Sequential Nonparametric Testing with the Law of the Iterated Logarithm

Akshay Balsubramani
University of California, San Diego

Aaditya Ramdas
University of California, Berkeley

Abstract

We propose a new algorithmic framework for sequential hypothesis testing with i.i.d. data, which includes A/B testing, nonparametric two-sample testing, and independence testing as special cases. It is novel in several ways: (a) it takes linear time and constant space to compute on the fly, (b) it has the same power guarantee (up to a small factor) as a non-sequential version of the test with the same computational constraints, and (c) it accesses only as many samples as are required – its stopping time adapts to the unknown difficulty of the problem. All our test statistics are constructed to be zero-mean martingales under the null hypothesis, and the rejection threshold is governed by a uniform non-asymptotic law of the iterated logarithm (LIL). For nonparametric two-sample mean testing, we also provide a finite-sample power analysis, and the first non-asymptotic stopping time analysis for this class of problems. We verify our predictions for type I and II errors and stopping times using simulations.

1 INTRODUCTION

Nonparametric statistical decision theory poses the problem of making a decision between a null (H_0) and alternate (H_1) hypothesis over a dataset with the aim of controlling both false positives and false negatives (in statistics terms, maximizing power while controlling type I error), all without making assumptions about the distribution of the data being analyzed. Such hypothesis testing is based on a "stochastic proof by contradiction" – the null hypothesis is thought of by default to be true, and is rejected only if the observed data are statistically very unlikely under the null.

There is increasing interest in solving such problems in a "big data" regime, in which the sample size N can be huge. We present a sequential testing framework for these problems that is particularly suitable for two related scenarios prevalent in many applications:

1) The dataset is extremely large and high-dimensional, so even a single pass through it is prohibitive.
2) The data is arriving as a stream, and decisions must be made with minimal storage.

Sequential tests have long been considered strong in such settings. Such a test accesses the data in an online/streaming fashion, assessing after every new datapoint whether it *then* has enough evidence to reject the null hypothesis. However, prior work tends to be univariate or parametric or asymptotic, while we are the first to provide non-asymptotic guarantees on multivariate nonparametric problems.

To elaborate on our motivations, suppose we have a gigantic amount of data from each of two unknown distributions, enough to detect even a minute difference in their means $\mu_1 - \mu_2$ if it exists. Further suppose that, unknown to us, deciding whether the means are equal is actually statistically easy ($|\mu_1 - \mu_2|$ is large), meaning that one can conclude $\mu_1 \neq \mu_2$ with high confidence by just considering a tiny fraction of the dataset. Can we take advantage of this, despite our ignorance of it?

A naive solution would be to discard most of the data and run a batch (offline) test on a small subset. However, we do not know how hard the problem is, and hence do not know how large a subset will suffice — sampling too little data might lead to incorrectly not rejecting the null, and sampling too much would unnecessarily waste computational resources. If we somehow knew $\mu_1 - \mu_2$, we would want to choose the fewest number of samples (say n^*) to reject the null while controlling type I error at some target level.

1.1 OVERVIEW OF OUR APPROACH

Our sequential test solves the problem by automatically stopping after seeing about n^* samples, while still controlling type I and II errors almost as well as the equivalent linear-time batch test. Without knowing the true problem difficulty, we are able to detect it with virtually no computational or statistical penalty. We devise and formally analyze a sequential algorithm for a variety of problems, starting with a basic test of the bias of a coin, and developing this to nonparametric two-sample mean testing, with further extensions to general nonparametric two-sample and independence testing.

Our proposed procedure only keeps track of a single scalar test statistic, which we construct to be a zero-mean random walk under the null hypothesis. It is used to test the null hypothesis each time a new data point is processed. A major statistical issue is dealing with the apparent multiple hypothesis testing problem – if our algorithm observes its first rejection of the null at time t, it might raise suspicions of being a false rejection, because $t - 1$ hypothesis tests were already conducted and the t-th may have been rejected purely by chance. Applying some kind of multiple testing correction, like the Bonferroni or Benjamini-Hochberg procedure, is exceedingly conservative and produces very suboptimal results over a large number of tests. However, since the random walk moves only a relatively small amount every iteration, the tests are far from independent.

Formalizing this intuition requires adapting a classical probability result, the law of the iterated logarithm (LIL), with which we control for type I error (when H_0 is true). The LIL can be described as follows. Imagine tossing a fair coin, assigning $+1$ to heads and -1 to tails, and keeping track of the sum S_t of t coin flips. The LIL asserts that asymptotically, S_t always remains bounded between $\pm\sqrt{2t \ln \ln t}$ (and this "envelope" is tight).

When H_1 is true, we prove that the sequential algorithm does not need the whole dataset as a batch algorithm would, but automatically stops after processing just "enough" data points to detect H_1, depending on the unknown difficulty of the problem being solved. The near-optimal nature of this adaptive type II error control (when H_1 is true) is again due to the remarkable LIL.

As alluded to earlier, all of our test statistics can be thought of as random walks, which behave like S_t under H_0. The LIL then characterizes how such a random walk behaves under H_0 – our algorithm will keep observing new data since the random walk values will simply bounce around within the LIL envelope. Under H_1, the random walk is designed to have nonzero mean, and hence will eventually stray outside the LIL envelope, at which point the process stops and rejects the null hypothesis.

For practically applying this argument to finite samples and reasoning about type II error and stopping times, we cannot use the classical asymptotic form of the LIL typically stated in textbooks such as Feller (1950), instead adapting a finite-time extension of the LIL by Balsubramani (2015). As we will see, the technical contribution is necessary to investigate the stopping time, and control type I and II errors non-asymptotically *and* uniformly over all t.

