The Annals of Statistics
2013, Vol. 41, No. 1, 125–142
DOI: 10.1214/12-AOS1076

AN ALGORITHM TO COMPUTE THE POWER OF MONTE CARLO TESTS WITH GUARANTEED PRECISION[1]

BY AXEL GANDY AND PATRICK RUBIN-DELANCHY

Imperial College London and University of Bristol

This article presents an algorithm that generates a conservative confidence interval of a specified length and coverage probability for the power of a Monte Carlo test (such as a bootstrap or permutation test). It is the first method that achieves this aim for almost any Monte Carlo test. Previous research has focused on obtaining as accurate a result as possible for a fixed computational effort, without providing a guaranteed precision in the above sense. The algorithm we propose does not have a fixed effort and runs until a confidence interval with a user-specified length and coverage probability can be constructed. We show that the expected effort required by the algorithm is finite in most cases of practical interest, including situations where the distribution of the p-value is absolutely continuous or discrete with finite support. The algorithm is implemented in the R-package *simctest*, available on CRAN.

1. Introduction.

Let p be a random variable taking values in $[0, 1]$ with unknown cumulative distribution function (CDF) F. For some $\alpha \in (0, 1)$, we want to approximate $\beta = F(\alpha)$ by Monte Carlo simulation. Assume that we cannot sample from F directly, but that it is possible to generate a collection of random variables $(X_j^i : i \in \mathbb{N}, j \in \mathbb{N})$, where $X_1^i, X_2^i, \ldots \sim \text{Bernoulli}(p_i)$ independently and $p_1, p_2, \ldots$ are unobserved independent copies of p, that is, $p_1, p_2, \ldots \sim F$ independently.

This problem comes about when computing the power or level of a Monte Carlo test, such as a bootstrap or permutation test, or in general a test that rejects on the basis of simulations under the (potentially estimated) null hypothesis. In this context, p is the (random) p-value, α the nominal level of the test and β its power. In this situation $X_1^i, X_2^i, \ldots$ are generated as follows: simulate a dataset (thus implicitly generating p_i), compute the observed test statistic and then, for $j = 1, 2, \ldots,$ use a sampling technique (such as bootstrapping or permutation) on the observed dataset to get a (re)sampled realization of the test statistic under the null hypothesis. Define X_j^i as the indicator that the (re)sampled test statistic is at least as extreme as the observed test statistic.

Received March 2012; revised August 2012.

[1] Supported by EPSRC Grant EP/H040102/1.

MSC2010 subject classifications. Primary 62-04, 62L12; secondary 62L15, 62F40.

Key words and phrases. Monte Carlo testing, significance test, power, algorithm.

A typical approach is to choose $N, M \in \mathbb{N} = \{1, 2, \dots\}$ and estimate β by

$$\hat{\beta}_{\text{naïve}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\left(\frac{1}{M} \sum_{j=1}^{M} X_j^i \right) \leq \alpha \right],$$

where $\mathbb{I}$ is the indicator function. A problem of this approach is that the bias of $\hat{\beta}_{\text{naïve}}$ is unknown. For example, using [1], equation (2), it can be shown that no matter how large N and M are chosen,

$$\sup_{\mathrm{P} \in \mathcal{P}} |\mathrm{E}\hat{\beta}_{\text{naïve}} - \beta| \geq 0.5,$$

where $\mathcal{P}$ is the set of all probability distributions on [0, 1]. Better bounds are available under the assumption that $\mathrm{E}\hat{\beta}_{\text{naïve}}$ is concave in α, see [3], Section 4.2.5. However, this would usually not be known in a given application.

More advanced estimation methods have been proposed. For instance, Oden [10] has investigated how to choose the relative sizes of N (controlling the variance) and M (controlling the bias), to minimize the total estimation error for certain distributions of p. [1] partially correct the bias by extrapolation.

However, existing procedures do not provide a formal, finite-sample guarantee on the accuracy of $\hat{\beta}$ for a general test. This is partly because the problem has always been approached with the principle of finding as accurate an estimate as possible for a fixed computational effort.

We approach the problem with the priorities reversed: we make an exact probabilistic statement about the result, allowing the computational effort to be random.

The algorithm that we propose is guaranteed to provide a conservative confidence interval (CI) for β of a given coverage probability. This interval will, after a finite expected number of samples, reach any desired length, provided that F is Hölder continuous in a neighborhood of α with exponent $\xi > 0$. This is satisfied if, for example, in a neighborhood of α, p is absolutely continuous with respect to Lebesgue measure with bounded density. In this case $\xi = 1$.

For practical use, the inner workings of the algorithm can be ignored. Users only need to provide the required precision (maximum CI length and minimum coverage probability) and a mechanism for generating the X_j^i. The algorithm is implemented in the R-package *simctest*, available on CRAN.

The article is structured as follows. In Section 2 we describe the basic algorithm. Theorem 2.1 demonstrates that, under very mild conditions, the algorithm terminates in finite expected time. Sections 3 and 4 present additional methodology to reduce the computational effort, some details of which are in supplementary material [6]. Section 5 contains a simulation study. In Section 6, we suggest an adaptive rule which ensures that the computational effort is only high if the estimate is in a region of interest. In Section 7 we demonstrate the use of our algorithm on a simple permutation test example. Proofs and auxiliary lemmas are in the Appendix. Within these, Lemma A.1 confirms an observation made in [5], main text page 1507 and Figure 4, about the distance between certain stopping boundaries.

2. The basic algorithm.

2.1. *Description.* We use the notation introduced in the first paragraph of the Introduction. For every $i \in \mathbb{N}$, we call the Bernoulli sequence $(X_j^i)_{j\in\mathbb{N}}$ a *stream*. The algorithm will use a fixed number N of these streams.

For each stream i, our algorithm aims to decide if $p_i \leq \alpha$ or if $p_i > \alpha$. We use the sequential algorithm of [5] for this purpose.

To simplify notation, we often drop the stream index i when referring to a generic stream; for example, we write X_j, p instead of X_j^i, p_i. Furthermore, we use a subscript to indicate the probability distribution of such a stream conditional on a specific value of p, that is, $\mathrm{P}_q(\cdot) = \mathrm{P}(\cdot|p = q)$ for some $q \in [0, 1]$.

The procedure in [5] defines two deterministic sequences, an upper boundary $(U_t : t \in \mathbb{N})$ and a lower boundary $(L_t : t \in \mathbb{N})$. While the partial sum $S_t = \sum_{j=1}^t X_j$ has hit neither boundary, the stream is *unresolved*. The procedure terminates at the hitting time

$$\tau = \inf\{t : S_t \geq U_t \text{ or } S_t \leq L_t\}.$$

If the upper boundary is hit, we decide $p > \alpha$ and report a *negative outcome* (p is not significant at level α). If the lower boundary is hit we decide $p \leq \alpha$ and report a *positive outcome* (p is significant at level α).

The boundaries are constructed to give a desired uniform bound $\varepsilon > 0$ on the probability of a wrong decision, that is,

$$\begin{aligned} &\mathrm{P}_p(S_\tau = U_\tau) \leq \varepsilon \qquad \text{for } p \leq \alpha, \\ &\mathrm{P}_p(S_\tau = L_\tau) \leq \varepsilon \qquad \text{for } p > \alpha. \end{aligned} \tag{2.1}$$

Figure 1 shows an example of U_t and L_t with $\varepsilon = 0.01$ and $\alpha = 0.05$.

