Risk-limiting Financial Audits via Weighted Sampling without Replacement (Supplementary Material)

Shubhanshu Shekhar[1] **Ziyu Xu**[1] **Zachary Lipton**[2, 3] **Pierre Liang**[3] **Aaditya Ramdas**[1, 2]

[1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[3]Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

ORGANIZATION

We discuss some additional background material in Appendix A, and then formulate the proofs omitted from the body of the paper for Proposition 2 and Proposition 3 in Appendix B. In Appendix C, we provide Hoeffding and empirical-Bernstein style CSs that are an alternative to the betting CSs we provide in the paper. We then use the Hoeffding CS constructed in the aforementioned to section as an example of the fact that previous CSs for unweighted mean estimation with uniform sampling from Waudby-Smith and Ramdas [17, 18] can be recovered by our method for a specific choice of (λ_t) in Appendix D. Empirical results from simulations are shown for these CSs in Appendix E, in which we compare the Hoeffding and empirical-Bernstein CSs to the betting CS discussed in the main body of the paper. Finally, we end by presenting the results of applying our methods on a real-world housing dataset in Appendix F.

The code for reproducing the results of this paper is available here: `https://github.com/sshekhar17/WeightedWoRConfSeq`.

A ADDITIONAL BACKGROUND

A.1 RELATED WORK ON CONFIDENCE SEQUENCES (CS)

Confidence sequences (CSs) are a fundamental tool in sequential analysis, and were introduced into this literature by Robbins and coauthors in a series of papers starting with [2]. Some other important early works in this area include [10] and [7]. More recently, there has been a resurgence of interest in confidence sequences, particularly motivated by its applications in anytime valid inference. In anytime valid inference, data samples are received sequentially, and the goal is to derive statistical guarantees that are valid even if one stops sampling and performs inference at a data-dependent time [8, 9, 5, 4]. Thus, confidence sequences can also be employed in the multi-armed bandit setting (where one can sample adaptively from multiple streams of data) to enable best arm identification [6]. Most of the papers mentioned above rely on making certain moment assumptions on the data-generating distributions. An important line of recent work aims to relax these assumptions, by constructing confidence sequences for heavy-tailed data or contaminated data [15, 1, 11, 16].

A.2 BETTING-BASED CS CONSTRUCTION

The betting-based approach for constructing confidence sequences builds upon the idea of *testing-by-betting*, popularized recently by Shafer [12]. This principle states that we can refute a claim (equivalently, a null hypothesis H_0) about the probability distribution generating some data stream, if we can increase our wealth by repeatedly betting on the observations with the restriction that the betting payoffs are *fair* under H_0. The restriction of fair payoff implies that that bettor is not expected to make large gains under H_0, irrespective of the betting strategy employed (formally, the wealth process is a non-negative supermartingale). Consequently, if the bettor ends up making a large profit by betting on the observations, this can be considered as evidence against H_0; with the relative growth in wealth provide a precise measure of the strength of

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

evidence.

To use the above principle for constructing CSs, we simultaneously play a continuum of betting games, indexed by $m \in [0,1]$, each with an initial wealth of \$1. For every $m \in [0,1]$, we bet against the claim $H_{0,m}$ that the true misspecified fraction m^* is equal to m. We design the payoff functions of this betting game, such that the resulting wealth process, $\{W_t(m) : t \geq 1\}$, is a non-negative martingale if $H_{0,m}$ were true, but grows at an exponential rate otherwise. Due to this property, the process at m^*, denoted by $\{W_t(m^*) : t \geq 1\}$, is actually a *test martingale*; that is, a nonnegative martingale with an initial value 1. Hence, Ville's inequality (recalled below in Fact 1) implies that with probability at least $1-\alpha$, the process $(W_t(m^*))$ never exceeds the value $1/\alpha$. This fact, suggests a natural definition of a CS for m^*, consisting of sets $C_t = \{m : W_t(m) < 1/\alpha\}$, since these sets contain m^* for all $t \geq 1$, with probability at least $1-\alpha$. To conclude, the betting-based approach breaks the task of constructing confidence sequences into two smaller tasks:

1. Choosing a sequence of payoff functions that are fair under $H_{0,m}$: we achieve this by using the idea of importance weighting.
2. Developing a betting strategy that ensures fast growth of the wealth process for all $m \neq m^*$: we use the `ApproxKelly` strategy for this in our construction.

An additional design choice that is unique to the problem studied in this paper, is that of the *sampling strategy* to select the transaction indices. We discuss several strategies in Section 3, whose performance is determined by the availability and accuracy of side-information.

