Integrating Reward Maximization and Population Estimation: Sequential Decision-Making for Internal Revenue Service Audit Selection*

Peter Henderson[1] PHEND@CS.STANFORD.EDU
Ben Chugg[1] BENCHUGG@LAW.STANFORD.EDU
Brandon Anderson[2] BRANDON.M.ANDERSON@IRS.GOV
Kristen Altenburger[1] KALTENBURGER@STANFORD.EDU
Alex Turk[2] ALEXANDER.H.TURK@IRS.GOV
John L. Guyton[2] JOHN.GUYTON@IRS.GOV
Jacob Goldin[1] JSGOLDIN@LAW.STANFORD.EDU
Daniel E. Ho[1] DHO@LAW.STANFORD.EDU
Stanford University[1]
Internal Revenue Service[2]

Abstract

We introduce a new setting, *optimize-and-estimate structured bandits*. Here, a policy must select a batch of arms, each characterized by its own context, that would allow it to both maximize reward and maintain an accurate (ideally unbiased) population estimate of the reward. This setting is inherent to many public and private sector applications and often requires handling delayed feedback, small data, and distribution shifts. We demonstrate its importance on real data from the United States Internal Revenue Service (IRS). The IRS performs yearly audits of the tax base. Two of its most important objectives are to identify suspected misreporting and to estimate the "tax gap" — the global difference between the amount paid and true amount owed. Based on a unique collaboration with the IRS, we cast these two processes as a unified optimize-and-estimate structured bandit. We provide a novel mechanism for unbiased population estimation that achieves rewards comparable to baseline approaches. This approach has the potential to improve audit efficacy, while maintaining policy-relevant estimates of the tax gap. This has important social consequences given that the current tax gap is estimated at nearly half a trillion dollars. We suggest that this problem setting is fertile ground for further research and we highlight its interesting challenges. The results of this and related research are currently being incorporated into the continual improvement of the IRS audit selection methods.[1]

Keywords: Bandits, Active Learning, Sequential Decision-Making, Dual Objectives, Internal Revenue Service

[*] Full version available at `https://arxiv.org/pdf/2204.11910.pdf`

[1] The views expressed here are those of the authors and do not necessarily reflect the views or the official position of the Internal Revenue Service. All data work for this project involving confidential taxpayer information was done on IRS computers. At no time was confidential taxpayer data outside of the IRS computing environment. JG, DEH are IRS employees under an agreement made possible by the Intragovernmental Personnel Act of 1970 (5 U.S.C. 3371-3376). PH is a Student Volunteer with the IRS.

1. Introduction

Sequential decision-making algorithms, like bandit algorithms and active learning, have been used across a number of domains: from ad targeting to clinical trial optimization (Bouneffouf and Rish, 2019). In the public sector, these methods are not yet widely adopted, but could improve the efficiency and quality of government services if deployed with care (Henderson et al., 2021). Many administrative enforcement agencies in the United States (U.S.) face the challenge of allocating scarce resources for auditing regulatory non-compliance. But these agencies must also balance additional constraints and objectives simultaneously. In particular, they must maintain an accurate estimate of population non-compliance to inform policy-making. In this paper, we focus on the potential of unifying audit processes of the Internal Revenue Service (IRS) with these multiple objectives under a bandit-like framework. We call our setting *optimize-and-estimate structured bandits*. This framework is useful in practical settings, challenging, and has the potential to bring together methods from survey sampling, bandits, and active learning. It poses an interesting and novel challenge for the machine learning community and can benefit many public and private sector applications. It is critical to many U.S. federal agencies that are bound *by law* to balance enforcement priorities with population estimates of improper payments (Henderson et al., 2021; Office of Management and Budget, 2018, 2021).

2. Setting

The IRS selects taxpayers to audit every year to detect under-reported tax liability. Improving audit selection could yield 10:1 returns in revenue and help fund socially beneficial programs (Sarin and Summers, 2019). But the agency must also provide an accurate assessment of the tax gap (the projected amount of tax under-reporting if all taxpayers were audited). Based on a unique multiyear collaboration with the IRS, we were provided with full micro data access to masked audit data to research how machine learning could improve audit selection. We benchmark several sampling approaches and examine the trade-offs between them. We identify several interesting results and challenges using historical taxpayer audit data in collaboration with the IRS. The results of this and related research are currently being incorporated into the continual improvement of the IRS audit selection methods. Related work and more background on the IRS and the data we used are in Appendices A and B.

Notation. We formulate our models in a modified version of the structured bandit framework (Mersereau et al., 2009). At timestep t, there is a set $\mathcal{A}_t$ of N_t arms to choose from and a budget of K_t arms that can be chosen. After K_t arms are chosen, their rewards are revealed. The goal is to maximize the average reward of the chosen batch of arms at a given timestep. However, the agent has an additional requirement that it be able to yield an accurate (ideally unbiased) estimate of the average reward across all arms – even those that have not been chosen (the population reward). As in the structured bandit setting, we assume that the reward can be modeled by a function $r_t^a = f_{\theta^*}(X_t^a)$ where X_t^a are some set of features associated with arm a at timestep t, and θ^* are the parameters of the true reward function.

In our IRS setting each arm (a_t) represents a taxpayer which the model could select for a given year (t). The associated features (X_t^a) are the 500 covariates in our data for the tax return. The reward (r_t^a) is the adjustment recorded after the audit. The population average reward that the agent seeks to accurately model is the average adjustment (summing together would instead provide the tax gap).

3. Methods

We focus on three methods: (1) ϵ-greedy; (2) optimism-based approaches; (3) ABS sampling. Since (1) and (2) are standard bandit algorithms, we relegate a discussion of those approaches to Appendix C. For all methods we use a random forest regressor as the underlying model due to non-linearity of the data and a small data regime. See full version for more discussion on this point. We also compare an LDA baseline, which predicts whether the true reward is greater than \$200 (our no-change cutoff). Arms are selected based on increasing likelihood that they are part of the \$200+ reward class. This is included both as context to our broad modeling decisions, and as an imperfect stylized proxy for one component of the current risk-based selection approach used by the IRS.

Algorithm 1 ABS (Logistic)

```
Input: $\alpha$, $H$, $\zeta$, $K$, $(X_0, r_0)$
Train model $f_{\hat{\theta}}$ on initial data $(X_0, r_0)$.
for $t = 1, \ldots, T$ do
    Receive observations $X_t$
    Predict rewards $\hat{r}_a = f_{\hat{\theta}}(x_a)$.
    Sample top $\zeta$ predictions.
    $\forall_a \ \hat{\rho}_a \leftarrow (1 + \exp(-\alpha(\hat{r}_a - \kappa))^{-1}$
    Construct strata $S_1, \ldots, S_H$ by solving (1).
    Form distribution $\{\pi_h\}$ over strata via
    $\pi_h = \frac{\lambda_h}{\sum_{h'} \lambda_{h'}}$.
    repeat
        $h \sim (\pi_1, \ldots, \pi_H)$
        Sample arm uniformly at random from $S_h$.
    until $K - \zeta$ samples drawn
    Compute $\hat{\mu}_{\mathrm{HT}}$ once true rewards are collected.
    Retrain model $\hat{f}$ on $(\cup_i^t X_i, \cup_i^t r_i)$.
end for
```

ABS Sampling. Adaptive Bin Sampling guarantees statistically unbiased population estimates, while enabling an explicit trade-off between reward and the variance of the estimate. Pseudocode is given in Algorithm 1. Fix timestep t and let K be our budget. Let $\hat{r}_a = f_{\hat{\theta}}(X_t^a)$ be the predicted risk for return X_t^a. First we sample the top ζ returns. To make the remaining $K - \zeta$ selections, we parameterize the predictions with a logistic function, $\hat{\rho}_a = \frac{1}{1+\exp(-\alpha(\hat{r}_a - \kappa))}$ or an exponential function $\hat{\rho}_a = \exp(\alpha \hat{r}_a)$. κ is the value of the K-th largest value amongst reward predictions $\{\hat{r}_t^a\}$. As α decreases, $\{\hat{\rho}_t^a\}$ approaches a uniform distribution which results in lower variance for $\hat{\mu}(t)$ but lower reward. As α increases, the variance of $\hat{\mu}(t)$ increases but so too does the reward. Figure 2 provides a visualization.