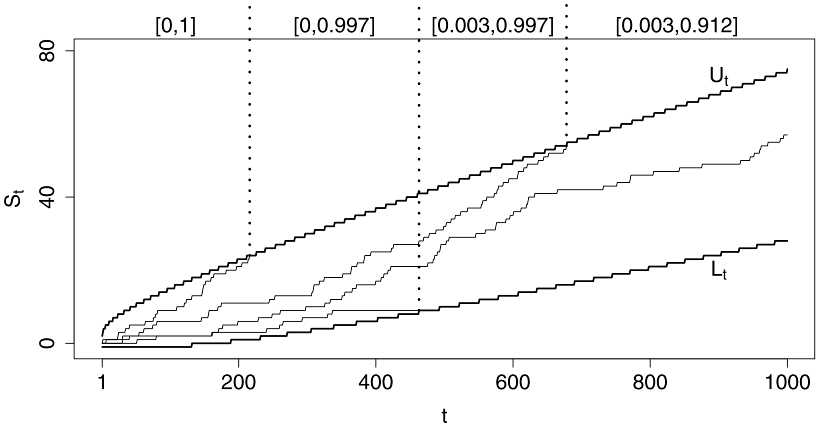

FIG. 1. *Confidence intervals generated by the algorithm using $N = 4$, $\varepsilon = 0.01$, $\alpha = 0.05$, $\varepsilon_t = \varepsilon t/(1000 + t)$ and $\gamma = 0.05$.*

To be more precise, the boundaries are constructed recursively using a *spending sequence* (ε_t) with $0 \le \varepsilon_t \nearrow \varepsilon$ as $t \to \infty$. The spending sequence governs how quickly the error probability ε is spent, guaranteeing

$$\begin{aligned} \mathrm{P}_p(S_\tau = U_\tau, \tau \le t) \le \varepsilon_t \qquad &\text{for } p \le \alpha, \\ \mathrm{P}_p(S_\tau = L_\tau, \tau \le t) \le \varepsilon_t \qquad &\text{for } p > \alpha. \end{aligned}$$

The precise recursive construction is given in (A.1), in the Appendix.

Our algorithm runs N streams in parallel until enough have been resolved to meet the required precision. More formally, it operates as follows:

ALGORITHM 1 (Basic algorithm).

```
Let $t = 0$; $R_0 = 0$; $A_0 = 0$; $\mathcal{U}_0 = \{1, \dots, N\}$, $S_0^1 = 0, \dots, S_0^N = 0$
while $|I(R_t, A_t, |\mathcal{U}_t|; \gamma)| > \Delta$
    Let $t = t + 1$, $R_t = R_{t-1}$, $A_t = A_{t-1}$, $\mathcal{U}_t = \mathcal{U}_{t-1}$
    for $i \in \mathcal{U}_t$
        Let $S_t^i = S_{t-1}^i + X_t^i$
        If $S_t^i \ge U_t$ let $A_t = A_t + 1$, $\mathcal{U}_t = \mathcal{U}_t \setminus \{i\}$
        If $S_t^i \le L_t$ let $R_t = R_t + 1$, $\mathcal{U}_t = \mathcal{U}_t \setminus \{i\}$
Report $I(R_t, A_t, |\mathcal{U}_t|; \gamma)$ as confidence interval for $\beta$.
```

$\mathcal{U}_t$ is a set containing the indices of unresolved streams at time t. $|\cdot|$ denotes the size of finite sets as well as the length of intervals. R_t and A_t count, respectively, the number of positive and negative outcomes.

$I(R_t, A_t, |\mathcal{U}_t|; \gamma)$ is a conservative confidence interval for β based on R_t, A_t and $|\mathcal{U}_t|$. It is constructed as follows. Because of (2.1), the probability that a stream has a positive outcome is in the interval $[(1-\varepsilon)\beta, (1-\varepsilon)\beta + \varepsilon]$. Therefore, if all streams were resolved, the following interval would be a conservative confidence interval for β with coverage probability $1 - \gamma$:

$$\mathcal{I}_\infty = \mathcal{I}_\infty(R_\infty, A_\infty; \gamma) = \left[\frac{\beta_-^* - \varepsilon}{1 - \varepsilon}, \frac{\beta_+^*}{1 - \varepsilon}\right],$$

where R_∞ (A_∞) denotes the number of positive (negative) outcomes and $[\beta_-^*, \beta_+^*]$ is the Clopper–Pearson confidence interval [2] with coverage probability $1 - \gamma$ for the success probability of a Binomial random variable observed to be R_∞ after $R_\infty + A_\infty$ trials.

The subscript in $\mathcal{I}_\infty$ represents that this is the interval that would be obtained by our algorithm if it were run until all streams were resolved.

To obtain a conservative confidence interval $\mathcal{I}_t$ while there are unresolved streams, we take the union of all confidence intervals that could be obtained after observing the outcomes of the remaining streams, that is, we let $\mathcal{I}_t =$

$I(R_t, A_t, |\mathcal{U}_t|; \gamma)$ where

$$I(r, a, u; \gamma) = \bigcup_{r_\infty = r}^{r+u} \mathcal{I}_\infty(r_\infty, r + a + u - r_\infty; \gamma). \tag{2.2}$$

By construction, $\mathcal{I}_1 \supseteq \mathcal{I}_2 \supseteq \cdots \supseteq \mathcal{I}_\infty$ and

$$\mathrm{P}[\beta \in \mathcal{I}_1 \cap \cdots \cap \beta \in \mathcal{I}_t \cap \cdots \cap \beta \in \mathcal{I}_\infty] \geq 1 - \gamma.$$

Figure 1 illustrates the algorithm in a toy example with only $N = 4$ streams. The thin lines depict the 4 corresponding partial sum sequences, S_t^i. When S_t^i hits one of the boundaries the stream is stopped, causing a retraction of the confidence interval for β (annotated at the top of the graph).

2.2. *Expected time.* A simpler algorithm than Algorithm 1 would be to run N streams until all are resolved. N can be chosen such that the CI length is at most Δ for all outcomes. However, this algorithm is unusable in practice as it typically requires an infinite expected effort. Indeed, from [5], page 1506, if the CDF F of p has a nonzero derivative at α, a very common case, then $\mathrm{E}[\tau_i] = \infty$, where τ_i denotes the hitting time of the ith stream. This makes the overall expected effort infinite.

We now show that with our algorithm we can choose N and (ε_t) such that the expected effort is finite. The key is to make N large enough that not all streams have to be resolved.

The effort of Algorithm 1, as measured by the number of X_t^i used, is

$$e = \sum_{i=1}^{N} \min\{\tau_i, \tau_{(N-k)}\}, \tag{2.3}$$

where k is the number of unresolved streams when the algorithm finishes and $\tau_{(1)} \leq \cdots \leq \tau_{(N)}$ denote the order statistics of $\tau_1, \dots, \tau_N$.

By choosing N large enough and ε small enough, we can ensure $k \geq \kappa$ for any given $\kappa \geq 1$. The effort is then bounded above by $\tau_{(N-\kappa)} N$. Thus to ensure that $\mathrm{E}[e]$ is finite, it suffices to prove $\mathrm{E}[\tau_{(N-\kappa)}] < \infty$ for some κ. The following theorem shows that in many cases κ can be taken as small as 2.