In summary, our sequential testing framework has the following properties:

(A) Under H_0, it controls type I error, using a finite-time LIL computable in terms of empirical variance.

(B) Under H_1, and with type II error controlled at a target level, it automatically stops after seeing the same number of points as the corresponding computationally-constrained oracle batch algorithm.

(C) Each update takes $O(d)$ time and constant memory.

In later sections, we develop formal versions of these statements. The statistical observations, particularly the stopping time, follow from the finite-time LIL through simple concentration of measure arguments that extend to very general sequential testing settings, but have seemingly remained unobserved in the literature for decades because of the finite-time LIL necessary to make them.

We begin by describing a sequential test for the bias of a coin in Section 2. We then provide a sequential test for nonparametric two-sample *mean* testing in Section 3. We run extensive simulations in Section 4 to bear out predictions of our theory, followed by a comparison to the extensive existing literature on the subject. We also include extensions to general nonparametric two-sample and independence testing problems, in the appendices. All proofs (and code for experiments) are deferred to the full version (Balsubramani and Ramdas (2015)).

2 DETECTING THE BIAS OF A COIN

This section will illustrate how a simple sequential test can perform statistically as well as the best batch test in hindsight, while automatically stopping essentially as soon as possible. We will show that such early stopping can be viewed as quite a general consequence of concentration of measure. Just for this section, let K represent a constant that may take different values on each appearance, but is always absolute.

Consider observing i.i.d. binary flips $A_1, A_2, \cdots \in \{-1, +1\}$ of a coin, which may be fair or biased towards

```
Fix $N$ and compute $p_N$
if $S_N > p_N$ then
    Reject $H_0$
else
    Fail to reject $H_0$
```

```
Fix $N$
for $n = 1$ to $N$ do
    Compute $q_n$
    if $S_n > q_n$ then
        Reject $H_0$ and return
Fail to reject $H_0$
```

Figure 1: Batch (left) and sequential (right) tests.

$+1$, with $P(A_i = +1) = \rho$. We want to test for fairness, detecting unfairness as soon as possible. Formulated as a hypothesis test, we wish to test, for $\delta \in (0, \frac{1}{2}]$:

$$H_0 : \ \rho = \frac{1}{2} \quad \text{vs.} \quad H_1(\delta) : \ \rho = \frac{1}{2} + \delta$$

For any sample size n, the natural test statistic for this problem is $S_n = \sum_{i=1}^{n} A_i$. S_n is a (scaled) simple mean-zero random walk under H_0. A standard hypothesis testing approach to our problem is a basic *batch* test involving S_N, which tests for deviations from the null for a fixed sample size N (Fig. 1, left). A basic Hoeffding bound shows that

$$S_N \leq \sqrt{\frac{N}{2} \ln \frac{1}{\alpha}} =: p_N$$

with probability $\geq 1 - \alpha$ under the null, so type I error is controlled at level α :

$$P_{H_0}(\text{reject } H_0) = P_{H_0}(S_N > p_N) \leq e^{-2p_N^2/N} = \alpha.$$

2.1 A SEQUENTIAL TEST

The main test we propose will be a sequential test as in Fig. 1. It sees examples as they arrive one at a time, up to a large time N, the maximum sample size we can afford. The sequential test is defined with a sequence of positive thresholds $\{q_n\}_{n \in [N]}$. We show how to set q_n to justify statements (A) and (B) in Section 1.1.

Type I Error. Just as the batch threshold p_N is determined by controlling the type I error with a concentration inequality, the sequential test also chooses $q_1, \dots, q_N$ to control the type I error at α:

$$P_{H_0}(\text{reject } H_0) = P_{H_0}\left(\exists n \leq N : S_n > q_n\right) \leq \alpha \quad (1)$$

This inequality concerns the uniform concentration over infinite tails of S_n, but what $\{q_n\}_{n \in [N]}$ satisfies it? Asymptotically, the answer is governed by a foundational result, the LIL:

Theorem 1 (Law of the iterated logarithm (Khinchin (1924))). *With probability* 1, $\limsup_{n \to \infty} \frac{S_n}{\sqrt{n \ln \ln n}} = \sqrt{2}$.

The LIL says that q_n should have a $\sqrt{n \ln \ln n}$ asymptotic dependence on n, but does not specify its α dependence.

Our sequential testing insights rely on a stronger non-asymptotic LIL proved in (Balsubramani (2015), Theorem 2): with probability at least $1 - \alpha$, we have $|S_n| \leq \sqrt{Kn \ln\left(\frac{\ln n}{\alpha}\right)} =: q_n$ simultaneously for *all* $n \geq K \ln(\frac{4}{\alpha}) := n_0$. This choice of q_n satisfies (1) for $n_0 \leq n \leq N$, and specifies the sequential test as in Fig. 1. (Choosing q_n this way is unimprovable in all parameters up to absolute constants (Balsubramani (2015))).

Type II Error. For practical purposes, $\sqrt{\ln \ln n} \leq \sqrt{\ln \ln N}$ can be treated as a small constant (even when $N = 10^{20}, \sqrt{\ln \ln N} < 2$). Hence, $q_N \approx p_N$ (more discussion in the appendices), and the power is:

$$P_{H_1(\delta)}\left(\exists n \leq N : S_n > q_n\right) \geq P_{H_1(\delta)}\left(S_N > q_N\right) \quad (2)$$
$$\approx P_{H_1(\delta)}\left(S_N > p_N\right) \quad (3)$$

So the sequential test is essentially as powerful as a batch test with N samples (and similarly the n^{th} round of the sequential test is like an n-sample batch test).

