Risk-limiting Financial Audits via Weighted Sampling without Replacement

Abstract

We introduce the notion of a risk-limiting financial auditing (RLFA): given N transactions, the goal is to estimate the total misstated monetary fraction (m^*) to a given accuracy ϵ, with confidence $1 - \delta$. We do this by constructing new confidence sequences (CSs) for the weighted average of N unknown values, based on samples drawn without replacement according to a (randomized) weighted sampling scheme. Using the idea of importance weighting to construct test martingales, we first develop a framework to construct CSs for arbitrary sampling strategies. Next, we develop methods to improve the quality of CSs by incorporating side information about the unknown values associated with each item. We show that when the side information is sufficiently predictive, it can directly drive the sampling. Addressing the case where the accuracy is unknown *a priori*, we introduce a method that incorporates side information via control variates. Crucially, our construction is adaptive: if the side information is highly predictive of the unknown misstated amounts, then the benefits of incorporating it are significant; but if the side information is uncorrelated, our methods learn to ignore it. Our methods recover state-of-the-art bounds for the special case when the weights are equal, which has already found applications in election auditing. The harder weighted case solves our more challenging problem of AI-assisted financial auditing.

1 INTRODUCTION

Consider the following scenario: in a given year, a company has N recorded financial transactions with reported monetary values $M(i) \in (0, \infty)$ for each $i \in [N] \coloneqq \{1, \ldots, N\}$. As required by law, an external auditor is required to attest with "reasonable assurance" about whether the financial records as a whole are free from "material misstatement." For example, the company has cash receipts for sales of products, and it wants to ensure that the reported monetary value matches the true amount that was made on the sales according to prescribed accounting rules as some receipts may actually represent past sales or future deliveries. This can be done, for instance, by manually examining the entire sales process to determine the true sales amount against the the amount recorded by the company. Since the task of *auditing* each transaction can be complex requires substantial human labor it can be prohibitively expensive to perform a comprehensive audit of a company's records.

Suppose that the auditor has built an AI system for "automated auditing", i.e., this AI system can output predictions about the accuracy of a transaction value, based on receipts, OCR (optical character recognition), databases, etc. Such systems are in a state of active development and deployment, and the high level of industry demand is unsurprising given the remarkable predictive capabilities of modern machine learning algorithms. But there's a catch: because the system is trained and deployed on differently distributed data, its accuracy on a new set of records in a new time period is unknown *a priori*. Even if anecdotally, the AI system seems to perform reasonably well on data collected from a variety of companies, we cannot make statistically certifiable conclusions based solely on the output of the AI system on a new company and/or in a new time period. Thus we can think of AI systems in deployment as black boxes for which we have (reasonable) hopes of high accuracy but lack formal guarantees.

The auditor's goal is to minimize the amount of manual auditing that must be done by a person, while accurately estimating the true monetary amount of those transactions that have not manually audited. When the AI system is accurate, we want to reduce the amount of human auditing effort required. More importantly, we want a statistically rigorous conclusion regardless of the AI system accuracy. Hence, our method should interpolate between using predictions to reduce its uncertainty rapidly when the system is accurate, and the most efficient AI-free strategy when the system is inaccurate.

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

Problem setup and notation. Denote the unknown misstated fraction of the ith transaction as $f(i) \in [0,1]$, for each $i \in [N]$. In other words, if $M^*(i)$ denotes the true value of the transaction i, and $M(i)$ is the reported value, then[1] $f(i) = |M^*(i) - M(i)|/M(i)$. We can normalize the reported transaction values by the sum over all transaction values to get a weight $\pi(i) := M(i)/(\sum_{i=1}^{N} M(i))$ for each $i \in [N]$, where $\sum_{i=1}^{n} \pi(i) = 1$. The auditor wishes to obtain an estimate of $m^* = \sum_{i=1}^{N} \pi(i) f(i)$, the fraction of the total monetary value that is misstated, up to an accuracy $\varepsilon \in [0,1]$. By $S(i)$, we denote the *side information*, a score for the ith transaction that (ideally) predicts $f(i)$. In our setup, the side information can be generated through any method, e.g., through an AI system that automatically analyzes the documents a human auditor would use, may also be available to the auditor. Each transaction can be evaluated by the auditor to reveal $M^*(i)$ (or equivalently, $f(i)$). Thus, *given an $\varepsilon > 0$, in what order should the transactions be audited to estimate m^* within ε additive accuracy, using the fewest number of calls to the auditor?*

If we allow for no uncertainty, i.e., we want to produce a confidence interval (CI) for m^* with 100% confidence, then the best strategy is to audit the transactions in decreasing order of their reported value, and stop when the remaining transactions constitute smaller than an ε fraction of the total. However, we can show that if we want to provide an estimate of m^* that is ε-accurate with probability at least $1-\delta$, for a tolerance level δ (e.g., 0.01), there exist strategies based on randomized sampling WoR that allow us to stop much earlier. In other words, for each $t \in [N]$, we adaptively construct a sampling distribution q_t over the remaining $N - t + 1$ unaudited transactions, and sample I_t, the index of the tth transaction to audit, according to q_t. We then obtain $f(I_t)$ through manual auditing, and incorporate this new information to update our estimate of m^*. If our residual uncertainty is sufficiently small (i.e., smaller than ε), we stop sampling. Otherwise, we continue the process by drawing the next index, I_{t+1}, according to an appropriately chosen distribution q_{t+1}.

Before presenting the technical details, we note that we use $(X_t)_{t\in\mathbb{I}}$ to denote a sequence of objects indexed by a set $\mathbb{I}$, and the tth object is X_t. We drop the indexing subscript if it is clear from context. For any $t \in [N]$, we use $\mathcal{F}_t := \sigma(\{I_i\}_{i\in[t]})$ to denote the sigma-algebra over our query selections for the first t queries.