THEOREM 2.1. *Suppose that $\varepsilon \leq 1/4$ and there exist constants $\lambda > 0$, $q > 1$ and $T \in \mathbb{N}$ such that $\varepsilon_t - \varepsilon_{t-1} \geq \lambda t^{-q}$ for all $t \geq T$. Further, suppose that in a neighborhood of α the CDF F of p is Hölder continuous with exponent ξ. Then* $\mathrm{E}[\tau_{(i)}] < \infty$ *for $i \leq N - \lfloor 2/\xi \rfloor$. In particular, if $\xi = 1$ (the CDF is Lipschitz continuous in a neighborhood of α), then* $\mathrm{E}[\tau_{(N-2)}] < \infty$.

F is Hölder continuous with exponent ξ in a neighborhood of α if there exists an open interval U containing α for which there exists a $c > 0$ such that for all $x, y \in U$, $|F(x) - F(y)| \leq c|x - y|^\xi$.

The conditions on ε and (ε_t) are, for example, satisfied by $\varepsilon_t = \varepsilon t/(1000+t)$ and any $\varepsilon \le 1/4$ with $\lambda = 1$ and $q = 2$. This spending sequence (ε_t) is the default spending sequence in the R-package *simctest*.

The conditions on F are mild. For example, if F has a bounded density in a neighborhood of α, then $\xi = 1$. If the distribution of p is discrete and has finite support (e.g., in a permutation test), then $\xi = 1$ if $\mathrm{P}[p = \alpha] = 0$. Otherwise, it is in principle possible to find $\alpha' > \alpha$ such that

$$\beta = \mathrm{P}[p \le \alpha] = \mathrm{P}[p \le \alpha'], \qquad \mathrm{P}[p = \alpha'] = 0.$$

Applying the algorithm to α' instead of α, we again have $\xi = 1$.

Henceforward the conditions of Theorem 2.1 are assumed to be satisfied with $\xi = 1$. The algorithm will meet the user-specified precision requirements with a finite expected effort if it will terminate by time $\tau_{(N-2)}$ with probability one, or if $\mathrm{P}[|\mathcal{I}_{\tau_{(N-2)}}| > \Delta] = 0$. As can be verified, with $N-2$ of N streams resolved the largest possible CI length occurs when there are $\lfloor (N-2)/2 \rfloor$ positive outcomes. N must therefore satisfy $|I(\lfloor (N-2)/2\rfloor, \lceil (N-2)/2 \rceil, 2; \gamma)| \le \Delta$. We shall call the minimal such N the *blind minimal* N, $N_{\mathcal{B}}$.

3. Choosing the number of streams.

The computational effort of Algorithm 1 can be large; see Section 5. In this section we introduce two improvements concerning the choice of N: a *pilot sample* that can allow a smaller N than $N_{\mathcal{B}}$, $N_{\mathcal{P}}$, and an estimate of the optimal $N \ge N_{\mathcal{P}}$, using information from the pilot.

3.1. *Reducing the simple minimum N.*

Before running the main algorithm, we propose to first obtain a *pilot sample*, where n streams are run and stopped at a maximum number of steps $t_{\max}$, obtaining a preliminary confidence interval $\mathcal{I}_{\mathcal{P}} = I(R_{\mathcal{P}}, A_{\mathcal{P}}, |\mathcal{U}_{\mathcal{P}}|; \gamma_{\mathcal{P}})$, where I is defined in (2.2), $\gamma_{\mathcal{P}}$ is some pre-specified value (substantially) less than γ and $R_{\mathcal{P}}$, $A_{\mathcal{P}}$, $|\mathcal{U}_{\mathcal{P}}|$ are the number of positive outcomes, negative outcomes and unresolved streams.

In the main run the following interval can then be reported

$$\mathcal{I}_t^{(\mathcal{P})} = I(R_t, A_t, |\mathcal{U}_t|; \gamma - \gamma_{\mathcal{P}}) \cap \mathcal{I}_{\mathcal{P}}. \tag{3.1}$$

This respects the coverage probability $1-\gamma$, since a Bonferroni correction was used. We call the minimal N such that for all $r \in \{0, 1, \ldots, N-2\}$:

$$\bigl|I(r, N-2-r, 2; \gamma - \gamma_{\mathcal{P}}) \cap \mathcal{I}_{\mathcal{P}}\bigr| \le \Delta$$

the *pilot-based minimal* N denoted by $N_{\mathcal{P}}$. Given $\mathcal{I}_{\mathcal{P}}$ it can be determined by a computational search.

For $N \ge N_{\mathcal{P}}$ the confidence interval will always reach the desired length if at most 2 streams are unresolved. $N_{\mathcal{P}}$ can be much smaller than $N_{\mathcal{B}}$. Indeed, after $N-2$ of N streams in the main run are resolved, the maximum CI length achievable is for a number of positive outcomes r that satisfies $r/(N-2) \in \mathcal{I}_{\mathcal{P}}$. As

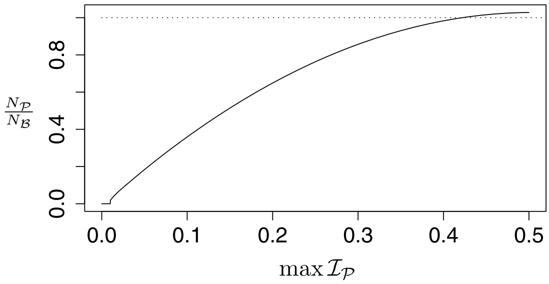

FIG. 2. *Ratio of the pilot-based minimum* N, $N_{\mathcal{P}}$, *over the blind version*, $N_{\mathcal{B}}$ *as a function of the rightmost point* $\max I_{\mathcal{P}}$ *of the pilot sample interval, with* $\Delta = 0.01$, $\varepsilon = 0.0001$, $\gamma = 0.01$, $\gamma_{\mathcal{P}} = \gamma/10$. *Here*, $N_{\mathcal{B}} = 68{,}311$.

demonstrated for pilot intervals $\mathcal{I}_{\mathcal{P}}$ to the left of 0.5 in Figure 2, the minimum number of streams needed in the main run can be reduced substantially, in particular, if $\mathcal{I}_{\mathcal{P}}$ lies far to the left (or right) of 0.5.

Heuristically, the disadvantage of a small increase in the coverage probability from $1 - \gamma$ to $1 - (\gamma - \gamma_{\mathcal{P}})$ can be outweighed by being able to exclude large intervals centered around 0.5.

3.2. *Approximation of the optimal number of streams.* In this section, we choose N within the range of possible Ns ($N \geq N_{\mathcal{P}}$) in order to minimize $\mathrm{E}(e)$, where e is defined in (2.3). We use a heuristic approach, which we only sketch briefly. Details can be found in the supplementary material [6].

From the pilot sample, we obtain an estimate of the probability of a stream stopping before $t_{\max}$, its expected stopping time under this event, and a preliminary estimate of β.

The expected stopping time of streams finishing after $t_{\max}$ is predicted on the basis of the approximation $\mathrm{P}[\tau_i > t | \tau_i > t_{\max}] \approx c\sqrt{\log(t)/t}$. This appears to be appropriate (for a large enough $t_{\max}$) when the p-value distribution is sufficiently "well behaved" around α.