The distribution of transformed predictions $\{\hat{\rho}_a\}$ is then stratified into H non-intersecting strata $S_1, \ldots, S_H$. We choose strata in order to minimize intra-cluster variance, such that there are at least $K - \zeta$ points per bin:

$$\begin{aligned} \min_{S_1,\ldots,S_H} \quad & \sum_h \sum_{\hat{\rho} \in S_h} \|\hat{\rho} - \lambda_h\|^2, \\ \text{s.t.} \quad & |S_h| \geq K - \zeta, \end{aligned} \tag{1}$$

where $\lambda_h = |S_h|^{-1} \sum_{\hat{\rho} \in S_h} \hat{\rho}$ is the average value of the points in bin b. We place a distribution (π_h) over the bins by averaging the risk in each bin: $\pi_h = \frac{\lambda_h}{\sum_{h'} \lambda_{h'}}$. To make our selection,

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased Methods	ABS-1	**\$41.5M***	**0.4** ✓	31.0	**37.6%**
	ϵ-only	\$41.3M*	4.3✓	37.4	38.3%
	ABS-2	\$40.5M*	0.6✓	24.5	38.3%
	Random	\$12.7M	1.5✓	**14.7**	53.1%
Biased Methods	Greedy	**\$43.6M***	16.4 ✗	8.8	**36.5%**
	UCB-1	\$42.4M*	15.3 ✗	9.4	38.6%
	ϵ-Greedy	\$41.3M*	**6.1** ✗	**7.5**	38.3%
	UCB-2	\$40.7M*	15.6 ✗	10.21	40.7%

Table 1: Best settings with overlapping CIs (*) on R. R is cumulative reward. μ_{PE} is average percent difference of the population estimate across seeds. σ_{PE} is standard deviation of the percent difference across seeds. μ_{NR} is the no change rate. See Appendix G.1 for hyperparameters. Biased methods with no guarantees are highly undesirable (✗). ϵ-only is the same as ϵ-Greedy, but population estimation uses only the ϵ sample as a random sample. Random is where the full, 600 arm, sample is random.

we sample $K - \zeta$ times from $(\pi_h, \ldots, \pi_H)$ to obtain a bin, and then we sample *uniformly* within that bin to choose the return. We do not recalculate $(\pi_1, \ldots, \pi_H)$ after each selection, so while we are sampling without replacement at the level of returns (we cannot audit the same taxpayer twice), we are sampling with replacement at the level of bins. Like with other HT-based methods Potter (1990); Alexander et al. (1997), to reduce variance we also add an option for a minimum probability of sampling a bin, which we call the trim %. See Appendix F for more details and proofs.

4. Evaluation Protocol

Here we provide the metrics used to judge the performance of the algorithms. The details of the experimental protocol are discussed in Appendix D. We use three metrics to evaluate our models: cumulative reward, percent difference of the population estimate, and the no-change rate. **Cumulative reward (R)** is simply the total reward of all arms selected by the agent across the entire time series. It represents the total amount of under-reported tax revenue returned to the government after auditing. This is averaged across seeds and denoted as R. **Percent difference (μ_{PE}, σ_{PE})** is the difference between the estimated population average and the true population average: $100\% * (\hat{\mu} - \mu^*)/\mu^*$. μ_{PE} is absolute mean percent difference across seeds (bias). σ_{PE} is the standard deviation of the percent difference across random seeds. **No-change rate (μ_{NR})** is the percent of arms that yield no reward (we round down such that any reward <\$200 is considered no change). NR is of some importance. An audit that results in no adjustment can be perceived as unfair, because the taxpayer did not commit any wrongdoing (Lawsky, 2008). It can have adverse effects on future compliance (Beer et al., 2015; Lederman, 2018). μ_{NR} is the average NR across seeds.

5. Results

We highlight several key findings. A more robust discussion can be found in Appendix H, and further results and sensitivity analyses can be found in the full version of the paper.

5.1 Unbiased population estimates are possible with little impact to reward

As seen in Table 1, ABS sampling can achieve similar returns to the best performing methods in terms of audit selection, while yielding an unbiased population estimate. Conversely, greedy, ϵ-greedy, and UCB approaches – which use a model-based population estimation method – do not achieve unbiased population estimates. ABS reduces variance by 16% over the best ϵ-only method, yielding even better reward.

5.2 Greedy is not all you need

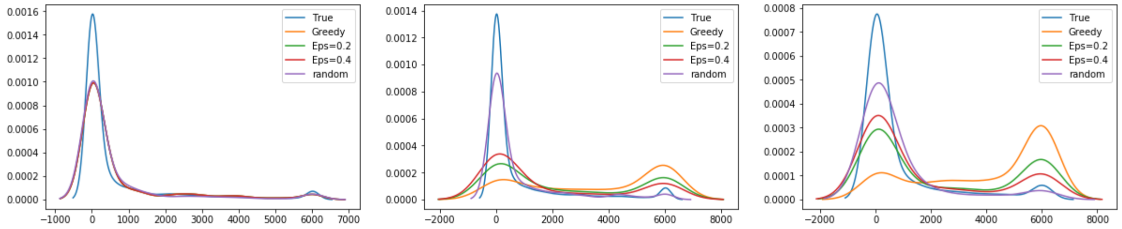


Figure 1: A kernel density plot of the distribution of sampled arms by reward from 2007, 2010, and 2014 (left to right). "True" refers to the density of all returns for the given year. X-axis is true reward. Y-axis is sampling distribution density.

Greedy surprisingly achieves more optimal reward compared to all other methods (see Table 1). This aligns with prior work suggesting that a purely greedy approach in contextual bandits might be enough to induce sufficient exploration under highly varied contexts (Bietti et al., 2018; Bastani et al., 2021). Figure 1 demonstrates greedy sampling's implicit exploration for one random seed. As the years progress, greedy is (correctly) more biased toward sampling arms with high rewards. Nonetheless, it yields a large number of arms that are the same as a random sample would yield. This inherent exploration backs the hypothesis that the test sample is highly stochastic, leading to implicit exploration. The key difference from our result and that of Bietti et al. (2018) and Bastani et al. (2021) is our additional population estimation objective. The greedy policy has a significant bias when it comes to model-based population estimation. Even if greedy can be optimal for a high-variance contextual bandit, it is not optimal for the optimize-and-estimate setting. ϵ-greedy achieves a compromise between variance that may be more acceptable in settings when some bias is permitted, but bias is not desirable in most public sector settings.

