Adaptive Monte Carlo Multiple Testing via Multi-Armed Bandits

Martin J. Zhang [1] **James Zou** [1 2 3] **David Tse** [1]

Abstract

Monte Carlo (MC) permutation test is considered the gold standard for statistical hypothesis testing, especially when standard parametric assumptions are not clear or likely to fail. However, in modern data science settings where a large number of hypothesis tests need to be performed simultaneously, it is rarely used due to its prohibitive computational cost. In genome-wide association studies, for example, the number of hypothesis tests m is around 10^6 while the number of MC samples n for each test could be greater than 10^8, totaling more than nm=10^{14} samples. In this paper, we propose Adaptive MC multiple Testing (AMT) to estimate MC p-values and control false discovery rate in multiple testing. The algorithm outputs the same result as the standard full MC approach with high probability while requiring only $\tilde{O}(\sqrt{n}m)$ samples. This sample complexity is shown to be optimal. On a Parkinson GWAS dataset, the algorithm reduces the running time from 2 months for full MC to an hour. The AMT algorithm is derived based on the theory of multi-armed bandits.

1. Introduction

Monte Carlo (MC) permutation testing is considered the gold standard for statistical hypothesis testing. It has the broad advantage of estimating significance non-parametrically, thereby safeguarding against inflated false positives (Dwass, 1957; Davison et al., 1997; Boos & Zhang, 2000; Lehmann & Romano, 2006; Phipson & Smyth, 2010). It is especially useful in cases where the distributional assumption of the data is not apparent or likely to be violated.

A good example is genome-wide association study (GWAS), whose goal is to identify associations between the genotypes (single nucleotide polymorphisms or SNPs) and the phenotypes (traits) (Visscher et al., 2017). For testing the association between a SNP and the phenotype, the p-value is often derived via closed-form methods like the analysis of variance (ANOVA) or the Pearson's Chi-squared test (Purcell et al., 2007). However, these methods rely on certain assumptions on the null distribution, the violation of which can lead to a large number of false positives (Yang et al., 2014; Che et al., 2014). MC permutation test does not require distributional assumption and is preferable in such cases from a statistical consideration (Gao et al., 2010). However, the main challenge of applying MC permutation test to GWAS is *computational*.

MC permutation test is a special type of MC test where the p-values are estimated by MC sampling from the null distribution — permutation test computes such MC samples by evaluating the test statistic on the data points but with the responses (labels) randomly permuted. Let T^{obs} be the observed test statistic and $T_1^{\text{null}}, T_2^{\text{null}}, \cdots, T_n^{\text{null}}$ be n independently and identically distributed (i.i.d.) test statistics randomly generated under the null hypothesis. The MC p-value is written as

$$P^{\text{MC}(n)} \overset{\text{def}}{=} \frac{1}{n+1}\left(1+\sum_{j=1}^{n}\mathbb{I}\{T_j^{\text{null}} \geq T^{\text{obs}}\}\right), \qquad (1)$$

which conservatively estimates the ideal p-value $P^{\infty} \overset{\text{def}}{=} \mathbb{P}(T^{\text{null}} \geq T^{\text{obs}})$. In addition, $P^{\text{MC}(n)}$ converges to the ideal p-value P^{∞} as the number of MC samples $n \to \infty$.

GWAS is an example of large-scale multiple testing: each SNP is tested for association with the phenotype, and there are many SNPs to test. For performing m such tests simultaneously, the data is collected and each of the m null hypotheses is associated with an ideal p-value (Fig.1a). A common practice, as visualized in Fig.1b, is to first compute an MC p-value for each test using n MC samples and then apply a multiple testing procedure to the set of MC p-values $\{P_i^{\text{MC}(n)}\}$ to control the false positives, e.g., using the Bonferroni procedure (Dunn, 1961) or the Benjamini-Hochberg procedure (BH) (Benjamini & Hochberg, 1995). Here, as folklore, the number of MC samples n is usually chosen to

[1] Department of Electrical Engineering, Stanford University [2] Department of Biomedical Data Science, Stanford University [3] Chan-Zuckerberg Biohub. Correspondence to: James Zou <jamesyzou@gmail.com>, David Tse <dntse@stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

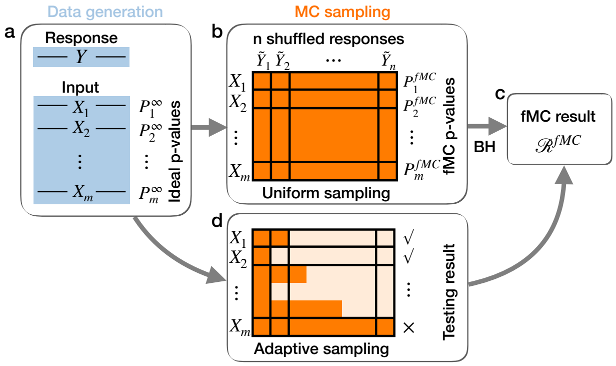


Figure 1. Workflow. (a) N data samples are collected and the kth data point has a response (label) $Y^{(k)}$ and m inputs (features) $(X_1^{(k)}, \cdots, X_m^{(k)})$. There are m null hypotheses to test; the ith null hypothesis corresponds to no association between the ith input $\mathbf{X}_i$=$(X_i^{(1)}, \cdots, X_i^{(N)})$ and the response $\mathbf{Y}$=$(Y^{(1)}, \cdots, Y^{(N)})$. (b) The standard fMC workflow is to: 1) compute an MC p-value for each test i using n MC samples $T_{i1}^{\text{null}}, \cdots, T_{in}^{\text{null}}$; 2) apply the BH procedure on the set of MC p-values to control FDR. In this example, T_{ij}^{null} is the correlation between the ith input $\mathbf{X}_i$ and a randomly permuted response $\mathbf{Y}_{\sigma_j}$. (c) The fMC result is to make discovery (claim association) for a subset of inputs. (d) AMT directly estimates the fMC p-values by adaptive MC sampling and recovers the fMC testing result with high probability.

be at least 10 or 100 times[1] of the number of tests m. In GWAS, there are around 10^6 SNPs to be examined simultaneously via multiple testing and n is recommended to be at least 10^8 (Johnson et al., 2010). The total number of MC samples is nm=10^{14}, infeasible to compute.

