Comparing Methods of Opportunistic Risk Limiting Audits

Eric Chen

Rohith Raghavan

January 2025

Abstract

Auditing elections is an important part of preserving faith in the electoral system and verifying the accuracy of the reported results of an election. Conventional election audits involve taking a set number or percentage of ballots and checking if the samples match the reported winner. However, these methods are unreliable for close races and excessive for races with a wide margin. Risk-limiting audits use statistical tests in order to assign a certain risk limit, the maximum probability that the results are incorrect, by sampling ballots one at a time until the risk limit is achieved. Our research explores opportunistic auditing, the ability to audit multiple races at the same time, and attempts to determine what strategies are most effective for opportunistic auditing. We examine complex multi-state and strata races that are audited using the ALPHA (Stark) supermartingale and test different sampling strategies across drifts and margins to answer the core question: how can existing auditing tests/martingales provide useful risk guarantees over multiple simultaneous races?

1 Introduction

Elections are often viewed as an objective reflection of the truth, but upon closer inspection are not immune from inaccuracies. Whether through deliberate attacks on voting hardware [2] or mistakes in transportation or hand-counting, election results may not be perfectly accurate. This uncertainty in results leads to decreased trust in the reported results from elections. Recently, we have seen the extremes of this distrust. During the 2020 US presidential election cycle, doubt towards voting security peaked as claims of voter fraud spread across social media. Now, more than ever, improvements to audit protocol are essential to increase public trust in our government.

One naive approach to auditing elections is to sample an arbitrary number of ballots and observe which candidate receives the majority in that sample.

For example, in Pennsylvania, county election boards must recount either 2000 or 2% of all ballots [5]. Performing these simple audits provide some level of confidence on the winner of the election, but have notable flaws. We do not take advantage of our reported proportion of votes for the winning candidate. In elections with large margins, resources may be wasted on running audits with large numbers of ballots. Conversely, even after an audit is run in a close race, we may not have enough information to conclude that reported results are correct.

Instead, risk-limiting audits (RLAs) take advantage of statistical principles to run efficient audits. RLAs take the form of hypothesis tests. The null hypothesis is that the reported winner received the miority of the votes. Once the probability that the null is correct falls below an arbitrary risk limit, usually 0.05, there is now convincing evidence that the reported winner actually received a majority of the votes. As a result, the audit naturally adjusts itself based on the shares of votes. If there is a large margin of votes between the candidates, the audit will wrap up quickly.

Research on RLAs tends to focus primarily on improving audits for individual electoral races. In practice, however, elections are rarely performed in a vacuum. Votes for multiple positions or issues may occur at the same time. Additionally, votes may for one race may reside in multiple different locations. Although it is possible to perform audits on each race individually, the audit can be made more efficient by auditing every election together. We assume that the votes for all simultaneous races are recorded on one ballot. By allowing one ballot to contribute to multiple races at once, we can lower the total number of ballots required to audit all races. Auditing multiple levels of races this way is known as *opportunistic auditing*.

Our research explores the efficiency and practicality of opportunistic audits. Specifically, we aim to address a few questions about the deployment of opportunistic audits. Are the efficiency gains from opportunistic audits worth the logistical overhead? What are the most effective strategies for performing opportunistic audits?

2 Background

Elections are run across multiple different jurisdictions in order to capture local opinions and out of convenience. *Strata* are groups of voters in one jurisdiction. We define the *levels* of an election to be races in which different sets of strata vote. For example, a person from Boston might vote in three levels of election: one for president, one for governor, and one for mayor. *Global* races are elections which poll the entire population.

Simple RLAs produce confidence intervals across each stratum independently. These p-values, which measure our confidence that the reported winner received the majority of the votes, cannot be trivially combined across strata. Techniques like Fisher's combined probability test, which can combine arbitrary p-values, add a significant amount of uncertainty. As a result, we cannot feasi-

bly conduct opportunistic audits by running simple RLAs across relevant strata without a signficant increase in ballots required to ensure confidence.

Many challenges face optimizing opportunistic auditing. There are many different goals that auditing strategies may try to reach. Audits can be allocated a set number of ballots across all strata, and attempt to minimize the largest p-value in any race. Strategies may also try to audit all races to completion while drawing the fewest ballots possible. These different goals for optimization make opportunistic auditing an open space for experimentation. Opportunistic audits must compromise between auditing more races and increasing confidence in audited strata.

