RiLACS: Risk-limiting audits via confidence sequences

Ian Waudby-Smith[1], Philip Stark[2], and Aaditya Ramdas[1]

Carnegie Mellon University[1]
University of California, Berkeley[2]

ianws@cmu.edu, stark@stat.berkeley.edu, aramdas@cmu.edu

May 28, 2021

Abstract

Accurately determining the outcome of an election is a complex task with many potential sources of error, ranging from software glitches in voting machines to procedural lapses to outright fraud. Risk-limiting audits (RLA) are statistically principled "incremental" hand counts that provide statistical assurance that reported outcomes accurately reflect the validly cast votes. We present a suite of tools for conducting RLAs using confidence sequences — sequences of confidence sets which uniformly capture an electoral parameter of interest from the start of an audit to the point of an exhaustive recount with high probability. Adopting the SHANGRLA [1] framework, we design nonnegative martingales which yield computationally and statistically efficient confidence sequences and RLAs for a wide variety of election types.

1 Introduction

The reported outcome of an election may not match the validly cast votes for a variety of reasons, including software configuration errors, bugs, human error, and deliberate malfeasance. Trustworthy elections start with a trustworthy paper record of the validly cast votes. Given access to a trustworthy paper trail of votes, a risk-limiting audit (RLA) can provide a rigorous probabilistic guarantee:

1. If an initially announced assertion $\mathcal{A}$ about an election is *false*, this will be corrected by the audit with high probability;
2. If the aforementioned assertion $\mathcal{A}$ is *true*, no "correction" will take place, and the original assertion will be confirmed (with probability one).

Here, an electoral assertion $\mathcal{A}$ is simply a claim about the aggregated votes cast (e.g. "Alice received more votes than Bob"). An auditor may wish to audit several claims: for example, whether the reported winner is correct or whether the margin of victory is as large as announced.

From a statistical point of view, efficient risk-limiting audits can be implemented as sequential hypothesis tests. Namely, one tests the null hypothesis H_0: "the assertion $\mathcal{A}$ is false," versus the alternative H_1: "the assertion $\mathcal{A}$ is true". Imagine then observing a random sequence of voter-cast ballots $X_1, X_2, \ldots, X_N$, where N is the total number of ballots. A sequential hypothesis test is represented by a sequence $(\phi_t)_{t=1}^N$ of binary-valued functions:

$$\phi_t := \phi(X_1, \ldots, X_t) \mapsto \{0, 1\},$$

where $\phi_t = 1$ represents rejecting H_0 (typically in favor of H_1), and $\phi_t = 0$ means that H_0 has not yet been rejected. The sequential test (and thus the RLA) stops as soon as $\phi_t = 1$ or once all N ballots are observed, whichever comes first. The "risk-limiting" property of RLAs states that if the assertion is false (in other words, if H_0 holds), then

$$\mathbb{P}_{H_0}\left(\exists t \in \{1, \ldots, N\} : \phi_t = 1\right) \leq \alpha,$$

which is equivalent to type-I error control of the sequential test. Another way of interpreting the above statement is as follows: if the assertion is incorrect, then with probability at least $(1 - \alpha)$, $\phi_t = 0$ for every $t \in \{1, \ldots, N\}$ and hence all N ballots will eventually be inspected, at which point the "true" outcome (which is the result of the full hand count) will be known with certainty.

1.1 SHANGRLA reduces election auditing to sequential testing

Designing the sequential hypothesis test $(\phi_t)_{t=1}^N$ depends on the type of vote, the aggregation method, or the social choice function for the election, and thus past works have constructed a variety of tests. Some works have designed $(\phi_t)_{t=1}^N$ in the context of a particular type of election [2, 3, 4]. On the other hand, the "SHANGRLA" (**S**ets of **H**alf-**A**verage **N**ulls **G**enerate **RLA**s) framework unifies many common election types including plurality elections, approval voting, ranked-choice voting, and more by reducing each of these to a simple hypothesis test of whether a finite collection of finite lists of bounded numbers has mean $\mu^\star$ at most $1/2$ [1]. Let us give an illustrative example to show how SHANGRLA can be used in practice.

Suppose we have an election with two candidates, Alice and Bob. A ballot may contain a vote for Alice or for Bob, or it may contain no valid vote, e.g., because there was no selection or an overvote. It is reported that Alice and Bob received N_A and N_B votes respectively with $N_A > N_B$ and that there were a total of N_I invalid ballots for a total of $N = N_A + N_B + N_I$ voters. We encode votes for Alice as "1", votes for Bob as "0" and invalid votes as "1/2", to obtain a set of numbers $\{x_1, x_2, \ldots, x_N\}$. Crucially, Alice indeed received more votes than Bob if and only if $\mu^\star := \frac{1}{N}\sum_{i=1}^N x_i > 1/2$. In other words, *the report that Alice beat Bob can be translated into the assertion that* $\mu^* \in (1/2, 1]$.

SHANGRLA proposes to audit an assertion by testing its complement: rejecting that "complementary null" is affirmative evidence that the assertion is indeed true. In other words, if one can ensure that $X_1, X_2, \ldots, X_N$ is a random permutation of $\{x_1, \ldots, x_N\}$ by sampling ballots without replacement (each ballot is chosen uniformly amongst remaining ballots), then we can concern ourselves with designing a hypothesis test $(\phi_t)_{t=1}^N$ to test the null $H_0 : \mu^\star \leq 1/2$ against the alternative $H_1 : \mu^\star > 1/2$.

One of the major benefits of SHANGRLA is the ability to reduce a wide range of election types to a testing problem of the above form. This permits the use of powerful statistical techniques which were designed specifically for such testing problems (but may not have been designed with RLAs in mind). Throughout this paper, we adopt the SHANGRLA framework, and while we return to the example of plurality elections for illustrative purposes, all of our methods can be applied to any election audit which has a SHANGRLA-like testing reduction [1].

1.2 Confidence sequences

In the fixed-time (i.e. non-sequential) hypothesis testing regime, there is a well-known duality between hypothesis tests and confidence intervals for a parameter $\mu^\star$ of interest. We describe this briefly for $\mu^\star \in [0,1]$ for simplicity. For each $\mu \in [0,1]$, suppose that $\phi^\mu \equiv \phi^\mu(X_1, \ldots, X_n) \mapsto \{0,1\}$ is a level-α nonsequential, fixed-sample test for the hypothesis $H_0 : \mu^\star = \mu$ versus $H_1 : \mu^\star \neq \mu$. Then, a nonsequential, fixed-sample $(1-\alpha)$ confidence interval for $\mu^\star$ is given by the set of all $\mu \in [0,1]$ for which ϕ^μ does not reject, that is $\{\mu \in [0,1] : \phi^\mu = 0\}$.

