{T}_b) \geq (1+\eta)b\} = \bigcup_{k \in [K]} \left\{|\lambda^k(\widetilde{T}_b - 1)| < b\right\} \bigcap \{\lambda^{j_0}(\widetilde{T}_b) \geq (1+\eta)b\}.$$

Therefore, due to Boole's inequality, it suffices to show that for every $k \in C$ and $j \notin C$ we have

$$\begin{aligned}&\mathsf{P}_C\left(\lambda^k(\widetilde{T}_b - 1) \leq b,\ \lambda^{j_0}(\widetilde{T}_b) \geq (1+\eta)b\right) \to 0, \\ &\mathsf{P}_C\left(-\lambda^j(\widetilde{T}_b - 1) \leq b,\ \lambda^{j_0}(\widetilde{T}_b) \geq (1+\eta)b\right) \to 0.\end{aligned} \tag{14}$$

Fix $k \in C$ and $j \notin C$. Then, from assumptions (6)–(7) we have

$$\sum_{n=1}^{\infty} \mathsf{P}_C\left(\left|\frac{1}{n}\lambda^k(n) - \mathcal{I}^k\right| > \epsilon\right) < \infty, \quad \sum_{n=1}^{\infty} \mathsf{P}_C\left(\left|\frac{1}{n}\lambda^j(n) + \mathcal{I}^j\right| > \epsilon\right) < \infty.$$

If $k = j_0$, trivially $\mathcal{I}^k = \mathcal{I}^{j_0}$; otherwise, $k \in A$, in which case we have

$$\mathcal{I}^k \geq L_1(A) \geq L(A) = \mathcal{I}^{j_0},$$

where the first two inequalities follow from the definition of $L_1(A)$ and $L(A)$ in (13), and the last from the definition of j_0. On the other hand, since $j \notin C$ implies $j \notin A$, again from (13) we have

$$\mathcal{I}^j \geq L_0(A) \geq L(A) = \mathcal{I}^{j_0}.$$

Thus, from Lemma 3 in the Appendix it follows that (14) holds, which completes the proof. □

4 GAP RULE

In this section we assume that we know a priori that there are exactly m signals, for some $1 \leq m \leq K-1$, and we focus on the "*gap rule*", $(\widehat{T}_b, \widehat{D}_b)$, which stops when the gap between the m-th and the $(m+1)$-th top log-likelihood ratio statistics is larger than some positive threshold b, and selects the alternative hypothesis in the m streams with the top log-likelihood ratios upon stopping. Formally,

$$\begin{aligned}\widehat{T}_b &:= \inf\left\{n \geq 1 : \lambda^{(m)}(n) - \lambda^{(m+1)}(n) \geq b\right\},\\ \widehat{D}_b &:= \{i_1(\widehat{T}_b), \ldots, i_m(\widehat{T}_b)\},\end{aligned}$$

where $\lambda^{(1)}(n) \geq \ldots \geq \lambda^{(K)}(n)$ are the ordered log-likelihood ratio statistics at time n, and $i_1(n), \ldots, i_K(n)$ are the corresponding stream indices, i.e., $\lambda^{(k)}(n) = \lambda^{i_k(n)}(n)$ for every $k \in [K]$. For the estimation of the misclassification probability of the gap-rule

$$\alpha_{gap}(b) := \mathsf{P}_A(\widehat{D}_b \neq A)$$

for some $A \subset [K]$ such that $|A| = m$, we propose an importance sampling approach based on a change of measure from P_A to a uniform mixture over $\{\mathsf{P}_C : |C \setminus A| = |A \setminus C| = 1\}$, i.e.,

$$\widehat{\mathsf{P}} := \frac{1}{m(K-m)} \sum_{k \in A} \sum_{j \notin A} \mathsf{P}_{(A \setminus \{k\}) \cup \{j\}}, \tag{15}$$

Specifically, we suggest estimating $\alpha_{gap}(b)$ by averaging independent realizations of

$$\widehat{\alpha}(b) := \widehat{\Lambda}_b^{-1}\, \mathbb{I}\{\widehat{D}_b \neq A\},$$

where $\widehat{\Lambda}_b$ is the likelihood ratio of $\widehat{\mathsf{P}}$ versus P_A when both measures are restricted to the σ-algebra generated by all observations up to time $\widehat{T}_b$, i.e.,

$$\widehat{\Lambda}_b := \frac{1}{m(K-m)} \sum_{k \in A} \sum_{j \notin A} \exp\left\{\lambda^j(\widehat{T}_b) - \lambda^k(\widehat{T}_b)\right\}.$$

We will show that, under the same conditions as the ones we imposed in the previous section, this is a logarithmically efficient estimator of $\alpha_{gap}(b)$.