Using these quantities we can approximate the expected effort for each N. The optimum N, denoted by $N_{\mathcal{O}}$, is found by searching over a sensible range $N_{\mathcal{P}} \leq N \leq N_{\max}$.

4. Stopping based on joint information.

We now describe a testing procedure that analyzes the current set of unresolved streams as a whole and allows the algorithm to stop with more unresolved streams. It reports a lower bound r_t (a_t) on the number of positive (negative) outcomes from the remaining streams if both of the following hypotheses are rejected,

$$H_0^+ : \left|\{i \in \mathcal{U}_t : p_i \leq \alpha\}\right| < r_t, \qquad H_0^- : \left|\{i \in \mathcal{U}_t : p_i > \alpha\}\right| < a_t,$$

where $r_t, a_t \geq 0$ and $r_t + a_t \leq |\mathcal{U}_t|$. The choice of r_t and a_t is discussed later.

The hypotheses will be rejected for large values of the test statistics,

$$T^+ = \sum_{i=r_t}^{|\mathcal{U}_t|} \mathbb{I}[G_t^\alpha(S_t^{(i)}) \le \eta], \qquad T^- = \sum_{i=1}^{|\mathcal{U}_t|-a_t+1} \mathbb{I}[G_t^\alpha(S_t^{(i)}) \ge 1-\eta],$$

where $S_t^{(1)} \le \cdots \le S_t^{(|\mathcal{U}_t|)}$ are the *ordered* partial sums corresponding to the unresolved streams, η is a chosen (small) positive value and

$$G_t^\alpha(x) = \mathrm{P}_\alpha[S_t \le x | \tau > t],$$

that is, G_t^α is the CDF of a cumulative sum of t Bernoulli variables with success probability α, conditional on not having hit either boundary by time t. This function is computed recursively.

The random variable X is said to be *stochastically smaller* than the random variable Y, denoted $X \le_{st} Y$, if $\mathrm{P}(X \le x) \ge \mathrm{P}(Y \le x)$ for all $x \in \mathbb{R}$.

THEOREM 4.1. *Under H_0^+, $T^+ \le_{st} B^+$ and under H_0^-, $T^- \le_{st} B^-$, where B^+ and B^- are Binomial variables with success probability η and size $|\mathcal{U}_t| - r_t + 1$ and $|\mathcal{U}_t| - a_t + 1$, respectively.*

H_0^+ and H_0^- can therefore be rejected *conservatively* when T^+ and T^- are significantly large for the corresponding Binomial variables.

Using Bonferroni correction, a minimum coverage probability of $1-\gamma$ is guaranteed if for all t we compute a confidence interval

$$\mathcal{I}_t^{\mathcal{J}} = I(\tilde{R}_t, \tilde{A}_t, |\tilde{\mathcal{U}_t}|; \gamma - \gamma_{\mathcal{P}} - \gamma_{\mathcal{J}}) \cap \mathcal{I}_{\mathcal{P}},$$

where $(\tilde{R}_t, \tilde{A}_t, |\tilde{\mathcal{U}_t}|) = (R_t + r_t, A_t + a_t, |\mathcal{U}_t| - r_t - a_t)$ if the test rejects, $(R_t, A_t, |\mathcal{U}_t|)$ otherwise, and $\gamma_{\mathcal{J}} < \gamma - \gamma_{\mathcal{P}}$ is an upper bound on the overall probability of wrongly rejecting either hypothesis at any point in time. To guarantee this bound, at each time t, each hypothesis is tested at level $\xi_t/2$, where $\xi_1, \xi_2, \ldots \ge 0$ are constants satisfying $\sum_{i=1}^\infty \xi_i = \gamma_{\mathcal{J}}$.

r_t and a_t are chosen such that $|\mathcal{I}_t^{\mathcal{J}}| \le \Delta$ if both tests reject, so that the algorithm can stop immediately if this occurs.

The procedure is mostly useful when the number of resolutions required, $r_t + a_t$, is small compared to the number of remaining streams $|\mathcal{U}_t|$. As an extreme example, suppose that $r_t = 1, a_t = 0$ and $|\mathcal{U}_t| = 100$. In this case, it can be possible to conclude with virtual certainty that *at least* 1 of the 100 streams has a p-value less than α, when concluding the same about any individual stream could require many more samples.

In this procedure there are a number of free parameters that we set somewhat heuristically. From a small simulation study we established that choosing $\eta = 0.05$ gave good results. As for r_t and a_t, they are chosen to be equal and then as small as possible subject to the algorithm terminating if the hypotheses can be rejected,

TABLE 1
Average effort (in millions) of our adaptive methods ("No test" and "With test") compared with the minimum N and the optimal N

	$\beta = \mathbf{0.05}$		$\beta = \mathbf{0.7}$		$\beta = \mathbf{0.9}$		$\beta = \mathbf{0.99}$	
	Av.	**(S.E.)**	**Av.**	**(S.E.)**	**Av.**	**(S.E.)**	**Av.**	**(S.E.)**
Optimal N	12.3	(0.14)	3329	(35)	539	(8.4)	16.2	(0.08)
Min. N	12.5	(0.16)	8498	(296)	548	(9.2)	16.1	(0.08)
No test	10.5	(0.22)	3324	(41)	568	(7.9)	10.4	(0.10)
With test	8.0	(0.19)	1541	(18)	317	(5.2)	10.4	(0.09)

since for simple p-value distributions it is likely that the unresolved p-values would be roughly evenly distributed around α.

In the simulation studies that follow and in the R-implementation, $\gamma_{\mathcal{J}} = \gamma/10$, ξ_t is only positive when $t = t_i = 2i \times 10^5$ for $i \in \mathbb{N}$ and $\sum_1^{t_i} \xi_t = \gamma_{\mathcal{J}} \times 20/(20+i)$.

5. Simulations. This simulation study illustrates the effort required by our algorithm and the effect of the improvements in Sections 3 and 4. For all experiments we set $\alpha = 0.05$, $\Delta = 0.02$, $1-\gamma = 0.99$, $\varepsilon = 0.0001$ and $\varepsilon_t = \varepsilon 1000/(1000+t)$. Four p-value distributions were considered, Beta$(1, x)$ with x such that $\mathrm{P}[p \le \alpha] = \alpha, 0.7, 0.9, 0.99$, that is, $x = 1$ (a uniform distribution) and roughly $x = 23.5$, $x = 44.9$ and $x = 89.8$, respectively. As before, the effort is measured by the total number of samples generated.

Table 1 shows the average effort based on 100 replicated runs in the left subcolumns. In the right subcolumns we report the estimated standard error of the corresponding estimate, that is, the standard deviation of the sample divided by $\sqrt{100}$.

The first two rows report the average effort for the optimal N (which would not be available in practice) and the minimum N, $N_{\mathcal{B}}$, when using Algorithm 1 without any of the improvements suggested in Sections 3 and 4. These were computed by resampling from 10^6 pre-simulated replicates of the tuple (stopping-time, outcome), for each distribution, from which we emulated the operation of the algorithm. (Finding the optimal N would otherwise have taken too much time.)

The third row illustrates the improvements of Section 3, which concern the choice of N, setting $\gamma_{\mathcal{P}} = 0.1\gamma$. In the fourth row we additionally implemented the test on joint information, described in Section 4, with $\gamma_{\mathcal{J}} = 0.1\gamma$. In both of these rows each value represents the average effort observed from actually running the algorithm 100 times. Each run used its own pilot sample consisting of 1000 streams forced to terminate after 1000 steps. The effort of the pilot is included in the average effort.