Early Stopping. The standard motivation for using sequential tests is that they often require few samples to reject statistically distant alternatives. To investigate this with our working example, suppose N is large and the coin is actually biased, with a fixed unknown $\delta > 0$. Then, if we somehow had full knowledge of δ when using the batch test and wanted to ensure a desired type II error $\beta < 1$, we would use just enough samples $n_\beta^*(\delta)$ (written as n^* in context):

$$n_\beta^*(\delta) = \min\left\{n : P_{H_1(\delta)}\left(S_n \leq p_n\right) \leq \beta\right\} \quad (4)$$

so that for all $n \geq n_\beta^*(\delta)$, since $p_n = o(n)$,

$$\beta \geq P_{H_1(\delta)}\left(S_n \leq p_n\right) = P_{H_1(\delta)}\left(S_n - n\delta \leq p_n - n\delta\right)$$
$$\geq P_{H_1(\delta)}\left(S_n - n\delta \leq -Kn\delta\right) \quad (5)$$

Examining (5), note that $S_n - n\delta$ is a mean-zero random walk. Therefore, standard lower bounds for the binomial tail tell us that $n_\beta^*(\delta) \geq \frac{K \ln(1/\beta)}{\delta^2}$ suffices, and no test can statistically use much less than $n^*_\beta(\delta)$ samples under $H_1(\delta)$ to control type II error at β.

How many samples does the sequential test use? The quantity of interest is the test's stopping time τ, which is $< N$ when it rejects H_0 and N otherwise. In fact, the expected stopping time is close to n^* under any alternate hypothesis:

Theorem 2. *For any δ and any $\beta > 0$, there exist absolute constants K_1, K_2 such that*

$$\mathbb{E}_{H_1}[\tau] \leq \left(1 + \frac{K_1 \beta^{K_2}}{\ln \frac{1}{\beta}}\right) n^*_\beta(\delta)$$

Theorem 2 shows that the sequential test stops roughly as soon as we could hope for, under any alternative δ, despite our ignorance of δ! We will revisit these ideas when presenting our two-sample sequential test later in Section 3.1.

2.2 DISCUSSION

Before moving to the two-sample testing setting, we note the generality of these ideas. Theorem 2 is proved for biased coin flips, but it uses only basic concentration of measure ideas: upper and lower bounds on the tails of a statistic that is a cumulative sum incremented each timestep. Many natural test statistics follow this scheme, particularly those that can be efficiently updated on the fly. Our main sequential two-sample test in the next section does also.

Theorem 2 is notable for its uniformity over δ and β. Note that q_n (and therefore the sequential test) are independent of both of these – we need only to set a target type I error bound α. Under any alternative $\delta > 0$, the theorem holds for all β simultaneously. As β decreases, $n^*_\beta(\delta)$ of course increases, but the leading multiplicative factor $\left(1 + \frac{K_1 \beta^{K_2}}{\ln \frac{1}{\beta}}\right)$ decreases. In fact, with an increasingly stringent $\beta \to 0$, we see that $\frac{\mathbb{E}_{H_1}[\tau]}{n^*} \to 1$; so the sequential test in fact stops closer to n^*, and hence τ is almost *deterministically* best possible. Indeed, the proof of Theorem 2 also shows that $P_{H_1}(\tau \geq n) \leq e^{-Kn\delta^2}$, so the probability of lasting n steps falls off exponentially in n, and is therefore quite sharply concentrated near the optimum $n^*_\beta(\delta)$.

We formalize this precise line of reasoning completely non-asymptotically in an even stronger high-dimensional setting, in the analysis of our main two-sample test in the next section.

3 TWO-SAMPLE MEAN TESTING

In this section, we present our main sequential two-sample test. Assume that we have samples $X_1, \ldots, X_n, \cdots \sim P$ and $Y_1, \ldots, Y_n, \cdots \sim Q$, with P, Q being unknown arbitrary continuous distributions on $\mathbb{R}^d$ with means $\mu_1 = \mathbb{E}_{X \sim P}[X], \mu_2 = \mathbb{E}_{Y \sim Q}[Y]$, and we need to test

$$H_0 : \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1 : \mu_1 \neq \mu_2 \tag{6}$$

Denote covariances of P, Q by Σ_1, Σ_2 and $\Sigma := \frac{1}{2}(\Sigma_1 + \Sigma_2)$. Define $\delta := \mu_1 - \mu_2$ so that $\delta = 0$ under H_0. Let $\Phi(\cdot)$ denote the standard Gaussian CDF, and $[\ln\ln]_+(x) := \ln\ln[\max(x, e^e)]$.

3.1 A LINEAR-TIME SEQUENTIAL TEST

Our sequential test follows the scheme in Fig. 1, so we only need to specify a sequence of rejection thresholds q_n. To do this, we denote

$$h_i = (X_{2i-1} - Y_{2i-1})^\top (X_{2i} - Y_{2i}).$$

and define our sequential test statistic as the following *stochastic process* evolving with n:

$$T_n = \sum_{i=1}^n h_i.$$

Under H_0, $\mathbb{E}[h_i] = 0$, and T_n is a zero-mean random walk.

Proposition 1. $\mathbb{E}[T_n] = \mathbb{E}[h] = n\|\delta\|^2$, *and*

$$\mathrm{var}(T_n) = n\,\mathrm{var}(h) = n(4\,\mathrm{tr}(\Sigma^2) + 4\delta^\top \Sigma \delta) =: nV_0.$$

We assume – for now – that our data are bounded, i.e.

$$\|X\|, \|Y\| \leq 1/2,$$

so that by the Cauchy-Schwarz inequality, w.p. 1,

$$|T_n - T_{n-1}| = |(X_{2n-1} - Y_{2n-1})^\top (X_{2n} - Y_{2n})| \leq 1$$

Since T_n has bounded differences, it exhibits Gaussian-like concentration under the null. We examine the cumulative variance process of T_n under H_0,

$$\sum_{i=1}^n \mathbb{E}\left[(T_i - T_{i-1})^2 \mid h_{1:(i-1)}\right] = \sum_{i=1}^n \mathrm{var}(h_i) = nV_0$$

Using this, we can control the behavior of T_n under H_0.