Lemma 2 For any $b > 0$ we have $\widehat{\mathsf{P}}(\widehat{T}_b < \infty) = 1$. Moreover,

$$\limsup_{b \to \infty} \frac{\log \alpha_{gap}(b)}{b} \leq -1, \quad \liminf_{b \to \infty} \frac{|\log \widehat{\mathsf{E}}\left[\widehat{\alpha}^2(b)\right]|}{b} \geq 2. \tag{16}$$

Proof. Fix $b > 0$. A similar argument as in the proof of Lemma 1 can be used to show that the gap rule terminates almost surely under $\widehat{\mathsf{P}}$. On the event $\{\widehat{D}_b \neq A\}$ there exist $k_0 \in A$ and $j_0 \notin A$ such that $\lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b) \geq b$, and consequently

$$\widehat{\Lambda}_b \geq \frac{1}{m(K-m)} e^b.$$

Therefore, $\widehat{\alpha}_b \leq m(m-K)e^{-b}$, which clearly implies

$$\alpha_{gap}(b) = \widehat{\mathsf{E}}\left[\widehat{\alpha}(b)\right] \leq m(m-K)e^{-b}, \qquad \widehat{\mathsf{E}}\left[\widehat{\alpha}^2(b)\right] \leq (m(m-K))^2\, e^{-2b}. \tag{17}$$

Taking logarithms, dividing by b and letting $b \to \infty$ in these two inequalities implies the asymptotic upper bounds in (16). □

Theorem 2 Assume that (6)–(7) hold. The importance distribution $\widehat{\mathsf{P}}$, defined in (15), is logarithmically efficient, that is,

$$\lim_{b\to\infty} \frac{|\log \widehat{\mathsf{E}}\left[\widehat{\alpha}^2(b)\right]|}{|\log \alpha_{gap}^2(b)|} = 1.$$

Proof. From (16), and by similar argument as in Remark 1, it suffices to show that

$$\liminf_{b\to\infty} \frac{1}{b}\log \alpha_{gap}(b) \geq -1.$$

Fix $b > 0$ and recall the definitions of $L_0(A)$ and $L_1(A)$ in (13). Then, there exist $k_0 \in A$ and $j_0 \notin A$ such that $L_0(A) = \mathcal{I}^{j_0}$ and $L_1(A) = \mathcal{I}_1^{k_0}$. We set $C = (A \setminus \{k_0\}) \cup \{j_0\}$. Clearly, $|C| = m$. Then, for any $\eta > 0$ we have

$$\begin{aligned}\alpha_{gap}(b) = \mathsf{P}_A(\widehat{D}_b \neq A) &\geq \mathsf{P}_A(\widehat{D}_b = C) \\ &= \mathsf{E}_C\left[\exp\{-(\lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b))\};\ \widehat{D}_b = C\right] \\ &\geq \mathsf{E}_C\left[\exp\{-(\lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b))\};\ \widehat{D}_b = C,\ \lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b) < (1+\eta)b\right] \\ &\geq e^{-(1+\eta)b}\left[1 - \mathsf{P}_C(\widehat{D}_b \neq C) - \mathsf{P}_C\left(\lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b) \geq (1+\eta)b\right)\right].\end{aligned}$$

From (17) it follows that $\mathsf{P}_C(\widehat{D}_b \neq C) \to 0$ as $b \to \infty$. Since η is arbitrary, by a similar argument as in the proof of Theorem 1, it suffices to show that as $b \to \infty$ we have

$$\mathsf{P}_C\left(\lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b) \geq (1+\eta)b\right) \to 0.$$