We end this discussion by recalling a statement of Ville's inequality [14].

Fact 1 (Ville's Inequality)**.** *Suppose* $\{M_t : t \geq 0\}$ *denotes a nonnegative supermartingale adapted to a filtration* $\{\mathcal{F}_t : t \geq 0\}$*. Then, for any* $\alpha > 0$*, we have*

$$\mathbb{P}\left(\exists t \geq 0 : M_t \geq 1/\alpha\right) \leq \frac{\mathbb{E}[M_0]}{\alpha}.$$

A.3 WORKING WITH MINIBATCHES

In this paper, we have developed our methodology under the assumption that the transactions are sampled and sent to the human auditor, one at a time. In practical scenarios, it may be preferable for the human auditor to evaluate a minibatch of transactions a time, rather than querying the transactions one-by-one. This generalization can be easily handled by updating the wealth process with the averaged payoff (over the minibatch in each round). More specifically, we can proceed as follows, for any $m \in [0,1]$, and for $t = 1, 2, \ldots$:

- Calculate the next sampling distribution, q_t.
- Sample the next batch of transactions, $\mathcal{B}_t := \{I_t^{(1)}, \ldots, I_t^{(B)}\}$, with $I_t^{(j)}$ is drawn according to q_t restricted on $\mathcal{N}_t \setminus \{I_t^{(1)}, \ldots, I_t^{(j-1)}\}$. Recall that, we now have $\mathcal{N}_t = [N] \setminus \left(\cup_{s=1}^{t-1} \mathcal{B}_s\right)$.
- Obtain the true f values, $\{f(I_t^{(i)} : 1 \leq i \leq B\}$ from the oracle (i.e., the human auditor).
- Update the wealth process:

$$W_t(m) = W_{t-1}(m) \times \left(1 + \frac{1}{B}\sum_{i=1}^{B} \lambda_t^{(i)} \left(Z_t^{(i)} - \mu_t^{(i)}(m)\right)\right), \quad \text{for } m \in [0,1].$$

  Here $\lambda_t^{(i)}$ denotes the bet based on $\cup_{s=1}^{t-1}\mathcal{B}_t \cup \{X_t^{(1)}, \ldots, X_t^{(i-1)}\}$, and $Z_t^{(i)} - \mu_t^{(i)}$ denotes the analogous payoff function, as defined in Section 2.

- Update the CS, as $C_t = \{m \in [0,1] : W_t(m) < 1/\alpha\}$.

B PROOFS

B.1 PROOF OF PROPOSITION 2

Recall that $\mathbb{E}_{I_n\sim q}[Z_n] = \mu_n(m^*)$ for any sampling distribution $q \in \Delta^{\mathcal{N}_n}$. Now, we note the following equivalencies

$$\begin{aligned}
&\mathbb{E}_{I_n\sim q}[B_n(\lambda, m)] \\
&= \mathbb{E}_{I_n\sim q}[\lambda(Z_n - \mu_n(m)) - \lambda^2(Z_n - \mu_n(m))^2] \\
&= \mathbb{E}_{I_n\sim q}[\lambda(Z_n - \mu_n(m))] - \mathbb{E}_{I_n\sim q}[\lambda^2(Z_n - \mu_n(m))^2] \\
&= \lambda(m^* - m) - \lambda^2\mathbb{E}_{I_n\sim q}[(Z_n - \mu_n^*(m) - \mu_n(m) + \mu_n^*(m))^2] \\
&= \lambda(m^* - m) - \lambda^2(m^* - m)^2 + 2\lambda^2\mathbb{E}_{I_n\sim q}[Z_n - \mu_n(m^*)](m^* - m) - \lambda^2\mathbb{E}_{I_n\sim q}[(Z_n - \mu_n^*(m))^2] \\
&= \lambda(m^* - m) - \lambda^2(m^* - m)^2 - \mathbb{V}_{I_n\sim q}[Z_n].
\end{aligned}$$

The above equivalencies show that $q_n^* = \operatorname{argmin}_{q\in\Delta^{\mathcal{N}_n}} \mathbb{V}_{I_n\sim q}[Z_n]$ by definition of q_n^* in (**??**), and since λ, m^*, m are fixed in the optimization problem. Consequently, the minimizer of $\mathbb{V}_{I_n\sim q}[Z_n]$ is when the distribution of Z_n has support on only a single value and the variance is 0. This is achieved when $q_n(i) \propto \pi(i)f(i)$ for each $i \in \mathcal{N}_n$. Hence, we have shown our desired result.