Theorem 3 (Balsubramani (2015))**.** *Take any $\xi > 0$. Then with probability $\geq 1 - \xi$, for all n simultaneously,*

$$|T_n| < C_0(\xi) + \sqrt{2C_1 nV_0 [\ln\ln]_+(nV_0) + C_1 nV_0 \ln\left(\frac{4}{\xi}\right)}$$

where $C_0(\xi) = 3(e-2)e^2 + 2\left(1+\sqrt{\frac{1}{3}}\right)\ln\left(\frac{8}{\xi}\right)$*, and* $C_1 = 6(e-2)$.

Unfortunately, we cannot use the theorem directly to get computable deviation bounds for type I error control, because the covariance matrix Σ is unknown a priori. nV_0 must instead be estimated on the fly as part of the sequential test, and its estimate must be concentrated tightly and *uniformly over time*, so as not to present a statistical bottleneck if the test runs for a long time. We prove such a result, necessary for sequential testing, relating nV_0 to the empirical variance process $\widehat{V}_n = \sum_i h_i^2$.

Lemma 4. *With probability* $\geq 1-\xi$*, for all* n *simultaneously, there is an absolute constant* C_3 *such that*

$$nV_0 \leq C_3(\widehat{V}_n + C_0(\xi))$$

Its proof uses a self-bounding argument and is in the Appendix. Now, we can combine these to prove a novel uniform *empirical* Bernstein inequality to (practically) establish concentration of T_n under H_0.

Theorem 5 (Uniform Empirical Bernstein Inequality for Random Walks)**.** *Take any* $\xi > 0$*. Then with probability* $\geq 1-\xi$*, for all* n *simultaneously,*

$$|T_n| < C_0(\xi) + \sqrt{2\widehat{V}_n^*\left([\ln\ln]_+\widehat{V}_n^* + \ln\left(\frac{4}{\xi}\right)\right)}$$

where $\widehat{V}_n^* := C_3(\widehat{V}_n + C_0(\xi))$*,* $C_0(\xi) = 3(e-2)e^2 + 2\left(1+\sqrt{\frac{1}{3}}\right)\ln\left(\frac{8}{\xi}\right)$ *and* C_3 *is an absolute constant.*

Its proof follows immediately from a union bound on Thm. 3 and Lem. 4. Thm. 5 depends on $\widehat{V}_n$, which is easily calculated by the algorithm on the fly in constant time per iteration. Ignoring constants for clarity, Thm. 5 effectively implies that our sequential test from Figure 1 controls type I error at α by setting

$$q_n \propto \ln\left(\frac{1}{\alpha}\right) + \sqrt{2\widehat{V}_n \ln\left(\frac{\ln \widehat{V}_n}{\alpha}\right)} \tag{7}$$

Practically, we suggest using the above threshold with a constant of 1.1 to guarantee type-I error approximately α (this is all one often wants anyway, since any particular choice of $\alpha = 0.05$ is anyway arbitrary). This is what we do in our experiments, with excellent success in simulations. For exact or conservative control, consider using a small constant multiple of the above threshold, such as 2.

The above sequential threshold is remarkable, because within the practically useful and simple expression lies a deep mathematical result – the uniform Bernstein LIL effectively involves a union bound for the error probability over an infinite sequence of times. Any other naive attempt to union bound the error probabilities for a possibly infinite sequential testing procedure will be too loose and hence too conservative. Furthermore, the classical LIL is known to be asymptotically tight including constants, and our non-asymptotic LIL is also tight up to small constant factors.

This type-I error control with an implicit infinite union bound surprisingly does not lead to a loss in power. Indeed, our statistic possesses essentially the same power as the corresponding linear-time batch two sample test, and also stops early for easy problems. We make this precise in the following two subsections.

3.2 A LINEAR-TIME BATCH TEST

Here we study a simple linear-time batch two-sample mean test, following the template in Fig. 1. Consider the linear-time statistic $T_N = \sum_{i=1}^{N} h_i$, where, as before, $h_i = (x_{2i-1} - y_{2i-1})^\top(x_{2i} - y_{2i})$. Note that the h_is are also i.i.d., and T_N relies on $2N$ data points from each distribution.

Let V_{N0}, V_{N1} be $\operatorname{var}(T_N) = N\operatorname{var}(h)$ under H_0, H_1 respectively. Recalling Proposition 1:

$$V_{N0} := NV_0 := 4N\operatorname{tr}(\Sigma^2),$$
$$V_{N1} := NV_1 := N(4\operatorname{tr}(\Sigma^2) + 4\delta^\top\Sigma\delta).$$

Then since T_N is a sum of i.i.d. variables, the central limit theorem (CLT) implies that (where $\xrightarrow{d}$ is convergence in distribution)

$$\frac{T_N}{\sqrt{V_{N0}}} \xrightarrow{d}_{H_0} \mathcal{N}(0,1) \tag{8a}$$

$$\frac{T_N - N\|\delta\|^2}{\sqrt{V_{N1}}} \xrightarrow{d}_{H_1} \mathcal{N}(0,1) \tag{8b}$$

Based on this information, our test rejects the null hypothesis whenever

$$T_N > \sqrt{V_{N0}}\, z_\alpha, \tag{9}$$

where z_α is the $1-\alpha$ quantile of the standard normal distribution. So Eq. (8a) ensures that

$$P_{H_0}\left(\frac{T_N}{\sqrt{V_{N0}}} > z_\alpha\right) \leq \alpha,$$

giving us type I error control under H_0.