At time $\widehat{T}_b - 1$, there exist $k \in C, j \notin C$ such that $\lambda^k(\widehat{T}_b - 1) - \lambda^j(\widehat{T}_b - 1) < b$. Thus, by Boole's inequality it suffices to show that for every $k \in C$ and $j \notin C$ we have

$$\mathsf{P}_C\left(\lambda^k(\widehat{T}_b - 1) - \lambda^j(\widehat{T}_b - 1) < b, \lambda^{j_0}(\widehat{T}_b) - \lambda^{k_0}(\widehat{T}_b) \geq (1+\eta)b\right) \to 0. \tag{18}$$

Fix $k \in C$ and $j \notin C$. Due to assumption (6) we have

$$\sum_{n=1}^{\infty} \mathsf{P}_C\left(\left|\frac{1}{n}(\lambda^k(n) - \lambda^j(n)) - (\mathcal{I}^k + \mathcal{I}^j)\right| \geq \epsilon\right) < \infty,$$
$$\sum_{n=1}^{\infty} \mathsf{P}_C\left(\left|\frac{1}{n}(\lambda^{j_0}(n) - \lambda^{k_0}(n)) - (\mathcal{I}^{j_0} + \mathcal{I}^{k_0})\right| \geq \epsilon\right) < \infty.$$

If $k \neq k_0$ and $j \neq j_0$, then $k \in A$ and $j \notin A$, and from (13) we have

$$\mathcal{I}^k + \mathcal{I}^j \geq L_1(A) + L_0(A) = \mathcal{I}^{k_0} + \mathcal{I}^{j_0}$$

Clearly, if either $k = k_0$ or $j = j_0$, then $\mathcal{I}^k + \mathcal{I}^j \geq \mathcal{I}^{k_0} + \mathcal{I}^{j_0}$ still holds. Thus, by Lemma 3 in the Appendix we have that (18) holds, which completes the proof. □

5 A SIMULATION STUDY

Our goal in this section is to illustrate the theoretical results of the previous two sections with a simulation study. To this end, suppose that, for each $k \in [K]$, $\{X_n^k : n \in \mathbb{N}\}$ is a sequence of independent random variables with common density f^k relative to some σ-finite measure μ^k. In this context, for each stream $k \in [K]$ the hypothesis testing problem (1) takes the form

$$H_0^k \ : f^k = f_0^k \quad \text{versus} \quad H_1^k \ : f^k = f_1^k,$$

where f_0^k and f_1^k are densities with common support. Further, the log-likelihood ratio process λ^k, defined in (2), becomes a random walk. Thus, assumption (6) is satisfied if and only if $\lambda^k(1)$ has a finite *second* moment under both P_0^k and P_1^k, and $\mathcal{I}_0^k$ and $\mathcal{I}_1^k$ reduce to the Kullback-Leibler numbers between f_0^k and f_1^k, i.e.,

$$\mathcal{I}_0^k = \int \log\left(\frac{f_0^k}{f_1^k}\right) f_0^k d\mu^k \quad \text{and} \quad \mathcal{I}_1^k = \int \log\left(\frac{f_1^k}{f_0^k}\right) f_1^k d\mu^k.$$

The assumption (7) requires that the two hypotheses are symmetric, in the sense that $\mathcal{I}_0^k = \mathcal{I}_1^k$ for every $k \in [K]$. This assumption is satisfied in the fundamental problem of testing the drifts of Gaussian random walks, that is when $f_0^k = \mathcal{N}(\theta_0^k, \sigma_k^2)$ and $f_1^k = \mathcal{N}(\theta_1^k, \sigma_k^2)$ for some $\sigma_k > 0$ and real numbers $\theta_0^k \neq \theta_1^k$. In this case, the distribution of λ^k under P_1^k is the same as the distribution of $-\lambda^k$ under P_0^k, and

$$\mathcal{I}_0^k = \frac{(\theta_1^k - \theta_0^k)^2}{2\sigma_k^2} = \mathcal{I}_1^k.$$