Risk-limiting financial audit (RLFA). Formally, a (ε, δ)*-risk-limiting financial audit (RFLA)* is a procedure that outputs an interval $\mathcal{C}$ where $|\mathcal{C}| \le \varepsilon$ and $\mathcal{C}$ contains the true misstated fraction, m^*, with probability at least $1 - \delta$. This is a natural generalization of risk-limiting audits that are used to ensure statistically valid election auditing [17, 18, 12] to the financial setting, where each transaction is weighted by its reported monetary value (as opposed to uniform weighting for all votes in the election setting). We also consider other possible definitions of an RLFA in Appendix B. Our goal is to produce $\mathcal{C}$ that satisfies the conditions of an RLFA with as few audits, i.e, queries of f, as possible. To produce such an interval, we propose a framework for building RLFAs by constructing confidence sequences, which we introduce next.

Confidence sequences for estimation. Let $T \in [N]$ be a random stopping time, that is, a random variable for which the event $\{T = t\}$ belongs to $\mathcal{F}_t$ for each $t \in [N]$, and let $\mathcal{T}$ denote the universe of all such stopping times. *Confidence sequences* [11, 9] (CSs), or time-uniform confidence intervals, are sequences of intervals, $(\mathcal{C}_t)_{t\in[N]}$, that satisfy

$$\sup_{T\in\mathcal{T}} \mathbb{P}\left(m^* \notin \mathcal{C}_T\right) \le \delta \Leftrightarrow \mathbb{P}\left(\exists t \in [N] : m^* \notin \mathcal{C}_t\right) \le \delta,$$

where $\delta \in (0,1)$ is a fixed error level. Ramdas et al. [13] showed the equivalence above, i.e., that any sequence of intervals $(\mathcal{C}_t)$ that satisfies one side of the implication will immediately satisfy the other as well.

Using this equivalence, we can define a simple (ε, δ)-RLFA procedure: construct a CS for m^*, denoted by $(\mathcal{C}_t)$, and produce $\mathcal{C}_\tau$ where τ is the following stopping time:

$$\tau = \tau(\varepsilon, \delta) := \min\{t \ge 1 : |\mathcal{C}_t| \le \varepsilon\}. \tag{1}$$

The width of all nontrivial CSs converges to zero as $t \to N$, and thus the above stopping time is well-defined, and is usually smaller than N.

Note that the only source of randomness in this problem is the randomized sampling strategy $(q_t)_{t\in[N]}$, used to select transactions for manual evaluation. Hence, $(q_t)_{t\in[N]}$ is another design choice for us to make. To summarize, our goal in this paper is to **(i)** design sampling strategies (q_t), and **(ii)** develop methods of aggregating the information so collected with any available side information, in order to construct CSs for m^* whose width decays rapidly to 0.

Among existing works in literature, the recent papers by Waudby-Smith and Ramdas [22, 21] are the most closely related to our work. In these works, the authors considered the problem of estimating the average value of N items via WoR sampling—however, they considered only uniform sampling, and estimating only the unweighted mean of the population. Our methods work with any sampling scheme, and can estimate any weighted mean; we recover their existing results in Appendix D.

WoR confidence intervals for a fixed sample size. Most existing results on concentration inequalities for observations drawn via WoR sampling focus on the fixed sample size setting, starting with Hoeffding [8], who bounded the probability of deviation of the unweighted empirical mean with WoR sampling in terms of the range of the observations.

[1] We are primarily concerned with estimating the downside that arises from misstatement, e.g., $M(i)$ represents the money that should have been received for a sale, and $M^*(i)$ represents the actual money received. In this scenario, we may lose at most $M(i)$ amount of money if $M^*(i) = 0$. Hence, we assume $f(i) \in [0,1]$.

In particular, Hoeffding [8] showed that for observations $X_{I_1}, \ldots, X_{I_n} \in [a, b]$ drawn uniformly WoR from N values $(X_i)_{i\in[N]}$, we have

$$\mathbb{P}\left(\frac{\sum_{t=1}^{n} X_{I_t}}{n} - \frac{\sum_{t=1}^{N} X_i}{N} > \varepsilon\right) \leq \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right). \quad (2)$$

In WoR sampling, as the sample size n approaches N, the total number of items, we expect the empirical estimate to approximate the true average very accurately. This observation, not captured by the above bound, was made formal by Serfling [15], who showed that the n in (2) can be replaced by $\frac{n}{1-(n-1)/N}$, thus highlighting the significant improvement possible for larger n values. Ben-Hamou et al. [3] prove a Hoeffding style concentration inequality on the unweighted sample mean to its own expectation, which is a different estimand than the weighted population mean. Finally, in the unweighted case, Bardenet and Maillard [2] obtained variance adaptive Bernstein and empirical-Bernstein variants of Serfling's results, that are tighter in cases where the variance of the observations is small. These results were uniformly dominated by those of Waudby-Smith and Ramdas [21, 22], that have found successful application to auditing elections [23].

1.1 CONTRIBUTIONS

We introduce the concept of (ε, δ)-RLFA that generalizes the notion of a risk-limiting audit introduced by Stark [17] for election auditing. Unlike risk-limiting audits, where the main concern is testing an announced result, the objective of an RLFA is to precisely estimate the misstated monetary fraction of the reported financial transactions. To accomplish this, we make the following key technical contributions:

1. *New CSs for weighted means with non-uniform sampling.* To design an (ε, δ)-RLFA procedure, we construct novel CSs for m^* that are based on a betting method that was pioneered in [22] in Section 4, as well as Hoeffding and empirical-Bernstein CSs in Appendix C (which are looser but have a simple analytical form). Our results generalize previous methods in two ways: **(i)** they can estimate the weighted mean of N items, and **(ii)** they work with adaptive, data-dependent, sampling strategies. In particular, our betting CSs, which we show empirically are the most powerful in Appendix E) are based on simultaneously playing gambling games with an aim to disprove the possibility that $m^* = m$, for each $m \in [0, 1]$. Values for m, where we accumulate much wealth are eliminated from the CS. Consequently, we develop a simple, lucrative betting strategy for this setting (`ApproxKelly`), which is equivalent to formulating narrower CSs.
2. *Adaptive sampling strategies that minimize CS width.* In addition to designing CSes that are intrinsically narrow, we are also able to change the sampling distribution of the transactions at each time step, and develop a sampling strategy that will minimize CS width in concert with any valid CS construction. We propose two sampling strategies, `prop-M` and `prop-MS`, the latter of which can incorporate approximately accurate scores $(S(i))_{i\in[N]}$ to improve the sample efficiency of our CSs. This is accomplished by choosing the sampling distribution, at each time step, that maximizes the wealth accumulated by the betting strategies that underlie our CSs. We find that this is approximately equivalent to choosing the sampling distribution with the minimal variance, and we show that our sampling strategies result in a noticeable improvement over uniform sampling through simulations in Section 5.
3. *Robust use of side information to tighten CSs.* Finally, in Section 4, we develop a principled way of leveraging any available side information, inspired by the idea of control variates used for variance reduction in Monte Carlo sampling. Interestingly, our method adapts to the quality of the side information—if $(S(i))_{i\in[N]}$ and $(f(i))_{i\in[N]}$ are highly correlated, the resulting CSs are tighter, while in the case of uncorrelated $(S(i))$, we simply learn to discard the side information.

2 BETTING-BASED CS CONSTRUCTION

We derive our CSs by designing sequential tests to simultaneously check the hypotheses that $m^* = m$, for all $m \in [0, 1]$. By the principle of *testing by betting* [16], this is equivalent to playing repeated gambling games aimed at disproving the null $m^* = m$, for each $m \in [0, 1]$. Formally, for all $m \in [0, 1]$, we construct a process $(W_t(m))_{t\in[N]}$ (the wealth process), such that **(i)** if $m = m^*$, then $(W_t(m))$ is a *test martingale*, i.e., a nonnegative martingale with initial value 1, and **(ii)** if $m \neq m^*$, then $W_t(m)$ grows at an exponential rate. Recall that a process $(W_t)_{t\in[N]}$ adapted to $(\mathcal{F}_t)_{t\in[N]}$ is a supermartingale iff $\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] \leq W_{t-1}$ for all $t \in [N]$, and a martingale if the inequality is replaced with an equality. Assuming we can construct such a process, we define the confidence set at any time t as the set of those $m \in [0, 1]$ for which $(W_t(m))$ is 'small', because a nonnegative martingale is unlikely to take large values.

As mentioned earlier, this approach requires us to design sampling distributions (q_t), and a method for constructing a CS $(\mathcal{C}_t)$ from the queried indices. We begin by formally defining a sampling strategy.

Definition 1 (Sampling Strategy)**.** *A sampling strategy consists of a sequence $(q_t)_{t\in[N]}$, where q_t is a probability distribution on the set $\mathcal{N}_t := [N] \setminus \{I_1, \ldots, I_{t-1}\}$. Here I_j denotes the index drawn according to the predictable (i.e., $\mathcal{F}_{j-1}$-measurable) distribution q_j.*

A natural baseline sampling strategy is to set q_t to be uniform over $\mathcal{N}_t$ for all $t \in [N]$. We will develop other, more powerful, sampling strategies that are more suited to our problem in Section 3.

We now describe how to construct the wealth process for an arbitrary sampling strategy. First, define the following:

$$Z_t := f(I_t)\frac{\pi(I_t)}{q_t(I_t)}, \text{ and } \mu_t(m) := m - \sum_{j=1}^{t-1} \pi(I_j) f(I_j).$$

Note that $\mu_t(m)$ is the remaining misstated fraction after accounting for the first $t-1$ queries to f if m is truly the total misstated fraction. Now, we can define the *wealth process*:

$$W_t(m) = W_{t-1}(m) \times (1 + \lambda_t(m)\,(Z_t - \mu_t(m)))\,,$$

with $W_0 = 1$. $(\lambda_t(m))_{t\in[N]}$ is a predictable sequence with values in $[0, 1/u_t(m)]$, and $u_t(m)$ is the largest value in the support of $Z_t - \mu_t(m)$, for each $t \in [N]$. Note that this constraint on $(\lambda_t(m))$ ensures that $W_t(m)$ is nonnegative for each $t \in [N]$. We also let $W_0(m) = 1$ for all $m \in [0, 1]$. If we view the wealth process as the wealth we earn from gambling on the outcome of $Z_t - \mu_t(m)$, then $(\lambda_t(m))$ represents a betting strategy, i.e., how much money to gamble each turn. Hence, we refer to $(\lambda_t(m))$ as a *betting strategy*.

It is easy to verify that $(W_t(m^*))$ is a nonnegative martingale for any sampling strategy (q_t) and betting strategy $(\lambda_t(m^*))$. Hence, it is unlikely to take large values, as we describe next.

Proposition 1. *For any sampling and betting strategies (q_t) and $(\lambda_t(m^*))$, the following holds:*

$$\mathbb{P}\left(\exists t \geq 1 : W_t(m^*) \geq 1/\delta\right) \leq \delta.$$

This is a consequence of Ville's inequality, first obtained by Ville [20], which is a time-uniform version of Markov's inequality for nonnegative supermartingales. This result immediately implies that for any sampling strategy, and any betting strategy, the term m^* must lie in the set

$$\mathcal{C}_t = \{m : W_t(m) < 1/\delta\} \tag{3}$$

with probability at least $1-\delta$, making $(\mathcal{C}_t)$ a $(1-\delta)$-CS.

Theorem 1. *$(\mathcal{C}_t)$ is an $(1-\delta)$-CS, where $\mathcal{C}_t$ defined by (3). Hence, a procedure that outputs $\mathcal{C}_\tau$ is an (ε, δ)-RLFA, for any sampling strategy (q_t) and betting strategies $(\lambda_t(m))$ for each $m \in [0, 1]$. Recall that the τ is defined in (1) as the first time where $|\mathcal{C}_t| \leq \varepsilon$.*

This methodology gives us flexible framework for constructing different $(\mathcal{C}_t)$ that result in different RLFAs. Now, we can turn our attention to finding betting strategies $(\lambda_t(m))$ that reduces the CS width quickly and minimizes τ.