5.3 A more focused approach audits returns with larger under-reporting, which correlates with higher cumulative total positive income

A key motivator for our work is that inefficiently-allocated randomness (random samples that carry little new information over time) in audit selection will not only be suboptimal for the government, but could impose unnecessary burdens on taxpayers (Lawsky, 2008; Davis-Nozemack, 2012). Although we do not take a normative position on the precise contours of a fair distribution of audits, we examine how alternative models shape the income distribution of audited taxpayers. As shown in Figure 6a, we find that as methods become more optimal we see an increase in the total positive income (TPI) of the individuals selected for audit (RF Greedy selects between \$1.8M and \$9.4M more cumulative TPI than LDA Greedy, effect size 95% CI matched by seed). In Figure 6b we show the distribution of ABS hyperparameter settings we sampled. As the settings are more likely to increase reward and decrease no change rates, the cumulative TPI increases. This indicates that taxpayers with lower TPI are less likely to be audited as models are more likely to sample in the higher range of the risk distribution.

5.4 Errors are heteroskedastic, causing difficulties in using model-based optimism methods

We also find that, surprisingly, our optimism-based approach audits higher-income taxpayers more (\$1.2M to \$5.8M million cumulative TPI more than RF Greedy) despite yielding similar returns as the greedy approach. We believe this is because adjustments and model errors are heteroskedastic. Though TPI is correlated with the adjustment amount (Pearson $r = 0.49, p < 10^{-5}$), all errors across model fits were heteroskedastic according to a Breusch–Pagan test ($p < 10^{-5}$). A potential source of large uncertainty estimates in the high income range could be because: (1) there are fewer datapoints in that part of the feature space; (2) audits may not give an accurate picture of misreporting at the higher part of the income space, resulting in larger variance and uncertainty (Guyton et al., 2021); or (3) additional features are needed to improve precision in part of the state space. This makes it difficult to use some optimism-based approaches since there is a confound between aleatoric and epistemic uncertainty.

6. Discussion

We have introduced the optimize-and-estimate structured bandit setting. The setting is motivated by common features of public sector applications (e.g., multiple objectives, batched selection), where there is wide applicability of sequential decision making, but, to date, limited understanding of the unique methodological challenges. Our proposed ABS approach enables parties to explicitly trade off population estimation and reward maximization. We have shown how this framework addresses longstanding concerns in the real-world setting of IRS detection of tax evasion. It could shift audits toward tax returns with larger understatements and recover more revenue than the status quo, while maintaining an unbiased population estimate. Though there are other real-world objectives to consider, such as the effect of audit policies on tax evasion, our results suggest that unifying audit selection with estimation may help ensure that processes are as fair, optimal, and robust as possible. We hope that the methods we describe here are a starting point for both additional research into sequential decision-making in public policy and new research into optimize-and-estimate structured bandits.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Charles H Alexander, Scot Dahl, and Lynn Weidman. Making estimates from the american community survey. In *Annual Meeting of the American Statistical Association (ASA), Anaheim, CA*, 1997.

James Andreoni, Brian Erard, and Jonathan Feinstein. Tax compliance. *Journal of economic literature*, 36(2):818–860, 1998.

Elliott Ash, Sergio Galletta, and Tommaso Giommoni. A machine learning approach to analyze and support anti-corruption policy. 2021.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Sebastian Beer, Matthias Kasper, Erich Kirchler, and Brian Erard. Audit impact study. Technical report, National Taxpayer Advocate, 2015.

Andrew Belnap, Jeffrey L Hoopes, Edward L Maydew, and Alex Turk. Real effects of tax audits: Evidence from firms randomly selected for irs examination. *Available at SSRN 3437137*, 2020.

Jeremy Bertomeu, Edwige Cheynel, Eric Floyd, and Wenqiang Pan. Using machine learning to detect misstatements. *Review of Accounting Studies*, 26(2):468–519, 2021.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021.

Niladri S Chatterji, Aldo Pacchiano, Peter L Bartlett, and Michael I Jordan. On the theory of reinforcement learning with once-per-episode feedback. *arXiv preprint arXiv:2105.14363*, 2021.

Ben Chugg and Daniel E Ho. Reconciling risk allocation and prevalence estimation in public health using batched bandits. *NeurIPS workshop on Machine Learning in Public Health*, 2021.

Congressional Budget Office. Trends in the internal revenue service's funding and enforcement. 2020. URL `https://www.cbo.gov/publication/56467#_idTextAnchor002`.

Karie Davis-Nozemack. Unequal burdens in EITC compliance. *Law & Ineq.*, 31:37, 2012.

Daniel de Roux, Boris Perez, Andrés Moreno, Maria del Pilar Villamil, and César Figueroa. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 215–222, 2018.

Jason DeBacker, Bradley T Heim, Anh Tran, and Alexander Yuskavage. The effects of IRS audits on EITC claimants. *National Tax Journal*, 71(3):451–484, 2018.

Gabe Dickey, S Blanke, and L Seaton. Machine learning in auditing. *The CPA Journal*, pages 16–21, 2019.

Madalina M Drugan and Ann Nowé. Scalarization based pareto optimal set of arms identification algorithms. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2690–2697. IEEE, 2014.

Brian Erard and Jonathan S Feinstein. The individual income reporting gap: what we see and what we don't. In *IRS-TPC Research Conference on New Perspectives in Tax Administration*, 2011.

Akram Erraqabi, Alessandro Lazaric, Michal Valko, Emma Brunskill, and Yun-En Liu. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pages 709–717. PMLR, 2017.

Jessica Esteban, Ronald E McRoberts, Alfredo Fernández-Landa, José Luis Tomé, and Erik Næsset. Estimating forest volume and biomass and their changes using random forests and remotely sensed data. *Remote Sensing*, 11(16):1944, 2019.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

Government Accountability Office. New compliance research effort is on track, but important work remains. `https://www.gao.gov/assets/gao-02-769.pdf`, 2002. United States General Accounting Office: Report to the Committee on Finance, U.S. Senate. Online; Accessed Jan 10, 2022.

Government Accountability Office. Irs is implementing the national research program as planned. `https://www.gao.gov/assets/gao-03-614.pdf`, 2003. United States General Accounting Office: Report to the Committee on Finance, U.S. Senate. Online; Accessed Jan 10, 2022.

John Guyton, Kara Leibel, Dayanand S Manoli, Ankur Patel, Mark Payne, and Brenda Schafer. The effects of EITC correspondence audits on low-income earners. Technical report, National Bureau of Economic Research, 2018.

John Guyton, Patrick Langetieg, Daniel Reck, Max Risch, and Gabriel Zucman. Tax evasion by the wealthy: Measurement and implications. In *Measuring and Understanding the Distribution and Intra/Inter-Generational Mobility of Income and Wealth.* University of Chicago Press, 2020.