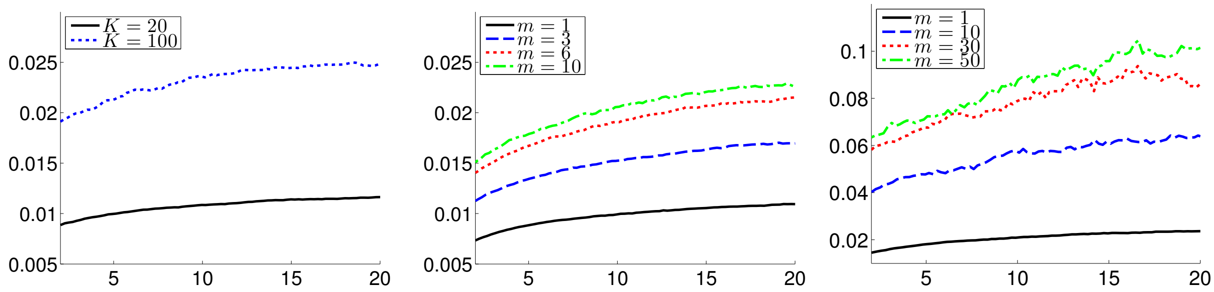
We will illustrate our theoretical results in this context. Although this is not needed for the logarithmic efficiency of the proposed algorithms, we will further assume that the hypotheses are homogeneous and set $\theta_0^k = \theta_0 = 0$, $\theta_1^k = \theta_1 = 0.5$ and $\sigma_k = \sigma = 1$. This is a convenient assumption, because in this case the misclassification probability of the intersection rule, $\alpha_{int}(b)$, does not depend on the true subset of signals, A. The same is true for the relative error of the proposed estimator of $\alpha_{int}(b)$, i.e., the standard deviation of the estimator divided by the estimate itself. In Figure 1a we plot this relative error against the estimate of $|\log_{10} \alpha_{int}(b)|$ for different values of b when $K = 20$ and $K = 100$.

The symmetry and homogeneity of this testing problem also guarantees that the misclassification probability of the gap rule, $\alpha_{gap}(b)$, is the same whenever the true subset of signals has size m and $K - m$. Therefore, for a given (even) K it suffices to simulate the gap rule for $m = \{1, \ldots, K/2\}$. In Figures 1b and 1c we plot the relative error of the proposed importance sampling estimator against the estimate of $|\log_{10} \alpha_{gap}(b))|$ for $K = 20$ and $K = 100$, respectively.

From Figure 1, we can see that even when the misclassification probability of the intersection rule is as small as 10^{-20}, the relative error of the proposed estimator is roughly 1% when $K = 20$, and 2.5% when $K = 100$. On the other hand, when the misclassification probability of the gap rule is as small as 10^{-20}, the maximum relative error of the proposed estimator is achieved when $m = K/2$ and is roughly 2.5% when $K = 20$ and 10% when $K = 100$. Therefore, it is clear that the difficulty of the estimation problem increases as the number streams increases.

6 DISCUSSION

It is easy to see that the logarithmic efficiency still holds if we employ *non-uniform* mixtures over $\{\mathsf{P}_C : |C \triangle A| = 1\}$, and $\{\mathsf{P}_C : |C \setminus A| = |A \setminus C| = 1\}$ for the estimation of $\alpha_{int}(b)$ and $\alpha_{gap}(b)$, respectively, and it would be interesting to understand the potential gains of such a non-uniform mixture. One direction of future work is to consider the computation of alternative error probabilities, such as the familywise type-I and type-II errors that have been considered by De and Baron (2012b), Bartroff and Song (2014), Song and Fellouris (2016). Finally, it is interesting to extend the results of the current paper to a more general framework where the processes $\{\lambda^k\}$ are not necessarily log-likelihood ratios.



(a) Intersection rule (b) Gap rule: $K = 20$ (c) Gap rule: $K = 100$

Figure 1: The x-axis is the estimate of $|\log_{10} \alpha_{int}(b)|$ in (a) and $|\log_{10} \alpha_{gap}(b)|$ in (b) and (c). The y-axis is the relative error, that is the standard deviation of the proposed estimator divided by the estimate itself. Each curve is computed based on $100,000$ realizations.

7 ACKNOWLEDGMENTS

This work was supported by the Division of Computing and Communication Foundations through the U.S. National Science Foundation under Grant 1514245 and in part by the Simons Foundation under Grant C3663.