As we discuss further in Section 2, an analogous duality holds for sequential hypothesis tests and *confidence sequences*. We first give a brief preview of the results to come. Consider a family of sequential hypothesis tests $\{(\phi_t^\mu)_{t=1}^N\}_{\mu \in [0,1]}$, meaning that for each μ, $(\phi_t^\mu)_{t=1}^N$ is a sequential test for μ. Then, the set of all μ for which $\phi_t^\mu = 0$,

$$C_t := \{\mu \in \mathbb{R} : \phi_t^\mu = 0\}$$

forms a $(1-\alpha)$ *confidence sequence* for $\mu^\star$, meaning that

$$\mathbb{P}(\exists t \in [N] : \mu^\star \notin C_t) \leq \alpha,$$

where $[N]$ is used to denote the set $\{1, 2, \ldots, N\}$. Since C_t is typically an interval $[L_t, U_t]$, we call the lower endpoint $(L_t)_{t=1}^N$ as a lower confidence sequence (and similarly for upper).

In particular, given the sequential hypothesis testing problem that arises in SHANGRLA, we can cast the RLA as a sequential estimation problem that can be solved by developing confidence sequences. As we will see in Section 2, our confidence sequences provide added flexibility and an intuitive visualizable interpretation for SHANGRLA-compatible election audits, without sacrificing any statistical efficiency.

1.3 Contributions and outline

The contributions of this work are twofold. First, we introduce confidence sequences to the election auditing literature as intuitive and flexible ways of interpreting and visualizing risk-limiting audits. Second, we present algorithms for performing RLAs based on confidence sequences by deriving statistically and computationally efficient nonnegative martingales. There is nothing lost in this perspective: confidence sequences can also be used as hypothesis tests, yielding RLAs which generalize and often outperform the current state-of-the-art.

In Section 2, we show how confidence sequences generate risk-limiting audits, how they relate to more familiar RLAs based on sequentially valid p-values, and how they can be used to audit multiple contests. Section 3 derives novel confidence sequence-based RLAs and compares them to past RLA methods via simulation. Finally, Section 4 discusses how all of the aforementioned results apply to risk-limiting tallies for coercion-resistant voting schemes.

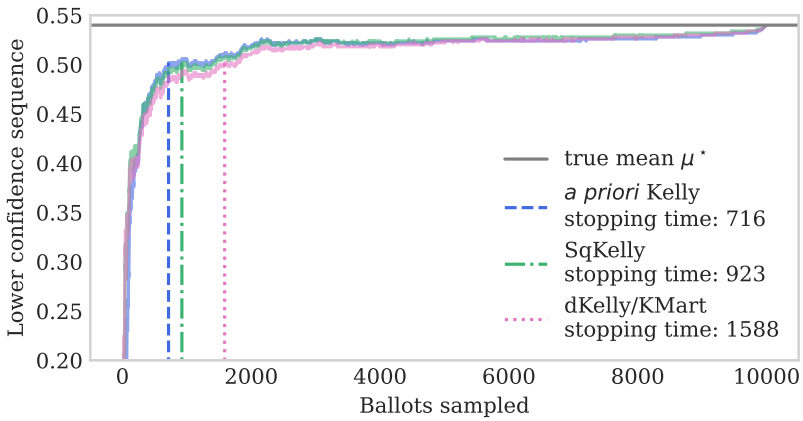

Figure 1: 95% Lower confidence sequences for the margin of a plurality election between Alice and Bob for three different auditing methods. Votes for Alice are encoded by "1" and those for Bob are encoded by "0". The parameter of interest is then the average of these votes, which in this particular example is 54% (given by the horizontal grey line). The outcome is verified once the lower confidence sequence exceeds 1/2. The time at which this happens is given by the vertical blue, green, and pink lines.

2 Confidence sequences are risk-limiting

Consider an election consisting of N ballots. Following SHANGRLA [1], suppose that these can be transformed to a set of $[0, u]$-bounded real numbers $x_1, \ldots, x_N \in [0, u]$ with mean $\mu^\star := \frac{1}{N}\sum_{i=1}^{N} x_i$ for some known $u > 0$. Suppose that electoral assertions can be made purely in terms of $\mu^\star$. A classical $(1-\alpha)$ confidence interval CI_n for $\mu^\star$ is an interval computed from data $X_1, X_2, \ldots, X_n$ with the guarantee that

$$\forall n \in [N],\ \mathbb{P}(\mu^\star \in \mathrm{CI}_n) \geq 1 - \alpha.$$

In contrast, a $(1-\alpha)$ *confidence sequence* for $\mu^\star$ is a sequence of confidence sets, $C_1, C_2, \ldots, C_N$ which all simultaneously capture $\mu^\star$ with probability at least $(1-\alpha)$. That is,

$$\underbrace{\mathbb{P}(\forall t \in [N],\ \mu^\star \in C_t) \geq 1 - \alpha}_{\text{simultaneous coverage probability}}, \quad \text{or equivalently} \quad \underbrace{\mathbb{P}(\exists t \in [N] : \mu^\star \notin C_t) \leq \alpha}_{\text{error probability}}.$$

The two probabilistic statements above are equivalent, but provide a different way of interpreting α and the corresponding guarantee.

If we have access to a $(1-\alpha)$ confidence sequence for $\mu^\star$, we can audit any assertion about the election outcome made in terms of $\mu^\star$ with risk limit α. Here, we use $\mathcal{A} \subseteq [0, u]$ to denote an assertion. For example, SHANGRLA typically uses assertions of the form "$\mu^\star$ is greater than 1/2", in which case $\mathcal{A} = (1/2, u]$.

Algorithm .1: Risk limiting audits via confidence sequences (RiLACS)

Input: Assertion $\mathcal{A} \subseteq [0, u]$, risk limit $\alpha \in (0, 1)$.
  for $t \in [N]$ **do**
    Compute $C_t \equiv C(X_1, \ldots, X_t)$ at level α.
    if $\mathcal{A} \subseteq C_t$ **then**
      Certify the assertion $\mathcal{A}$ and stop if desired.
    end if
  end for

The following theorem summarizes the risk-limiting guarantee of the above algorithm.