John Guyton, Patrick Langetieg, Daniel Reck, Max Risch, and Gabriel Zucman. Tax evasion at the top of the income distribution: Theory and evidence. Technical report, National Bureau of Economic Research, 2021.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

Peter Henderson, Ben Chugg, Brandon Anderson, and Daniel E. Ho. Beyond Ads: Sequential Decision-Making Algorithms in Law and Public Policy. *arXiv preprint arXiv:2112.06833*, 2021.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Ben Howard, Luci Lykke, David Pinski, and Alan Plumley. Can machine learning improve correspondence audit case selection? 2020.

Kuan-Hao Huang and Hsuan-Tien Lin. Linear upper confidence bound algorithm for contextual bandit problem with piled rewards. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 143–155. Springer, 2016.

Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pages 507–523. Springer, 2011.

Internal Revenue Service. Tax year 2006 tax gap estimate – summary of estimation methods. `https://www.irs.gov/pub/irs-soi/06rastg12methods.pdf`, 2012. Online; Accessed Jan 10, 2022.

Internal Revenue Service. Federal tax compliance research: Tax gap estimates for tax years 2008–2010, 2016.

Internal Revenue Service. Federal tax compliance research: Tax gap estimates for tax years 2011–2013, 2019.

Internal Revenue Service. IRS Update on Audits. `https://www.irs.gov/newsroom/irs-update-on-audits`, 2021. Online; Accessed Jan 10, 2022.

Andrew Johns and Joel Slemrod. The distribution of income tax noncompliance. *National Tax Journal*, 63(3):397, 2010.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Efficient adaptive experimental design for average treatment effect estimation. 2020.

Paul Kiel. It's getting worse: The IRS now audits poor americans at about the same rate as the top 1%. `https://www.propublica.org/article/irs-now-audits-poor-americans-at-about-the-same-rate-as-the-top-1-percent`, 2019. Online; Accessed Jan 10, 2022.

Sotiris Kotsiantis, Euaggelos Koumanakos, Dimitris Tzelepis, and Vasilis Tampakas. Predicting fraudulent financial statements with machine learning techniques. In *Hellenic Conference on Artificial Intelligence*, pages 538–542. Springer, 2006.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Benjamin Lansdell, Sofia Triantafillou, and Konrad Kording. Rarely-switching linear bandits: optimization of causal effects for the real world. *arXiv preprint arXiv:1905.13121*, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Sarah B Lawsky. Fairly random: On compensating audited taxpayers. *Conn. L. Rev.*, 41: 161, 2008.

Leandra Lederman. Does enforcement reduce voluntary tax compliance. *BYU L. Rev.*, page 623, 2018.

Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pages 161–168, 2014.

Victor L. Lowe. Statement Before the Subcommittee on Oversight House Committee on Ways and Means on How the Internal Revenue Service Selects and Audits Individual Income Tax Returns, 1976.

Chuck Marr and Cecile Murray. Irs funding cuts compromise taxpayer service and weaken enforcement. *http://www. cbpp. org/sites/default/files/atoms/files/6-25-14tax. pdf. Last accessed August*, 29:2016, 2016.

Adam J Mersereau, Paat Rusmevichientong, and John N Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12): 2787–2802, 2009.

Shekhar Mittal, Ofir Reich, and Aprajit Mahajan. Who is bogus? using one-sided labels to identify fraudulent firms from tax returns. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–11, 2018.

Subhojyoti Mukherjee, Ardhendu Tripathy, and Robert Nowak. Generalized chernoff sampling for active learning and structured bandit algorithms. *arXiv preprint arXiv:2012.08073*, 2020.

Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. In *International Conference on Machine Learning*, pages 3720–3729. PMLR, 2018.

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2018.

Office of Management and Budget. Requirements for payment integrity improvement. `https://www.whitehouse.gov/wp-content/uploads/2018/06/M-18-20.pdf`, 2018. Executive Office of the President. Online; Accessed Jan 10, 2022.

Office of Management and Budget. Requirements for payment integrity improvement. `https://www.whitehouse.gov/wp-content/uploads/2021/03/M-21-19.pdf`, 2021. Executive Office of the President. Online; Accessed Jan 10, 2022.

Frank J Potter. A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, volume 225230. American Statistical Association Washington, DC, 1990.

Anna N Rafferty, Huiji Ying, and Joseph Jay Williams. Bandit assignment for educational experiments: Benefits to students versus statistical power. In *International Conference on Artificial Intelligence in Education*, pages 286–290. Springer, 2018.

Natasha Sarin and Lawrence H Summers. Shrinking the tax gap: approaches and revenue potential. Technical report, National Bureau of Economic Research, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=H1aIuk-RW`.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. On the bias, risk, and consistency of sample means in multi-armed bandits. *SIAM Journal on Mathematics of Data Science*, 3(4):1278–1300, 2021.

Rafet Sifa, Anna Ladi, Maren Pielka, Rajkumar Ramamurthy, Lars Hillebrand, Birgit Kirsch, David Biesner, Robin Stenzel, Thiago Bell, Max L"ubbering, et al. Towards automated auditing with machine learning. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–4, 2019.

Dennis Soemers, Tim Brys, Kurt Driessens, Mark Winands, and Ann Nowé. Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Taxpayer Advocate Service. Improper earned income tax credit payments: Measures the IRS takes to reduce improper earned income tax credit payments are not sufficiently proactive and may unnecessarily burden taxpayers. `https://www.taxpayeradvocate.irs.gov/wp-content/uploads/2020/07/ARC18_Volume1_MSP_06_ImproperEarnedIncome.pdf`, 2018. 2018 Annual Report to Congress — Volume One. Online; Accessed Jan 10, 2022.

Cem Tekin and Eralp Turğay. Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing*, 66(14):3799–3813, 2018.

Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, pages 255–263. Hillsdale, NJ, 1993.

US Treasury. The american families plan tax compliance agenda. *Department of Treasury, Washington, DC*, 2021.

Eralp Turgay, Doruk Oner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, pages 1673–1681. PMLR, 2018.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The AI economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Acknowledgments

We would like to thank Emily Black, Hadi Elzayn, Robert Gillette, Tom Hertz, Dan Jurafsky, Mansheej Paul, Alan Plumley, and Evelyn Smith for helpful discussions. This work was supported by the Hoffman Yee program at Stanford's Institute for Human-Centered Artificial Intelligence and Arnold Ventures. PH is supported by the Open Philanthropy AI Fellowship. This work was conducted while BA was at Stanford University.

Appendix A. Related Work

There is growing interest in the application of ML to detecting fraudulent financial statements (Dickey et al., 2019; Bertomeu et al., 2021). Previous methods have included unsupervised outlier detection (de Roux et al., 2018), decision trees (Kotsiantis et al., 2006), and analyzing statements with NLP (Sifa et al., 2019). Closer to our methodology is a bandit approach is used by Soemers et al. (2018) to detect fraudulent credit card transactions. Meanwhile, Zheng et al. (2020) propose reinforcement learning to learn optimal tax policies, but do not focus on enforcement. Finally, some work has investigated the use of machine learning for improved audit selection in various settings (Howard et al., 2020; Ash et al., 2021; Mittal et al., 2018). None of these approaches takes into account population estimation and some do not use sequential decision-making. Some prior work has investigated general multi-objective optimization in the context of bandits Drugan and Nowé (2014); Tekin and Turğay (2018); Turgay et al. (2018). And a large body of work seeks to improve hypothesis testing efficiency via an active learning or structured bandit process (Kato et al., 2020; Mukherjee et al., 2020). The closest work to our own is that of Liu et al. (2014), Erraqabi et al. (2017), Rafferty et al. (2018), and Chugg and Ho (2021). In those works, the authors trade off maximizing some reward objective with estimation of a metric. Like other adaptive experimentation literature, these works are in the multi-armed bandit (MAB) setting (Lai and Robbins, 1985). Our work on the other hand deals with the unique challenges of the IRS setting, requiring the use of optimize-and-estimate *structured* bandits – where arms must be selected in batches, each with their own context, and rewards are delayed. Several works are close to the structured bandit approach we examine here (without population estimation). In those, the problem is framed as a contextual bandit where rewards are delivered in piles at some point later in time (Huang and Lin, 2016; Lansdell et al., 2019; Chatterji et al., 2021). To our knowledge the optimize-and-estimate structured bandit setting has not been proposed as we describe it here.