B.2 PROOF OF PROPOSITION 3

For any $t \geq 1$, introduce the random variable $D_t = t \times \left(\widetilde{A}_t - A_t\right) = \sum_{i=1}^t \beta_i U_i$. By construction of the term U_i, we know that for any $t \geq 1$, we have

$$\mathbb{E}[D_t|\mathcal{F}_{t-1}] = \sum_{i=1}^{t-1} \beta_i U_i + \beta_t \mathbb{E}[D_t|\mathcal{F}_{t-1}] = D_{t-1}.$$

Thus, $\{D_t : t \geq 1\}$ is a martingale process. Furthermore, since both β_t and U_t lie in the set $[-1, 1]$, the martingale process $\{D_t : t \geq 1\}$ has bounded differences. Hence, by using the time-uniform deviation inequality for martingales with bounded differences [5, Eq. (11)], we have

$$\mathbb{P}\left(\exists t \leq n : |D_t| > 1.7\sqrt{t\left(\log\log(2t) + 0.72\log(10.4/\delta)\right)}\right) < \delta.$$

Since $|\widetilde{A}_t - A_t|/t = |D_t|/t$, the result follows.

C HOEFFDING AND EMPIRICAL-BERNSTEIN CONFIDENCE SEQUENCES

In this section, we present a different approach for constructing confidence sequences, that are based on nonnegative supermartingales (NSMs), instead of nonnegative martingales used by the betting based CS (Section 2). While these CSs are typically looser than the betting CS defined in (4), they are computationally inexpensive and can be derived analytically. We will introduce two such CSs, the Hoeffding CS and empirical-Bernstein CS, and each will have boundaries that take on an explicit form. Thus, only constant time is needed to compute the boundaries for each new sample. In contrast, the betting CS computes its boundaries through a root finding procedure which requires $O(t)$ computations to derive updated boundaries after receiving the tth sample. We provide simulations comparing the Hoeffding and empirical-Bernstein CSs with the betting CS in Appendix E. Before defining our CSs, we first introduce the following quantities:

$$\widehat{m}_t \coloneqq \frac{\pi(I_t)}{q_t(I_t)}f(I_t) + \sum_{i=1}^{t-1}\pi(I_i)f(I_i), \quad \widehat{\mu}_t \coloneqq \frac{\sum_{i=1}^t \widehat{m}_i}{t}, \qquad \widehat{\mu}_t(\lambda_1^t) \coloneqq \frac{\sum_{i=1}^t \lambda_i\widehat{m}_i}{\sum_{i=1}^t \lambda_i}.$$

Note that $\widehat{\mu}_t(\lambda_1^t)$ where $\lambda_1 = \ldots = \lambda_t = 1$ is equivalent to $\widehat{\mu}_t$.

C.1 HOEFFDING CS

To define the Hoeffding CS, we first define the nonnegative supermartingale (NSM) associated with it. As noted in prior work [5, 18], let the following be a CGF-like function for Hoeffding:

$$\psi_{\mathrm{H}}^{c}(\lambda) \coloneqq \frac{\lambda^2 c^2}{8},$$

for any fixed $c > 0$. Now we define the following Hoeffding NSM as follows:

$$M_t^{\mathrm{H}}(m) \coloneqq \exp\left(\sum_{i=1}^{t} \lambda_i (Z_i - \mu_i(m)) - \psi_{\mathrm{H}}^{c_i}(\lambda_i)\right) = \exp\left(\sum_{i=1}^{t} \lambda_i (\widehat{m}_i - m) - \psi_{\mathrm{H}}^{c_i}(\lambda_i)\right),$$

where $c_t \geq \max_{i \in \mathcal{U}_t} \pi(i)/q_t(i)$, and both (λ_t) and (c_t) are predictable w.r.t. $(\mathcal{F}_t)$.

Proposition 1. *$(M_t^{\mathrm{H}}(m^*))_{t \in [N]}$ is an NSM.*

Proof. First, note that $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \mu_t(m)$ since we are assuming the null $H_{0,m}$ is true. Second, note that $Z_t \in [0, \max_{i \in \mathcal{U}_t} \pi(i)/q_t(i)]$ is bounded. Thus, the desired statement follows directly from the MGF bound on bounded random variables i.e. if $X \in [\ell, u]$ is a random variable, then $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\lambda^2 (u-\ell)^2/8)$ for any $\lambda \in \mathbb{R}$. □

Consequently, we can derive the following Hoeffding CS:

$$C_t^{\mathrm{H}} \coloneqq \left(\widehat{\mu}_t(\lambda_1^t) \pm \frac{\log(2/\alpha) + \sum_{i=1}^{t} \psi_{\mathrm{H}}^{c_i}(\lambda_i)}{\sum_{i=1}^{t} \lambda_i}\right) \cap [0,1],$$

where $c_t \geq \max_{i \in \mathcal{U}_t} \pi(i)/q_t(i)$, and both (λ_t) and (c_t) are predictable w.r.t. $(\mathcal{F}_t)$.