Theorem 1. *Let $(C_t)_{t=1}^N$ be a $(1-\alpha)$ confidence sequence for $\mu^\star$. Let $\mathcal{A} \subseteq [0, u]$ be an assertion about the electoral outcome (in terms of $\mu^\star$). The audit mechanism that certifies $\mathcal{A}$ as soon as $C_t \subseteq \mathcal{A}$ has risk limit α.*

Proof. We need to prove that if $\mu^\star \notin \mathcal{A}$, then $\mathbb{P}(\exists t \in [N] : C_t \subseteq \mathcal{A}) \leq \alpha$. First, notice that if $C_t \subseteq \mathcal{A}$, then we must have that $\mu^\star \notin C_t$ since $\mu^\star \notin \mathcal{A}$. Then,

$$\begin{aligned}\mathbb{P}(\exists t \in [N] : C_t \subseteq \mathcal{A}) &\leq \mathbb{P}(\exists t \in [N] : \mu^\star \notin C_t) \\ &\leq \alpha,\end{aligned}$$

where the second inequality follows from the definition of a confidence sequence. This completes the proof. □

Let us see how this theorem can be used in an example. Consider an election with two candidates, Alice and Bob, and a total of N cast ballots. Let $\{x_1, \ldots, x_N\}$ be the list of numbers that result from encoding votes for Alice as 1, votes for Bob as 0, and ballots that do not contain a valid vote as 1/2. Let $(C_t)_{t=1}^N$ be a $(1-\alpha)$ confidence sequence for $\mu^\star := \frac{1}{N}\sum_{i=1}^N x_i$. If we wish to audit the assertion that "Alice beat Bob", then $u = 1$ and $\mathcal{A} = (1/2, 1]$. We can sequentially sample $X_1, X_2, \ldots, X_N$ without replacement, certifying the assertion once $C_t \subseteq \mathcal{A}$. By Theorem 1, this limits the risk to level α.

2.1 Relationship to sequential hypothesis testing

The earliest work on RLAs did not use anytime p-values [5, 6], but since about 2009, most RLA methods have used anytime p-values to conduct sequential hypothesis tests [7, 8, 3, 1, 9]. An anytime p-value is a sequence of p-values $(p_t)_{t=1}^N$ with the property that under some null hypothesis H_0,

$$\mathbb{P}_{H_0}(\exists t \in [N] : p_t \leq \alpha) \leq \alpha. \tag{1}$$

The anytime p-values $p_t \equiv p_t(\mu)$ are typically defined implicitly for each null hypothesis $H_0 : \mu^\star = \mu$ and yield a sequential hypothesis test $\phi_t^\mu := \mathbb{1}(p_t(\mu) \leq \alpha)$. As alluded to in Section 1.2, this immediately recovers a confidence sequence:

$$C_t := \{\mu \in [0, u] : \phi_t^\mu = 0\}.$$

Notice in Figure 2 that the times at which nulls are rejected (or "stopping times") are the same for both confidence sequences and the associated p-values. Thus, nothing is lost by basing the RLA on

confidence sequences rather than anytime p-values. Confidence sequences benefit from being visually intuitive and are arguably easier to interpret than anytime p-values.

For example, consider conducting an RLA for a simple two-candidate election between Alice and Bob with no invalid votes. Suppose that it is reported that Alice won, i.e., $\mu^\star := \frac{1}{N}\sum_{i=1}^{N} x_i > 1/2$ where $x_i = 1$ if the ith ballot is for Alice, 0 if for Bob, and 1/2 if the ballot does not contain a valid vote for either candidate. A sequential RLA in the SHANGRLA framework would posit a null hypothesis $H_0 : \mu^\star \leq 1/2$ (the complement of the announced result: Bob actually won or the outcome is a tie), sample random ballots sequentially, and stop the audit (confirming the announced result) if and when H_0 is rejected at significance level α. If H_0 is not rejected before all ballots have been inspected, the true outcome is known.[1]

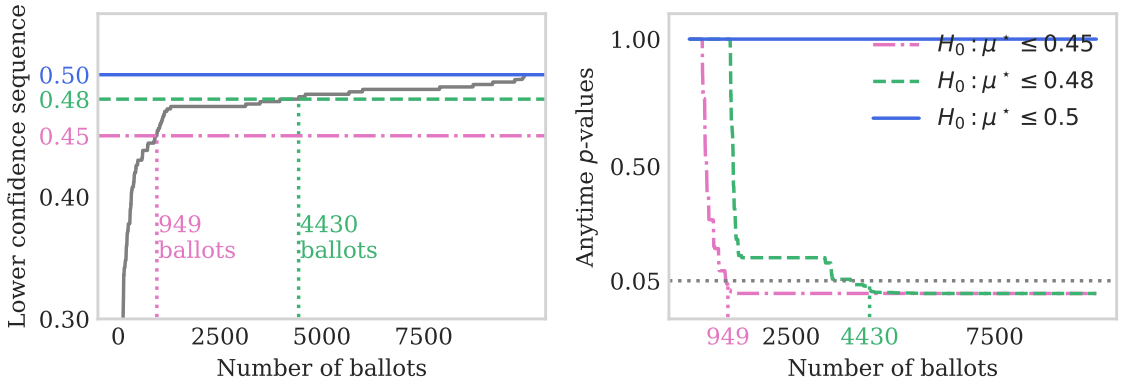

Figure 2: The duality between anytime p-values and confidence sequences for three nulls: $H_0 : \mu^\star \leq \mu_0$ for $\mu_0 \in \{0.45, 0.48, 0.5\}$. The p-value for $H_0 : \mu^\star \leq 0.45$ (pink dash-dotted line) drops below 5% after 949 samples, exactly when the 95% *lower* confidence sequence exceeds 0.45. However, the p-value for $H_0 : \mu^\star \leq 0.5$ never reaches 0.05 and the 95% confidence sequence never excludes 0.5, the true value of $\mu^\star$.

On the other hand, a ballot-polling RLA [2] based on confidence sequences proceeds by computing a lower $1-\alpha$ confidence bound for the fraction $\mu^\star$ of votes for Alice. The audit stops, confirming the outcome, if and when this lower bound is larger than 1/2. If that does not occur before the last ballot has been examined, the true outcome is known. In this formulation, there is no need to define a null hypothesis as the complement of the announced result and interpret the resulting p-value, and so on. The approach also works for comparison audits using the "overstatement assorter" approach developed in [1], which transforms the problem into the same canonical form: testing whether the mean of any list in a collection of nonnegative, bounded lists is less than 1/2.