2.1 POWERFUL BETTING STRATEGIES

Besides validity, we also want the size of the CS to shrink rapidly. This depends on how quickly the values of $W_t(m)$ for $m \neq m^*$ grow with t. One such criterion is to consider the *growth rate*, i.e., the expected logarithm of the outcome of each bet. We can define the *one-step growth rate* D_n, for each $n \in [N]$ as follows:

$$D_n(m, \lambda) := \log(1 + \lambda(Z_t - \mu_t(m))).$$

We are interested in maximizing the expected logarithm of the wealth process [7, 16], since it is equivalent to minimizing the expected time for a wealth process to exceed a fixed threshold (asymptotically, as the threshold grows larger) [5]. Thus, *in the context of the auditing problem, maximizing $\mathbb{E}[D_t(\lambda, m) \mid \mathcal{F}_{t-1}]$, approximately minimizes $\mathbb{E}[\tau]$*. The one-step growth rate is a broadly studied objective known as the "Kelly criterion" [10]. In general, finding the best sequence of bets $\lambda_t(m)$ for different values of n is non-tractable. Instead we consider the approximation $\log(1+x) \geq x - x^2$ for $|x| \leq 1/2$, and define the best constant bet λ_n^* in hindsight, as

$$B_t(m, \lambda) := \lambda\,(Z_t - \mu_t(m)) - \lambda^2\,(Z_t - \mu_t(m))^2\,, \tag{4}$$

$$\lambda_n^* := \operatorname*{argmax}_{\lambda\in[\pm 1/2c]} \frac{1}{n}\sum_{t=1}^{n} B_t(m, \lambda),$$

where $c = \max\{|Z_t - \mu_t(m)| : t \in [n]\}$. We get the following result on λ_n^* for each $n \in [N]$:

$$\lambda_n^* \propto \frac{\sum_{t=1}^{n} Z_t - \mu_t(m)}{\sum_{t=1}^{n}(Z_t - \mu_t(m))^2} := \frac{A_n}{V_n}.$$

Since λ_n^* depends on the nth sample itself, Z_n, we cannot use this strategy in our CS construction. Instead, at any $n \in [N]$, we can use a predictable approximation of this strategy, that we shall refer to as the `ApproxKelly` betting strategy. This strategy sets $\lambda_t(m)$ as follows:

$$\lambda_t(m) = c_t \frac{A_{t-1}}{V_{t-1}}, \qquad \texttt{(ApproxKelly)}$$

where the (predictable) factor c_t is selected to ensure that $\lambda_t(m) \times (Z_t - \mu_t(m)) \in (-1, \infty)$, i.e., to satisfy the nonnegativity constraint of $(W_t(m))$.

We note that we can use more accurate approximations of the term $\log(1+x)$, such as those derived by Fan et al. [6], Waudby-Smith and Ramdas [22], Ryu and Bhatt [14], to get better betting strategies. However, for simplicity, we focus on the `ApproxKelly` strategy in this paper.

2.2 LOGICAL CS

Irrespective of the choice of the sampling and betting strategies, we can construct a CS that contains m^* with probability 1, based on purely logical considerations. After sampling t transactions, we know that m^* is lower bounded by quantities derived from the the misstatement fraction accumulated in the items we have sampled already. Hence, we can derive the following deterministic bounds:

$$L_l(t) := \sum_{j=1}^{t} \pi(I_j) f(I_j) \leq m^*$$

$$U_l(t) := L_l(t) + \sum_{i\in\mathcal{U}_t} \pi(i) \geq m^*.$$

Note that $L_l(t)$ (resp. $U_l(t)$) values are obtained by noting that all the remaining unknown f values must be larger than 0 (resp. smaller than 1). Additionally, due to the

time-uniform nature of confidence sequences,we can intersect the logical CS with a 'probabilistic' CS constructed in (3), and obtain the following CS:

$$\widetilde{\mathcal{C}}_t := \mathcal{C}_t \cap [L_\ell(t), U_\ell(t)] \cap \widetilde{\mathcal{C}}_{t-1}, \tag{5}$$

where $\widetilde{C}_0 := [0, 1]$. Note that we may take the running intersection of a CS since it remains a CS, simply by definition. Consequently, the combined CS in (5) dominates the probabilistic CS.

3 SAMPLING STRATEGIES

The choice of the sampling strategy, (q_t), is also critical to reducing uncertainty about m^* quickly. Recall that q_t is a probability distribution on the remaining indices $\mathcal{N}_t$ for each $t \in [N]$. To motivate the choice of our sampling strategy, we first consider the following question: *what is the randomized sampling strategy that leads to the fastest reduction in uncertainty about m^*?*

In general, it is difficult to characterize this strategy in closed form (other than the computational aspect of the strategy being the solution of a multistage optimization problem). Thus, we consider a simplified question, that of finding the sampling strategy that maximizes the expectation of the one-step growth rate, $D_n(\lambda, m)$, for each $n \in [N]$. We seek to maximize the lower bound, $B_n(\lambda, m)$, introduced in (4):

$$q_n^* := \operatorname*{argmax}_{q \in \Delta^{\mathcal{N}_n}} \mathbb{E}_{I_n \sim q} \left[B_n(\lambda, m) \right], \tag{6}$$

where $\Delta^{\mathcal{N}_n}$ is the universe of distributions supported on $\mathcal{N}_n$. Our next result presents a closed-form characterization of q_n^*.