First consider the difference between the third and fourth rows of Table 1. The testing procedure can reduce the effort substantially, namely by 24%, 54%, 44% in the first three cases, although in the last case the reduction is not significant.

For the Uniform and Beta distribution with power 0.99, the optimal N and $N_{\mathcal{B}}$ turn out to be equal. Hence, the reduction of the effort seen in the third row over the first two rows is mostly due to the intersection method described in Section 3.1, which has allowed a smaller choice of N, $N_{\mathcal{P}}$.

For the Beta distribution with power 70%, the effort for the minimal N, in the second row, is over 2.5 times larger than for the optimal N, in the first row. As result, in this example it was crucial to estimate this optimum, by the procedure described in Section 3.2. The difference between the effort for the optimal N (unknown in practice) and the adaptively chosen $N_{\mathcal{O}}$ is not significant (although in this example enough simulations would show that the optimal N still performed better). As previously mentioned, introducing the testing procedure in this example further reduces the effort by a considerable margin, as demonstrated in the fourth row. It is of some comfort that the best improvements from the methodology of Sections 3 and 4 were found in the computationally most demanding scenario.

In the third row, for the Beta distribution with power 90%, adaptively choosing N actually increased the effort, although not substantially. The average $N_{\mathcal{O}}$ chosen is roughly 10,000, whereas $N_{\mathcal{B}}$ in the second row is 17,055 (for this distribution it is also the optimal N). We would expect to reduce the effort on this basis. However, this does not appear to completely compensate for the effort of the pilot and the error in coverage probability lost in computing the pilot-based CI. However, with the test we reduce the effort by 40% and improve on both efforts reported in the first two rows for this distribution.

Overall, from these experiments it seems that our suggested improvements reduce the expected effort substantially, as is best summarized in the difference between the bottom row and either of the first two.

For future reference, the default settings of our algorithm are those of the bottom row, namely: $\varepsilon = \Delta/200$, $\varepsilon_t = \varepsilon 1000/(1000+t)$, $\gamma_{\mathcal{P}} = \gamma_{\mathcal{J}} = 0.1\gamma$ and a pilot sample of 1000 streams terminated at $t_{\max} = 1000$.

6. Adaptive CI length.

When one resampling step is computationally demanding, the expected efforts listed in Table 1 may appear prohibitive. In this case, we recommend relaxing the fixed requirements on Δ, that is, allowing Δ to depend on the "location" of the confidence interval. This can reduce the expected effort of the algorithm substantially.

As a rule of thumb, the closer the power is to 0.5 the higher the expected effort (compare, e.g., the efforts for $\beta = 0.05$ and $\beta = 0.7$ in Table 1): first, because the p-value distribution tends to have more mass around α, meaning that each stream in the algorithm has a higher expected running-time, and second because the length of the confidence interval is largest when there are the same number of positive and negative outcomes.

On the other hand, we anticipate that if the power is around 0.5 or for that matter anywhere in the interval $[0.1, 0.9]$, say, the user will often only require a small enough confidence interval to conclude that β is not close α or 1. Indeed, a typical reason why one needs the power of a test is to check that the probability of rejection under the null hypothesis is close to α (which is typically small) or that under an alternative hypothesis β is close to 1.

Let $C = \{\beta \in [0, 1]^2 : \beta_1 \leq \beta_2\}$ denote the set of all possible confidence intervals for β. We allow the analyst to pre-specify a subset of C, A, say, such that if the current confidence interval is an element of A the algorithm terminates immediately.

It is reasonable to enforce that A satisfy the following three properties:

(i) A is closed.
(ii) $\{(\beta, \beta)^T : \beta \in [0, 1]\} \subseteq A$ (CIs of length 0 are allowed).
(iii) $\forall \beta \in A : \forall \alpha \in C : \beta_1 \leq \alpha_1 \leq \alpha_2 \leq \beta_2 \Rightarrow \alpha \in A$ (a subinterval of an allowed CI is allowed).

The next result shows that specifying A is equivalent to specifying the maximum CI length allowed as a function of the CI's midpoint.

LEMMA 6.1. *If $A \subseteq C$ satisfies* (i)–(iii), *then there exists a function $\Delta : [0, 1] \to [0, 1]$ such that for all $\beta \in C : \beta \in A \Leftrightarrow \beta_2 - \beta_1 \leq \Delta(\frac{\beta_1+\beta_2}{2})$.*

All of the theory we have presented in Sections 2–4 can be incorporated unaltered into an algorithm with adaptive Δ, with the single exception that finding $N_{\mathcal{P}}$ requires a brute-force search—one must ensure that $\Delta(M)$ will be met after $N - 2$ streams have stopped, for any possible CI midpoint M arising from all the possible outcomes of $N - 2$ streams.

The effort of our recommended method for fixed Δ is repeated from the fourth row of Table 1 to the first row of Table 2. These results are equivalent to a case where for all $M \in [0, 1]$, $\Delta(M) = 0.02 = \Delta_0(M)$. In the next rows of Table 2 we present the average effort of the algorithm for three other functions of the midpoint, all of which are illustrated in Figure 3. Depending on what is easiest to present, the rule is described through Δ or by the equivalent A.

TABLE 2
Average effort (in millions) for different functions of the CI midpoint

	$\beta = 0.05$		$\beta = 0.7$		$\beta = 0.9$		$\beta = 0.99$	
Function	**Av.**	**(S.E.)**	**Av.**	**(S.E.)**	**Av.**	**(S.E.)**	**Av.**	**(S.E.)**
Δ_0	8.0	(0.19)	1541	(18)	317	(5.2)	10.4	(0.09)
Δ_1	7.8	(0.20)	185	(3.2)	131	(2.3)	26.2	(0.77)
Δ_2	8.4	(0.46)	17.1	(0.46)	9.0	(0.06)	5.5	(0.08)
Δ_3	8.4	(0.46)	0.7	(<0.01)	0.6	(<0.01)	0.5	(<0.01)

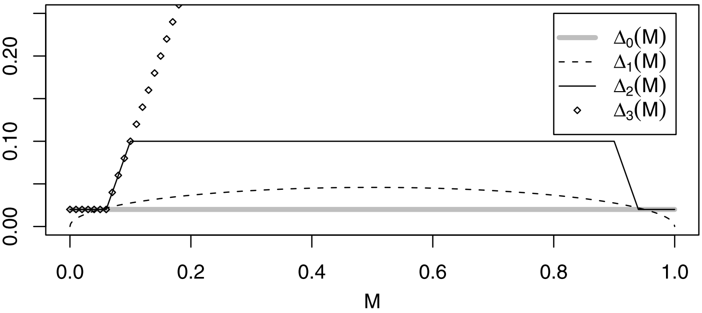

FIG. 3. *The four midpoint functions Δ_i used in Table 2.*

(1) $\Delta_1(M) = 0.02\sqrt{M(1-M)}/(\sqrt{0.05 \cdot 0.95})$. A function that allows roughly the same number of streams to remain unresolved for any β. Because the CI midpoint cannot be 0 or 1 exactly the fact that $\Delta(0) = \Delta(1) = 0$ is not problematic.