2.2 Auditing multiple contests

It is known that RLAs of multi-candidate, multi-winner elections can be reduced to several pairwise contests without adjusting for multiplicity [2]. This is accomplished by testing whether every single

[1] At any point during the sampling, an election official can choose to abort the sampling and perform a full hand count for any reason. This cannot increase the risk limit: the chance of failing to correct an incorrect reported outcome does not increase.

reported winner beat every single reported loser, and stopping once each of these tests rejects their respective nulls at level $\alpha \in (0,1)$. For example, suppose it is reported that a set of candidates $\mathcal{W}$ beat a set of candidates $\mathcal{L}$ in a k-winner plurality contest with K candidates in all (that is, $|\mathcal{W}| = k$ and $|\mathcal{L}| = K - k$). For each reported winner $w \in \mathcal{W}$ and each reported loser $\ell \in \mathcal{L}$, encode votes for candidate w as "1", votes for ℓ as "0" and ballots with no valid vote in the contest or with a vote for any other candidate as "1/2" to obtain the population $\{x_1^{w,\ell}, \ldots, x_N^{w,\ell}\}$. Then as before, candidate w beat candidate ℓ if and only if $\mu^\star_{w,\ell} := \frac{1}{N}\sum_{i=1}^N x_i^{w,\ell} > 1/2$. In a two-candidate plurality election we would have proceeded by testing the null $H_0^{w,\ell} : \mu^\star_{w,\ell} \leq 1/2$ against the alternative $H_1^{w,\ell} : \mu^\star_{w,\ell} > 1/2$. To use the decomposition of a single winner or multi-winner plurality contest into a set of pairwise contests, we test each null $H_0^{w,\ell} : \mu^\star_{w,\ell} \leq 1/2$ for $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$. The audit stops if and when *all* $k(K-k)$ *null hypotheses* are rejected. Crucially, if candidate $w \in \mathcal{W}$ did not win (i.e. $\mu^\star_{w,\ell} \leq 1/2$ for some $\ell \in \mathcal{L}$), then

$$\mathbb{P}(\text{reject all } H_{0,w,\ell} : w \in \mathcal{W}, \ell \in \mathcal{L}) \leq \min_{w \in \mathcal{W}, \ell \in \mathcal{L}} \mathbb{P}(\text{reject } H_{0,w,\ell}) \leq \alpha.$$

The same technique applies when auditing with confidence sequences. Let $\{(C_t^{w,\ell})_{t=1}^N\}$ be $(1-\alpha)$ confidence sequences for $\{\mu^*_{w,\ell}\}$, $w \in \mathcal{W}$, $\ell \in \mathcal{L}$. We verify the electoral outcome of every contest once $C_t^{w,\ell} \subseteq (1/2, u]$ for all $w \in \mathcal{W}$, $\ell \in \mathcal{L}$. Again, if $\mu^\star_{w,\ell} \leq 1/2$ for some $w \in \mathcal{W}$, and $\ell \in \mathcal{L}$, then

$$\begin{aligned} &\mathbb{P}(\forall w \in \mathcal{W}, \forall \ell \in \mathcal{L},\ C_t^{w,\ell} \subseteq (1/2, u]) \\ \leq &\min_{w \in \mathcal{W}, \ell \in \mathcal{L}} \mathbb{P}(C_t^{w,\ell} \subseteq (1/2, u]) \leq \alpha. \end{aligned}$$

This technique can be generalized to handle audits of any number of contests from the same audit sample, as explained in [1]. For the sake of brevity, we omit the derivation, but it is a straightforward extension of the above.

3 Designing powerful confidence sequences for RLAs

So far we have discussed how to conduct RLAs from confidence sequences for the parameter $\mu^\star$. In this section, we will discuss how to derive powerful confidence sequences for the purposes of conducting RLAs as efficiently as possible. For mathematical and notational convenience in the following derivations, we consider the case where $u = 1$. Note that nothing is lost in this setup since any population of $[0, u]$-bounded numbers can be scaled to the unit interval $[0, 1]$ by dividing each element by u.

As discussed in Section 2.1, we can construct confidence sequences by "inverting" sequential hypothesis tests. In particular, given a sequential hypothesis test $(\phi_t^\mu)_{t=1}^N$, the sequence of sets,

$$C_t := \{\mu \in [0,1] : \phi_t^\mu = 0\}$$

forms a $(1-\alpha)$ confidence sequence for $\mu^\star$. Consequently, order to develop powerful RLAs via confidence sequences, we can simply focus on carefully designing sequential tests $(\phi_t^\mu)_{t=1}^N$. [2]

[2]Notice that it is not always feasible to compute the set of all $\mu \in [0,1]$ such that $\phi_t^\mu = 0$ since $[0,1]$ is uncountably infinite. However, all confidence sequences we will derive in this section are intervals (i.e. convex), and thus we can find the endpoints using a simple grid search or standard root-finding algorithms.

To design sequential hypothesis tests, we start by finding *martingales* which translate to powerful tests. To this end, define $M_0(\mu) := 1$ and consider the following process for $t \in [N]$,

$$M_t(\mu) := \prod_{i=1}^{t} (1 + \lambda_i(X_i - \mathcal{C}_i(\mu))) \tag{2}$$

where $\lambda_i \in \left[0, \frac{1}{\mathcal{C}_i(\mu)}\right]$ is a tuning parameter depending only on $X_1, \ldots, X_{i-1}$, and

$$\mathcal{C}_i(\mu) := \frac{N\mu - \sum_{j=1}^{i-1} X_j}{N - i + 1}$$

is the conditional mean of $X_i \mid X_1, \ldots, X_{i-1}$ if the mean of $\{x_1, \ldots, x_N\}$ were μ.

Following [10, Section 6], the process $(M_t(\mu^\star))_{t=0}^N$ is a nonnegative martingale starting at one. Formally, this means that $M_0(\mu^\star) = 1$, $M_t(\mu^\star) \geq 0$, and

$$\mathbb{E}(M_t(\mu^*) \mid X_1, \ldots, X_{t-1}) = M_{t-1}(\mu^*)$$

for each $t \in [N]$. Importantly for our purposes, nonnegative martingales are unlikely to ever become very large. This fact is known as *Ville's inequality* [11, 12], which serves as a generalization of Markov's inequality to nonnegative (super)martingales, and can be stated formally as

$$\mathbb{P}\left(\exists t \in [N] : M_t(\mu^\star) \geq 1/\alpha\right) \leq \alpha M_0(\mu^\star) = \alpha, \tag{3}$$

where $\alpha \in (0, 1)$, and the equality follows from the fact that $M_0(\mu^\star) = 1$. As alluded to in Section 2, $(M_t(\mu^\star))_{t=0}^N$ can be interpreted as the reciprocal of an anytime p-value:

$$\mathbb{P}\left(\exists t \in [N] : \frac{1}{M_t(\mu^\star)} \leq \alpha\right) \leq \alpha,$$

which matches the probabilistic guarantee in (1). As a direct consequence of Ville's inequality, if we define the test $\phi_t^\mu := \mathbb{1}(M_t(\mu) \geq 1/\alpha)$, then

$$\mathbb{P}(\exists t \in [N] : \phi_t^{\mu^\star} = 1) \leq \alpha,$$

and thus $(\phi_t^\mu)_{t=1}^N$ is a level-α sequential hypothesis test. We can then invert $(\phi_t^\mu)_{t=1}^N$ and apply Theorem 1 to obtain confidence sequence-based RLAs with risk limit α.