Appendix B. Background

Institutional Background. The IRS maintains two distinct categories of audit processes. National Research Program (NRP) audits enable at population estimation of non-compliance while Operational (Op) audits are aimed at collecting taxes from non-compliant returns. The NRP is a core measurement program for the IRS to regularly evaluate tax non-compliance (Government Accountability Office, 2002, 2003). The NRP randomly selects, via a stratified random sample, around 15k tax returns each year for research audits (Andreoni et al., 1998; Johns and Slemrod, 2010; Internal Revenue Service, 2016, 2019), although this has been decreasing in recent years and there is pressure to reduce it further (Marr and Murray, 2016; Congressional Budget Office, 2020). These audits are used to identify new areas of noncompliance, estimate the overall tax gap (Internal Revenue Service, 2016, 2019), and estimate improper payments of certain tax credits. Given the most recent gross tax gap estimate of $441 billion (Internal Revenue Service, 2019), even minor increases in efficiency can yield large returns.

In addition to its use for tax gap estimation, NRP serves as a training set for certain Op audit programs like the Discriminant Function (DIF) System,[2] which is based on a modified Linear Discriminant Analysis (LDA) model (Lowe, 1976). DIF also incorporates other measures and policy objectives that we do not consider here. We instead focus on the stylized setting of only population estimation and reward maximization. Tax returns that have a high likelihood of a significant adjustment, as calculated by the DIF model, are have a higher probability of being selected for Op audits. It is important to highlight that Op data is not used for estimating the DIF risk model and is not used for estimating the tax gap (specifically, the individual income misreporting component of the tax gap). Though NRP audits are jointly used for population estimates of non-compliance and risk model training, the original sampling design was not optimized for both revenue maximization and estimator accuracy for tax non-compliance. Random NRP audits have been criticized for burdening compliant taxpayers and for failing to target known areas of known non-compliance (Lawsky, 2008). We show that some randomness in auditing yields important benefits, but demonstrate how a unified process can efficiently maximize revenue while maintaining accurate estimates of the tax gap. To some extent, the current process already exists as an informal sequential decision-making system. NRP strata are informed by the Op distribution, and are adjusted year-to-year. We posit that by formalizing the current IRS system in the form of a sequential decision-making problem, we can incorporate more methods to improve the efficiency, accuracy, and fairness of the system.

Data. The data used throughout this work is from the NRP's random sample (Andreoni et al., 1998; Johns and Slemrod, 2010; Internal Revenue Service, 2016, 2019), which we will treat as the full population of audits, since they are collected via a stratified random

[2] `https://www.irs.gov/irm/part4/irm_04-001-002`

sample and represent the full population of taxpayers. The NRP sample is formed by dividing the taxpayer base into activity classes based on income and claimed tax credits, and various strata within each class. Each stratum is weighted to be representative of the national population of tax filers. Then a stratified random sample is taken across the classes. NRP audits seek to estimate the correctness of the whole return via a close to line-by-line examination (Belnap et al., 2020). This differs from Op audits, which are narrower in scope and focus on specific issues. Given the added expensive nature of NRP audits, NRP sample sizes are relatively small at around 15,000 audits per year (Guyton et al., 2018). The IRS uses these audits to estimate the tax gap and average non-compliance, with some statistical adjustments to compensate for some limitations to the NRP sampling approach, as well as for naturally occurring variation in the depth of audit (Guyton et al., 2020; Internal Revenue Service, 2012, 2016, 2019). These statistical adjustments will also compensate for taxpayer misreporting that is difficult to find via auditing, upweighting certain types of misreporting via a multiplier (Erard and Feinstein, 2011). For the goals of this work we ignore those adjustments.

Legal requirements for the accuracy of these estimates exist (Taxpayer Advocate Service, 2018). The 2018 Office of Management and Budget (OMB) guidelines, for instance, recommended that the estimates for these values be "statistically valid" (unbiased estimates of the mean) and have "±3% or better margin of error at the 95% confidence level for the improper payment percentage estimate" (Office of Management and Budget, 2018). Later OMB guidelines have provided more discretion to programs for developing feasible point estimates and confidence intervals (CIs) (Office of Management and Budget, 2021). Unbiasedness remains a policy priority.

Our NRP stratified random audit sample covers from 2006 to 2014. We use 500 covariates as inputs to the model. The covariates we use include every value reported by a taxpayer on a tax return. The full paper contains summary statistics of the NRP research audits conducted on a yearly basis. Since NRP audits are stratified, the unweighted means represent the average adjustment made by the IRS to that year's return for all audited taxpayers in the sample. The weighted mean takes into account stratification weights for each sample. One can think of the weighted mean as the average taxpayer misreporting across all taxpayers in the United States, while the unweighted mean is the average taxpayer misreporting in the NRP sample.[3]

Appendix C. ϵ-greedy and Optimism Algorithms

ϵ-greedy. Here we choose to sample randomly with probability ϵ. Otherwise, we select the observation with the highest predicted reward according to a fitted model $f_{\hat{\theta}}(X_t^a)$, where $\hat{\theta}$ indicates fitted model parameters. To batch sample, we repeat this process K times. The underlying model is then trained on the received observation-reward pairs, and we repeat. For population estimation, we use a model-based approach (see, e.g., Esteban et al. 2019). After the model receives the true rewards from the sampled arms, the population estimate is

[3] The IRS makes a number of additional statistical adjustments for reporting the tax gap that are outside the scope of this work.

predicted as: $\hat{\mu}(t) = \frac{1}{\sum_a w_a} \sum_{a \in \mathcal{A}_t} w_a f_{\hat{\theta}}(X_t^a)$, where w_a is the NRP sample weight[4] from the population distribution.

Optimism. We refer readers to Lattimore and Szepesvári (2020) for a general introduction to Upper Confidence Bound and optimism-based methods. We import an optimism-based approach into this setting as follows. Consider a random forest with B trees $T_1, T_2, \ldots, T_B$. We form an optimistic estimate of the reward for each arm according to: $\hat{\rho}_t^a = \frac{1}{B} \sum_b T_b(X_t^a) + Z\text{Var}_b(T_b(X_t^a))$, where Z is an exploration parameter based on the variance of the tree-based predictions, similar to Hutter et al. (2011). We select the K returns with the largest optimistic reward estimates. We shorthand this approach as UCB and use the same model-based population estimation method as ϵ-greedy.