Proposition 2. *Note that $q_n^* = \mathit{argmin}_{q \in \Delta^{\mathcal{N}_n}} \mathbb{V}_{I_n \sim q}[Z_n]$, which implies that $q_n^*(i) \propto \pi(i) f(i)$. Hence, for any valid betting strategy (λ_t) and sampling strategy (q_t), we have $\mathbb{E}_{I \sim q_t}[B_t(\lambda_t, m)] \leq \mathbb{E}_{I \sim q_t^*}[B_t(\lambda_t, m)]$.*

We defer the proof to Appendix A.1, which proceeds by showing that maximizing the lower bound on the one-step growth rate is equivalent to minimizing the variance of Z_n. It turns out that $q_n^*(i) \propto \pi(i) f(i)$ is the minimum (in fact, zero) variance sampling distribution, and thus, (q_t^*) dominates any other sampling strategy w.r.t. maximizing the expected bound on the one-step growth rate.

Remark 1. *The oracle strategy in Proposition 2 can be considered as a solution of an alternative question: suppose there is an oracle who knows the true values of $f(i)$, and needs to convince an observer that the value m^* is within an interval of width ε with probability at least $1-\delta$. The oracle wishes to do so by revealing as few $f(i)$ values to the observer as possible. Clearly, any deterministic sampling strategy from the oracle will lead to skepticism from the observer (i.e., the observer will only be convinced once the $\pi(i)$ corresponding to the unrevealed $f(i)$ sum to ε). Hence, the sampling strategy used by the oracle must be random, and according to Proposition 2, it should draw transactions with probability $\propto \pi(i) \times f(i)$.*

Sampling without side information. Since the $(f(i))$ values are unknown by definition of the problem, we cannot use (q_t^*) in practice. Instead, we consider a sampling strategy that selects a index $i \in \mathcal{N}_t$ in proportion to its $\pi(i)$ value — we refer to this strategy as the `prop-M` strategy. This strategy is also known as "sampling proportional to size" in auditing literature [4], and is similar to the best deterministic strategy, which queries indices in descending order w.r.t. $\pi(i)$.

$$q_t(i) = \frac{\pi(i)}{\sum_{j \in \mathcal{N}_t} \pi(j)}, \tag{prop-M}$$

for each $i \in \mathcal{N}_t$. Sampling with `prop-M` minimizes the "worst case" support range, and max value, of Z_t. This allows for the largest possible choice of λ_t, i.e., our bet.

Using accurate side information for sampling. Proposition 2 motivates a natural sampling strategy in situations where we have access to side information $(S(i))$ that is known to be a high-fidelity approximation of the true $(f(i))$ values—draw indices proportional to $\pi(i) \times S(i)$. We will refer to this strategy as the `prop-MS` strategy:

$$q_t(i) = \frac{\pi(i) S(i)}{\sum_{j \in \mathcal{N}_t} \pi(j) S(j)}. \tag{prop-MS}$$

Under certain relative accuracy guarantees on the side information, we can characterize the performance achieved by the `prop-MS` strategy as compared to the optimal strategy of Proposition 2, as we state next.

Corollary 1. *Assume that the side information, $(S(i))$, is an accurate prediction of $(f(i))$, i.e., there exists a known parameter $a \in [0, 1)$, such that*

$$S(i)/f(i) \in [1 \pm a] \tag{7}$$

for all $i \in [N]$. With the `prop-MS` strategy for (q_t), we can ensure $\mathbb{E}_{I_t \sim q_t}[B_t(\lambda_t, m)] \geq \mathbb{E}_{I_t \sim q_t^*}[B_t(\lambda_t, m)] \left(\frac{1}{1+a}\right)^2$, where (q_t^*) is the optimal sampling strategy of Proposition 2.

In many cases, we do not have the accuracy guarantees on the side information required by Corollary 1, and we develop an approach to properly incorporate such side information in the next section.

4 USING POSSIBLY INACCURATE SIDE INFORMATION

Often, we do not have a uniform guarantee on accuracy on $(S(i))$ as we assumed in the previous section. In such cases, we cannot continue to use the `prop-MS` strategy, as it requires knowledge of the range of $f(i)/S(i)$ in order to select the betting fractions that ensure non-negativity of the process $(W_t(m))$. Nevertheless, we develop new techniques in this section that can exploit the side information without the uniform accuracy guarantees, provided that the side information is correlated with the unknown $(f(i))$ values. In particular,

the method developed in this section for incorporating the side information is orthogonal to the choice of the sampling strategy; and thus, it can be combined with any sampling strategy that ensures the non-negativity of the process $(W_t(m))$.

Our approach is based on the idea of control variates [1, § V.2] that are used to reduce the variance of Monte Carlo (MC) estimates of an unknown quantity, using some correlated side information whose expected value is known. More specifically, let $\widehat{m}$ denote an unbiased estimate of an unknown parameter m, and let $\widehat{v}$ denote another (possibly correlated to $\widehat{m}$) statistic with zero mean. Then, the new statistic, $\widehat{m}_\beta = \widehat{m} + \beta\widehat{v}$ is also an unbiased estimate of m, for all $\beta \in \mathbb{R}$. Furthermore, it is easy to check that $\mathbb{V}(\widehat{m}_\beta) = \mathbb{V}(\widehat{m}) + \beta^2\mathbb{V}(\widehat{v}) + 2\beta\mathrm{Cov}(\widehat{m}, \widehat{v})$, which implies that the variance of this new estimate is minimized at $\beta = \beta^* := -(\mathrm{Cov}(\widehat{m}, \widehat{v})/\mathbb{V}(\widehat{v}))$. Finally, note that the variance of $\widehat{m}_{\beta^*}$ cannot be larger than the variance of the original estimate $\widehat{m}$, since $\mathbb{V}(\widehat{m}_{\beta^*}) \le \mathbb{V}(\widehat{m}_0) = \mathbb{V}(\widehat{m})$ by the definition of β^*.