In practice, we may not know V_{N0}, so we standardize the statistic using the empirical variance – since we assume N is large, these scalar variance estimates do not change the effective power analysis. For non-asymptotic type I error control, we can use an empirical Bernstein inequality (Maurer and Pontil, 2009, Thm. 11), based on an unbiased estimator of V_N. Specifically, the empirical variance of h_is ($\widehat{V}_N$) can be used to reject the null whenever

$$T_N > \sqrt{2\widehat{V}_N \ln(2/\alpha)} + \frac{7N\ln(2/\alpha)}{3(N-1)}. \qquad (10)$$

Ignoring constants for clarity, the empirical Bernstein inequality effectively suggests that the batch test from Figure 1 will have type I error control of α on setting threshold

$$p_N \;\propto\; \ln\left(\frac{1}{\alpha}\right) + \sqrt{2\widehat{V}_N \ln\left(\frac{1}{\alpha}\right)} \qquad (11)$$

For immediate comparison, we copy below the expression for q_n from Eq. (7):

$$q_n \;\propto\; \ln\left(\frac{1}{\alpha}\right) + \sqrt{2\widehat{V}_n\left(\ln\frac{\ln\widehat{V}_n}{\alpha}\right)}.$$

This similarity explains the optimal power and stopping time properties, detailed in the next subsection.

One might argue that if N is large, then $\widehat{V}_N \approx V_N$, and in this case we can simply derive the (asymptotic) power of the batch test given in Eq.(9) as

$$\begin{aligned}
&P_{H_1}\left(\frac{T_N}{\sqrt{V_{N0}}} > z_\alpha\right) \qquad (12)\\
&= P_{H_1}\left(\frac{T_N - N\|\delta\|^2}{\sqrt{V_{N1}}} > z_\alpha\sqrt{\frac{V_{N0}}{V_{N1}}} - \frac{N\|\delta\|^2}{\sqrt{V_{N1}}}\right)\\
&= \Phi\left(\frac{\sqrt{N}\|\delta\|^2}{\sqrt{8\,\mathrm{tr}(\Sigma^2) + 8\delta^\top\Sigma\delta}} - z_\alpha\sqrt{\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}(\Sigma^2) + \delta^\top\Sigma\delta}}\right)
\end{aligned}$$

Note that the second term is a constant less than z_α. As a concrete example, when $\Sigma = \sigma^2 I$, and we denote the signal-to-noise ratio as $\Psi := \frac{\|\delta\|}{\sigma}$, then the power of the linear-time batch test is at least $\Phi\left(\frac{\sqrt{N}\Psi^2}{\sqrt{8d+8\Psi^2}} - z_\alpha\right)$.

3.3 POWER AND STOPPING TIME OF SEQUENTIAL TEST

The striking similarity of Eq. (11) and Eq. (7), mentioned in the previous subsection, is not coincidental. Indeed, both of these arise out of non-asymptotic versions of CLT-like control and LIL-like control, and we know that in the asymptotic regime for Bernoulli coin-flips, CLT thresholds and LIL threshold differ by just $\propto \sqrt{\ln\ln n}$ factors. Hence, it is not surprising to see the empirical Bernstein LIL match empirical Bernstein thresholds up to $\propto \sqrt{\ln\ln\widehat{V}_n}$ factors. Since the power of the sequential test is *at least* the probability of rejection at the very last step, and since $\sqrt{\ln\ln n} < 2$ even for $n = 10^{20}$, the power of the linear-time sequential and batch tests is essentially the same. However, a sequential test that rejects at the last step is of little practical interest, bringing us to the issue of early stopping.

Early Stopping. The argument is again identical to that Section 2, proving that $\mathbb{E}_{H_1}[\tau]$ is nearly optimal, and arbitrarily close to optimal as β tends to zero. Once more note that the "optimal" above refers to the performance of the oracle linear-time batch algorithm that was informed about the right number of points to subsample and use for the one-time batch test. Formally, let $n^*_\beta(\delta)$ denote this minimum sample size for the two-sample mean testing *batch* problem to achieve a power β, the * indicating that this is an oracle value, unknown to the user of the batch test. From Eq. (12), it is clear that for $N \;\geq\; \frac{8Tr(\Sigma^2)+8\delta^T\Sigma\delta}{\|\delta\|^4}(z_\beta \;+\; z_\alpha)^2$, the power becomes at least β. In other words,

$$n^*_\beta(\delta) \leq \frac{Tr(\Sigma^2) + \delta^T\Sigma\delta}{\|\delta\|^4} 8(z_\beta \;+\; z_\alpha)^2 \qquad (13)$$

Theorem 6. *Under H_1, the sequential algorithm of Fig. 1 using q_n from Eq.* (7) *has expected stopping time $\propto n^*_\beta(\delta)$.*

For clarity, we simplify (7) and (11) by dropping the initial $\ln\left(\frac{1}{\alpha}\right)$ additive term since it is soon dominated by the second term and does not qualitatively affect the conclusion.

3.4 DISCUSSION

This section's arguments have given an illustration of the flexibility and great generality of the ideas we used to test the bias of the coin. In the two-sample setting, we simply design the statistic $T_N = \sum_{i=1}^n h_i$ to be a mean-zero random walk under the null. As in the coin's case, the LIL controls type I error, and the remaining arguments are identical because of the common concentration properties of all random walks.