(2) A_2 is the largest set of confidence intervals that satisfies (i)–(iii) and that satisfies $\forall\beta \in A_2 : \beta_2 - \beta_1 \leq 0.1$ and $\forall\beta \in A_2$ with ($\beta_1 \leq 0.05$ or $\beta_2 \geq 0.95$): $\beta_2 - \beta_1 \leq 0.02$—a CI length of 0.02 is needed for high or low powers, but a CI length of 0.1 is admissible otherwise.

(3) A_3 is the largest set of confidence intervals that satisfies (i)–(iii) and that satisfies $\forall\beta \in A_3$ with $\beta_1 \leq 0.05$: $\beta_2 - \beta_1 \leq 0.02$. A precise estimate is only required if the confidence interval is at least partly to the left of α and any interval is admissible otherwise.

For the Uniform distribution, since all rules have $\Delta(0.05) = 0.02$, we would expect the effort to be comparable, as is observed. On the other hand, we see a dramatic reduction of the effort in other columns where the rule has allowed less precision. Overall, if we consider for example the effort for Δ_2, we hope that with this compromise the algorithm can be used in practice for moderately complicated tests.

7. Example: Permutation test.

Using exactly the example of [1], we computed the power of a permutation test on the difference of the means of two Gaussian samples, with sizes $K = 4$ and $L = 8$, identical standard deviation σ and standardized differences $(\mu_{\mathcal{G}} - \mu_{\mathcal{C}})/\sigma = 0.5, 1, 1.5$ and 2. We used a fixed $\Delta = 0.01$ and coverage probability 0.99. Our other parameters were set to the defaults listed at the end of Section 5.

The results are presented in Table 3. In three of the four cases our confidence interval excludes the corresponding estimate in [1] (although not after adding or subtracting two of their standard errors). Of course, our computational effort is considerably larger—but our key contribution is in providing a mechanism that *guarantees* the precision of the result.

In this simple example it is in fact possible to compute the p-value of each dataset exactly by evaluating all 495 permutations. Because of this the power can

TABLE 3
Power of the permutation test for the difference of means

$\boldsymbol{\Delta/\sigma}$	**0.5**	**1.0**	**1.5**	**2.0**
Truth	${}_{0.183}0.184_{0.185}$	${}_{0.441}0.442_{0.443}$	${}_{0.728}0.729_{0.730}$	${}_{0.912}0.912_{0.913}$
Our method	${}_{0.182}0.185_{0.192}$	${}_{0.440}0.443_{0.450}$	${}_{0.726}0.729_{0.736}$	${}_{0.910}0.914_{0.920}$
Boos and Zhang	0.175 (0.006)	0.439 (0.008)	0.731 (0.007)	0.921 (0.005)

be estimated by standard methodology with a Binomial-based confidence interval. In each case, a very accurate estimate of β was obtained by generating 10^6 datasets and computing the p-value for each exactly. The resulting estimates are presented in the first row of Table 3, using the convention ${}_{a}x_{b}$ to mean that the estimate is x, and the confidence interval is $[a, b]$. In the second row we present the results of our algorithm, using a fixed $\Delta = 0.01$ and coverage probability 0.99. In all cases, the "true" power falls within our estimated confidence interval, as would be expected. For the convenience of the reader, the third row presents the estimated powers and standard errors computed in [1].

8. Conclusions. We have proposed an open-ended algorithm to compute a conservative confidence interval for β, (almost) without any assumptions on the distribution of the p-value (Theorem 2.1). In practice, the method can be computationally expensive. However, various improvements (Sections 3 and 4) reduce the computational effort for fixed Δ by a sizeable margin. An adaptive Δ (Section 6) can ensure that the effort is only large if the estimated power is in a region where a high precision is required.

There remain areas of potential improvement: for instance the balance between the error spent on ε, the pilot and the testing procedure could be explored in more depth, as well as the choice of the spending sequences ε_t and ξ_t. The test for stopping based on joint information in Section 4 is somewhat ad-hoc, and conceivably a more powerful test could be derived. Finally, of course, the computational effort could also potentially be reduced by making additional assumptions on the p-value distribution.

How conservative is the confidence interval? From a few simple experiments, we found the length to be roughly twice as large as it needs to be for the nominal coverage probability. Although we have been conservative in many aspects of the algorithm, this disparity appears to be almost entirely due to the contribution from unresolved streams in (2.2). This is effectively the price of making almost no assumptions on the distribution of the p-values.

APPENDIX A: FINITE EXPECTED STOPPING TIME

The proof of Theorem 2.1 requires preliminary lemmas and the following recursive definition of the stopping boundaries from [5]:

$$
\begin{aligned}
U_t &= \min\{j \in \mathbb{N} : \mathrm{P}_\alpha(\tau \ge t, S_t \ge j) + \mathrm{P}_\alpha(\tau < t, S_\tau \ge U_\tau) \le \varepsilon_t\},\\
L_t &= \max\{j \in \mathbb{Z} : \mathrm{P}_\alpha(\tau \ge t, S_t \le j) + \mathrm{P}_\alpha(\tau < t, S_\tau \le L_\tau) \le \varepsilon_t\}.
\end{aligned}
\tag{A.1}
$$

LEMMA A.1. *If there exist constants $\lambda > 0$, $q > 0$ and $T \in \mathbb{N}$ such that $\varepsilon_t - \varepsilon_{t-1} \ge \lambda t^{-q}$ for all $t \ge T$, then, for all $t \ge T$,*

$$U_t \le \lceil t\alpha + \sqrt{t(q\log t - \log\lambda)/2}\rceil, \qquad L_t \ge \lfloor t\alpha - \sqrt{t(q\log t - \log\lambda)/2}\rfloor.$$

The square root is well defined since $1 \ge \varepsilon_t - \varepsilon_{t-1} \ge \lambda t^{-q}$.

PROOF. We will show $\mathrm{P}_\alpha(\tau \ge t, S_t \ge U_t^*) + \mathrm{P}_\alpha(\tau < t, S_\tau \ge U_\tau) \le \varepsilon_t$ for $t \ge T$. By (A.1) this implies $U_t \le \lceil t\alpha + \sqrt{t(q\log t - \log\lambda)/2}\rceil =: U_t^*$.

First, (A.1) implies

$$
\begin{aligned}
\mathrm{P}_\alpha(\tau < t, S_\tau \ge U_\tau) &= \mathrm{P}_\alpha(\tau \ge t-1, S_{t-1} \ge U_{t-1}) + \mathrm{P}_\alpha(\tau < t-1, S_\tau \ge U_\tau)\\
&\le \varepsilon_{t-1}.
\end{aligned}
$$

Furthermore, by Hoeffding's inequality [7],

$$
\begin{aligned}
\mathrm{P}_\alpha(\tau \ge t, S_t \ge U_t^*) &\le \mathrm{P}_\alpha(S_t \ge U_t^*) = \mathrm{P}_\alpha(S_t/t - \alpha \ge U_t^*/t - \alpha)\\
&\le \exp\{-2t(U_t^*/t - \alpha)^2\} \le \lambda t^{-q} \le \varepsilon_t - \varepsilon_{t-1},
\end{aligned}
$$

finishing the proof of $U_t \le U_t^*$. The bound for L_t can be shown similarly. □

The above formally confirms the observation in [5], main text, page 1507 and Figure 4, that $U_t - L_t$ appears to be proportional to $\sqrt{t\log t}$ for large t. Indeed, the spending sequence used, $\varepsilon_t = \varepsilon t/(1000 + t)$, satisfies the conditions of the lemma with $\lambda = 1$ and $q = 2$ (if one chooses $\varepsilon \le 1/4$).