This work considers the standard full MC (fMC) workflow, as shown in Fig.1b, of first computing p-values with MC sampling and then controlling the false discovery rate (FDR) by applying the BH procedure to the set of MC p-values. The aim is to reduce the number of MC samples while obtaining the same fMC testing result. The focus of the present paper is on solving a *computational* problem, i.e., accelerating the standard fMC workflow, rather than a *statistical* problem, e.g., improving the power of the test. An alternative goal may be to recover the BH discoveries on the ideal p-values $\{P_i^\infty\}$, which is an ill-posed problem that may take unrealistically many MC samples. Recovering the fMC result, however, takes at most nm samples and any improvement over the complexity nm of uniform sampling represents an improvement over the standard workflow.

Contribution. We propose Adaptive MC multiple Testing (AMT) to compute the fMC testing result via adaptive MC sampling. While recovering the fMC result with high probability, it effectively improves the sample complexity from nm to $\tilde{O}(\sqrt{n}m)$ under mild assumptions that encompass virtually all practical cases, where $\tilde{O}$ hides logarithmic factors. A matching lower bound is provided. In a GWAS dataset on the Parkinson's disease, it improves the computational efficiency by 2-3 orders of magnitude, reducing the running time from 2 months to an hour. We note that AMT is not specific to MC permutation test; it can be used for MC tests in general.

The fMC procedure computes n MC samples for each of the m null hypotheses. For each null hypothesis, a randomly selected subset of the MC samples can provide an estimate of its fMC p-value, whereas the size of this subset determines the estimation accuracy. Intuitively, to recover the fMC result, we only need to estimate how each fMC p-value compares with the corresponding BH threshold; hypotheses with p-values far away from the threshold can be estimated less accurately, thus requiring fewer MC samples. AMT turns this pure computational fMC procedure into a statistical estimation problem, where adaptive sampling can be used.

The specific adaptive sampling procedure is developed via a connection to the pure exploration problem in multi-armed bandits (MAB) (Audibert & Bubeck, 2010; Jamieson et al., 2014). Specifically, the top-k identification problem (Kalyanakrishnan et al., 2012; Chen et al., 2017; Simchowitz et al., 2017) aims to identify the best k arms via adaptive sampling. For AMT, we can think of the m null hypotheses as arms, fMC p-values as arm parameters, and MC samples as observations for each arm. Then recovering the fMC result corresponds to identifying a subset of best arms with small p-values. The difference is that the size of this subset is not known ahead of time — it is a function of the fMC p-values that needs to be learned from data. Nonetheless, the techniques in MAB is borrowed to develop AMT.

1.1. Background

Permutation test. Consider testing the association between input X and response Y using N data samples, i.e., the input vector $\mathbf{X} \in \mathbb{R}^N$ and the response vector $\mathbf{Y} \in \mathbb{R}^N$. A reasonable test statistic can be the Pearson's correlation $\rho(\mathbf{X}, \mathbf{Y})$. Let σ be a permutation on $\{1, \ldots, N\}$ and $\mathcal{S}$ be the set of all possible permutations. The permutation test statistic by permuting the response with σ can be written as $\rho(\mathbf{X}, \mathbf{Y}_\sigma)$. Under the null hypothesis that the response $\mathbf{Y}$ is exchangeable among N samples, the rank of the observed test statistic $\rho(\mathbf{X}, \mathbf{Y})$ among all permutation test statistics is uniformly distributed. Hence, the permutation p-value $p^{\text{Perm}} \overset{\text{def}}{=} \frac{1}{|\mathcal{S}|} \sum_{\sigma \in \mathcal{S}} \mathbb{I}\{\rho(\mathbf{X}, \mathbf{Y}_\sigma) \geq \rho(\mathbf{X}, \mathbf{Y})\}$ follows a uniform distribution over the support $\{\frac{1}{|\mathcal{S}|}, \frac{2}{|\mathcal{S}|}, \cdots, 1\}$. In most cases, the sample size N is too large for computing all

[1] For a hypothesis with ideal p-value P^∞, the relative error for the MC p-value with n MC samples is $1/\sqrt{nP^\infty}$ and therefore, choosing e.g. $n = 100/P^\infty$ gives a relative error of 0.1. Since in multiple testing the p-values we are interested in can be as small as $1/m$, it is recommended to set $n = 100m$.

possible permutations; MC permutation test is used where the permutations are uniformly sampled from $\mathcal{S}$.

FDR control. For simultaneously testing m null hypotheses with p-values $P_1, \cdots, P_m$, a common goal is to control FDR, defined as the expected proportion of false discoveries

$$\text{FDR} \overset{\text{def}}{=} \mathbb{E}\left[\frac{\text{Number of false discoveries}}{\text{Number of discoveries}}\right]. \qquad (2)$$

The most widely-used FDR control algorithm is the BH procedure (Benjamini & Hochberg, 1995). Let $P_{(i)}$ be the ith smallest p-value. The BH procedure rejects hypotheses $P_{(1)}, \cdots, P_{(r^*)}$, where r^* is the critical rank defined as $r^* \overset{\text{def}}{=} \max\left\{r : P_{(r)} \leq \frac{r}{m}\alpha, r \in \{1, 2, \cdots, m\}\right\}$. The BH procedure controls FDR under the assumption that the null p-values are independent and stochastically greater than the uniform distribution.

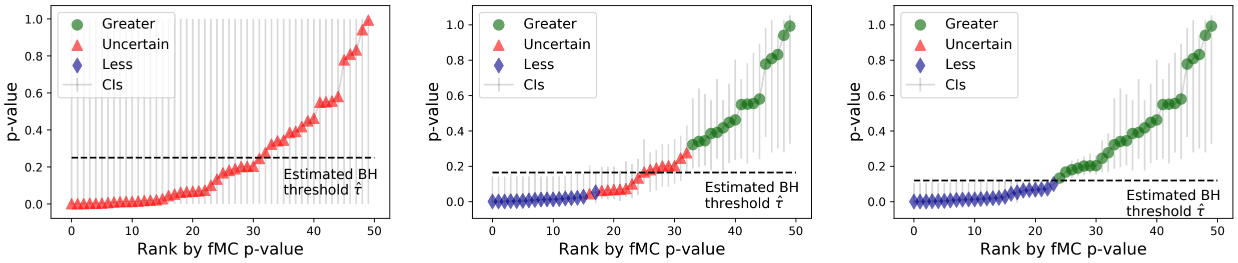
1.2. Related works

The idea of algorithm acceleration by converting a computational problem into a statistical estimation problem and designing the adaptive sampling procedure via MAB has witnessed a few successes. An early example of such works is the Monte Carlo tree search method (Chang et al., 2005; Kocsis & Szepesvári, 2006) to solve large-scale Markov decision problems, a central component of modern game playing systems like AlphaZero (Silver et al., 2017). More recent examples include adaptive hyper-parameter tuning for deep neural networks (Jamieson & Talwalkar, 2016; Li et al., 2016) and medoid computation (Bagaria et al., 2018a). The latter work gives a clear illustration of the power of such an approach. The medoid of a set of n points is the point in the set with the smallest average distance to other points. The work shows that by adaptively *estimating* instead of exactly *computing* the average distance for each point, the computational complexity can be improved from n^2 of the naive method to almost linear in n. This idea is further generalized in AMO (Bagaria et al., 2018b) that considers optimizing an arbitrary objective function over a finite set of inputs. In all these works, the adaptive sampling is by standard best-arm identification algorithms. This present work also accelerates the fMC procedure by turning it into a statistical estimation problem. However, no MAB algorithm is readily available for this particular problem.