C.2 EMPIRICAL-BERNSTEIN CS

Define the following CGF-like function for empirical-Bernstein:

$$\psi_{\mathrm{E}}^{c} \coloneqq \frac{-\log(1 - c\lambda) - c\lambda}{c^2},$$

for any $c > 0$. Now we define the following empirical-Bernstein NSM:

$$\begin{aligned} M_t^{\mathrm{EB}}(m) &\coloneqq \exp\left(\sum_{i=1}^{t} \lambda_i (Z_i - \mu_i(m)) - (Z_i - \widehat{\mu}_{i-1})^2 \psi_{\mathrm{E}}^{c_i}(\lambda_i)\right) \\ &= \exp\left(\sum_{i=1}^{t} \lambda_i (\widehat{m}_t - m) - (Z_i - \widehat{\mu}_{i-1})^2 \psi_{\mathrm{E}}^{c_i}(\lambda_i)\right), \end{aligned}$$

where $\lambda_t \in [0, 1/c_t)$, $c_t \geq \widehat{\mu}_{t-1}$, and both (λ_t) and (c_t) are predictable w.r.t. $(\mathcal{F}_t)$.

Constructing the upper CS through mirroring. M_t^{EB} can be used to construct a CS that lower bounds m^*, but naively constructing an analog NSM (i.e., by negating $Z_t - \mu_t(m)$ into $\mu_t(m) - Z_t$) results in a loose construction for the upper CS, since c_t would need to lower bound $\widehat{\mu}_{t-1} - Z_t$. Z_t (and hence c_t) be quite large depending on the sampling probabilities q_t. Thus, we use the fact that $m^* \in [0,1]$ and hence $1 - m^* \in [0,1]$ to construct a "mirroring" lower CS for $1 - m^* = \sum_{i=1}^{N} \pi(i)(1 - f(i))$. The lower CS for $1 - m^*$ is based upon the following NSM:

$$\begin{aligned} {M'}_t^{\mathrm{EB}}(m) &\coloneqq \exp\left(\sum_{i=1}^{t} \lambda_i \left(\widetilde{Z}_i - \left(1 - m + \sum_{j=1}^{i-1} \pi(I_j)(1 - f(I_j))\right) - (\widetilde{Z}_i - \widetilde{\mu}_{i-1})^2 \psi_{\mathrm{E}}^{c_i}(\lambda_i)\right)\right) \\ &= \exp\left(\sum_{i=1}^{t} \lambda_i (\widetilde{m}_t - (1-m)) - (\widetilde{Z}_i - \widetilde{\mu}_{i-1})^2 \psi_{\mathrm{E}}^{c_i}(\lambda_i)\right). \end{aligned}$$

Here, define $\widetilde{Z}_t \coloneqq \frac{\pi(I_t)}{q_t(I_t)}(1 - f(I_t))$, and let $\widetilde{m}_t$, and $\widetilde{\mu}_t$ be counterparts of $\widehat{m}_t$ and $\widehat{\mu}_t$ where $f(i)$ is replaced with $1 - f(i)$ in the respective definitions for each $i \in [N]$.

Proposition 2. *$(M_t^{\mathrm{EB}}(m^*))_{t\in[N]}$ and $(M_t'^{\mathrm{EB}}(m^*))_{t\in[N]}$ are both NSMs.*

To prove Proposition 2, we introduce the following key lemma from Fan et al. [3].

Lemma 1 (Fan et al. [3, Lemma 4.1])**.** *Let ξ be a number bounded from below i.e. satisfies $\xi \geq -c$, where $c \in \mathbb{R}^+$ is a fixed constant. Let the following be true: $\lambda \in [0, 1/c)$. Then,*

$$1 + \lambda\xi \geq \exp(\lambda\xi + \xi^2(\log(1 - c\lambda) + c\lambda)).$$

Proof. The proof revolves around the following function h:

$$h(x) \coloneqq \frac{\log(1 + x) - x}{x^2/2}, \qquad x > -1.$$

Note that f is increasing in its domain. Then, $\lambda\xi \geq -c\lambda > -1$ by definition of λ and ξ. Thus,

$$h(\lambda\xi) \geq h(-c\lambda) \Leftrightarrow \frac{\log(1 + \lambda\xi) - \lambda\xi}{\xi^2} \geq \frac{\log(1 - c\lambda) + c\lambda}{c^2}.$$