3.1 Designing martingales and tests from reported vote totals

So far, we have found a process $(M_t(\mu))_{t=0}^N$ which is a nonnegative martingale when $\mu = \mu^\star$, but what happens when $\mu \neq \mu^\star$? This is where the tuning parameters $(\lambda_t)_{t=1}^N$ come into the picture. Recall that an electoral assertion $\mathcal{A}$ is certified once $C_t \subseteq \mathcal{A}$. Therefore, to audit assertions quickly, we want C_t to be as tight as possible. Since C_t is defined as the set of $\mu \in [0, 1]$ such that $M_t(\mu) < 1/\alpha$, we can make C_t tight by making $M_t(\mu)$ as *large* as possible. To do so, we must carefully choose $(\lambda_t)_{t=1}^N$. This choice will depend on the type of election as well as the amount of information provided prior to the audit. First consider the case where reported vote totals are given (in addition to the announced winner).

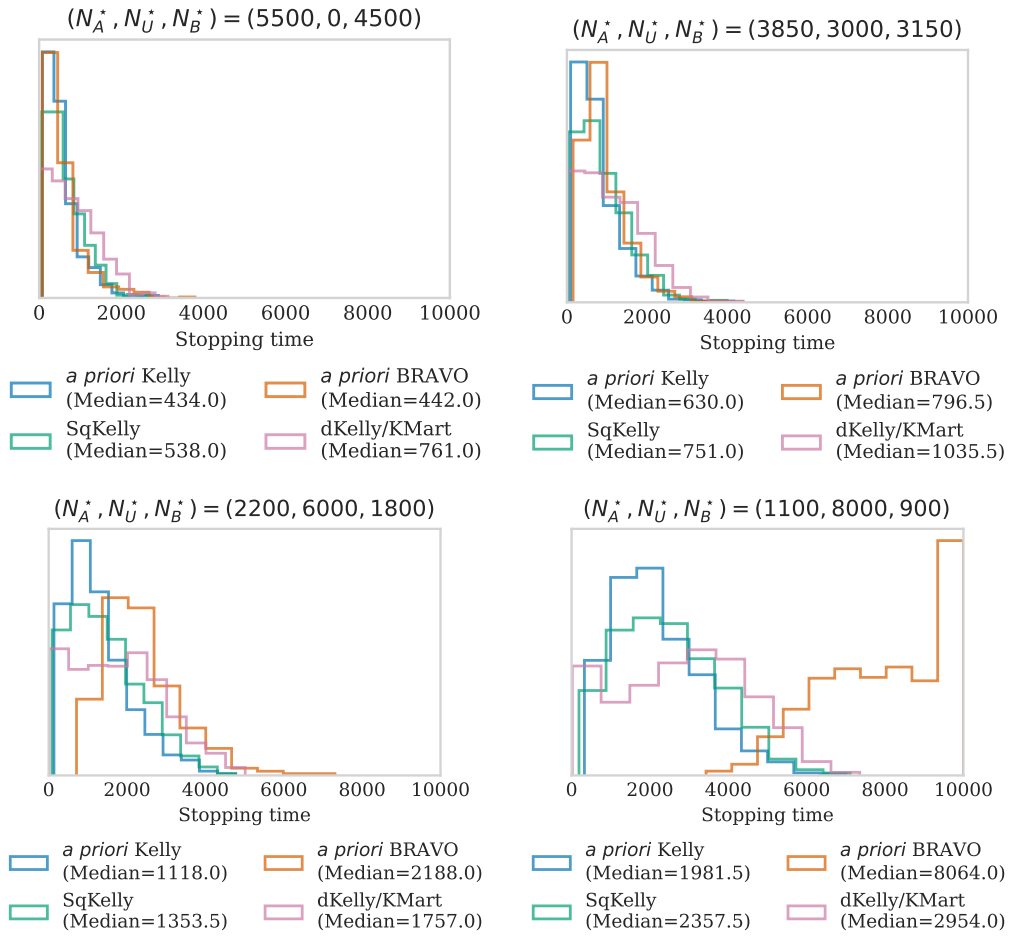

Figure 3: Ballot-polling audit stopping time distributions under four possible outcomes of a two-candidate plurality election. The first example considers an outcome where Alice and Bob received 5500 and 4500 votes respectively, and no ballots were invalid, for a margin of 0.1. The second, third, and fourth examples have the same margin, but with increasing numbers of invalid or "nuisance" ballots represented by $N_U^\star$. Notice that in the case with no nuisance ballots, *a priori* Kelly and BRAVO have an edge, while in the setting with many nuisance ballots, *a priori* Kelly vastly outperforms BRAVO. On the other hand, neither SqKelly nor dKelly require tuning based on the reported outcomes, but SqKelly outperforms dKelly in all four scenarios.

For example, recall the election between Alice and Bob of Section 2, and suppose that $\{x_1, \ldots, x_N\}$ is the list of numbers encoding votes for Alice as 1, votes for Bob as 0, and ballots with no valid vote for either candidate as 1/2. Recall that Alice beat Bob if and only if $\mu^\star := \frac{1}{N}\sum_{i=1}^N x_i > 1/2$, so we are interested in testing the null hypothesis

$$H_0 : \mu^\star \leq 1/2 \text{ against the alternative } H_1 : \mu^\star > 1/2.$$

Suppose it is reported that Alice beat Bob with N'_A votes for Alice, N'_B for Bob, and N'_U nuisance votes (i.e. either invalid or for another party). If the reported outcome is *correct*, then for any fixed λ, we know the exact value of

$$\prod_{i=1}^{N}(1+\lambda(x_i-1/2)), \tag{4}$$

which is an inexact but reasonable proxy for $M_N(1/2)$, the final value of the process $(M_t(1/2))_{t=0}^N$. We can then choose the value of λ' which maximizes (4). Some algebra reveals that the maximizer of (4) is given by

$$\lambda' := 2\frac{N'_A - N'_B}{N'_A + N'_B}. \tag{5}$$

We then truncate λ' to obtain

$$\lambda_t^{\mathrm{apK}} := \min\left\{\lambda', \frac{1}{\mathcal{C}_t(\mu^\star)}\right\}, \tag{6}$$

ensuring that it lies in the allowable range $[0, 1/\mathcal{C}_t(\mu)]$. We call this choice of λ_t^{apK} ***a priori* Kelly** due to its connections to Kelly's criterion [13, 10] for maximizing products of the form (4). This choice of λ_t^{apK} also has the desirable property of yielding convex confidence sequences, which we summarize below.