A APPENDIX

In this Appendix we state and prove a general lemma that was used in order to obtain asymptotic lower bounds on the error probabilities of interest.

Lemma 3 Let $\{\xi_i(n) : n \in \mathbb{N}\}$ $(i = 1, 2)$ be two stochastic processes on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$. The two processes can have arbitrary dependence structure. Suppose that there are positive constants $\mathcal{I}_1 \geq \mathcal{I}_2 > 0$ such that for every $\epsilon > 0$ we have

$$\sum_{n=1}^{\infty} \mathsf{P}\left(\left|\frac{1}{n}\xi_1(n) - \mathcal{I}_1\right| \geq \epsilon\right) < \infty, \quad \text{and} \quad \mathsf{P}\left(\lim_{n\to\infty} \frac{1}{n}\xi_2(n) = \mathcal{I}_2\right) = 1.$$

Then, for any random time T with $\mathsf{P}(T < \infty) = 1$ and any $q > 0$ we have

$$\lim_{b\to\infty} \mathsf{P}\left(\xi_1(T) \leq b,\ \xi_2(T+1) \geq (1+q)b\right) = 0.$$

Proof. Fix any $c \in (0, q)$, and let $n_b = \lceil b(1+c)/\mathcal{I}_2 \rceil$ be the smallest integer $\geq b(1+c)/\mathcal{I}_2$. Notice that $\mathsf{P}\left(\xi_1(T) \leq b,\ \xi_2(T+1) \geq (1+q)b\right)$ is upper bounded by $\mathrm{I}_b + \mathrm{II}_b$, where

$$\mathrm{I}_b := \mathsf{P}\left(\xi_1(T) \leq b,\ T \geq n_b\right), \quad \mathrm{II}_b := \mathsf{P}\left(\xi_2(T+1) \geq (1+q)b,\ T < n_b\right).$$

Thus it's sufficient to show as $b \to \infty$, $\mathrm{I}_b \to 0$ and $\mathrm{II}_b \to 0$. For the first term, we notice that for any $n \geq n_b$,

$$\frac{b}{n} \leq \frac{b}{n_b} \leq \frac{\mathcal{I}_2}{1+c} \leq \frac{\mathcal{I}_1}{1+c} < \mathcal{I}_1.$$

Let $\epsilon = \frac{c}{1+c}\mathcal{I}_1 > 0$, then

$$\begin{aligned} \mathrm{I}_b &= \sum_{n\geq n_b} \mathsf{P}\left(\xi_1(n) \leq b,\ T = n\right) \leq \sum_{n\geq n_b} \mathsf{P}\left(\frac{1}{n}\xi_1(n) \leq \frac{b}{n}\right) \leq \sum_{n\geq n_b} \mathsf{P}\left(\frac{1}{n}\xi_1(n) \leq \frac{\mathcal{I}_1}{1+c}\right) \\ &= \sum_{n\geq n_b} \mathsf{P}\left(\frac{1}{n}\xi_1(n) - \mathcal{I}_1 \leq -\epsilon\right) \leq \sum_{n\geq n_b} \mathsf{P}\left(\left|\frac{1}{n}\xi_1(n) - \mathcal{I}_1\right| \geq \epsilon\right). \end{aligned}$$

As $b \to \infty$, we have $n_b \to \infty$. Since by assumption ξ_1 convergences completely, we have $\mathrm{I}_b \to 0$.

For the second term, since $c \in (0, q)$, there exists $\epsilon' > 0$ such that for large enough b we have

$$\frac{(1+q)b}{n_b} = \frac{(1+q)b}{\lceil b(1+c)/\mathcal{I}_2 \rceil} \geq (1+\epsilon')\mathcal{I}_2.$$

As a result,

$$\mathrm{II}_b \leq \mathsf{P}\left(\max_{1 \leq n \leq n_b} \xi_2(n) \geq (1+q)b \right) \leq \mathsf{P}\left(\frac{1}{n_b} \max_{1 \leq n \leq n_b} \xi_2(n) \geq (1+\epsilon')\mathcal{I}_2 \right).$$

By Lemma A.1 of Fellouris and Tartakovsky (2016), we have as $n_b \to \infty$,

$$\mathsf{P}\left(\frac{1}{n_b} \max_{1 \leq n \leq n_b} \xi_2(n) \geq (1+\epsilon')\mathcal{I}_2 \right) \to 0,$$

which implies $\mathrm{II}_b \to 0$. Thus the proof is complete. □

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*, Volume 57. Springer Science & Business Media.