Appendix D. Experimental Protocol

Our evaluation protocol for all experiments follows the same pattern. For a given year we offer 80% of the NRP sample as arms for the agent to select from. We repeat this process across 20 distinct random seeds such that there are 20 unique subsampled datasets that are shared across all methods, creating a sub-sampled bootstrap for Confidence Intervals (more detail in full version Comparing methods seed-to-seed will be the same as comparing two methods on the same dataset. Each year, the agent has a budget of 600 arms to select from the population of 10k+ arms (description of budget selection in Appendix E). We delay the delivery of rewards for one year. This is because the majority of audits are completed and returned only after such a delay (DeBacker et al., 2018). Thus, the algorithm in year 2008 will only make decisions with the information from 2006. Because of this delay the first two years are randomly sampled for the entire budget (i.e., there is a warm start). After receiving rewards for a given year, the agent must then provide a population estimate of the overall population average for the reward (i.e., the average tax adjustment after audit). This process repeats until 2014, the final year available in our NRP dataset.

Appendix E. Budget Selection

At each timestep the IRS can select a limited budget of samples – for example the NRP sample in 2014 was 14,357 audits. This is a tiny fraction of audits as compared to the general population of taxpayers – and thus impossible to replicate when using the NRP sample to evaluate selection mechanisms. The goal of the NRP sample is to select a large enough sample to approximate the taxpayer base. The parallel in our experiments would be to ensure we select a sample which is smaller than the coreset needed to model the entire data.

Another way of thinking about the size of the budget allowed per year to approximate the NRP mechanism is by determining what is the minimum random sample to achieve a 3% margin of error with 95% confidence of the NRP sample population, per the 2018 OMB requirements for IPERIA reporting. Using an 80% (from subsampling) sample of the per-year average of 14102 NRP yearly samples, we are left with an average of 11282 arms per

[4] The returns in each NRP strata can be weighted by the NRP sample weights to make the sample representative of the overall population, acting as inverse propensity weights. We choose to resample by NRP weights to get close to true population estimate. See full version for more discussion.

year. We should need about a 975 arm budget for a random sampling mechanism (ignoring stratification) to achieve OMB specifications.

We then use the approach of Sener and Savarese (2018) to find a minimal coreset which a model could use to achieve a reasonable fit. We first fit a random forest to the entire dataset for a given year and calculate the residuals (or the mean squared error across the dataset). We then iteratively select batches of 25 samples according to the method presented by Sener and Savarese (2018). We use only the raw features to compute distance (contrasting the embedding space used by the authors). We refit a random forest with the same hyperparameters as the optimal fit on the smaller coreset sampler. Then we calculate the ratio of the mean squared error on the entire year's data to the optimal mean squared error. We find that the mean squared error is reduced to roughly 2x the overfit model at around 600 coreset samples, reducing very slowly after that point. We find that around 600 samples is the absolute minimum number required to reduce the mean squared error to a stable level at 2 times the optimal mean squared error. To simulate the small sample sizes of the NRP selection, we select this smaller budget of 600 as our main evaluation budget size, corresponding to roughly 4% of the 2014 NRP sample.

Note that in practice, IRS faces heterogeneous costs for audits. For the purposes of this work, we assume a fixed budget of arm/tax returns rather than a fixed budget of auditor hours.

Appendix F. ABS Sampling

In this section we provide further details on ABS, verify that the populate estimate is unbiased, and make some general remarks on the effects of various parameters on the variance of the estimate. Algorithm 1 gives an overview of ABS with logistic smoothing.

Fix a timestep t and let K be our budget. Let $\hat{r}_a = f_{\hat{\theta}}(X_t^a)$ be the predicted risk for return X_t^a. First we sample the top ζ returns. To make the remaining $K - \zeta$ selections, we parameterize the predictions with a logistic function,

$$\hat{\rho}_a = \frac{1}{1 + \exp(-\alpha(\hat{r}_a - \kappa))},$$

or an exponential function

$$\hat{\rho}_a = \exp(\alpha \hat{r}_a).$$

κ is the value of the K-th largest value amongst reward predictions $\{\hat{r}_t^a\}$. For the logistic we normalize such that $\hat{r}_a \in [-5, 5]$, and $\hat{r}_a \in [0, 1]$ for the exponential. The distribution of transformed predictions $\{\hat{\rho}_a\}$ is then stratified into H non-intersecting strata $S_1, \ldots, S_H$. We choose the strata in order to minimize intra-cluster variance, subject to the constraint of having at least $K - \zeta$ points per bin:

$$\min_{S_1,\ldots,S_H} \quad \sum_h \sum_{\hat{\rho} \in S_h} \|\hat{\rho} - \lambda_h\|^2, \quad \text{s.t.} \quad |S_h| \geq K - \zeta, \tag{2}$$

where $\lambda_h = |S_h|^{-1} \sum_{\hat{\rho} \in S_h} \hat{\rho}$ is the average value of the points in bin b. Note that $\sum_h \sum_{\hat{\rho} \in S_h} \|\hat{\rho} - \lambda_h\|^2 = \sum_h \sum_{\hat{\rho} \in S_h} |S_h| \text{Var} S_h$, so the quadratic program (1) is indeed minimizing intra-cluster variance.

We place a distribution (π_h) over the bins by averaging the risk in each bin, i.e,

$$\pi_h = \frac{\lambda_h}{\sum_{h'} \lambda_{h'}}. \tag{3}$$

To make our selection, we sample $K - \zeta$ times from $(\pi_h, \dots, \pi_H)$ to obtain a bin, and then we sample *uniformly* within that bin to choose the return. We do not recalculate $(\pi_1, \dots, \pi_H)$ after each selection, so while we are sampling without replacement at the level of returns (we cannot audit the same taxpayer twice), we are sampling with replacement at the level of bins. This is (i) because of computational feasibility, and (ii) in order to obtain an unbiased estimate of the mean via Horvitz-Thompson Sampling Horvitz and Thompson (1952).

In particular, note that the probability that arm a in stratum S_h is sampled is $p_a = (K - \zeta)\pi_h / N_h$ (see the next subsection for a derivation), where $N_h = |S_h|$ is the size of S_h. If $\mathcal{K}$ is the set of returns chosen for auditing and S_{H+1} contains those ζ points first sampled, then

$$\hat{\mu}_{\text{HT}}(t) = \frac{1}{\sum_a w_a} \bigg(\sum_{a \in \mathcal{K} \setminus S_{H+1}} \frac{w_a r_a}{p_a} + \sum_{a \in S_{H+1}} w_a r_a \bigg),$$

is an unbiased estimator of the true mean

$$\mu(t) = \frac{1}{\sum_a w_a} \sum_a w_a r_a.$$

To see this, let $\mathbf{1}_{a \in \mathcal{K}}$ be the random variable indicating whether arm a is sampled. Since $\mathbb{E}[\mathbf{1}_{a \in \mathcal{K}}] = p_a$, linearity of expectation gives that

$$\begin{aligned}
\mathbb{E}[\hat{\mu}_{\text{HT}}] &= \frac{1}{\sum_a w_a} \mathbb{E}\bigg[\sum_{a \in \mathcal{A} \setminus S_{H+1}} \frac{w_a r_a}{p_a} \mathbf{1}_{a \in \mathcal{K}} + \sum_{a \in S_{H+1}} w_a r_a \bigg] \\
&= \frac{1}{\sum_a w_a} \bigg(\sum_{a \in \mathcal{A} \setminus S_{H+1}} \frac{w_a r_a}{p_a} \mathbb{E}[\mathbf{1}_{a \in \mathcal{K}}] + \sum_{a \in S_{H+1}} w_a r_a \bigg) \\
&= \frac{1}{\sum_a w_a} \sum_{a \in \mathcal{A}} w_a r_a.
\end{aligned}$$