Our test statistic T_N is chosen with several considerations in mind. First, the batch test is linear-time in the sample complexity, so we are comparing algorithms with the *same computational budget*, on a fair footing. There exist batch tests using U-statistics that have higher power than ours (Reddi et al. (2015)) for a given N, but they use more computational resources ($O(N^2)$ rather than $O(N)$).

Also, the batch statistic is a sum of random increments, a common way to write many hypothesis tests, and one that can be computed on the fly in the sequential setting. Note that T_N is a scalar, so our arguments do not change with d, and we inherit the favorable high-dimensional statistical performance of the statistic; Reddi et al. (2015) has more relevant discussion. The statistic also has been shown to have powerful generalizations in the recent statistics literature, which we discuss in the appendices.

Though we assume data scaled to have norm $\frac{1}{2}$ for convenience, this can be loosened. Any data with bounded norm $B > \frac{1}{2}$ can be rescaled by a factor $\frac{1}{B}$ just for the analysis, and then our results can be used. This results in an empirical Bernstein bound like Thm. 5, but of order $O\left(C_0(\xi) + \sqrt{\widehat{V}_n \ln\left(\frac{\ln(B\widehat{V}_n)}{\xi}\right)}\right)$. The dependence on B is very weak, and is negligible even when $B = \text{poly}(d)$.

In fact, we only require control of the higher moments (e.g. by Bernstein conditions, which generalize boundedness and sub-Gaussianity conditions) to prove the non-asymptotic Bernstein LIL in Balsubramani (2015), exactly as is the case for the usual Bernstein concentration inequalities for averages (Boucheron et al. (2013)). Therefore, our basic arguments hold for unbounded increments h_i as well. In fact, the LIL itself, as well as the non-asymptotic LIL bounds of Balsubramani (2015), apply to martingales – much more general versions of random walks capable of modeling dependence on the past history. Our ideas could conceivably be extended to this setting to devise more data-dependent tests, which would be interesting future work.

4 EMPIRICAL EVALUATION

In this section, we evaluate our proposed sequential test on synthetic data, to validate the predictions made by our theory concerning its type I/II errors and the stopping time.

We simulate data from two multivariate Gaussians ($d = 10$), motivated by our discussion at the end of Section 3.2: each Gaussian has covariance matrix $\Sigma = \sigma^2 I_d$, one has mean $\mu_1 = \mathbf{0}^d$ and the other has $\mu_2 = (\delta, 0, 0, \ldots, 0) \in \mathbb{R}^d$ for some $\delta \geq 0$. We keep $\sigma = 1$ here to keep the scale of the data roughly consistent with the biased-coin example, though we find the scaling of the data makes no practical difference, as we discussed.

4.1 RUNNING THE TEST AND TYPE I ERROR

Like typical hypothesis tests, ours is designed to control type I error. When implementing our algorithmic ideas, it suffices to set q_n as in (7), where the only unknown parameters are proportionality constants C, C_0: $q_n \propto C_0 + \sqrt{C\widehat{V}_n\left(\ln \frac{\ln \widehat{V}_n}{\alpha}\right)}$. The theory suggests that C, C_0 are absolute constants, and prescribes upper bounds for them, which can conceivably be loose because of the analytic techniques used (as Balsubramani (2015) discusses). On the other hand, in the asymptotic limit the bounds become tight; the empirical $\widehat{V}_n$ converges quickly to its mean V_n, and we know from second-moment versions of the LIL that $C = \sqrt{2}$ and $C_0 = 0$ are correct. However, as we consider smaller finite times, that bound must relax (at the extremely low $t = 1$ or 2 when flipping a fair coin, for instance).

Nevertheless, we find that in practice, for even moderate sample sizes like the ones we test here, the same reasonable constants suffice in all our experiments: $C = \sqrt{2}$ and $C_0 = \ln(\frac{1}{\alpha})$, with C_0 following Thm. 5 and similar fixed-sample Bennett bounds (Boucheron et al. (2013); Balsubramani (2015); also see the appendices). The situation is exactly analogous to how the Gaussian approximation is valid for even moderate sample sizes in batch testing, making possible a huge variety of common tests that are asymptotically and empirically correct with reasonable constants to boot.

To be more specific, consider the null hypothesis for the example of the coin bias testing given earlier; these fair coin flips are the most *anti*-concentrated possible bounded steps, and render our empirical Bernstein machinery ineffective, so they make a good test case. We choose C and C_0 as above, and plot the cumulative probability of type I violations $\Pr_{H_0}(\tau \leq n)$ up to time n for different α (where τ is the stopping time of the test), with the results in Fig. 2. To control type I error, the curves need to be asymptotically upper-bounded by the desired α levels (dotted lines). This does not appear true for our recommended settings of C, C_0, but the figure still indicates that type I error is controlled even for very high n with our settings. A slight further raise in C beyond $\sqrt{2}$ suffices to guarantee much stronger control.

Fig. 2 also seems to contain linear plots, which we cannot fully explain. We conjecture it is related to the standard proof of the classical LIL, which divides time into epochs of exponentially growing size (Feller (1950)). For more on provable correctness with low C, see the appendices.

4.2 TYPE II ERROR AND STOPPING TIME

Now we verify the results at the heart of the paper – uniformity over alternatives δ of the type II error and stopping time properties.