Proposition 1. *Let $X_1, \ldots, X_N$ be a sequential random sample from $\{x_1, \ldots, x_N\}$ with $\mu^\star := \frac{1}{N}\sum_{i=1}^N x_i$. Consider $(\lambda_t^{\mathrm{apK}})_{t=1}^N$ from (6) and define the process $M_t(\mu) := \prod_{i=1}^t (1+\lambda_i^{\mathrm{apK}}(X_i - \mathcal{C}_i(\mu)))$ for any $\mu \in [0,1]$. Then the confidence set*

$$C_t^{\mathrm{apK}} := \{\mu \in [0,1] : M_t(\mu) < 1/\alpha\},$$

is an interval with probability one.

Proof. Notice that since $\lambda' \geq 0$, $\mathcal{C}_t(\mu) \geq 0$, and $X_i \geq 0$, we have that

$$\lambda_t^{\mathrm{apK}}(X_i - \mathcal{C}_t(\mu)) = \min\{\lambda' X_i, X_i/\mathcal{C}_t(\mu)\} - \min\{\lambda'\mathcal{C}_t(\mu), 1\},$$

is a nonincreasing function of μ for each $t \in [N]$. Consequently, $M_t(\mu)$ is a nonincreasing & quasiconvex function of μ, so its sublevel sets are convex. □

Note that *any* sequence $(\lambda_t)_{t=1}^N$ such that $\lambda_t \in [0, 1/\mathcal{C}_t(\mu)]$ would have yielded a valid nonnegative martingale, but we chose that which maximizes (4) so that the resulting hypothesis test $\phi_t := \mathbb{1}(M_t(1/2) > 1/\alpha)$ is powerful. In situations more complex than two-candidate plurality contests, the maximizer of (4) can still be found efficiently via standard root-finding algorithms. All of these methods are implemented in our software package (to be made public in the final version).

While audits based on *a priori* Kelly display excellent empirical performance (see Figure 3), their efficiency may be hurt when vote totals are erroneously reported. Small errors in reported vote totals seem to have minor adverse effects on stopping times (and in some cases can be slightly beneficial), but larger errors can significantly affect stopping time distributions (see Figure 4). If we wish to audit the reported winner of an election but prefer not to rely on (or do not have access to) exact reported vote totals, we need an alternative to *a priori* Kelly. In the following section, we describe a family of such alternatives.

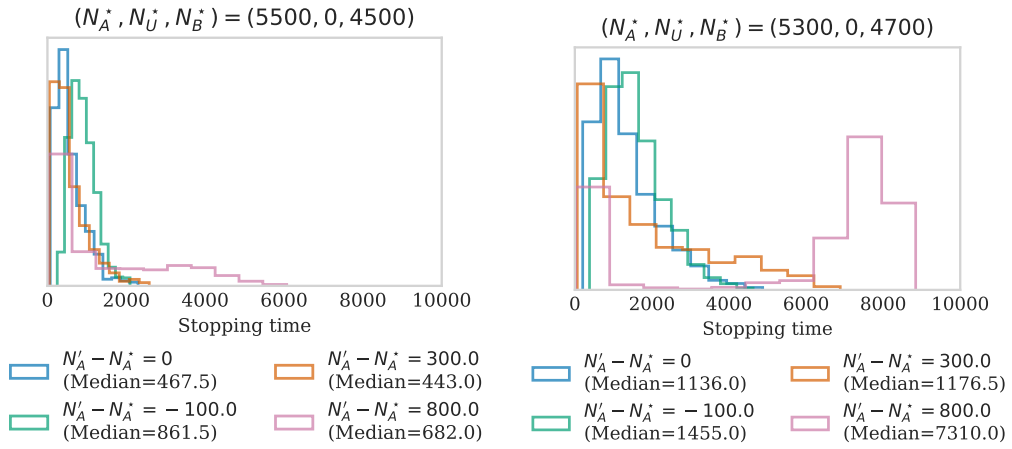

Figure 4: Stopping times for *a priori* Kelly under various degrees of error in reported outcomes. In the above legends, $N_A^\star$ refers to the *true* number of votes for Alice, while N'_A refers to the incorrectly reported number of votes. Notice that empirical performance is relatively strong for $N'_A - N_A^\star \in \{0, 300\}$ but is adversely affected when $N'_A - N_A^\star \in \{-100, 800\}$, especially in the right-hand side plot with a narrower margin.

3.2 Designing martingales and tests without vote totals

If the exact vote totals are not known, but we still wish to audit an assertion (e.g. that Alice beat Bob), we need to design a slightly different martingale that does not depend on maximizing (4) directly. Instead of finding an optimal λ', we will take $D \geq 2$ points evenly-spaced on the allowable range $[0, 1/\mathcal{C}_t(\mu)]$ and "hedge our bets" among all of these. Making this more precise, note that a convex combination of martingales (with respect to the same filtration) is itself a martingale [10], and thus for any $(\theta_1, \ldots, \theta_D)$ such that $\theta_d \geq 0$ and $\sum_{d=1}^D \theta_d = 1$, we have that

$$M_t^D(\mu^\star) := \sum_{d=1}^{D} \theta_d \prod_{i=1}^{t} \left(1 + \frac{d}{(D+1)\mathcal{C}_i(\mu^\star)}(X_i - \mathcal{C}_i(\mu^\star))\right) \tag{7}$$

forms a nonnegative martingale starting at one. Notice that we no longer have to depend on the reported vote totals to begin an audit. Furthermore, confidence sequences generated using sublevel sets of $M_t^D(\mu)$ are intervals with probability one [10, Proposition 4]. Nevertheless, choosing $(\theta_1, \ldots, \theta_D)$ is a nontrivial task. A natural — but as we will see, suboptimal — choice is to set $\theta_d = 1/D$ for each $d \in [D]$. Previous works [10] call this **dKelly** (for "diversified Kelly"), a name which we adopt here. In fact, this choice of $(\theta_1, \ldots, \theta_D)$ gives an arbitrarily close and computationally efficient approximation to the *Kaplan martingale* (**KMart**) [1] which can otherwise be expensive for large N.