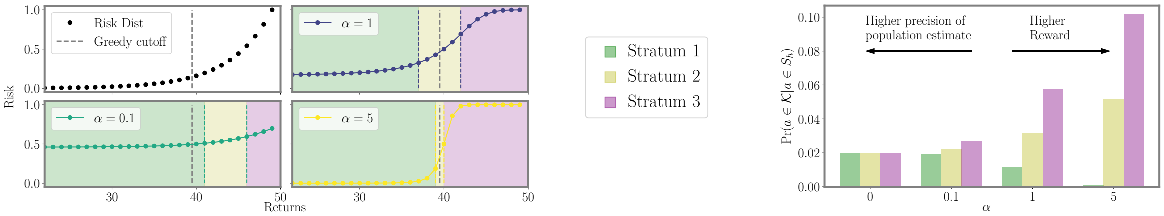
F.1 Variance of Population Estimate

Write the Horvitz-Thompson estimator as

$$\hat{\mu}_{\text{HT}} = \frac{1}{N} \sum_a \frac{r_a}{p_a} \mathbf{1}_{a \in \mathcal{K}},$$

where $\mathcal{K}$ is the set of selected arms and $p_a = \Pr(a \in \mathcal{K})$ is arm a's inclusion probability in $\mathcal{K}$. Then

$$\begin{aligned}
\text{Var}(\hat{\mu}_{\text{HT}}) &= \frac{1}{N^2} \sum_{a,b} \frac{r_a r_b}{p_a p_b} \text{Cov}(\mathbf{1}_{a \in \mathcal{K}} \mathbf{1}_{b \in \mathcal{K}}) \\
&= \frac{1}{N^2} \bigg(\sum_a \frac{r_a^2}{p_a}(1 - p_a) + \sum_a \sum_{b \neq a} \frac{r_a r_b}{p_a p_b}(p_{a,b} - p_a p_b) \bigg),
\end{aligned}$$



(a) Risk distribution and parameterizations

(b) Probability of sampling an individual

Figure 2: Illustration of ABS on 50 synthetic observations. (a) Hypothetical risk distribution and three parameterizations corresponding to different values of α: 0.1, 1, and 5. Greedy selection, represented by the dotted (gray) line in each panel would choose the $K = 10$ returns with the highest risk. The parameterized risk distributions are clustered into three strata (S_1, S_2, S_3), represented by the colored panels. As α varies, the cluster assignments change. (b) Probabilities of sampling a single individual from the three strata. As α increases, more weight is put onto the higher risk returns (Stratum 3).

where $p_{a,b} = \Pr(a, b \in \mathcal{K}) = p_{b,a}$ is the joint inclusion probability of arms a and b. Note that for the ζ arms in stratum S_{h+1}, $p_a = 1$ and $p_{a,b} = p_b$. Therefore, all terms involving such arms are zero and they do not contribute to the variance.

We can make this expression more specific to our case by rewriting the inclusion probabilities as functions of the strata. Fix an arm a and suppose it's in stratum S_h. Let $m = K - \zeta$ be the number of returns we're randomly sampling (i.e., discarding those ζ points greedily chosen from the top of the risk distribution). The law of total probability over the m trials gives

$$p_a = \sum_{\ell=0}^{m} \Pr(a \in \mathcal{K} || S_h \cap \mathcal{K}| = \ell) \Pr(|S_h \cap \mathcal{K}| = \ell).$$

The first term in the product is the probability that a is chosen as one of ℓ elements in a bucket of size $N_h = |S_h|$. The second term is the probability that S_h was selected precisely ℓ times and is distributed as a binomial. Therefore,

$$p_a = \sum_{\ell=0}^{m} \frac{\ell}{N_h} \binom{K}{\ell} \pi_h^{\ell} (1 - \pi_h)^{m-\ell} = \frac{m\pi_h}{N_h}.$$

Now consider $p_{a,b}$ for distinct arms a, b. Let $b \in S_g$. Conditioning on $b \in \mathcal{K}$ gives $\Pr(a \in \mathcal{K} | b \in \mathcal{K}) = \frac{(m-1)\pi_h}{N_h}$ if $g \neq h$ since there are now $m - 1$ trials to select a. If $g = h$, then $\Pr(a \in \mathcal{K} | b \in \mathcal{K}) = \frac{(K-1)\pi_h}{N_h - 1}$ since there are $m - 1$ trials to select a from a bin of size $N_h - 1$. Thus

$$\begin{aligned} p_{a,b} &= \Pr(a \in \mathcal{K} | b \in \mathcal{K}) \Pr(b \in \mathcal{K}) \\ &= \begin{cases} \frac{m(m-1)\pi_h \pi_g}{N_h N_g}, & \text{if } g \neq h, \\ \frac{m(m-1)\pi_h^2}{N_h (N_h - 1)}, & \text{if } g = h. \end{cases} \end{aligned}$$

Rewriting the variance as a summation over the strata, we see that the variance is the difference of two terms V_1 and V_2 where V_1 as a sum across all strata and V_2 includes cross-terms dependent on the relationship between strata.

$$\mathrm{Var}(\hat{\mu}_{\mathrm{HT}}) = \frac{1}{mN^2}(V_1 - V_2),$$

where

$$V_1 = \sum_{h=1}^{H}(N_h \pi_h^{-1} - m) \sum_{a \in S_h} r_a^2,$$

and

$$V_2 = \sum_{h=1}^{H} \sum_{a \in S_h} r_a \left(\frac{N_h - m}{N_h - 1} \sum_{b \in S_h \setminus a} r_b + \sum_{g \neq h} \sum_{b \in S_g} r_b \right).$$

We make a few remarks on the variance here, but leave a full analysis to future work. If the budget is small relative to the strata sizes (as is the case here), then $\frac{N_h - m}{N_h - 1} \approx 1$, and V_2 reduces to $\sum_a \sum_{b \neq a} r_a r_b$ which is independent of the strata. As α grows and we place more weight on those returns deemed higher risk by the model, $p_a \to 0$ for lower risk arms. This results in many arms clustered in a few strata with high N_h and low π_h, which increases V_1. Also, as ζ grows and we perform more greedy sampling, m decreases and the variance increases roughly proportionally.

Appendix G. Hyperparameter Tuning

The ideal approach would tune the hyperparameters for function approximators using cross-validation every time the model is fit in the active learning process. However, we found this approach to be extremely computationally expensive, a small grid of experiments requiring over a week to run. In the interest of reducing energy consumption – see, for example, discussion by Henderson et al. (2020) – we instead opt for a less computationally expensive proxy. We take train our function approximators on 2006 and then evaluate their RARE score and population estimate for the 2008 year, using 5-fold cross-validation across both years. We then run a grid search for all function approximators used. Finally we find the top point on the smoothed Pareto frontier between RARE and population estimation to find the optimal hyperparameters. To do this we rank the reward and population estimation criteria based on Since there is concept drift from year to year, we expect that these hyperparameters are sub-optimal and results may be even further improved with careful per-year hyperparameter tuning. However, this approach is sufficient for the purposes of our experiments.