Fig. 3 plots the power of the sequential test $P_{H_1(\delta)}(\tau < N)$ against the maximum runtime N using the Gaussian

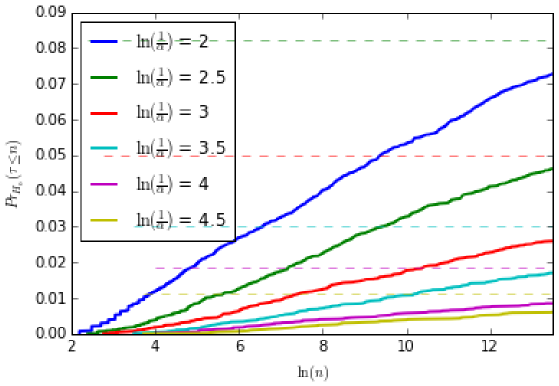

Figure 2: $\Pr_{H_0}(\tau \leq n)$ for different α, on biased coin. Dotted lines of corresponding colors are the target levels α.

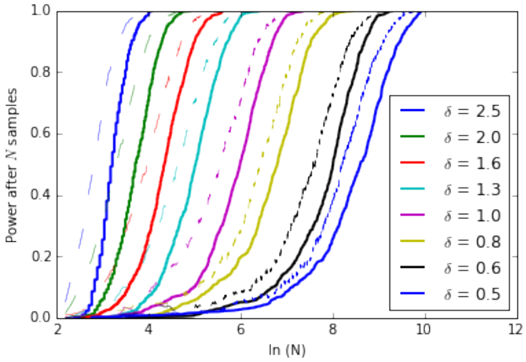

Figure 3: Power vs. $\ln(N)$ for different δ, on Gaussians. Dashed lines represent power of batch test with N samples.

data, at a range of different alternatives δ; the solid and dashed lines represent the power of the batch test (11) with N samples, and the sequential test with maximum runtime N. As we might expect, the batch test has somewhat higher power for a given sample size, but the sequential test consistently performs well compared to it. The role of N here is basically to set a desired tolerance for error; increasing N does not change the intermediate updates of the algorithm, but does increase the power by potentially running the test for longer. So each curve in Fig. 3 illustrates the statistical tradeoff inherent in hypothesis testing against a fixed simple alternative, but the great advantage of our sequential test is in achieving *all of them simultaneously with the same algorithm.*

To highlight this point, we examine the stopping time compared to the batch test for the Gaussian data, in Fig. 4. We see that the distributions of $\ln(\tau)$ are all quite concentrated, and that their medians (marked) fit well to a slope-4 line, showing the predicted $\frac{1}{\delta^4}$ dependence on δ. Some more experiments are in the appendices.

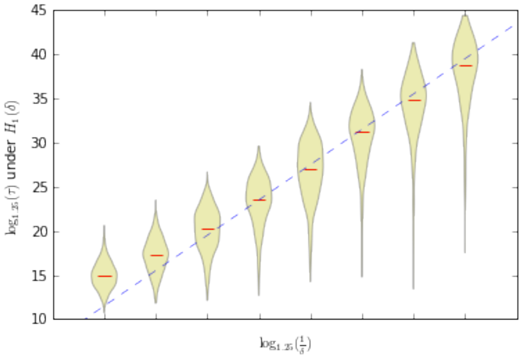

Figure 4: Distribution of $\log_{1.25}(\tau)$ for $\delta \in \{0.5(1.25)^c : c \in \{7, 6, \dots, 0\}\}$, so that the abscissa values $\{\log_{1.25}(\frac{1}{\delta})\}$ are a unit length apart. Dashed line has slope 4.

5 RELATED WORK

Parametric or asymptotic methods. Our statements about the control of type I/II errors and stopping times are very general, following up on early sequential analysis work. Most sequential tests operate in the Wald's framework expounded in Wald (1945). In a seminal line of work, Robbins and colleagues delved into sequential hypothesis testing in an asymptotic sense Robbins (1985). Apart from being asymptotic, their tests were most often for simple hypotheses (point nulls and alternatives), were univariate, or parametric (assuming Gaussianity or known density). That said, two of their most relevant papers are Robbins (1970) and Darling and Robbins (1967), which discuss statistical methods related to the LIL. They give an asymptotic version of the argument of Section 2, using it to design sequential Kolmogorov-Smirnov tests with power one. Other classic works that mention using the LIL for testing various simple or univariate or parametric problems include Darling and Robbins (1968a,b); Lai (1977); Lerche (1986). These all operate in the asymptotic limit in which the classic LIL can be used to set q_N.

For testing a simple null against a simple alternative, the sequential probability ratio test (SPRT) was proved to be optimal by the seminal work of Wald and Wolfowitz (1948), but this applies when both the null and alternative have a known parametric form. The same authors also suggested a univariate nonparametric two-sample test in Wald and Wolfowitz (1940), but presumably found it unclear how to combine these two lines of work.

Bernstein-based methods. Finite-time uniform LIL-type concentration tools from Balsubramani (2015) are crucial to our analysis, and we adapt them in new ways; but novelty in this respect is not our primary focus here, because less recent concentration bounds can also be used to yield similar results. It is always possible to use a weighted union bound (allocating failure probability ξ over time as $\xi_n \propto \frac{\xi}{n^2}$) over fixed-$n$ Bernstein bounds, resulting in a deviation bound of $O\left(\sqrt{V_n \ln \frac{n}{\xi}}\right)$. A more advanced "peeling" argument, dividing time n into exponentially growing epochs, improves the bound to $O\left(\sqrt{V_n \ln \frac{\ln n}{\xi}}\right)$ (e.g. in Jamieson et al. (2014)). This suffices in many simple situations, but in general is still arbitrarily inferior to our bound of $O\left(\sqrt{V_n \ln \ln \frac{V_n}{\xi}}\right)$, precisely in the case $V_n \ll n$ in which we expect the second-moment Bernstein bounds to be most useful over Hoeffding bounds. A yet more intricate peeling argument, demarcating the epochs by exponential intervals in V_n rather than n, can be used to achieve our iterated-logarithm rate, in conjunction with the well-known second-order uniform martingale bound due to Freedman (1975). This serves as a sanity check on the non-asymptotic LIL bounds of Balsubramani (2015), where it is also shown that these bounds have the best possible dependence on all parameters. However, it can be verified that even a suboptimal uniform concentration rate like $O\left(\sqrt{V_n \ln \frac{V_n}{\xi}}\right)$ would suffice for the optimal stopping time properties of the sequential test to hold, with only a slight weakening of the power.