Our work applies MAB to FDR control by building an efficient *computational* tool to run the BH procedure *given the data*. There are recent works that also apply MAB to FDR control but in a *statistical inference* setting where the *data collection* process itself can be made adaptive over the different tests. In these works, each arm also corresponds to a test, but each arm parameter takes on a value that corresponds to either null or alternative. Fresh data can be adaptively sampled for each arm and the goal is to select a subset of arms while controlling FDR (Yang et al., 2017; Jamieson & Jain, 2018). In such settings, each observation is a new data and the p-values for the alternative hypotheses can be driven to zero. This is different from AMT where the arm observations are MC samples simulated from the data. As a result, the fMC p-values themselves are the arm parameters and the goal is to *compute* them efficiently to perform BH. In an application like GWAS, where all the SNPs data are typically collected simultaneously via whole genome sequencing, adaptive data collection does not apply but overcoming the computational bottleneck of the full MC procedure is an important problem addressed by the present work. See more details of bandit FDR in Supp. Sec. 2.2.

In the broader statistical literature, adaptive procedures (Besag & Clifford, 1991; Gandy et al., 2017) or importance sampling methods (Yu et al., 2011; Shi et al., 2016) were developed to efficiently compute a single MC p-value. For testing multiple hypotheses with MC tests, interesting heuristic adaptive algorithms were proposed without formal FDR guarantee (Sandve et al., 2011; Gandy & Hahn, 2017); the latter (Gandy & Hahn, 2017) was developed via modifying Thompson sampling, another MAB algorithm. Asymptotic results were provided that the output of the adaptive algorithms will converge to the desired set of discoveries (Guo & Peddada, 2008; Gandy & Hahn, 2014; 2016). Specifically, the most recent work (Gandy & Hahn, 2016) provided a general result that incorporates virtually all popular multiple testing procedures. However, none of the above works provide a standard FDR control guarantee (e.g., FDR $\leq \alpha$) nor an analysis of the MC sample complexity; the MC sample complexity was analyzed in another work only for the case of using Bonferroni procedure (Hahn, 2015). In the present work, standard FDR control guarantee is provided along with upper and lower bounds on the MC sample complexity, establishing the optimality of AMT.

There are also works on fast MC test for GWAS or eQTL (expression quantitative trait loci) study (Pahl & Schäfer, 2010; Kimmel & Shamir, 2006; Browning, 2008; Jiang & Salzman, 2012; Zhang et al., 2012); they consider a different goal which is to accelerate the process of separately computing each MC p-value. In contrast, AMT accelerates the entire workflow of both computing MC p-values and applying BH on them, where the decision for each hypothesis also depends globally on others. The state-of-art method is the sequential Monte Carlo procedure (sMC) that is implemented in the popular GWAS package PLINK (Besag & Clifford, 1991; Purcell et al., 2007; Che et al., 2014). For each hypothesis, it keeps MC sampling until having observed s extreme events or hit the sampling cap n. Then BH is applied on the set of sMC p-values. Here we note that the sMC p-values are conservative so this procedure controls FDR. sMC is discussed and thoroughly compared against in the rest of the paper.



(a) Initialization (b) In progress (c) Termination

Figure 2. Progression of AMT. In this toy example, n=1000, m=50, and α=0.25. AMT maintains upper and lower CBs for each hypothesis (vertical grey bar). (a) At initialization, the estimated BH threshold is set to be maximum $\hat{\tau} = \alpha$ while all CBs cross $\hat{\tau}$. Thus, all hypotheses are in $\mathcal{U}$ and need to be further sampled (red triangle). (b) As the algorithm progresses, more MC samples narrow the confidence intervals and some hypotheses become certain to be greater (green circle) or less (blue diamond) than the estimated BH threshold. The estimated BH threshold also moves down accordingly. (c) At termination, there is no uncertain hypothesis.

2. Problem Formulation

Let m be the number of hypotheses and $P_1^\infty, \cdots, P_m^\infty$ be the ideal p-values. We use the standard notation $[m] \overset{\text{def}}{=} \{1, 2, \cdots, m\}$. For two numbers $a, b \in \mathbb{R}$, $a \wedge b$ means $\min(a, b)$ and $a \vee b$ means $\max(a, b)$.

For each hypothesis $i \in [m]$, we assume the MC samples are available of the form

$$\left[B_{i,1}, B_{i,2}, \cdots, B_{i,n} \middle| P_i^\infty = p_i^\infty\right] \overset{\text{i.i.d.}}{\sim} \text{Bern}(p_i^\infty). \quad (3)$$

Note that one can think of $B_{i,j} = \mathbb{I}\{T_{i,j}^{\text{null}} \geq T_i^{\text{obs}}\}$.

To contrast with adaptive MC sampling, we change the superscript from "MC(n)" to "fMC" for the fMC p-values. Specifically, the fMC procedure uniformly computes n MC samples for each hypothesis, yielding fMC p-values

$$P_i^{\text{fMC}} \overset{\text{def}}{=} \frac{1}{n+1}\left(1 + \sum_{j=1}^{n} B_{i,j}\right), \quad i \in [m]. \quad (4)$$

Here, the extra "1" in the brackets is to make the fMC p-value conservative under the null. We would like to point out that there are two sources of randomness. The first is from the data generation process corresponding to the ideal p-values $\{P_i^\infty\}$ while the second is from MC sampling; they correspond to panel a and panels b-c in Fig.1, respectively. The second source of randomness corresponding to MC sampling is of primary interest in the present paper.