Even with recent advancements to RLAs, there remain some limitations inherent to the auditing strategy. In elections with close margins, all risk limiting audits will struggle to reject the null hypothesis. All risk limiting audits, and all audits in general, will most likely have to complete a full manual recount in order to be confident in their reported winner. Additionally, most RLAs rely on pulling ballots one-by-one, and updating the test statistic after each vote is counted. This leads to additional overhead and costs, as drawing ballots individually is very inefficient. RLAs which take samples of ballots in batches have been studied [1] for individual races, but are less understood for multiple strata. In our implementation of opportunistic auditing, the stratum selector requires information from all past ballots in order to accurately suggest a stratum to pull from. As a result, we face the same challenge of reading each ballot individually.

Our protocol for opportunistic audits relies on a central body, which instructs different strata to sample one ballot. The central body uses the test statistics it calculates for each race to determine which stratum will be sampled from next.

We had to make some assumptions in order to analyze opportunistic audits. We assumed that it is not a security risk to sample ballots with replacement when performing audits. It is possible to reach similar conclusions when sampling without replacement, but it adds to the complexity of the model. We also assumed that audits are being performed in strata with cast vote records (CVRs). These record the locations of each ballot counted. Audits performed in districts with CVRs tend to require fewer ballots than non-CVR districts to reach the same risk limits. Like with our choice to sample with replacement, existing auditing strategies can analyze results from both CVR and non-CVR strata, but limiting our districts to ones with CVRs simplifies our modeling.

3 Related Works

3.1 BRAVO

BRAVO is a ballot-polling audit strategy that is based on Wald's Sequential Probability Ratio Test(SPRT) [4]. SPRT is a statistical test that affirms or rejects a hypothesis by sequentially sampling data and incorporating the sampled data into previous results; in a ballot polling audit, the items being sampled are ballots and the information used is the chosen candidate of the ballot. The

audit determines whether the winning candidate received more votes than the losing candidate; multiple BRAVO audits must be conducted simultaneously in a race with more than two candidates for every winner-loser pair. A test statistic T is initially set to 1 and is multiplied by a factor greater than 1 when a vote for the winner is sampled or a factor less than 1 if the vote is for the losing candidate. The value of T is left unchanged if the ballot is invalid or for another candidate. The factors for the winner and loser are equal to $2s_w$ and $2 - 2s_w$ respectively, where s_w is the reported proportion of votes that were for the winner overall ballots that were marked for the winner or the loser. In the context of opportunistic auditing, BRAVO is not an effective audit strategy to use because of its inefficiency and lack of ability to deal with stratification.

Using the reported relative margin is a weakness of BRAVO because, in the case of true electoral fraud, the reported results cannot be heavily relied on. Furthermore, the strategy is highly inefficient when the reported results differ from the actual results. BRAVO also lacks an effective mechanism to deal with elections in which ballots are stratified, making it hard to implement on a large scale. In spite of these drawbacks, BRAVO is remarkable for its simple algorithm and ease to use, allowing it to serve as an extremely practical way to audit elections. Our research will use BRAVO as a starting point to compare efficiency between different opportunistic voting strategies.

3.2 SUITE

SUITE is a method for running RLAs which supports stratified sampling [6]. The audit works by disproving the intersection of the null-hypothesis of each strata. The P-values of each hypothesis are combined using Fisher's combining function, resulting in a single maximum global P-value. If this global P-value is below the designated risk limit, the audit can stop.

One of the main goals of SUITE was to support stratifying counties with and without CVRs. CVRs track the location of each paper ballot, allowing individual votes to be compared between digital records and physical ballots. Therefore, in CVR counties, ballot-level comparison audits can be run, which verify that the system's interpretation of ballots is correct. Audits in CVR counties usually require less ballots in order to achieve the same risk limit as no-CVR counties. Therefore, by stratifying the samples of ballots in the two different types of counties, a global risk limit can be achieved more efficiently.

SUITE demonstrates the applicability of stratified auditing in practical situations. Although one of its focuses was on the differences between CVR and no-CVR strata, its approach to stratification is also relevant to combining similar strata.

3.3 ALPHA

ALPHA is a family of risk-limiting audits which generalize strategies such as BRAVO [9]. It uses betting martingales to lower the number of ballots required to reject the null. Although BRAVO is optimal when the reported vote share

is identical to the actual vote share, this is rarely the case, and fails to account for any errors which may occur in practice. ALPHA improves on BRAVO by constantly adapting its strategy as the audit is running. The audit also runs identically in ballot polling and comparison audits. As a result, ALPHA supports stratified sampling by simply multiplying the martingales of each strata.

ALPHA offers an alternative approach to stratified auditing which eliminates some of the uncertainty introduced by Fisher's combining function. However, it does not cover different strategies for stratum selection, as well as batch-level comparison and polling audits. These audits draw ballots in batches instead of individually, increasing the number of ballots required but decreasing the number of batches. We hope to examine the practicality of these different strategies of ALPHA.