The desired statement follows from expanding this inequality and rearranging terms. □

Proof of Proposition 2. We will only show the proof that $(M_t^{\mathrm{EB}}(m^*))$ is an NSM, since the proof that $(M'^{\mathrm{EB}}_t(m^*))$ is an NSM will follow a similar derivation. Let $Y_t = Z_i - \mu_i(m)$ and $\delta_t = \widehat{\mu}_{t-1} - \mu_i(m)$. Note that $Y_t - \delta_t = Z_t - \widehat{\mu}_{t-1}$. To prove our desired statement, it suffices to show the following is true:

$$\mathbb{E}\left[\exp(\lambda_t Y_t - (Y_t - \widehat{\mu}_{t-1})^2 \psi_{\mathrm{E}}^{\widehat{\mu}_{t-1}}) \mid \mathcal{F}_{t-1}\right] \leq 1. \tag{1}$$

We will now show that (1) is indeed true:

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda_t Y_t - (Y_t - \delta_t)^2 \psi_{\mathrm{E}}^{\widehat{\mu}_{t-1}}) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\exp(\lambda_t(Y_t - \delta_t) - (Y_t - \delta_t)^2 \psi_{\mathrm{E}}^{\widehat{\mu}_{t-1}}) \mid \mathcal{F}_{t-1}\right] \exp(\lambda_t\delta_t) \\
&\leq \mathbb{E}\left[1 + \lambda_t(Y_t - \delta_t) \mid \mathcal{F}_{t-1}\right] \exp(\lambda_t\delta_t) \\
&= \mathbb{E}\left[1 - \lambda_t\delta_t \mid \mathcal{F}_{t-1}\right] \exp(\lambda_t\delta_t) \leq 1.
\end{aligned}$$

The 1st inequality is by application of Lemma 1 with $\xi_t = Y_t - \delta_t$ and $c = \widehat{\mu}_{t-1}$ ($Z_t \geq 0$, so $\xi_t \geq -\widehat{\mu}_{t-1}$). The 2nd equality is the result of $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = (m^* - m_t^*) - (m^* - m_t^*) = 0$. The last inequality is by $1 - x \leq \exp(-x)$ for all $x \in \mathbb{R}$. Thus, we have proved (1) and our desired statement as a result. □

As a result, we can construct the following empirical-Bernstein CS:

$$C_t^{\mathrm{EB}} \coloneqq \left(\widehat{\mu}(\lambda'^t_1) - \frac{\log(2/\alpha) + \sum_{i=1}^t (Z_i - \widehat{\mu}_{i-1})^2 \psi_E^c(\lambda'_i)}{\sum_{i=1}^t \lambda'_i}, 1 - \widetilde{\mu}_t(\lambda_1^t) + \frac{\log(2/\alpha) + \sum_{i=1}^t (\widetilde{Z}_t - \widetilde{\mu}_{t-1})^2 \psi_E^c(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right) \cap [0, 1].$$

The mirroring trick of constructing a lower CS for $1 - m^*$ was originally employed to construct confidence bounds for off policy evaluation in contextual bandits [13, 19]. To the best of our knowledge, this is the first use of the mirroring trick simply for mean estimation.

For both the Hoeffding and empirical-Bernstein CSs, we show in Appendix D that we are able to recover the unweighted, uniform sampling versions introduced by Waudby-Smith and Ramdas [17] as a special case, i.e., when q_t is the uniform distribution over the remaining items and π is uniform over all items. Thus, our formulations of the Hoeffding and empirical-Bernstein CSs generalize the CSs for sampling without replacement in [17] to weighted sampling and estimation.

D CONNECTIONS WITH WAUDBY-SMITH AND RAMDAS [17, 18]

CSs for estimation of the unweighted mean through uniform sampling without replacement are shown in Waudby-Smith and Ramdas [17] for Hoeffding and empirical-Bernstein style CSs, and Waudby-Smith and Ramdas [18] for betting style CSs. In this section, we show these results are a special case of our results that also account for non-uniform sampling strategies and weighted means. For simplicity, we will show the Hoeffding case, and the results for the other CSs follow a similar argument. Following the notation of Waudby-Smith and Ramdas [17], let $(X(i))_{i\in[N]}$ be a finite population of values in $[0,1]$ (without loss of generality to arbitrary bounds on the $X(i)$). Let $X_1,\dots,X_N$ be random variables that are the the result of from sampling uniformly w/o replacement from this population. Waudby-Smith and Ramdas [17] construct the following NSM for $\mu \coloneqq \frac{1}{N}\sum_{i=1}^{N} X(i)$:

$$M_t^{\text{WS-R}}(m) \coloneqq \exp\left(\sum_{i=1}^{t}\lambda_i\left(X_i - m + \frac{1}{N-i+1}\sum_{j=1}^{i-1}(X_j-m)\right) - \psi_H(\lambda_i)\right),$$

and derive the CS

$$C_t^{\text{WS-R}} \coloneqq \left(\widehat{\mu}_t^{\text{WS-R}}(\lambda_1^t) \pm \frac{\log(2/\alpha) + \sum_{i=1}^{t}\psi_H(\lambda_i)}{\sum_{i=1}^{t}\lambda_i\left(1+\frac{i-1}{N-i+1}\right)}\right).$$

The center of this CS is defined as

$$\widehat{\mu}_t^{\text{WS-R}}(\lambda_1^t) \coloneqq \frac{\sum_{i=1}^{t}\lambda_i\left(X_i + \frac{1}{N-i+1}\sum_{j=1}^{i-1}X_j\right)}{\sum_{i=1}^{t}\lambda_i(1+\frac{i-1}{N-i+1})} = \frac{\sum_{i=1}^{t}\lambda_i\left(X_i + \frac{1}{N-i+1}\sum_{j=1}^{i-1}X_j\right)}{\sum_{i=1}^{t}\lambda_i\cdot\frac{N}{N-i+1}}.$$

In the weighted setting, $\pi(i) = 1/N$ and $f(i) = X(i)$ for each $i \in [N]$ implies that we are estimating the uniformly weighted average $m^* = \mu$. For each $t \in [N]$ and $i \in \mathcal{N}_t$, set $q_t(i) = 1/(N-t+1)$ to be the uniform distribution over the remaining items. This gets us the following estimate of the mean from our Hoeffding CS:

$$\widehat{\mu}_t({\lambda'_1}^t) = \frac{\sum_{i=1}^{t}\lambda'_i\left(\frac{N-i+1}{N}X_i + \frac{1}{N}\sum_{j=1}^{i-1}X_j\right)}{\sum_{i=1}^{t}\lambda'_i}.$$

By setting $\lambda'_i = \lambda_i N/(N-i+1)$, for each $i \in [t]$, we get that $\widehat{\mu}_t({\lambda'}_1^t) = \widehat{\mu}_t^{\text{WS-R}}(\lambda_1^t)$. To see that $C_t^H = C_t^{\text{WS-R}}$, we set $c_t = (N-t+1)/N$ for each $t \in [N]$. Note that that is minimum possible value that c_t can be since $\pi(i) = 1/N$ for each $i \in [N]$, and $q_t(i) = 1/(N-t+1)$ for each $i \in \mathcal{N}_t$. As a result, we are able to recover the Hoeffding CS from [17] as a special case of our Hoeffding CS.

E EXPERIMENTS COMPARING DIFFERENT CS CONSTRUCTIONS

In Figure 1, we compare the width of the Hoeffding, empirical-Bernstein, and betting CSs under the `prop-M` sampling strategy, i.e., under a weighted sampling strategy. We follow the same setup as Experiment 1 in Section 5. We see that the empirical-Bernstein CS is tighter in cases where N_{lg} is larger. In these cases, the support size, c_t, is large for N_{lg} transactions, with makes the Hoeffding CS looser as a result. On the other hand, empirical-Bernstein is able to take advantage of the low-variance from a large number of transactions having similar misstated fractions $f(I_t)$. However, when N_{lg} is small, most transactions will have a small support size, c_t, and the Hoeffding CS will be tighter than empirical-Bernstein as a result. The betting CS is tighter than both Hoeffding and empirical-Bernstein CSs in all our simulated setups. This trend is reflected in Figure 2 where we plot the histogram of the first time the CS reaches $\varepsilon = 0.2$ width, i.e., the empirical-Bernstein CS reaches ε width faster than the Hoeffding CS when N_{lg} is larger, and the betting CS is the faster than both Hoeffding and empirical-Bernstein CSs.