Returning to our problem, given some possibly inaccurate side information $(S(i))$, define the control variate (that is, an analog of the term $\widehat{v}$) as

$$U_t := S(I_t) - \mathbb{E}_{I' \sim q_t}[S(I')],$$

and let (β_t) denote a sequence of predictable terms taking values in $[-1, 1]$ used to weigh the effect of (U_t). Note that, similar to $\widehat{v}$, the term U_t has zero mean for each $t \in [N]$. We now define the wealth process with control variates, denoted by $(\widetilde{W}_t(m))$, and its corresponding CS as follows:

$$\widetilde{W}_t(m) := \prod_{t=1}^{n} \left(1 + \lambda_t(m)(Z_t + \beta_t U_t - \mu_t(m))\right),$$
$$\mathcal{C}_t = \{m \in [0,1] : \widetilde{W}_n(m) < 1/\alpha\}, \tag{8}$$

where $(\lambda_t(m))$ is a betting strategy for each $m \in [0, 1]$.

Theorem 2. *For any set of side information $(S(i))$, sequence (β_t), sampling strategy (q_t), and betting strategies $(\lambda_t(m))$, $(\mathcal{C}_t)$ as defined in (8) is an $(1-\delta)$-CS for m^*. Consequently, the procedure that produces $\mathcal{C}_\tau$ is an (ε, δ)-RLFA.*

The discussion above suggests that by a suitable choice of the parameters (β_t), we can reduce the variance of the first term. To see why this is desirable, recall that the optimal value of the approximate growth rate after n steps of the new wealth process satisfies the following:

$$\widetilde{B}_n(\lambda, m) := \lambda(Z_t + \beta_t U_t - \mu_t(m)) - \lambda^2(Z_t + \beta_t U_t - \mu_t(m))^2,$$
$$\max_\lambda \widetilde{B}_n(\lambda, m) \propto \frac{\sum_{t=1}^n Z_t + \beta_t U_t - \mu_t(m)}{\sum_{t=1}^n (Z_t + \beta_t U_t - \mu_t(m))^2}.$$

Note that by setting $\beta_t = 0$ for all $t \in [n]$, we recover $\widetilde{B}_n(\lambda, m) = B_n(\lambda, m)$, i.e., the wealth lower bound with no side information. Next, we observe that $\sum_{t=1}^n \beta_t U_t$ concentrates strongly around its mean (0).

Proposition 3. *For any $\delta \in (0, 1)$ and sequence (β_t), the following statement is simultaneously true for all $n \in [N]$ with probability at least $1 - \delta$*

$$\left|\frac{1}{n}\sum_{t=1}^n \beta_t U_t\right| = \mathcal{O}\left(\sqrt{\frac{\log(\log n/\delta)}{n}}\right).$$

This result, proved in Appendix A.2, implies that in order to select the parameters (β_t), we can focus on its effect on the second order term in the denominator. In particular, the best value of β for the first n observations, is the one that minimizes the denominator, and can be defined as follows:

$$\beta_n^* := \operatorname*{argmin}_{\beta \in [-1,1]} \sum_{t=1}^n (Z_t - \mu_t(m) + \beta U_t)^2 \propto -\frac{\sum_{t=1}^n (Z_t - \mu_t(m))U_t}{\sum_{t=1}^n U_t^2}.$$

The numerator of β_n^* varies with $\sum_{t=1}^n f(I_t)S(I_t)$—hence, the magnitude of β_t increases with the amount of correlation between $f(i)$ and $S(i)$. Since β_n^* is not predictable (it is $\mathcal{F}_n$ instead of $\mathcal{F}_{n-1}$ measurable), we will use the following strategy of approximating β_n^* at each $n \in [N]$:

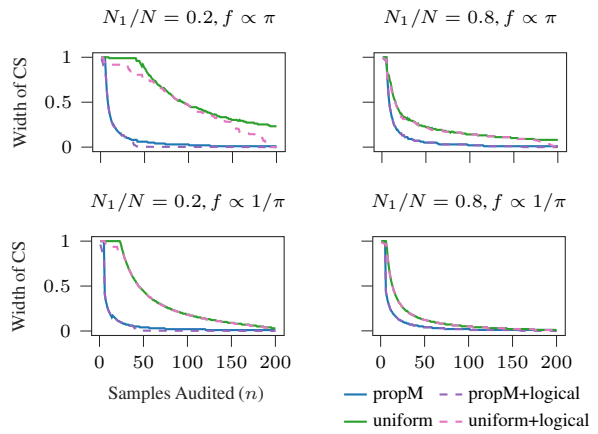
$$\beta_n \propto -\frac{\sum_{t=1}^{n-1} (Z_t - \mu_t(m))U_t}{\sum_{t=1}^{n-1} U_t^2}.$$

for $n \ge 2$ and we let $\beta_1 = 0$. This provides a principled way of incorporating side information even when the relationship between the side information and the ground truth is unclear.

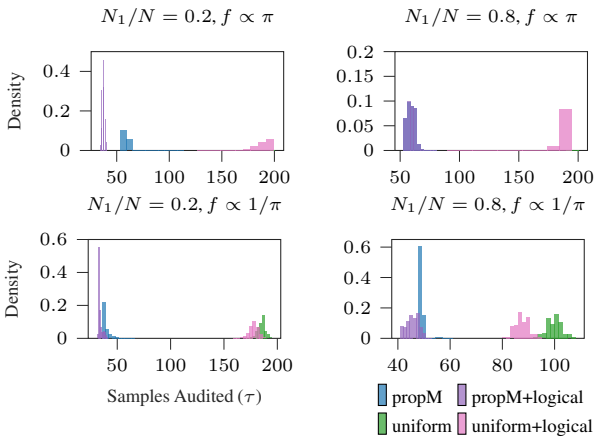
5 EXPERIMENTS

We conduct simulations of our RLFA methods on a variety of scenarios for π and f. For each simulation setup, we choose two positive integers N_{lg} and N_{sm} such that $N_{\text{lg}} + N_{\text{sm}} = N$. We generate the weight distribution π, consisting of N_{lg} 'large' values and N_{sm} 'small' values. The exact range of values taken by these terms are varied across experiments, but on an average the ratio of 'large' over 'small' π values lie between 10 and 10^3. We then generate the f values in one of two ways: (1) $f \propto \pi$, where indices with where large π values take f values in $[0.4, 0.5]$ and small π values take on f values in $[0.001, 0.01]$, or (2) $f \propto 1/\pi$, where the f value ranges are swapped for large and small values. The simulations in this section focus on the different sampling strategies as well as the efficacy of control variates — we provide additional experiments comparing the betting CS with other types of CS in Appendix E.