4 Preliminary Trials

Our initial work consisted of implementing BRAVO and developing a codebase in which we were able to run different audits using created ballots of sample election data. This step was novel in implementing risk-limiting audits as most other theoretical audit implementations used only the value on a ballot. However, opportunistic testing requires the ability to view a ballot and read the values from multiple races of the same ballot, leading to our creation and usage of a ballot object and associated helper classes to manage the ballot lists.

Initial testing with BRAVO was used to determine the impact additional ballots would have on a race already achieving a predetermined risk limit, which is crucial to understanding the incremental information gained from each ballot used to audit past completion. Each addition ballot had decreasing marginal returns until the risk limit was flat-lined. Furthermore, our testing with BRAVO clearly showed that slight changes in the real results compared to the reported results lead to vast inefficiencies, caused by BRAVO's inability to effectively adapt to incoming ballot data. Because of this, we decided to focus our testing on ALPHA.

The ALPHA supermartingale was previously implemented by Stark. However, the existing implementation did not use a voting object and class structure compatible with opportunistic auditing but rather mainly served as a tool to compute the run time and computation required to run an audit. The source code used to generate graphs and statistics for ALPHA did not actually simulate each ballot, but instead only looked at ballot counts for each candidate and race.

In order to model opportunistic audits, we needed finer control over which ballots were passed into the model. Because of this, the existing ALPHA code was extensively retooled to use a similar class structure of the ballots and ballot manifest from BRAVO, capable of interrupting audits to change strategy or switch sampling from multiple different stratum unlike the previously existing iterations of the code. Our implementation of opportunistic audits retained much of ALPHA's original source for calculating martingales and estimating the

proportion of winners in a given stratum, but was mostly rewritten to better suit our needs.

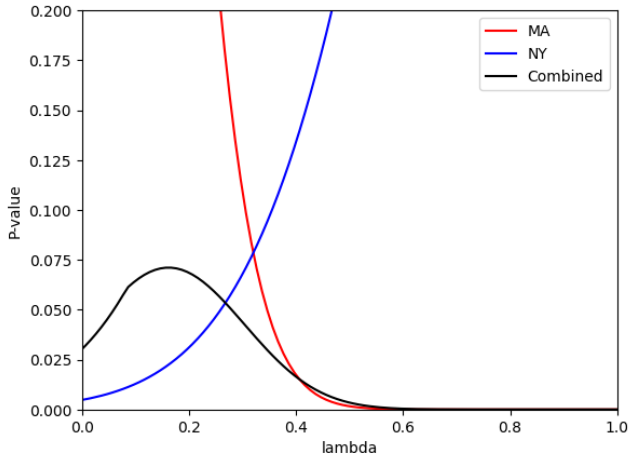
We first observed very simple elections. For example, we modeled an election where the proportion of votes which went to the winner were either 0.52% or 0.60%. Then, we tried multiple combinations of the size of the margin in the stratum. We especially examined what happens when the proportion of votes for the winner is different from in other strata. We summarize the data below.

Average ballots used	Presidential	Presidential first	State first
Big global and MA gov, tight NY gov	466.65	672.0	725.25
Big global, tight in both gov races	513.8	813.15	717.8
Big in MA gov / NY pres	611.55	758.7	771

In particular, we observe that for larger margins, our auditing protocol becomes more efficient. Less additional ballots are required to audit each stratum after auditing the global election.

To gain more insight on why simply combining p-values from different strata generated by BRAVO together, we observed trends in its performance. For example, in a 2 strata election between Massachusetts and New York, if we were to draw 70 ballots total, and vary the proportion of ballots drawn from Massachusetts, λ, we would observe the following curves.

Figure 1: P-values for varying λ for 70 ballots



While each individual race's p-value dips well below the risk limit of 0.05, the worst-case combined p-value sits noticeably above the global risk limit. This suggests that simply combining the p-values from two audits is far from precise enough to be used to improve existing audit protocol.

Due to these constraints, we focused on rewriting ALPHA to be more easily extensible and to give the author more control over the audit on an individual

ballot level as opposed to optimizing BRAVO further.

5 Strategy Overview

5.1 Defining Strategy

In this paper, we explore three major strategies in our testing of opportunistic auditing. These strategies are closely related if not equivalent to the stratum selectors proposed by Stark in the ALPHA martingale. Stratum selectors, including the round robin selector and the multinomial selector proposed by Stark, use data from previous ballots to select the next stratum to sample from in order to minimize the number of ballots required to audit the election. However, in the context of opportunistic auditing, information gained from a stratum includes data on non-global races and optimizing the number of ballots drawn also involves consideration for state and local races. Thus, we adopt the term *strategy* to denote a stratum selector function that utilizes information from all races and strata in order to determine the next strata to draw from. Notably, the creation of strategies faces the dilemma of not just optimizing but choosing what races should or should not be prioritized in an audit; the inevitable trade-off between workload and audit breadth means that different strategies may choose to value various aspects of opportunistic auditing differently.