LEMMA A.2. *Suppose that F is Hölder continuous with exponent ξ in a neighborhood of α, that the conditions of Lemma A.1 hold, and that $\varepsilon \le 1/4$. Then, for any $\eta \in (0, 1)$, there exist constants κ and $\tilde{T}$ such that*

$$\mathrm{P}(\tau > t) \le 2\mathrm{e}^{-2t^\eta} + \kappa t^{\xi(\eta-1)/2}, \qquad t \ge \tilde{T}.$$

Hence, $\mathrm{P}(\tau > t) = o(t^d)$ for any $d > -\xi/2$.

PROOF. Let F be the CDF of p. Then, for any $t \in \mathbb{N}$,

$$\mathrm{P}(\tau > t) = I\{[0, p_t^-]\} + I\{(p_t^-, p_t^+)\} + I\{[p_t^+, 1]\},$$

where $I\{A\} = \int_A \mathrm{P}_p(\tau > t)\,\mathrm{d}F(p)$ and $0 \le p_t^- < \alpha < p_t^+ \le 1$. When $0 \le p \le p_t^-$ and $L_t/t - p_t^- > 0$,

$$\mathrm{P}_p(\tau > t) \le \mathrm{P}_p(S_t > L_t) \le \mathrm{P}_{p_t^-}(S_t > L_t) \le \exp\{-2t(L_t/t - p_t^-)^2\},$$

using Hoeffding's inequality for the rightmost bound. Hence, letting

$$p_t^- = \max\{L_t/t - t^{(\eta-1)/2}, 0\}, \qquad t \in \mathbb{N}$$

for some $\eta \in \mathbb{R}$, we get

$$\mathrm{P}_p(\tau > t) \le \exp\{-2t^\eta\}, \qquad 0 \le p \le p_t^-, t \in \mathbb{N}.$$

Do we have $0 \le p_t^- < \alpha$? The lower bound is obvious. The upper bound also holds, since the proof of Theorem 2 in [5] shows that if $\varepsilon \le 1/4$, then $L_t/t < \alpha$ for all $t \in \mathbb{N}$.

Similarly we can define $p_t^+ = \min\{U_t/t + t^{(\eta-1)/2}, 1\}$, $t \in \mathbb{N}$, guaranteeing that $\alpha < p_t^+ \le 1$. Then, for any $\eta \in \mathbb{R}$,

$$\mathrm{P}_p(\tau > t) \le \exp(-2t^\eta), \qquad p_t^+ \le p \le 1, t \in \mathbb{N}.$$

We therefore have

$$I\{[0, p_t^-]\} + I\{[p_t^+, 1]\} \le 2\exp(-2t^\eta). \tag{A.2}$$

It remains for us to obtain a bound on $I\{(p_t^-, p_t^+)\}$. Using Theorem 1 in [5], $U_t - \alpha t = o(t)$, $\alpha t - L_t = o(t)$. Thus, by restricting $\eta < 1$, $p_t^- \to \alpha$, $p_t^+ \to \alpha$ and there exists a time T^* such that F is Hölder continuous over (p_t^-, p_t^+) for all $t \ge T^*$. It follows that for some constant $h > 0$,

$$I\{(p_t^-, p_t^+)\} \le \int_{(p_t^-, p_t^+)} \mathrm{d}F(p) \le F(p_t^+) - F(p_t^-) \le h(p_t^+ - p_t^-)^\xi, \qquad t \ge T^*.$$

Let $\tilde{T} = \max\{T, T^*, 2\}$, where T is defined in Lemma A.1. For $t \ge \tilde{T}$,

$$\begin{aligned} I\{(p_t^-, p_t^+)\} &\le h(p_t^+ - p_t^-)^\xi \\ &\le h\big[2t^{(\eta-1)/2} + 2\big[\sqrt{t(q\log t - \log\lambda)/2} + 1\big]/t\big]^\xi \\ &\le h\big[2t^{(\eta-1)/2} + 2\big[\sqrt{(q+a)/2}\sqrt{t\log t} + 1\big]/t\big]^\xi \\ &\le h\big[2t^{(\eta-1)/2} + b\sqrt{\log t/t}\big]^\xi \\ &\le h\big[(2+c)t^{(\eta-1)/2}\big]^\xi \qquad \text{(requiring } \eta > 0\text{)}, \end{aligned}$$

where $a = \max\{0, -\log\lambda/\log\tilde{T}\}$, $b = 2(\sqrt{(q+a)/2} + 1)$, $c = b\sqrt{\log t/t}|_{t=\tilde{T}}$. We needed $\tilde{T} \ge 2$ in the definition of a and used it in the third inequality ($1 < \sqrt{2\log 2}$). Using (A.2), the proof is complete after we take $\kappa = h(2+c)^\xi$. □

PROOF OF THEOREM 2.1. Using standard results for order statistics [4],

$$\mathrm{P}(\tau_{(N-k)} > t) = \sum_{j=0}^{N-k-1} \binom{N}{j} \mathrm{P}(\tau > t)^{N-j} \mathrm{P}(\tau \le t)^j \le c_1 \mathrm{P}(\tau > t)^{k+1}$$

for $t \ge 0$ and some $c_1 > 0$. Therefore, using Lemma A.2

$$\mathrm{E}(\tau_{(N-q)}) = \sum_{t=0}^{\infty} \mathrm{P}(\tau_{(N-k)} > t) \le 1 + \sum_{t=1}^{\infty} c_1 \mathrm{P}(\tau > t)^{k+1} \le 1 + \sum_{t=1}^{\infty} c_2 t^{(k+1)d}$$

for all $d > -\xi/2$, with c_2 chosen based on c_1 and d. The summation in the right-hand side is finite if the exponent of t is strictly smaller than -1. $\lfloor 2/\xi \rfloor$ is the smallest possibility for $k \in \mathbb{N}$ such that there exists a $d > -\xi/2$ with $(k+1)d < -1$. □

APPENDIX B: HYPOTHESIS TEST

The proof of Theorem 4.1 first requires the following lemma.

LEMMA B.1. *Suppose that X_j^1 and X_j^2 are two sequences of independent Bernoulli variables with success probabilities π_1 and π_2, respectively, where $0 \le \pi_1 \le \pi_2 \le 1$, and put $S_t^k = \sum_{j=1}^t X_t^k$ for $k = 1, 2$. Let $\{l_t : t \in \mathbb{N}\}$ and $\{u_t : t \in \mathbb{N}\}$ be two arbitrary integer sequences, and let*

$$\tau_k = \begin{cases} \infty, & \text{if } l_t < S_t^k < u_t \text{ for all } t \in \mathbb{N}, \\ \min\{j : S_j^k \le l_j \text{ or } S_j^k \ge u_j\}, & \text{otherwise}. \end{cases}$$