Applying the BH procedure to the fMC p-values yields a set of discoveries $\mathcal{R}^{\text{fMC}} \subset [m]$. Since the fMC p-values are stochastically greater than the uniform distribution under the null hypothesis (Phipson & Smyth, 2010), the set of fMC discoveries $\mathcal{R}^{\text{fMC}}$ has a FDR controlled below the nominal level α. Here, let $P_{(r)}^{\text{fMC}}$ represent the rth smallest p-value and define the critical rank as

$$r^* \overset{\text{def}}{=} \max\left\{r : P_{(r)}^{\text{fMC}} \leq \frac{r}{m}\alpha, r \in [m]\right\}. \quad (5)$$

The BH threshold can be written as $\tau^* \overset{\text{def}}{=} \frac{r^*}{m}\alpha$ while the set of fMC discoveries $\mathcal{R}^{\text{fMC}} \overset{\text{def}}{=} \{i : P_i^{\text{fMC}} \leq \tau^*\}$. The goal is to compute the fMC discoveries $\mathcal{R}^{\text{fMC}}$ with high probability while requiring minimum number of MC samples. Formally, we aim to minimize the number of MC samples for the algorithm such that the algorithm output $\mathcal{R} \subset [m]$ satisfies $\mathbb{P}(\mathcal{R} = \mathcal{R}^{\text{fMC}}) \geq 1 - \delta$, for some given $\delta > 0$.

3. Algorithm

AMT is described as in Algorithm 1. It adopts a top-down procedure by starting with an initial critical rank estimate $\hat{r} = m$ and gradually moving down until it reaches the true critical rank r^*. Specifically, it maintains upper and lower confidence bounds (CBs) for each hypothesis $(p_i^{\text{ub}}, p_i^{\text{lb}})$, the critical rank estimate $\hat{r}$, and the corresponding BH threshold estimate $\hat{\tau} = \frac{\hat{r}}{m}\alpha$. Based on the current estimate, the hypotheses can be categorized as:

$$\begin{aligned} &\text{Certain to be greater than } \hat{\tau}: && \mathcal{C}_{\text{g}} = \{i : p_i^{\text{lb}} > \hat{\tau}\} \\ &\text{Certain to be less than } \hat{\tau}: && \mathcal{C}_{\text{l}} = \{i : p_i^{\text{ub}} \leq \hat{\tau}\} \\ &\text{Uncertain:} && \mathcal{U} = \{i : p_i^{\text{lb}} \leq \hat{\tau} < p_i^{\text{ub}}\}. \end{aligned} \quad (6)$$

As shown in Fig.2a, at initialization the critical rank estimate $\hat{r}$ is set to be the largest possible value m and all hypotheses are uncertain as compared to the estimated BH threshold $\hat{\tau}$; they will be further sampled. In Fig.2b, as more MC samples narrow the confidence intervals, some hypotheses will become certain to be greater/less than $\hat{\tau}$; they will leave $\mathcal{U}$ and stop being sampled. At the same time, according to (5) the estimate $\hat{r}$ cannot be the true critical rank r^* if

Algorithm 1 The AMT algorithm.

```
Input: failure probability $\delta$, nominal FDR $\alpha$.
Initialization: $\frac{\delta}{2mL}$-CBs $\{p_i^{\text{lb}} = 0, p_i^{\text{ub}} = 1\}_{i \in [m]}$, critical rank estimate $\hat{r} = m$, BH threshold estimate $\hat{\tau} = \alpha$, hypothesis sets $\mathcal{C}_{\text{g}} = \emptyset, \mathcal{C}_{\text{l}} = \emptyset, \mathcal{U} = [m]$.
repeat
    Sample obtain the next batch of MC samples for each hypothesis in $\mathcal{U}$ and update their $\frac{\delta}{2mL}$-CBs (Sec. 3.1).
    Update reduce $\hat{r}$ one at a time and update $\mathcal{C}_{\text{g}}$ correspondingly until the following hold at the same time:
        $$\mathcal{C}_{\text{g}} = \left\{ i : p_i^{\text{lb}} > \frac{\hat{r}}{m}\alpha \right\}, \quad \hat{r} = m - |\mathcal{C}_{\text{g}}|.$$
    Update the estimated BH threshold $\hat{\tau} = \frac{\hat{r}}{m}\alpha$ and the hypothesis sets
        $$\mathcal{U} = \{ i : p_i^{\text{lb}} \leq \hat{\tau} < p_i^{\text{ub}} \}, \quad \mathcal{C}_{\text{l}} = \{ i : p_i^{\text{ub}} \leq \hat{\tau} \}.$$
until $\mathcal{U} \neq \emptyset$
Return: $\mathcal{R} = \mathcal{C}_{\text{l}}$.
```

more than $m - \hat{r}$ p-values are greater than the corresponding estimated BH threshold $\hat{\tau}$. Therefore, we can decrease $\hat{r}$ and update $\mathcal{C}_{\text{g}}$ until $m - \hat{r} = |\mathcal{C}_{\text{g}}|$. Note that the estimated BH threshold will be reduced correspondingly. The algorithm repeats such a sample-and-update step until the set $\mathcal{U}$ becomes empty as shown in Fig.2c. Then it outputs the discoveries.

For practical consideration, every time a batch of MC samples is obtained for each hypothesis in $\mathcal{U}$ instead of one, with batch sizes $[h_1, \cdots, h_L]$ prespecified as $h_l = \gamma^l$ for some $\gamma > 1$. Here, $\sum_{l=1}^{L} h_l = n$ and $L = \Theta(\log n)$. The batched sizes are chosen as a geometric series so that 1) after every batch, the confidence intervals for a hypothesis being sampled will shrink by roughly a constant factor; 2) the number of batches L is relatively small to save the computation on updating the estimated quantities. In the actual implementation, we chose $h_1 = 100$ and $\gamma = 1.1$ for all experiments.

3.1. Confidence bounds

Since the fMC p-values are themselves random, the CBs are defined conditional on the fMC p-values, where the MC samples for the ith hypothesis are drawn *uniformly and without replacement* from the set of all n MC samples $\{b_{i,j}\}_{j \in [n]}$. This gives finite population CBs whose uncertainty is 0 when n samples are obtained.

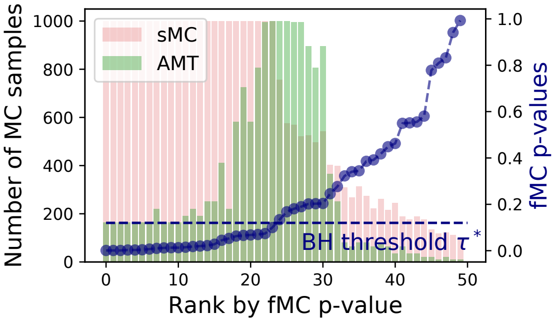


Figure 3. A toy example with n=1000 and m=50. sMC computes more MC samples for hypotheses with smaller p-values while AMT computes more MC samples for hypotheses with p-values closer to the BH threshold.