5.2 Strategies Tested

During our testing, we mainly utilized 3 different strategies applied for a two level system with a global presidential race and statewide governor races: the round robin(RR), "Lowest T," and "Average P Optimization" strategies. Each of these strategies prioritizes different aspects of an audit with varying benefits and drawbacks.

The round robin strategy is adapted from that used by Stark in ALPHA. In our context, round robin refers to selecting from every strata in an order that continually loops. In this manner, round robin is completely nonadaptive to information received opportunistically but rather ensures that every strata is sampled almost equally. Though the strategy is simple, round robin tends to be extremely effective given that it does not miss data from any strata. For our purposes, round robin serves as a control against which to measure our more advanced strategies.

The "Lowest T" strategy is one that greedily chooses the state whose governor race is the worst performing and has the lowest T value. The motivation for this strategy comes from a similar motivation to that of round robin: if the lowest T-value state is always chosen, then that state will always improve such that the T-values of all governor races will be roughly equal. In this way, the strategy allows for the selection of all strata while effectively weighing for margin. Furthermore, if the winner of a governor's race is incorrectly reported, then the strategy will pigeonhole itself onto that state and fail to complete, providing

another layer of security.

Finally, the "Average P Optimization" strategy seeks to minimize the expected value of the arithmetic mean of all p-values of state governor races. Specifically, using the margin of ballots that have been drawn so far and the current value of every governor race martingale, the strategy examines every stratum and calculates the expected decrease in p-value if a ballot were to be sampled from that stratum, finally choosing the stratum with the highest expected decrease. As p-values decrease, they move slower; because of this, the average P optimization will also favor worse-performing strata with higher p values but the greatest room to decrease and get closer to completion.

By comparing the nonadaptive round robin to two strategies that prioritize and adapt using state results, these 3 strategies highlight different mentalities toward opportunistic auditing, the testing of which will provide direction on which is more effective.

6 Drift Study

When results in the presidential stratum and governor line up in every state, opportunistic auditing poses less complexity; the presidential race will take significantly more ballots to audit than states due to the slowdown of stratification and thus a strategy that only looks at the presidential race will likely perform well in an opportunistic setting. However, the opposite occurs in the case of *drift,* where presidential results and governor results in a given state have a significant difference in their margin. For this reason, we analyze the effects of varying margins, which in our case refers to the winner's proportion of the vote, on the efficacy of our different strategies. For the purposes of our testing, we utilized a "point mass" approach to our audit simulations. Instead of viewing each ballot as a binary vote for either candidate, such as a 0 for Democrats and 1 for Republicans, their value was proportional to the reported results. For example, if Democrats received 40% of the vote in a stratum, the value of each ball ot in that stratum was set to 0.6. This reduced computational difficulty with real results perfectly matching reported results. Two strata of equal size were used in audits sampled with replacement, and within every test, the only margin varied was that of the presidential race in the first stratum. All races, statewide and presidential, were audited to completion.

First, a system with considerable drift was used. In stratum 2, the presidential margin, like in all trials, was 60%, and the governor margins were 55% in both strata. However, the presidential margin in the first strata was shifted from 42.5% to 67.5% in increments of 2.5 %, highlighting a broad range of large and small drifts above and below the statewide governor's race.

We can see that Lowest T and Round Robin perform equivalently, which is unsurprising given that the equal governor margins lead Lowest T to act in an equivalent manner to round robin. The Average P Optimization strategy performs far worse when the presidential margin is significantly lower then the state margin but the opposite when it greater in the stratum. This difference ex-

State Pres Margin	Lowest T Ballots Used	Average P Optimization	Round Robin
42.5%	3192	4219	3192
45%	1223	1261	1223
47.5%	610	343	610
50%	574	289	574
52.5%	574	289	574
55%	574	289	574
57.5%	574	289	574
60%	574	289	574
62.5%	574	289	574
65%	574	289	574
67.5%	574	289	574

Table 1: First Drift test, governor margins 55% and 55%

acerbates in the second test where all margins are the same except the governor race in strata 1 increases the winner's share to 60%.