No side information: uniform vs. `prop-M` sampling. In the first experiment, we compare the performance of the `prop-M` strategy with the uniform baseline. In addition to this, we also illustrate the significance of logical CS (introduced in Section 2.2) especially in cases when there are a few large π values. From the widths of the CSs plotted in Figure 1a, we can see that `prop-M` outperforms the uniform



(a) Width of CSs

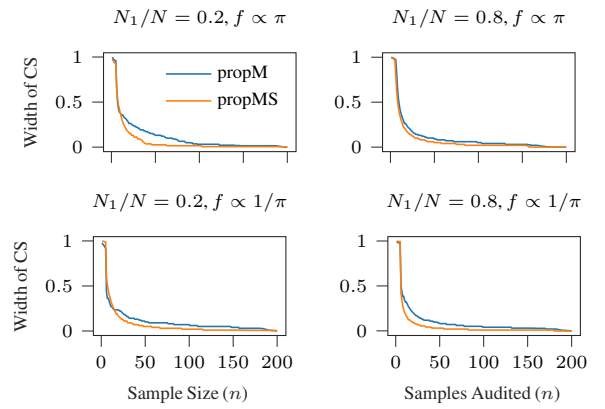


(b) Distribution of samples audited (τ). We omit the uniform (without logical CS) CS histograms, as they concentrated entirely at N.

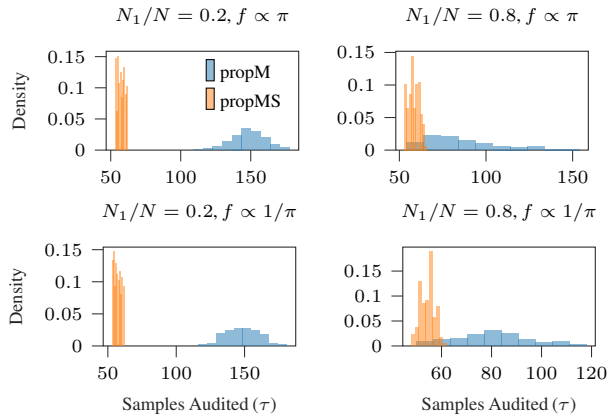
Figure 1: A comparison of `prop-M` vs. uniform sampling distributions, and the impact of intersecting with the logical CS (Section 2.2) where $\varepsilon = \delta = 0.05$. The `prop-M` strategy produces tighter CSs that results in fewer audited samples. Intersecting with the logical CS further reduces the width, particularly when few transactions are large ($N_{\text{lg}} = 0.2$).

baseline in all four cases. The gap in performance increases when N_{lg} is small since π deviates more significantly from the uniform weighting: it consists of a few large weights with the rest close to 0. On the other hand, when N_{lg} is large, the weights resemble the uniform distribution, leading to the competitive performance of the uniform baseline. The logical CSs are most useful in the case of small N_{lg}, especially with $f \propto \pi$. This is because for small N_{lg}, every query to an index with large π value leads to a significant reduction in the uncertainty about m^*.

Next, in Figure 1b, we plot the distribution of the stopping time τ for an RLFA with $\varepsilon = \delta = 0.05$, over 500 independent trials. The `prop-M` strategy leads to a significant reduction in the sample size requirement to obtain an ε-accurate estimate of m^* as compared to the uniform baseline, both with and without the logical CS. Furthermore, the distribution of τ with



(a) Width of CSs



(b) Distribution of samples audited (τ).

Figure 2: Comparison of the `prop-MS` vs. the `prop-M` strategy with accurate side information ($S(i)$), i.e., $S(i)/f(i) \in [0.9, 1.1]$ where $\varepsilon = \delta = 0.05$. We see that `prop-MS` outperforms `prop-M` in both CS width and sample efficiency.

the `prop-M` strategy often has less variability than the uniform strategy. Hence, `prop-M` has demonstrated itself empirically to be a better sampling strategy than simply sampling uniformly, as one would do when all the weights are equal.

Using `prop-MS` with accurate side information. In the second experiment, we study the benefit of incorporating accurate side information in the design of our CSs, by comparing the performance of `prop-MS` strategy with that of the `prop-M` strategy. We generate S randomly while ensuring $S(i)/f(i) \in [1 \pm a]$ (from (7)) for some $a \in (0, 1)$. Thus smaller values of a imply that the scores $S(i)$ are more accurate approximations of $f(i)$ for all $i \in [N]$.

In Figure 2a, we can see that the `prop-MS` strategy with accurate side information dominates the `prop-M` strategy. This is further reflected in the distribution of τ for an RLFA where $\varepsilon = \delta = 0.05$ in Figure 2b. Hence, in situations where we are confident in the accuracy of our side information, we should incorporate it directly into our sampling strategy to reduce the width of the CS.

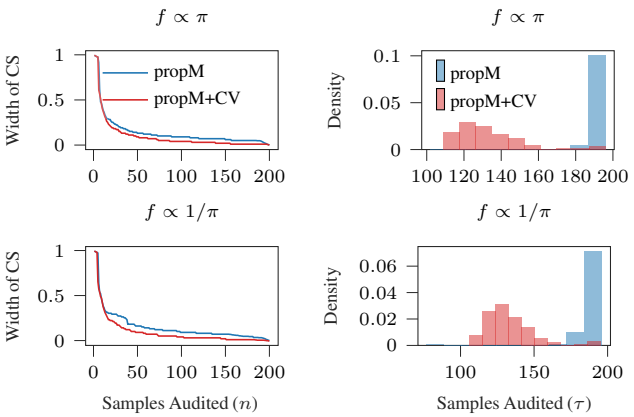


Figure 3: The plots above show the width of the CSs and the distribution of τ for the $f \propto \pi$ and the $f \propto 1/\pi$ cases, where $N_{\text{lg}}/N = 0.2$ and $c = 0.9$.