Specifically, for any $k \in [n]$, let $\tilde{B}_1, \tilde{B}_2, \cdots, \tilde{B}_k$ be random variables sampled uniformly and without replacement from the set $\{b_j\}_{j \in [n]}$ and $\hat{p}_k = \frac{1}{k}\left(1 \vee \sum_{i=j}^{k} \tilde{B}_j\right)$. The δ-CBs $p^{\text{ub}}, p^{\text{lb}}$ satisfy $\mathbb{P}(p \geq p^{\text{ub}}) \leq \delta$ and $\mathbb{P}(p \leq p^{\text{lb}}) \leq \delta$. For the analysis, we assume that the CBs take the form

$$p^{\text{ub}} = \hat{p}_k + \sqrt{\frac{\hat{p}_k c(\delta)}{k}}, \quad p^{\text{lb}} = \hat{p}_k - \sqrt{\frac{\hat{p}_k c(\delta)}{k}}, \tag{7}$$

where $c(\delta)$ is a constant depending only on the probability δ. Eq. (7) represents a natural form that most CBs satisfy with different $c(\delta)$. We consider this general form to avoid tying AMT up with a specific type of CB, for the adaptive procedure is independent of the choice of CB. All binomial confidence intervals can be used here for this sample-without-replacement case (Bardenet et al., 2015); we chose Agresti-Coull confidence interval (Agresti & Coull, 1998) for the actual implementation.

3.2. Comparison to sMC

In sMC, for each hypothesis, MC samples are obtained until either s extreme values (MC observation equal to 1) are observed, or n total permutations are computed with S total successes, where $S < s$. Let K be the number of MC samples obtained for the hypothesis. The sMC p-value is defined as

$$P^{\text{sMC}} \overset{\text{def}}{=} \begin{cases} \frac{s}{K} & K < n \\ \frac{S+1}{n+1} & K = n \end{cases}. \tag{8}$$

After this, BH is applied on the set of sMC p-values to obtain the testing result.

As shown in Fig.3, sMC computes more MC samples for hypotheses with smaller p-values while AMT computes more MC samples for hypotheses with p-values closer to the BH threshold, effectively addressing the hardness of recovering the fMC result, i.e., deciding how each fMC p-value compares with the BH threshold. See also Supp. Sec. 2.1 for how to choose the parameter s.

4. Theoretical Guarantee

We present the high probability recovery and FDR control result, the upper bound, and the lower bound in order. For the upper bound, we first state the $\tilde{O}(\sqrt{n}m)$ result in Proposition 1, which is a direct consequence of the main instance-wise upper bound as stated in Theorem 2.

4.1. Correctness

Theorem 1. *(Correctness)* `AMT` *recovers the fMC result with probability at least $1-\delta$, i.e.,*

$$\mathbb{P}(\mathcal{R}^{\mathrm{AMT}} = \mathcal{R}^{\mathrm{fMC}}) \geq 1 - \delta. \tag{9}$$

Moreover, `AMT` *controls FDR at level $\pi_0\alpha + \delta$, where π_0 is the null proportion.*

Remark 1. *A stronger version is actually proved for (9):* `AMT` *recovers the fMC result with probability at least $1-\delta$ conditional on any set of fMC p-values $\{P_i^{\mathit{fMC}}\} = \{p_i\}$, i.e.,*

$$\mathbb{P}\left(\mathcal{R}^{\mathrm{AMT}} = \mathcal{R}^{\mathrm{fMC}} \middle| \{P_i^{\mathit{fMC}}\} = \{p_i\}\right) \geq 1 - \delta. \tag{10}$$

This also corresponds to the δ-correctness definition in the lower bound Theorem 3. For the FDR control argument, δ is negligible as compared to α; δ is set to be a $o(1)$ term, e.g., $\delta = \frac{1}{m}$. Hence, $\pi_0\alpha + \delta \leq \alpha$ in most cases.

4.2. Upper bound

Without loss of generality, let us assume that the ideal p-values, corresponding to the generation of the data, are drawn i.i.d. from an unknown distribution $F(p)$, which can be understood as a mixture of the null distribution and the alternative distribution, i.e., $F(p) = \pi_0 p + (1-\pi_0)F_1(p)$, where π_0 is the null proportion and $F_1(p)$ is the alternative distribution. The following result shows that the sample complexity of `AMT` is $\tilde{O}(\sqrt{n}m)$ under mild assumptions of $F(p)$.

Proposition 1. *Assume that the ideal p-values are drawn i.i.d. from some unknown distribution $F(p)$ with density $f(p)$ that is either constant ($f(p) = 1$) or continuous and monotonically decreasing. With $\delta = \frac{1}{m\sqrt{n}}$, the total number of MC samples for* `AMT` *satisfies*

$$\mathbb{E}[N] = \tilde{O}(\sqrt{n}m), \tag{11}$$

where $\tilde{O}$ hides logarithmic factors with respect to m and n.

Remark 2. *The asymptotic regime is when $m \to \infty$ while $n = \Omega(m)$. This is because the number of MC samples n should always be larger than the number of hypothesis tests m. A more complete result including δ is $\tilde{O}\left(\sqrt{n}m \log\frac{1}{\delta} + \delta mn\right)$.*

For the assumption on the ideal p-value distribution $F(p)$, $f(p) = 1$ corresponds to the case where all hypotheses are true null while $f(p)$ being continuous and monotonically decreasing essentially assumes that the alternative p-values are stochastically smaller than uniform. Such assumption includes many common cases, e.g., when the p-value is calculated from the z-score $Z_i \sim \mathcal{N}(\mu, 1)$ with $\mu = 0$ under the null and $\mu > 0$ under the alternative (Hung et al., 1997).

A strictly weaker but less natural assumption is sufficient for the $\tilde{O}(\sqrt{n}m)$ result. Let $\tau^\infty = \sup_{[0,1]}\{\tau : \tau \leq F(\tau)\alpha\}$. It assumes that $\exists c_0, c_1 > 0$ s.t. $\forall p \in [\tau^\infty - c_0, 1]$, $f(p) \leq \frac{1}{\alpha} - c_1$. As shown in the proof, τ^∞ is the BH threshold in the limiting case and $f(\tau^\infty) < \frac{1}{\alpha}$ as long as $f(p)$ is strictly decreasing on $[0, \tau^\infty]$. Hence, this weaker assumption contains most practical cases and the $\tilde{O}(\sqrt{n}m)$ result holds generally. However, this weaker assumption involves the definition of τ^∞ which is technical. We therefore chose the stronger but more natural assumption in the statement of the corollary.