Better choices of $(\theta_d)_{d=1}^D$ exist for the types of elections one might encounter in practice. Recall that near-optimal values of λ are given by (5). However, setting $\theta_d = 1/D$ for each $d \in [D]$ implicitly treats each $d/((D+1)\mathcal{C}_i(\mu^\star))$ as equally reasonable values of λ. Elections with large values of $\mu^\star$ (e.g. closer to 1) are "easier" to audit, and the interesting or "difficult" regime is when $\mu^\star$ is close to (but

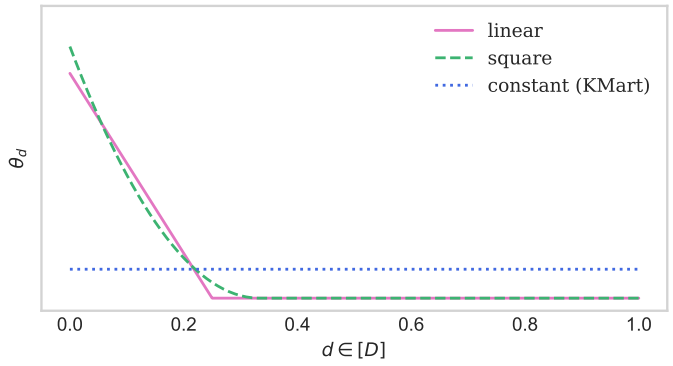

Figure 5: Various values of the convex weights $(\theta_1, \ldots, \theta_D)$ which can be used in the construction of the diversified martingale (7). Notice that the linear and square weights are largest for d near 0, and decrease as d approaches $1/4$, finally remaining at 0 for all large d. Smaller values of d are upweighted since they correspond to the bets in $M_t^D(\mu^\star)$ which are optimal for smaller (i.e. interesting) electoral margins. This is in contrast to the constant weight function which sets $\theta_d = 1/D$ for each $d \in [D]$. Linear and square weights perform well in practice (see Figure 3) but these can be tuned and tailored based on prior knowledge and the particular problem at hand.

strictly larger than) $1/2$. Therefore, we recommend designing $(\theta_1, \ldots, \theta_D)$ so that $(M_t^D(1/2))_{t=0}^N$ upweights optimal values of λ for margins close to 0, and downweights those for margins close to 1. Consider the following concrete examples. First, we have the truncated-linear weights,

$$\theta_d^{\text{linear}} := \frac{\gamma_d^{\text{linear}}}{\sum_{d=1}^D \gamma_d^{\text{linear}}}, \quad \text{where } \gamma_d^{\text{linear}} = \max\{0, 1 - 2d\},$$

and we normalize by $\sum_d \gamma_d^{\text{linear}}$ to ensure that $\sum_d \theta_d = 1$. Another compelling choice is given by the truncated-square weights,

$$\theta_d^{\text{square}} := \frac{\gamma_d^{\text{square}}}{\sum_{d=1}^D \gamma_d^{\text{square}}}, \quad \text{where } \gamma_d^{\text{square}} := (1/3 - x)^2 \mathbb{1}_{d \le 1/3}.$$

These values of θ_d^{linear} and θ_d^{square} are large for $d \approx 0$ and small for $d \gg 0$, and hence the summands in the martingale given by (7) are upweighted for implicit values of λ which are optimal for "interesting" margins close to 0, and downweighted for simple margins much larger than 0 (see Figure 5).

When M_t^D is combined with θ_d^{square}, we refer to the resulting martingales and confidence sequences as **SqKelly**, and compare their empirical performance against *a priori* Kelly, dKelly, and BRAVO in Figure 3. A hybrid approach is also possible: suppose we want to use reported outcomes or prior knowledge alongside these convex-weighted martingales. We can simply choose $(\theta_1, \ldots, \theta_D)$ so that M_t^D upweights values in a neighborhood of λ' (or some other value chosen based on prior knowledge[3]).

[3]The use of the word "prior" here should not be interpreted in a Bayesian sense. No matter what values of $(\theta_1, \ldots, \theta_D)$ are chosen, the resulting tests and confidence sequences have *frequentist* risk-limiting guarantees.

4 Risk-limiting tallies via confidence sequences

Rather than audit an already-announced electoral outcome, it may be of interest to determine (for the purposes of making a first announcement) the election winner with high probability, without counting all N ballots. Such procedures are known as risk-limiting tallies (RLTs) which were developed for coercion-resistant, end-to-end verifiable voting schemes [14]. For example, suppose a voter is being coerced to vote for Bob. If the final vote tally reveals that Bob received few or no votes, then the coercer will suspect that the voter did not comply with instructions. RLTs provide a way to mitigate this issue by providing high-probability guarantees on the true winner, leaving a large proportion of votes shrouded. In such cases, the voter is guaranteed plausible deniability, as they can claim to the coercer that their ballot is simply among the unrevealed ones.

While the motivations for RLTs are quite different from those for RLAs, the underlying techniques are similar. The same is true for confidence sequence-based RLTs. All methods introduced in this paper can be applied to RLTs (with the exception of "*a priori* Kelly" since it depends on the reported outcome) but with two-sided power. Consider the martingales we discussed in Section 3.2,

$$M_t^D(\mu^\star) := \sum_{d=1}^{D} \theta_d \prod_{i=1}^{t} \left(1 + \frac{d}{(D+1)\mathcal{C}_i(\mu^\star)}(X_i - \mathcal{C}_i(\mu^\star))\right). \tag{8}$$

where $(\theta_1, \ldots, \theta_D)$ are convex weights. Recall that our confidence sequences at a given time t were defined as those $\mu \in [0,1]$ for which $M_t^D(\mu) < 1/\alpha$. In other words, a given value μ is only excluded from the confidence set if $M_t^D(\mu)$ is large. However, notice that $M_t^D(\mu)$ will become large if the conditional mean $\mathcal{C}_t(\mu^\star) \equiv \mathbb{E}(X_t \mid X_1, \ldots, X_{t-1})$ is larger than the null conditional mean $\mathcal{C}_t(\mu)$, but the same cannot be said if $\mathcal{C}_t(\mu^\star) < \mathcal{C}_t(\mu)$. As a consequence, the resulting confidence sequences are all one-sided *lower* confidence sequences. To ensure that our bounds have non-trivial two-sided power, we can simply combine (8) with a martingale which also grows when $\mathcal{C}_t(\mu^\star) < \mathcal{C}_t(\mu)$.