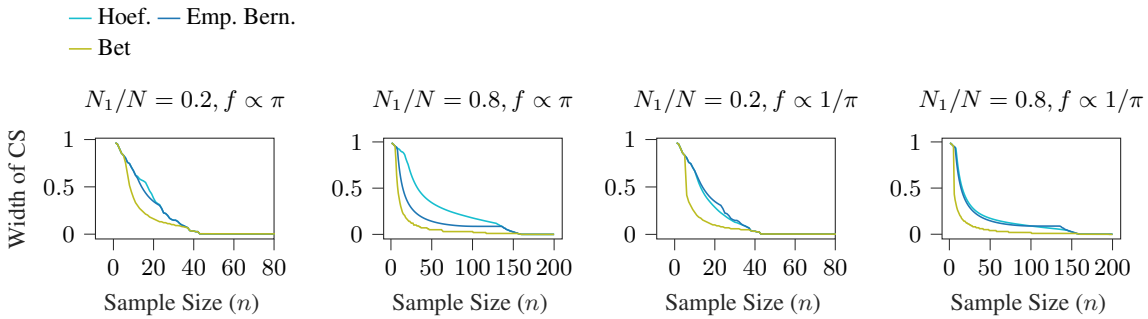

Figure 1: Plots showing the variation of the width of the betting, Hoeffding, and empirical-Bernstein CSs using `prop-M` sampling strategies in different data regimes with $N = 200$; all CSs here also are intersected with the logical CS of Section 2.2. We can see that empirical-Bernstein is tighter than Hoeffding in cases where the proportion of transactions with large weights is high (i.e., N_{lg}/N is large), and vice versa. Across the board, the betting CS is tighter than both Hoeffding and empirical-Bernstein CSs.

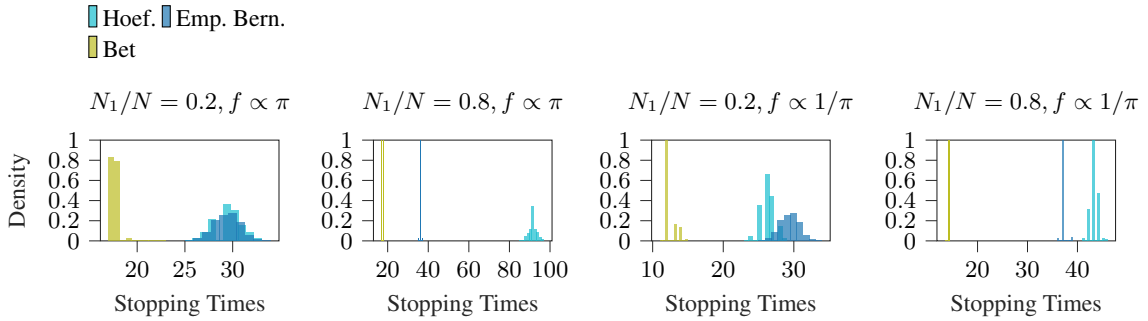

Figure 2: Histograms of the first time for each CS to reach a width of $\varepsilon = 0.2$ with $N = 200$. We choose a larger ε than in Experiment 1 for purposes of demonstrating the difference between the CSs, as all CSs converge to the logical CS and have nearly identical stopping time distribution for small values of ε. We can see that the empirical-Bernstein CS is reaches ε width earlier than Hoeffding CS when N_{lg} is large and the reverse is true when N_{lg} is small, and the betting CS reaches ε uniformly the fastest.

F EXPERIMENTS WITH HOUSING SALES DATA

We now apply our auditing scheme to the transactions in the 'House Sales in King County' dataset from Kaggle. The dataset consists of $21,616$ datapoints, each consisting of 21 features describing the house (such as the number of bedrooms, the number of bathrooms, square footage, floors, condition, etc), and one target variable `price`. We treat the `price` values as the 'reported monetary values' for our framework.

Creating the ground truth. To adapt the dataset to our problem, we first need to generate the 'ground truth', that is, the true f-values. To do this, we proceed in the following steps:

- We first select 10% of the dataset, and assign them some arbitrary f values in the range $(0, 0.7)$.
- Using this 'labelled' dataset, we train a random forest regressor with 200 trees, and mean-absolute error criterion.
- Finally, this trained regressor is then used to generate the ground truth for the remaining 90% of the dataset.

The reason for using this approach for generating the ground truth, is that we want it to be dependent on the additional features associated with each transaction.

Generating the side-information. Having obtained the M and f-values, we obtain the side-information (i.e., S-values) by using 10% of the remaining labelled data to train another predictor for generating side-information on the rest of the data. In our experiments, we either used a single decision-tree regressor, or a random forest with a small number of trees (fewer

than 50). Informally, we expect that increasing the capacity of the regressor should lead to increased correlation between the ground truth and the side-information.