Bernstein inequalities that only depend on empirical variance have been used for stopping algorithms in Hoeffding races (Loh and Nowozin (2013)) and other even more general contexts (Mnih et al. (2008)). This line of work uses the empirical bounds very similarly to us, albeit in the nominally different context of direct estimation of a mean. As such, they too require uniform concentration over time, but achieve it with a crude union bound (failure probability $\xi_n \propto \frac{\xi}{n^2}$), resulting in a deviation bound of $O\left(\sqrt{\widehat{V}_n \ln \frac{n}{\xi}}\right)$. Applying the more advanced techniques above, it may be possible to get our optimal concentration rate, but to our knowledge ours is the first work to derive and use uniform LIL-type empirical Bernstein bounds.

Practical Usage. To our knowledge, implementing sequential testing in practice has previously invariably relied upon CLT-type results patched together with heuristic adjustments of the CLT threshold (e.g. the widely-used scheme for clinical trials of Peto et al. (1977) has an arbitrary conservative choice of $q_n = 0.001$ through the sequential process and $q_N = 0.05 = \alpha$ at the last datapoint). These perform as loose functional versions of our uniform finite-sample LIL upper bound, though without theoretical guarantees. In general, it is unsound to use an asymptotically normal distribution under the null at stopping time τ – the central limit theorem (CLT) applies to any *fixed* time t, but it may not apply to a *random* stopping time τ (see the random-sum CLT of Anscombe (1952), and Gut (2012) and references). This has caused myriad practical complications in implementing such tests (see Lai et al. (2008), Section 4). One of our contributions is to rigorously derive a directly usable finite-sample sequential test, in a way we believe can be extended to a large variety of testing problems.

We emphasize that there are several advantages to our proposed framework and analysis which, taken together, are unique in the literature. We tackle the multivariate nonparametric (possibly even high-dimensional) setting, with composite hypotheses. Moreover, we not only prove that the power is asymptotically one, but also derive finite-sample rates that illuminate dependence of other parameters on β, by considering non-asymptotic uniform concentration over finite times. The fact that it is not provable via purely asymptotic arguments is why our optimal stopping property has gone unobserved for a wide range of tests, even as basic as the biased coin. In our more refined analysis, it can be verified (Thm. 2) that the stopping time diverges to ∞ when the required type II error $\to 0$, i.e. power $\to 1$.

6 CONCLUSION

We have presented a sequential scheme for multivariate nonparametric hypothesis testing against composite alternatives, which comes with a full finite-sample analysis in terms of on-the-fly estimable quantities. Its desirable properties include type I error control by considering finite-time LIL concentration; near-optimal type II error compared to linear-time batch tests, due to the iterated-logarithm term in the LIL; and most importantly, essentially optimal early stopping, uniformly over a large class of alternatives. We presented some simple applications in learning and statistics, but our design and analysis techniques are general, and their extensions to other settings are of continuing future interest.

Acknowledgments

AB was supported in part by NSF grant IIS-1162581, and AR in part by ONR MURI grant N000140911052. The authors are grateful to the very helpful reviewers, and to the ITA workshop, at which this collaboration started.

References

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 600–607. Cambridge Univ Press.

Balsubramani, A. (2015). Sharp uniform martingale concentration bounds. *arXiv preprint arXiv:1405.2639*.

Balsubramani, A. and Ramdas, A. (2015). Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486*.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Darling, D. and Robbins, H. (1967). Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1188–1192.

Darling, D. and Robbins, H. (1968a). Some further remarks on inequalities for sample sums. *Proceedings of the National Academy of Sciences of the United States of America*, 60(4):1175.

Darling, D. and Robbins, H. (1968b). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences of the United States of America*, 61(3):804.

Feller, V. (1950). *An Introduction to Probability Theory and Its Applications: Volume One*. John Wiley & Sons.

Freedman, D. A. (1975). On tail probabilities for martingales. *Ann. Probability*, 3:100–118.

Gut, A. (2012). Anscombe's theorem 60 years later. *Sequential Analysis*, 31(3):368–396.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*.

Khinchin, A. Y. (1924). über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*, 6:9–20.

Lai, T. L. (1977). Power-one tests based on sample sums. *The Annals of Statistics*, pages 866–880.

Lai, T. L., Su, Z., et al. (2008). *Sequential nonparametrics and semiparametrics: Theory, implementation and applications to clinical trials*. Institute of Mathematical Statistics.

Lerche, H. R. (1986). Sequential analysis and the law of the iterated logarithm. *Lecture Notes-Monograph Series*, pages 40–53.

Loh, P.-L. and Nowozin, S. (2013). Faster hoeffding racing: Bernstein races via jackknife estimates. In *Algorithmic Learning Theory*, pages 203–217. Springer.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Mnih, V., Szepesvári, C., and Audibert, J.-Y. (2008). Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679. ACM.

Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British journal of cancer*, 35(1):1.

Reddi, S. J., Ramdas, A., Póczos, B., Singh, A., and Wasserman, L. (2015). On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, pages 1397–1409.

Robbins, H. (1985). *Herbert Robbins Selected Papers*. Springer.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339.