State Pres Margin	Lowest T Ballots Used	Average P Optimization	Round Robin
42.5%	3810	4218	3194
45%	1381	1644	1224
47.5%	833	696	610
50%	526	408	574
52.5%	361	288	574
55%	361	288	574
57.5%	361	288	574
60%	361	288	574
62.5%	361	288	574
65%	361	288	574
67.5%	361	288	574

Table 2: Second Drift test, governor margins 60% and 55%

Notably, though the margins are more relaxed, the Average P Optimization performs worse in the low president margin but even better when the margin is greater. This seems to suggest that his strategy, which focuses heavily on optimizing state p values, proves better when the constraining factor is state elections, while it struggles when the presidential race requires more ballots as it does not optimize for that. The Lowest T strategy appears to be a middle ground between the other two, with slight deficiencies at the 42.5 % mark compared to round robin but benefits on the other side of the spectrum. For the final test, drift was isolated to just the first stratum as the second stratum governor's race margin was st 60 % and the first one returned to 55 %.

Once again, we see the same phenomena when comparing Average P optimization with round robin. While these tests do highlight round robin's continued effectiveness, they also emphasize the fact that the direction of the drift

State Pres Margin	Lowest T Ballots Used	Average P Optimization	Round Robin
42.5%	3194	4751	3194
45%	1224	1409	1224
47.5%	612	707	610
50%	574	371	574
52.5%	574	212	574
55%	574	182	574
57.5%	574	182	574
60%	574	182	574
62.5%	574	182	574
65%	574	182	574
67.5%	574	182	574

Table 3: Third Drift test, governor margins 55% and 60 %

is extremely important as to what strategy will be effective. We can conclude that in cases where drift leads to a closer presidential race, then strategies like round robin with no opportunistic favoring appear to perform better, whereas when state races are closer, strategies that incorporate state data outperform.

7 Explorations of 3+ Strata

We briefly explored working in systems with greater than 2 strata and adapted the ALPHA code to do so. Some preliminary tests run using 3 equally sized strata each with a governor's race and varying directions of drift were tested using a ballot polling simulation that averaged over 250 trials. Note that while the presidential was always audited, the state races were not forcibly audited in every strategy.

Strategy:	RR (Pres Only)	RR (All)	Lowest T-Value	Average P Optimization
Ballots Used:	472.08	905.07	1002.62	843.91
State Audit %:	41.3	100	56.3	100

Table 4: Three Strata Results

We can see that in this case, the Average P optimization performs just as well if not better than round robin when all races are being audited. However, the comparison between a round robin strategy that terminates at presidential completion and one that continues till state completion highlights the added workload that opportunistic auditing incurs.

8 Conclusion and Future Work

In an increasingly polarized political landscape, public trust is essential for governments and organizations. Our research on strategies in election auditing

provides insight on how to effectively scale and run wide-scale audits across multiple simultaneous races. We observed that a variety of strategies and protocols are all feasible and have their own tradeoffs between ballot count required and contributions to higher level elections. Specifically, in our examinations of drift, we can reasonably conclude that the direction of drift is extremely important toward the relative workload of strategies.

We believe that testing a greater number of strategies as well as with more strata is critical to building a greater understanding of the intricacies of opportunistic auditing and the various tradeoffs faced when designing strategy. We hope that our work can begin the conversations that these tradeoffs truly ask: what do we value the most in securing our electoral systems, and how can we design strategies and audits that meet those values?

9 Acknowledgments

We would like to thank the MIT PRIMES program for making this project possible as well as the coordinators Dr. Slava Gerovitch, Professor Pavel Etingof, and Professor Srini Devadas for their support.

And lastly, we thank our mentor Mayuri Sridhar for her time, effort, and patience in guiding us through this project and helping us grow as researchers.

References

[1] Moni Naor Bar Karov. New algorithms and applications for risk-limiting audits.

[2] Ariel J. Feldman, J. Alex Halderman, and Edward W. Felten. Security analysis of the diebold accuvote-ts voting machine. 2007.

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

[4] Mark Lindeman, Philip B. Stark, and Vincent S. Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *2012 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE 12)*, Bellevue, WA, August 2012. USENIX Association.

[5] Pennsylvania Department of State. Risk limiting audit directive, 2022.

[6] Kellie Ottoboni, Philip B. Stark, Mark Lindeman, and Neal McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (suite), 2018.

[7] Jacob V. Spertus and Philip B. Stark. Sweeter than suite: Supermartingale stratified union-intersection tests of elections, 2022.

[8] Philip B. Stark. Risk-limiting audits by stratified union-intersection tests of elections (suite). `https://github.com/pbstark/CORLA18`, 2019.

[9] Philip B. Stark. Alpha: Audit that learns from previously hand-audited ballots, 2022.