Proposition 2. *For nonnegative vectors $(\theta_1^+, \ldots, \theta_D^+)$ and $(\theta_1^-, \ldots, \theta_D^-)$ that each sum to one, define the processes*

$$M_t^{D+}(\mu) := \sum_{d=1}^{D} \theta_d^+ \prod_{i=1}^{t} \left(1 + \frac{d}{(D+1)\mathcal{C}_i(\mu^\star)}(X_i - \mathcal{C}_i(\mu^\star))\right),$$
$$M_t^{D-}(\mu) := \sum_{d=1}^{D} \theta_d^- \prod_{i=1}^{t} \left(1 - \frac{d}{(D+1)(1 - \mathcal{C}_i(\mu^\star))}(X_i - \mathcal{C}_i(\mu^\star))\right).$$

Next, for $\beta \in [0,1]$, define their mixture

$$M_t^{D\pm}(\mu) := \beta M_t^{D+}(\mu) + (1-\beta) M_t^{D-}(\mu).$$

Then, $M_t^{D\pm}(\mu^\star)$ is a nonnegative martingale starting at one. Consequently,

$$C_t^{\pm} := \{\mu \in [0,1] : M_t^{D\pm}(\mu) < 1/\alpha\}$$

forms a $(1-\alpha)$ confidence sequence for $\mu^\star$.

Proof. This follows immediately from the fact that both $M_t^{D+}(\mu^\star)$ and $M_t^{D-}(\mu^\star)$ are martingales with respect to the same filtration, and that convex combinations of such martingales are also martingales. □

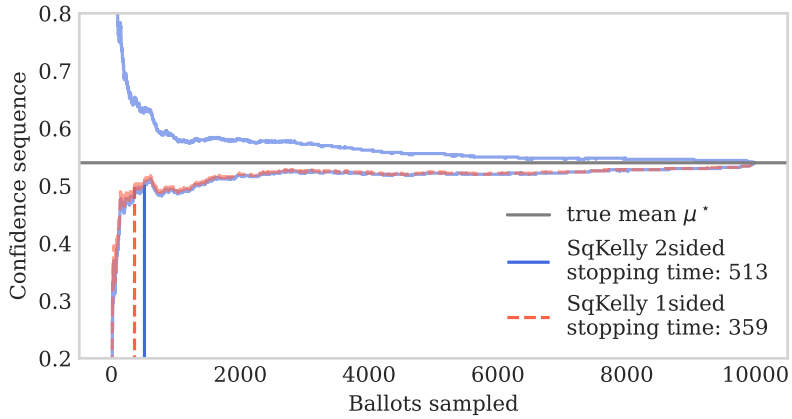

Figure 6: Confidence sequence-based risk-limiting tally for a two-candidate election. Unlike RLAs, RLTs require two-sided confidence sequences so that the true winner can be determined (with high probability) without access to an announced result. Notice that testing the same null $H_0 : \mu^\star \leq 0.5$ is less efficient in an RLT than in an RLA. This is a necessary sacrifice for having nontrivial power against other alternatives.

With this setup and notation in mind, M_t^D as defined in Section 3.2 is a special case of $M_t^{D\pm}$ with $\beta = 1$. As noted by [14], RLTs of multiple contests *do* require correction for multiple testing, unlike RLAs. The same is true for confidence sequence-based RLTs (and hence the tricks of Section 2.2 do not apply). It suffices to perform a simple Bonferroni correction by constructing $(1 - \alpha/K)$ confidence sequences to tally K simultaneous contests.

5 Summary

This paper presented a general framework for conducting risk-limiting audits based on confidence sequences, and derived computationally and statistically efficient martingales for computing them. We showed how *a priori* Kelly takes advantage of the reported vote totals (if available) to stop ballot-polling audits significantly earlier than extant ballot-polling methods, and how alternative martingales such as SqKelly also provide strong empirical performance in the absence of reported outcomes. Finally, we demonstrated how a simple tweak to the aforementioned algorithms provide two-sided confidence sequences which can be used to perform risk-limiting tallies. Confidence sequences and these martingales can be applied to ballot-level comparison audits and batch-level comparison audits as well, using "overstatement assorters" [1], which reduce comparison audits to the same canonical statistical problem: testing whether the mean of any list in a collection of non-negative bounded lists is less than 1/2. We hope that this new perspective on RLAs and its associated software will aid in making election audits simpler, faster, and more transparent.

References

[1] Philip B Stark. Sets of half-average nulls generate risk-limiting audits: Shangrla. In *International Conference on Financial Cryptography and Data Security*, pages 319–336. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 11, 14

[2] Mark Lindeman, Philip B Stark, and Vincent S Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *EVT/WOTE*, 2012. 2, 6

[3] Kellie Ottoboni, Matthew Bernhard, Alex Halderman, Ronald Rivest, and Philip Stark. Bernoulli ballot polling: a manifest improvement for risk-limiting audits. In *International Conference on Financial Cryptography and Data Security*, pages 226–241. Springer, 2019. 2, 5

[4] Ronald L Rivest. ClipAudit: A simple risk-limiting post-election audit. *arXiv preprint arXiv:1701.08312*, 2017. 2

[5] Philip B Stark et al. Conservative statistical post-election audits. *The Annals of Applied Statistics*, 2(2):550–581, 2008. 5

[6] Philip B Stark. Cast: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security*, 4(4):708–717, 2009. 5

[7] Philip B Stark. Risk-limiting postelection audits: Conservative p-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4(4):1005–1014, 2009. 5

[8] Kellie Ottoboni, Philip Stark, Mark Lindeman, and Neal McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In *Intl. Joint Conference on Electronic Voting*, pages 174–188. Springer, 2018. 5

[9] Zhuoqun Huang, Ronald L Rivest, Philip B Stark, Vanessa J Teague, and Damjan Vukcevic. A unified evaluation of two-candidate ballot-polling election auditing methods. In *International Joint Conference on Electronic Voting*, pages 112–128. Springer, 2020. 5

[10] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2021. 8, 10, 11

[11] Jean Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939. 8

[12] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. 8

[13] JL Kelly Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4): 917–926, 1956. 10

[14] Wojciech Jamroga, Peter B Roenne, Peter YA Ryan, and Philip B Stark. Risk-limiting tallies. In *International Joint Conference on Electronic Voting*, pages 183–199. Springer, 2019. 13, 14