Proposition 1 is based on an instance-wise upper bound conditional on the fMC p-values $\{P_i^{\mathrm{fMC}}\} = \{p_i\}$, stated as follows.

Theorem 2. *Conditioning on any set of fMC p-values $\{P_i^{\mathrm{fMC}}\} = \{p_i\}$, let $p_{(i)}$ be the ith smallest p-value and $\Delta_{(i)} = |p_{(i)} - \frac{i \vee r^*}{m}\alpha|$. For the CBs satisfying (7), the total number of MC samples N satisfies*

$$\mathbb{E}\left[N \middle| \{P_i^{\mathrm{fMC}}\} = \{p_i\}\right] \leq \sum_{i=1}^{r^*} n \wedge \left(\frac{4(1+\gamma)^2 c\left(\frac{\delta}{2mL}\right)\tau^*}{\Delta_{(i)}^2}\right)$$
$$+ \sum_{i=r^*+1}^{m} n \wedge \left(\max_{k \geq i} \frac{4(1+\gamma) c\left(\frac{\delta}{2mL}\right) p_{(k)}}{\Delta_{(k)}^2}\right) + \delta mn.$$

Remark 3. *Note that $L = \log_\gamma n$ and for common CBs, $c(\delta) = \log\frac{1}{\delta}$. By setting $\delta = \frac{1}{m}$ and $\gamma = 1.1$, we have*

$$\mathbb{E}\left[N \middle| \{P_i^{\mathrm{fMC}}\} = \{p_i\}\right] \leq \sum_{i=1}^{r^*} n \wedge \left(\frac{18\log(50m^2\log n)\tau^*}{\Delta_{(i)}^2}\right)$$
$$+ \sum_{i=r^*+1}^{m} n \wedge \left(\max_{k \geq i} \frac{9\log(50m^2\log n)p_{(k)}}{\Delta_{(k)}^2}\right) + n.$$

The terms in the summations correspond to the number of MC samples for each hypothesis test. The denominator $\Delta_{(i)}^2$ represents the hardness for determining if to reject each hypothesis while the hypothesis-dependent numerator (τ^* in the first summation and $p_{(k)}$ in the second) represents a natural scaling of the binomial proportion confidence bound. The max *in the second term corresponds to the specific behavior of the top-down approach; it is easy to construct examples where this is necessary. The factor $\log(50m^2\log n)$ corresponds to the high probability bound which is log in m and loglog in n. This is preferable since n may be much larger than m. Overall, the bound is conjectured to be tight except improvements on the $\log m$ term (to perhaps $\log\log m$).*

Table 1. Recovery of the fMC result.

Failure prob. δ	Avg. MC samples per hypothesis (±std)	Prop. of success recovery
0.001	1128±73	100%
0.01	1033±72	100%
0.1	930±70	100%

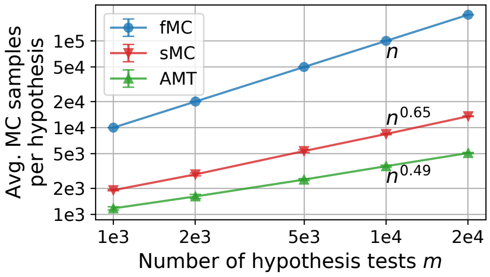


Figure 4. Average number of MC samples per hypothesis test for different algorithms while increasing the number of hypothesis tests m and letting n=10m.

4.3. Lower bound

We provide a matching lower bound for the $\tilde{O}(\sqrt{n}m)$ upper bound. Here, we define a δ-correct algorithm to be one that, conditional on any set of fMC p-values $\{P_i^{\text{fMC}}\} = \{p_i\}$, recovers the fMC result with probability at least $1 - \delta$.

Theorem 3. *Assume that the ideal p-values are drawn i.i.d. from some unknown distribution $F(p)$ with null proportion $\pi_0 > 0$. $\exists \delta_0 > 0$, s.t. $\forall \delta < \delta_0$, any δ-correct algorithm satisfies*

$$\mathbb{E}[N] = \tilde{\Omega}(\sqrt{n}m), \tag{12}$$

where $\tilde{\Omega}$ hides logarithmic factors with respect to m and n.

Remark 4. *In practical settings most hypotheses are true null and therefore $\pi_0 > 0$.*

5. Empirical Results

5.1. Simulated data

Setting. In the default setting, we consider m=1000 hypothesis tests, out of which 200 are true alternatives. The p-values are generated from z-scores $Z_i \sim \mathcal{N}(\mu, 1)$, where the effect size μ=0 under the null and μ=2.5 under the alternative. The number of fMC samples per hypothesis is set to be n=10,000 while the nominal FDR is α=0.1. We investigate the performance of AMT by varying different parameters. The performance of sMC is also reported for comparison, where we set its parameter s=100 according to the discussion which we postpone to Supp. Sec. 2.1. We also tested s=50, which shows a similar result and is hence omitted.

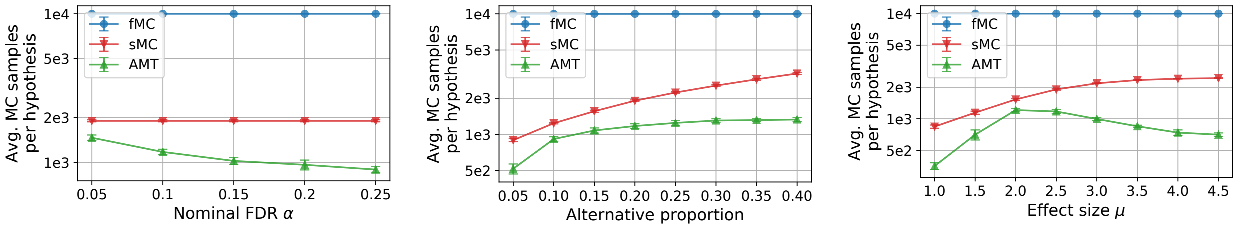
Reliability. We first investigate the reliability of AMT by varying δ, upper bound of the failure probability, where the probability of the CBs is set to be $\frac{\delta}{2m \log n}$. For each value of δ the experiment is repeated 10,000 times. In each repetition, a different set of data is generated and the AMT result is compared to the fMC result while fixing the random seed for MC sampling. As shown in Table 1, AMT recovers the fMC result in all cases while having a 10x gain in sample efficiency. We also found the AMT is rather stable that in practice a larger value of δ may be used; empirically, AMT starts to fail to recover the fMC result when δ exceeds 100.