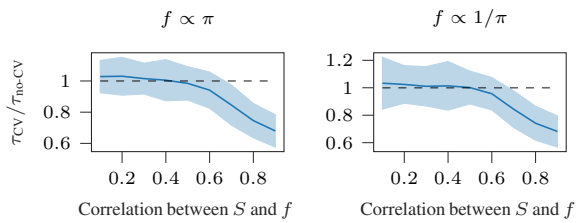


Figure 4: The figures plot the variation of the reduction in τ for an RLFA with $\varepsilon = 0.025, \delta = 0.05$, when the CS is constructed with and without using control variates . The x-axis denotes the parameter $c \in [0.1., 0.9]$, and thus controls the amount of correlation between S and f. As the amount of correlation between S and f increases, the CS with control variates decreases takes a decreasing fraction of the time it would take the CS w/o control variates.

Control variates from possibly inaccurate side information. Finally, we consider the case in which we do not have prior information about the accuracy of the side information. Thus, using the `prop-MS` strategy in this scenario directly can lead to very conservative CSs (this is because in the absence of tight guarantees on the range of the S/f ratio, we will have to use the worst case range). Instead, we compare the performance of the `prop-M` strategy, with and without using control variates described in Section 4. In this case, we set $S(i) = c \times f(i) + (1-c) \times R_i$ for $c \in (0, 1)$, where $(R_i)_{i \in [N]}$ are i.i.d.random variables distributed uniformly over $[0, 1]$. The parameter c controls the level of correlation between f and S values, with small c values indicating low correlation.

We generate the data with $N_{\text{lg}} = 40$ and $N = 200$. In Figure 3, we compare the CSs and the distribution of τ for an RLFA (with $\varepsilon = \delta = 0.05$) for the `prop-M` strategy with and without control variates, when the side information is generate with $c = 0.9$. Due to the high correlation, there is a significant decrease in the samples needed to reach an accuracy of ε, when using control variates.

Finally, in Figure 4, we study the variation in sample efficiency as the correlation between S and f changes (i.e., by varying c). In particular, for 9 linearly spaced c values in the range $[0.1, 0.9]$, we compute the τ for an RLFA without ($\tau_{\text{no-CV}}$) and with control variates (τ_{CV}) over 250 trials, and then plot the variation of the mean of their ratio, $\tau_{\text{CV}}/\tau_{\text{no-CV}}$.

Figure 4 highlights the key advantage of our CS construction using control variates — this method automatically adapts to the correlation between the side information and the f values. In cases where the side information is highly correlated (i.e., larger c values), the reduction in samples is large; whereas when the correlation is small, our approach automatically reduces the impact of the side information.

6 CONCLUSION

In this paper, we defined the concept of an (ε, δ)-RLFA and devised RLFA procedures from confidence sequences (CSs) for the weighted average of N terms (denoted by m^*), using adaptive randomized sampling WoR. First, for arbitrary randomized sampling strategies, we developed two methods of constructing CSs for m^* using test martingales. We then addressed the question of effectively incorporating side information, with or without guarantees on their accuracy, to improve the quality of the CSs constructed.

Our work opens up several interesting directions for future work. For instance, in Proposition 2, we characterized the sampling strategy that optimizes a lower bound on the one-step growth rate. Future work could investigate whether we can obtain a more complete characterization of the optimal policy, without relying on approximations like Proposition 2. Another interesting issue, not addressed in our paper is that of considering more general types of side information available to us. As described in Section 1, we have assumed that we have access to $[0, 1]$ valued side information that is supposed to be a proxy for the true (and unknown) f values. However, in practical auditing problems, the side information is usually available in terms of a collection of numeric, discrete and categorical features that are correlated with the unknown f values. Developing methods for incorporating these more realistic forms of side information into our framework for designing CSs is another important question for future work. Furthermore, another type of side information is any knowledge from a prior audit. For example, auditors may know before reviewing any data (transactions or AI-generated side-info) that for this year, some accounts are likely to have smaller or bigger f values than other accounts because of the specific performance incentives placed on the company managers by their supervisors or by the market conditions.

REFERENCES

[1] S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

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

[3] A. Ben-Hamou, Y. Peres, and J. Salez. Weighted sampling without replacement. *Brazilian Journal of Probability and Statistics*, 32(3):657–669, 2018.

[4] P. J. Bickel. Inference and auditing: the stringer bound. *International Statistical Review/Revue Internationale de Statistique*, pages 197–209, 1992.

[5] L. Breiman. Optimal Gambling Systems for Favorable Games. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 4.1:65–79, 1961.

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

[7] P. Grünwald, R. de Heide, and W. Koolen. Safe Testing. *arXiv:1906.07801*, 2020.

[8] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.

[9] 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.

[10] J. L. Kelly. A New Interpretation of Information Rate. *The Bell System Technical Journal*, page 10, 1956.

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

[12] M. Lindeman and P. B. Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10 (5):42–49, 2012.

[13] A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020.

[14] J. J. Ryu and A. Bhatt. On confidence sequences for bounded random processes via universal gambling strategies. *arXiv:2207.12382*, 2022.

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

[16] 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.

[17] P. B. Stark. Conservative statistical post-election audits. *The Annals of Applied Statistics*, 2(2):550–581, 2008.

[18] P. B. Stark. Cast: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security*, 4(4):708–717, 2009.

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

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

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

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

[23] I. Waudby-Smith, P. B. Stark, and A. Ramdas. Rilacs: Risk limiting audits via confidence sequences. In *International Joint Conference on Electronic Voting*, pages 124–139. Springer, 2021.

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