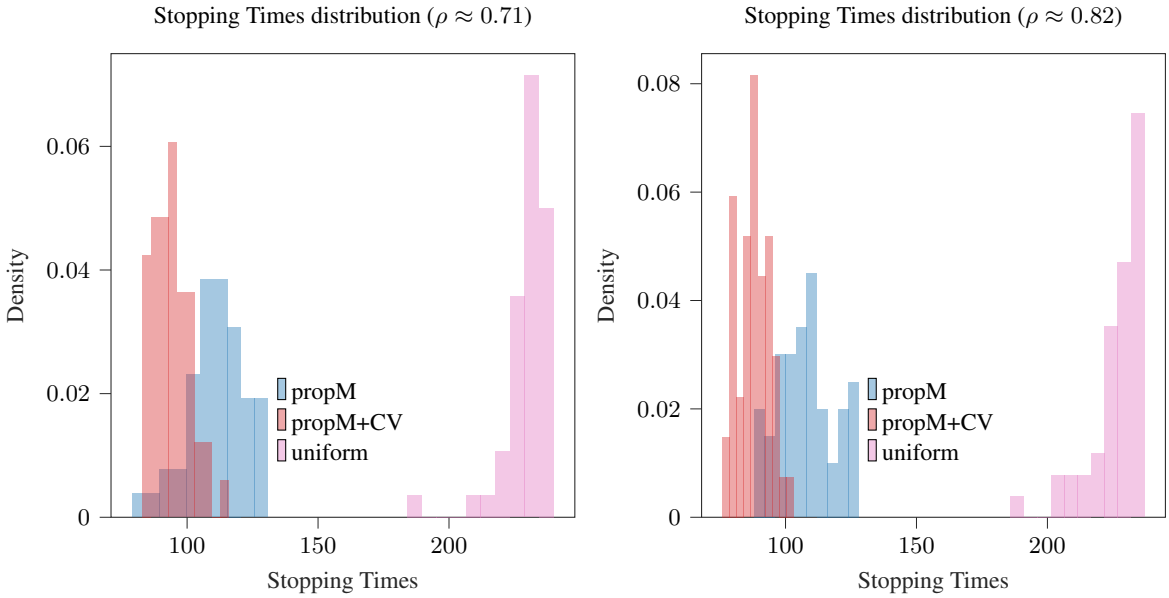

Figure 3: Histograms of the first time for each CS to reach a width of $\varepsilon = 0.05$ with $N = 250$. As expected, the `prop-M` strategy (both with and without control variates) is significantly more sample-efficient than the uniform baseline, and furthermore, the improvement by using control variates increases with increasing informativeness (i.e., ρ) of the side-information.

Experimental Results. In Figure 3, we consider two instances of this problem: (i) side-information generated by a decision-tree regressor, and (ii) side-information generated by a random-forest regressor, consisting of 10 trees. In the former case, the correlation between the side information and the ground-truth is approximately 0.71, while in the latter it is around 0.82. As shown in the plots, the `prop-M` based strategies (both with and without control variates) significantly outperform the uniform baseline strategy. Furthermore, the improvement by incorporating control variates increases with increasing correlation.

REFERENCES

[1] S. Bhatt, G. Fang, P. Li, and G. Samorodnitsky. Catoni-style confidence sequences under infinite variance. *arXiv preprint arXiv:2208.03185*, 2022.

[2] D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

[3] X. Fan, I. Grama, and Q. Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1–22, 2015.

[4] S. R. Howard and A. Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022.

[5] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[6] K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

[7] C. Jennison and B. W. Turnbull. Interim analyses: the repeated confidence interval approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 51(3):305–334, 1989.

[8] R. Johari, L. Pekelis, and D. J. Walsh. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*, 2015.

[9] R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[10] T. L. Lai. On confidence sequences. *The Annals of Statistics*, pages 265–280, 1976.

[11] P. Mineiro. A lower confidence sequence for the changing mean of non-negative right heavy-tailed observations with bounded mean. *arXiv preprint arXiv:2210.11133*, 2022.

[12] G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

[13] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-Confidence Off-Policy Evaluation. *AAAI Conference on Artificial Intelligence*, 2015.

[14] J. Ville. Etude critique de la notion de collectif. *Gauthier-Villars, Paris*, 1939.

[15] H. Wang and A. Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *arXiv preprint arXiv:2202.01250*, 2022.

[16] H. Wang and A. Ramdas. Huber-robust confidence sequences. In *International Conference on Artificial Intelligence and Statistics*, pages 9662–9679. PMLR, 2023.

[17] I. Waudby-Smith and A. Ramdas. Confidence sequences for sampling without replacement. *Neural Information Processing Systems*, 2020.

[18] I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2023 (forthcoming).

[19] I. Waudby-Smith, L. Wu, A. Ramdas, N. Karampatziakis, and P. Mineiro. Anytime-valid off-policy inference for contextual bandits. *arXiv:2210.10768*, 2022.