Then if $\mathrm{P}[\tau_k > t] > 0$ *for* $k = 1, 2$,

$$[S_t^1 | \tau_1 > t] \le_{st} [S_t^2 | \tau_2 > t].$$

PROOF. We will require a stronger form of stochastic ordering: for two discrete RVs X and Y, X is smaller than Y with respect to the likelihood ratio order, denoted $X \le_{lr} Y$, if

$$\frac{f_X(x)}{f_Y(x)} \downarrow x \qquad \text{on the support set of } Y, \tag{B.1}$$

where f_X and f_Y are the probability mass functions (PMFs) of X and Y [9], page 184. Further, a discrete RV Z has a log-concave distribution if [8]

$$f_Z(x)^2 \ge f_Z(x-1) f_Z(x+1), \qquad x \in \mathbb{N}. \tag{B.2}$$

$[S_1^1 | \tau_1 > 1]$ and $[S_1^2 | \tau_2 > 1]$ have log-concave distributions and $[S_1^1 | \tau_1 > 1] = X_1^1 \le_{lr} X_1^2 = [S_1^2 | \tau_1 > 1]$. Suppose the same holds true for $[S_t^1 | \tau_1 > t]$ and

$[S_t^2|\tau_2 > t]$. For $k = 1, 2$, $[S_{t+1}^k|\tau_k > t] = [S_t^k|\tau_k > t] + X_{t+1}^k$ is a convolution of two random variables with log-concave distributions, implying that it has itself a log-concave distribution [8], Lemma page 387. Using [8], Theorem 2.1(d)

$$\begin{aligned}[S_{t+1}^1|\tau_1 > t] &= [S_t^1|\tau_1 > t] + X_{t+1}^1 \leq_{lr} [S_t^2|\tau_2 > t] + X_{t+1}^1 \\ &\leq_{lr} [S_t^2|\tau_2 > t] + X_{t+1}^2 = [S_{t+1}^2|\tau_2 > t],\end{aligned}$$

using the properties assumed to be true at t and the log-concavity, likelihood ratio ordering and independence of X_{t+1}^1 and X_{t+1}^2.

For $k = 1, 2$, conditioning on $\tau_k > t + 1$ restricts the support of $[S_{t+1}^1|\tau_1 > t]$ and $[S_{t+1}^2|\tau_2 > t]$ to a same, smaller set, and (where supported) the new PMF is the old multiplied by a constant c_k. Therefore, directly from (B.1) and (B.2), we conclude that $[S_{t+1}^1|\tau_1 > t + 1] \leq_{lr} [S_{t+1}^2|\tau_2 > t + 1]$, and both distributions are log-concave.

By induction, these properties are true for all t. Likelihood ratio order implies the usual stochastic order [9], completing the proof. □

PROOF OF THEOREM 4.1. Let $n_t = |\mathcal{U}_t|$. T^+ can be bounded above by

$$T^+ \leq \sum_{i=r_t}^{n_t} \mathbb{I}[G_t^\alpha(\tilde{S}_t^{(i)}) \leq \eta] = \tilde{T}^+,$$

where $\{\tilde{S}_t^{(i)} : i = r_t, \dots, n_t\}$ are the partial sums corresponding to $p_{(r_t)} \leq p_{(r_t+1)} \leq \cdots \leq p_{(n_t)}$, the largest ordered p-values of the unresolved streams.

Under H_0^+, $p_{(i)} > \alpha$ for $i = r_t, \dots, n_t$. Let S_t^α be a partial sum generated by a p-value equal to α and let τ_α denote its stopping-time. By Lemma B.1,

$$[S_t^\alpha|\tau_\alpha > t] \leq_{st} [\tilde{S}_t^{(i)}|\tilde{\tau}_{(i)} > t],$$

where $\tilde{\tau}_{(i)}$ is the stopping time of $\tilde{S}_t^{(i)}$. Therefore, conditional on $\tau_\alpha, \tilde{\tau}_{(i)} > t$,

$$\mathbb{I}[G_t^\alpha(\tilde{S}_t^{(i)}) \leq \eta] \leq_{st} \mathbb{I}[G_t^\alpha(S_t^\alpha) \leq \eta] \leq_{st} X,$$

where X is a Bernoulli variable with success probability η. It follows that

$$\sum_{i=r_t}^{n_t} \mathbb{I}[G_t^\alpha(\tilde{S}_t^{(i)}) \leq \eta] \leq_{st} B^+,$$

where B^+ is a Binomial variable with success probability η and size $n_t - r_t + 1$. Therefore, $T^+ \leq \tilde{T}^+ \leq_{st} B^+$. The bound for T^- can be shown similarly. □

APPENDIX C: ON THE MIDPOINT RULE

PROOF OF LEMMA 6.1. Let $t \in [0,1]$ and define $\Delta(t) = \sup\{\beta_2 - \beta_1 : \frac{\beta_1+\beta_2}{2} = t, \beta \in A\}$. This is well defined because of (ii). The implication from left to right follows by the definition of Δ.

Let $\beta \in C : \beta_2 - \beta_1 \leq \Delta(\frac{\beta_1+\beta_2}{2})$. Let $t = \frac{\beta_1+\beta_2}{2}$. As A is compact and $D = \{\xi \in \mathbb{R}^2 : \xi_1 + \xi_2 = 2t\}$ is closed, $A \cap D$ is compact and thus $\{\beta_2 - \beta_1 : \frac{\beta_1+\beta_2}{2} = t, \beta \in A\}$ is compact also.

Hence, there exists a $\gamma \in A$ such that $(\gamma_2 + \gamma_1)/2 = t$ and $\gamma_2 - \gamma_1 = \Delta(t)$. This implies that $\beta \subseteq \gamma$ using (iii), implying that $\beta \in A$. □

SUPPLEMENTARY MATERIAL

Approximation of the optimal number of streams (DOI: 10.1214/12-AOS1076SUPP; .pdf). We describe a method that uses information from the pilot sample to approximate the expected effort of the algorithm as a function of the number N of streams. This method is used to choose N. Its performance is illustrated in a simulated experiment.

REFERENCES

[1] BOOS, D. and ZHANG, J. (2000). Monte Carlo evaluation of resampling-based hypothesis tests. *J. Amer. Statist. Assoc.* **95** 486–492.

[2] CLOPPER, C. and PEARSON, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26** 404–413.

[3] DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673

[4] EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events: For Insurance and Finance. Applications of Mathematics* (*New York*) **33**. Springer, Berlin. MR1458613

[5] GANDY, A. (2009). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *J. Amer. Statist. Assoc.* **104** 1504–1511. MR2750575

[6] GANDY, A. and RUBIN-DELANCHY, P. (2013). Supplement to "An algorithm to compute the power of Monte Carlo tests with guaranteed precision." DOI:10.1214/12-AOS1076SUPP.

[7] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[8] KEILSON, J. and GEBER, H. (1971). Some results for discrete unimodality. *J. Amer. Statist. Assoc.* **66** 386–389.

[9] KEILSON, J. and SUMITA, U. (1982). Uniform stochastic ordering and related inequalities. *Canad. J. Statist.* **10** 181–198. MR0691387

[10] ODEN, N. L. (1991). Allocation of effort in Monte Carlo simulation for power of permutation tests. *J. Amer. Statist. Assoc.* **86** 1074–1076.

DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE LONDON
SOUTH KENSINGTON CAMPUS
LONDON SW7 2AZ
UNITED KINGDOM
E-MAIL: a.gandy@imperial.ac.uk

SCHOOL OF MATHEMATICS
UNIVERSITY OF BRISTOL
UNIVERSITY WALK
BRISTOL BS8 1TW
UNITED KINGDOM
E-MAIL: patrick.rubin-delanchy@bristol.ac.uk