Scaling. Next we investigate the asymptotic MC sample complexity by increasing m while fixing n=10m. The experiment is repeated 5 times for each parameter and 95% confidence intervals are provided. The result is shown in Fig.4 where the number of MC samples per hypothesis scales sub-linearly with n. A simple linear fitting shows that the empirical scaling for AMT is $n^{0.49}$, validating the $\tilde{O}(\sqrt{n})$ scaling of average MC samples per hypothesis test as derived in Proposition 1 and Theorem 3. Empirically, sMC scales sub-linearly but with a higher rate of $n^{0.65}$.

Varying other parameters. Finally we vary other parameters including the nominal FDR α, alternative proportion, and the effect size μ, where the experiment for each parameter setting is repeated 5 times and 95% confidence intervals are provided. The results are shown in Fig.5. Here, sMC processes each hypothesis separately and computes more MC samples for hypotheses with smaller p-values, as discussed in Sec. 3.2. However, to obtain the BH result, the true difficulty is quantified by the closeness of the p-values to the BH threshold but zero — in other words, if a hypothesis test has a very small p-value while the BH threshold is large, it should not be hard to infer that this null hypothesis should be rejected. AMT captures this by adaptively computing more MC samples for null hypotheses with p-values closer to the BH threshold but zero, effectively adapting to different parameter settings and outperforms sMC in terms of the MC sample complexity.

5.2. GWAS on Parkinson's disease

We consider a GWAS dataset that aims to identify genetic variants associated with Parkinson's disease (Fung et al., 2006), which is known to be a complex disease and is likely to be associated with many different SNPs (Chang et al., 2017); FDR control via BH may yield more new discoveries that are interesting to the community. The dataset comprises 267 cases and 271 controls, each with genotype of 448,001 SNPs that are carefully designed to represent information about several million common genetic variants throughout the genome (Consortium et al., 2003).



(a) Nominal FDR α.

(b) Alternative proportion.

(c) Effect size μ.

Figure 5. Average number of MC samples per hypothesis test for different algorithms while varying different parameters.

The phenotype is binary disease/healthy while the genotype is categorical AA/Aa/aa/missing. SNPs with more than 5% missing values are removed to prevent discoveries due to the missing value pattern; this leaves 404,164 SNPs. The MC samples are based on the permutation test using the Pearson's Chi-squared test, where the phenotype is randomly permuted while keeping the same number of cases and controls. This experiment is run on 32 cores (AMD Opteron™ Processor 6378).

Small data. We first compare AMT with fMC on a smaller-scale data that consists of all 23,915 SNPs on chromosome 4 since fMC can not scale to the whole genome. The number of fMC samples is chosen to be n=$250,000$, yielding a total number of 6×10^9 MC samples that takes 34 mins to compute with 32 cores (4th row in Table 2). Most fMC p-values are similar to the p-values reported in the original paper (Fung et al., 2006) (Supp. Table 1). The slight difference is because the p-values in the original paper were computed using a different test (Pearsons Chi-squared test). FDR level α=0.1 yields 47 discoveries including all discoveries on chromosome 4 reported in the original paper; α=0.05 yields 25 discoveries. The AMT result is identical to the fMC result; it takes 123s and an average of 1,241 MC samples per hypothesis, representing a 17x gain in running time and 201x gain in MC sample efficiency. The same experiment is performed on other chromosomes (chromosome 1-3), which gives a similar result — AMT recovers the fMC result in all cases and as shown in Table 2, AMT has a gain of 17-39x in running time and 201-314x in MC sample efficiency. See also Supp. Table 2. for the fMC p-values.

Full data. We next consider the full dataset with 404,164 SNPs and set the number of fMC samples to be n=40,416,400, yielding a total number of 1.6×10^{13} MC samples. Since there is no internal adaptivity in the fMC procedure, it is reasonable to assume its running time to be proportional to the total number of MC samples, yielding an estimate of 2 months. It is noted that due to the computational cost, no full-scale permutation analysis has been performed on the dataset. The original paper performed permutation test on a subset of SNPs with theoretical p-values less than 0.05. However, such practice may cause outfitting since the same data is used for both hypothesis selection and testing.

Table 2. Small GWAS data. Average MC samples per hypothesis and running time for fMC and AMT. The same experiment is performed on chromosome 1-4 separately.

Chromosome	Avg. MC samples		Running time (s)	
(# of SNPs)	fMC	AMT	fMC	AMT
1 (31,164)	250,000	874 (286x)	3,148	100(31x)
2 (32,356)	250,000	797 (314x)	3,505	90 (39x)
3 (27,386)	250,000	964 (259x)	2,505	89 (28x)
4 (23,915)	250,000	1,241 (201x)	2,031	123 (17x)

We run AMT on this dataset with FDR level α=0.1, taking 1.1hr to finish and average 13,723 MC samples, representing a gain of 1500x in running time and 3000x in MC sample efficiency. We note that we should expect more computational gain for larger-scale problems since AMT scales linearly with $\sqrt{n}$ while fMC scales linearly with n. In addition, for larger-scale problems the MC samples are computed in larger batches which is more efficient, effectively closing the gap between the gain in actual running time and the gain in MC sample efficiency.

With a FDR level α=0.1, AMT made 304 discoveries, including 22/25 SNPs reported in the original paper. Among the three SNPs that are missing, rs355477 (p^{ub}=9.1e-5) and rs355464 (p^{ub}=1.8e-4) are borderline while rs11090762 (p^{lb}=5.9e-2) is likely to be a false positive. AMT has a different number of discoveries from the original paper since the original paper reports all SNPs with p-values $< 1e\text{-}4$ as discoveries instead of using the BH procedure. Also, we have not shown that the AMT discoveries are the same as the fMC discoveries here; we validate the correctness of AMT via the aforementioned small data experiment.

Code availability. The software is available at https://github.com/martinjzhang/AMT

Acknowledgements

We would like to thank Trevor Hastie, Jennifer Listgarten, Hantian Zhang, Vivek Bagaria, Eugene Katsevich, Govinda Kamath, Tavor Baharav, and Wesley Tansey for helpful discussions and suggestions. MZ is partially supported by Stanford Graduate Fellowship. JZ is supported by the Chan-Zuckerberg Initiative, NSF Grant CRII 1657155, NSF Grant AF 1763191, and NIH CEGS. MZ and DT are partially supported by the Stanford Data Science Initiative, NSF Grant CCF 0939370 (Center for Science of Information), NSF Grant CIF 1563098, and NIH Grant R01HG008164.