Bartroff, J., and J. Song. 2014. "Sequential Tests of Multiple Hypotheses Controlling Type I and II Familywise error rates". *Journal of Statistical Planning and Inference* 153:100–114.

Blanchet, J., and J. Liu. 2010. "Efficient Importance Sampling in Ruin Problems for Multidimensional Regularly Varying Random Walks". *Journal of Applied Probability* 47 (2): 301–322.

Chan, H. P., and T. L. Lai. 2005. "Importance Sampling for Generalized Likelihood Ratio Procedures in Sequential Analysis". *Sequential Analysis* 24 (3): 259–278.

Chan, H. P., and T. L. Lai. 2007. "Efficient Importance Sampling for Monte Carlo Evaluation of Exceedance Probabilities". *The Annals of Applied Probability* 17 (2): 440–473.

Collamore, J. 2002, 02. "Importance Sampling Techniques for the Multidimensional Ruin Problem for General Markov Additive Sequences of Random Vectors". *The Annals of Applied Probability* 12 (1): 382–421.

De, S. K., and M. Baron. 2012a. "Sequential Bonferroni Methods for Multiple Hypothesis Testing with Strong Control of Family-Wise Error Rates I and II". *Sequential Analysis* 31 (2): 238–262.

De, S. K., and M. Baron. 2012b. "Step-Up and Step-Down Methods for Testing Multiple Hypotheses in Sequential Experiments". *Journal of Statistical Planning and Inference* 142 (7): 2059–2070.

Erdos, P. 1949, 06. "On a Theorem of Hsu and Robbins". *The Annals of Mathematical Statistics* 20 (2): 286–291.

Fellouris, G., and A. G. Tartakovsky. 2016, January. "Multichannel Sequential Detection- Part I: Non-i.i.d. Data". *arXiv 1601.03379*.

Glasserman, P., and S. Juneja. 2008. "Uniformly Efficient Importance Sampling for the Tail Distribution of Sums of Random Variables". *Mathematics of Operations Research* 33 (1): 36–50.

Glasserman, P., and Y. Wang. 1997. "Counterexamples in Importance Sampling for Large Deviations Probabilities". *The Annals of Applied Probability* 7 (3): 731–746.

Hsu, P.-L., and H. Robbins. 1947. "Complete Convergence and the Law of Large Numbers". *Proceedings of the National Academy of Sciences* 33 (2): 25–31.

Siegmund, D. 1976. "Importance Sampling in the Monte Carlo Study of Sequential Tests". *The Annals of Statistics*:673–684.

Song, Y., and G. Fellouris. 2016. "Asymptotically Optimal, Sequential, Multiple Testing Procedures with Prior Information on the Number of Signals". *arXiv preprint arXiv:1603.02791*.

Tartakovsky, A., I. Nikiforov, and M. Basseville. 2014. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press.

Wald, A. 1945. "Sequential Tests of Statistical Hypotheses". *The Annals of Mathematical Statistics* 16 (2): 117–186.

AUTHOR BIOGRAPHIES

YANGLEI SONG is a graduate student in the Statistics Department of University of Illinois, Urbana-Champaign. He got his bachelor degree in Electrical Engineering in Tsinghua University, Beijing, China. His research interests include sequential analysis, multiple testing and applied probability. His e-mail address is ysong44@illinois.edu.

GEORGIOS FELLOURIS is an Assistant Professor in the Department of Statistics, and also affiliated with the Coordinated Science Lab, at the University of Illinois, Urbana–Champaign. He received the Ph.D. degree in Statistics from Columbia University in 2010, and the Diploma in Applied Mathematics in 2004 from the National Technical University of Athens, Greece. His research interests include sequential hypothesis testing, quickest change detection, decentralized decision making in sensor systems. His e-mail address is fellouri@illinois.edu.