References

Agresti, A. and Coull, B. A. Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

Audibert, J.-Y. and Bubeck, S. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pp. 13–p, 2010.

Bagaria, V., Kamath, G., Ntranos, V., Zhang, M., and Tse, D. Medoids in almost-linear time via multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 500–509, 2018a.

Bagaria, V., Kamath, G. M., and Tse, D. N. Adaptive monte-carlo optimization. *arXiv preprint arXiv:1805.08321*, 2018b.

Bardenet, R., Maillard, O.-A., et al. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3): 1361–1385, 2015.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test. *arXiv preprint arXiv:1807.05405*, 2018.

Besag, J. and Clifford, P. Sequential monte carlo p-values. *Biometrika*, 78(2):301–304, 1991.

Boos, D. D. and Zhang, J. Monte carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association*, 95(450):486–492, 2000.

Browning, B. L. Presto: rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies. *BMC bioinformatics*, 9(1):309, 2008.

Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold:model-xknockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Chang, D., Nalls, M. A., Hallgrímsdóttir, I. B., Hunkapiller, J., van der Brug, M., Cai, F., Kerchner, G. A., Ayalon, G., Bingol, B., Sheng, M., et al. A meta-analysis of genome-wide association studies identifies 17 new parkinson's disease risk loci. *Nature genetics*, 49(10):1511, 2017.

Chang, H. S., Fu, M. C., Hu, J., and Marcus, S. I. An adaptive sampling algorithm for solving markov decision processes. *Operations Research*, 53(1):126–139, 2005.

Che, R., Jack, J. R., Motsinger-Reif, A. A., and Brown, C. C. An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. *BioData mining*, 7(1):9, 2014.

Chen, L., Li, J., and Qiao, M. Nearly instance optimal sample complexity bounds for top-k arm selection. *arXiv preprint arXiv:1702.03605*, 2017.

Consortium, I. H. et al. The international hapmap project. *Nature*, 426(6968):789, 2003.

Davison, A. C., Hinkley, D. V., et al. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

Dunn, O. J. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Dwass, M. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pp. 181–187, 1957.

Fung, H.-C., Scholz, S., Matarin, M., Simón-Sánchez, J., Hernandez, D., Britton, A., Gibbs, J. R., Langefeld, C., Stiegert, M. L., Schymick, J., et al. Genome-wide genotyping in parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *The Lancet Neurology*, 5(11):911–916, 2006.

Gandy, A. and Hahn, G. Mmctesta safe algorithm for implementing multiple monte carlo tests. *Scandinavian Journal of Statistics*, 41(4):1083–1101, 2014.

Gandy, A. and Hahn, G. A framework for monte carlo based multiple testing. *Scandinavian Journal of Statistics*, 43 (4):1046–1063, 2016.

Gandy, A. and Hahn, G. Quickmmctest: quick multiple monte carlo testing. *Statistics and Computing*, 27(3): 823–832, 2017.

Gandy, A., Hahn, G., and Ding, D. Implementing monte carlo tests with p-value buckets. *arXiv preprint arXiv:1703.09305*, 2017.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., and Province, M. A. Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(1):100–105, 2010.

Guo, W. and Peddada, S. Adaptive choice of the number of bootstrap samples in large scale multiple testing. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

Hahn, G. Optimal allocation of samples to multiple hypothesis tests. *arXiv preprint arXiv:1502.07864*, 2015.

Hung, H. J., O'Neill, R. T., Bauer, P., and Kohne, K. The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, pp. 11–22, 1997.

Jamieson, K. and Jain, L. A bandit approach to multiple testing with false discovery control. *arXiv preprint arXiv:1809.02235*, 2018.

Jamieson, K. and Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pp. 240–248, 2016.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lilucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439, 2014.

Jiang, H. and Salzman, J. Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika*, 99(4):973–980, 2012.

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., and O'Brien, S. J. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC genomics*, 11(1):724, 2010.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pp. 655–662, 2012.

Kimmel, G. and Shamir, R. A fast method for computing high-significance disease association in large population-based studies. *The American Journal of Human Genetics*, 79(3):481–492, 2006.

Kocsis, L. and Szepesvári, C. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*, 2016.

Locatelli, A., Gutzeit, M., and Carpentier, A. An optimal algorithm for the thresholding bandit problem. *arXiv preprint arXiv:1605.08671*, 2016.

Pahl, R. and Schäfer, H. Permory: an ld-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics*, 26(17):2093–2100, 2010.

Phipson, B. and Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

Sandve, G. K., Ferkingstad, E., and Nygård, S. Sequential monte carlo multiple testing. *Bioinformatics*, 27(23): 3235–3241, 2011.

Serfling, R. J. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pp. 39–48, 1974.

Shi, Y., Kang, H., Lee, J.-H., and Jiang, H. Efficiently estimating small p-values in permutation tests using importance sampling and cross-entropy method. *arXiv preprint arXiv:1608.00053*, 2016.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Simchowitz, M., Jamieson, K., and Recht, B. The simulator: Understanding adaptive sampling in the moderate-confidence regime. *arXiv preprint arXiv:1702.05186*, 2017.

Storey, J. D., Taylor, J. E., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

Thulin, M. et al. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, 8(1):817–840, 2014.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

Xia, F., Zhang, M. J., Zou, J. Y., and Tse, D. Neuralfdr: Learning discovery thresholds from hypothesis features. In *Advances in Neural Information Processing Systems*, pp. 1541–1550, 2017.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. A framework for multi-a (rmed)/b (andit) testing with online fdr control. In *Advances in Neural Information Processing Systems*, pp. 5957–5966, 2017.

Yang, G., Jiang, W., Yang, Q., and Yu, W. Pboost: a gpu-based tool for parallel permutation tests in genome-wide association studies. *Bioinformatics*, 31(9):1460–1462, 2014.

Yu, K., Liang, F., Ciampa, J., and Chatterjee, N. Efficient p-value evaluation for resampling-based tests. *Biostatistics*, 12(3):582–593, 2011.

Zhang, M. J., Xia, F., and Zou, J. Adafdr: a fast, powerful and covariate-adaptive approach to multiple hypothesis testing. *bioRxiv*, pp. 496372, 2018.

Zhang, X., Huang, S., Sun, W., and Wang, W. Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study. *Genetics*, 190(4):1511–1520, 2012.