For handling hyperparameters of the sampling algorithms, we rely on sensitivity analyses rather than hyperparameter searches. This is in line with recent work that promotes reporting results over ranges of hyperparameters and random seeds, particularly for sequential decision-making systems (Henderson et al., 2018; Bouthillier et al., 2021; Agarwal et al., 2021). Hyperparameter grids were run as follows in Table 2.

For Table 1, the only place where we display single hyperparameter settings, we use the following selection protocol from the grid of hyperparameters above. Given the lack of longitudinal data, we rely on intra-year data sub-sampling for validation sets. We select 5 random seeds, corresponding to different validation subsampled datasets. First, we identify the top reward band by finding methods with overlapping confidence intervals on reward. See Joseph et al. (2018) for discussion on meritocratic fairness given overlapping

Method	Hyperparameter Grid
ϵ-Greedy	$\epsilon = 0.0, 0.1, 0.2, 0.4$ weighted, unweighted fit
UCB	$Z = 0.1, 1.0, 10.0, 100.0$ weighted, unweighted fit
ABS	mixing=exponential, logistic $\alpha = 0.1, 0.5, 1.0, 1.5, 2.0, 5, 10, 15$ $\zeta = 0.0, 0.2, 0.4, 0.6, 0.8$ trim=0%, 2.5%, 5% weighted, unweighted fit

Table 2: Hyperparameter grids

confidence intervals. Then, we order results by root mean squared error and select the top hyperaparameters for each method according to these validation seeds. We then use results from all train seeds to get results displayed in the paper.

G.1 Table 1 Hyperparameters

In Table 1 1, ABS-1 is a hyperparameter configuration that focuses slightly more on reward at the cost of population estimation variance. ϵ-Greedy uses an ϵ of 0.1, UCB-1 has $Z = 1$, UCB-2 has a larger exploration factor of $Z = 10$. ABS-1 uses an exponential mixing function with 80% greedy sample, $\alpha = 5$, and a 2.5% trim factor. ABS-2 uses a logistic mixing function, $\alpha = 0.5$, a 5% trim, and 80% greedy sample. Both ABS methods use an unweighted fit while all other approaches saw improved results with a weighted fit.

Appendix H. Results

H.1 Unbiased population estimates are possible with little impact to reward

As seen in Table 1, ABS sampling can achieve similar returns to the best performing methods in terms of audit selection, while yielding an unbiased population estimate. Conversely, greedy, ϵ-greedy, and UCB approaches – which use a model-based population estimation method – do not achieve unbiased population estimates. Others have noted that adaptively collected data often have negative bias (Nie et al., 2018; Neel and Roth, 2018), leading to biased models. In many public sector settings provably unbiased methods like ABS are *required*. For ϵ-greedy, using the ϵ-sample only would also achieve an unbiased estimate, yet due to its small sample size the variance is prohibitively high. ABS reduces variance by 16% over the best ϵ-only method, yielding even better reward. Trading off $1M over 9 years improves variance over ϵ-Greedy (ϵ-only) by 35%. It is possible to reduce this variance even further at the cost of some more reward (see Figure 3). Note, due to an extremely small sample size, though the ϵ sample is unbiased in theory, we see some minor bias in practice. Model-based estimates are significantly lower variance, but biased. This may be because models re-use information across years, whereas ABS does not. Future research could re-use information in ABS to reduce variance, perhaps with a model's assistance. Nonetheless, we emphasize that model-based estimates without unbiasedness guarantees are unacceptable for many public sector uses from a policy perspective.

H.2 ABS allows fine-grained control over variance-reward trade-off

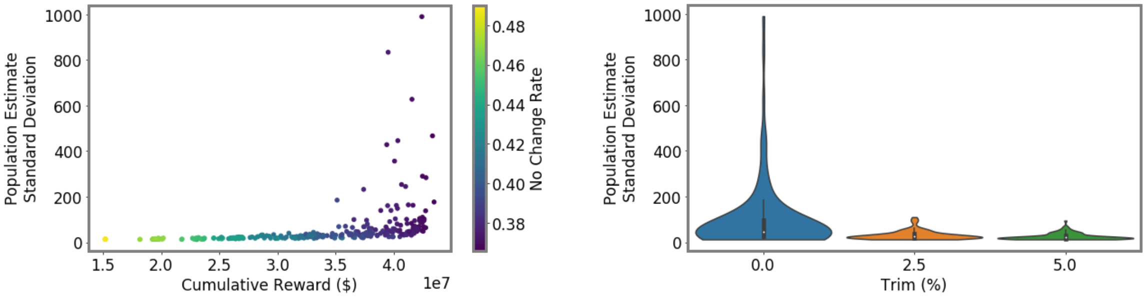


Figure 3: (Left) Population estimation empirical standard deviation versus reward for a grid of ABS hyperparameters. (Right) Population estimation variance as a function of the trimming factor.

We sample a grid of hyperparameters for ABS (see Appendix G). Figure 3 shows that more hyperparameter settings close to optimal rewards have higher variance in population estimates. We can control this variance with the trimming mechanism. This ensures that each bin of the risk distribution will be sampled some minimum amount. Figure 3 also shows that when we add trimming, we can retain large rewards and unbiased population estimates. Top configurations (Table 1) can keep variance down to only 1.7x that of a random sample, while yielding 3.2x reward. Overall, ABS allows for fine-grained control of the population estimation variance and reward trade-off.

H.3 Greedy is not all you need

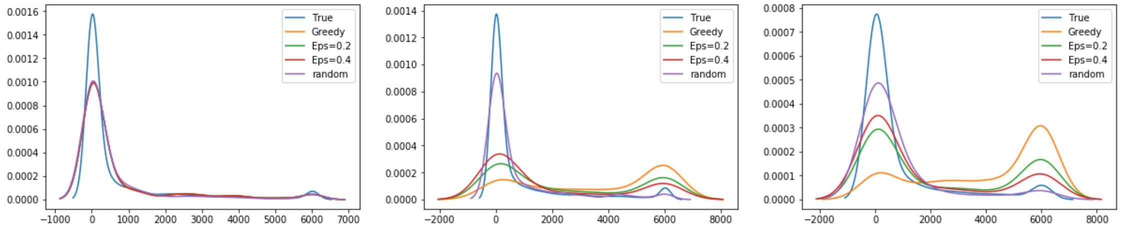


Figure 4: A kernel density plot of the distribution of sampled arms by reward from 2007, 2010, and 2014 (left to right). "True" refers to the density of all returns for the given year. X-axis is true reward. Y-axis is sampling distribution density.

Greedy surprisingly achieves more optimal reward compared to all other methods (see Table 1). This aligns with prior work suggesting that a purely greedy approach in contextual bandits might be enough to induce sufficient exploration under highly varied contexts (Bietti et al., 2018; Bastani et al., 2021). Here, there are several intrinsic sources of exploration that may cause this result: intrinsic model error, covariate drift (see full version for more discussion of covariate drift), differences in tax filing compositions, and the fact that our population of arms already come from a stratified random sample (changing in composition year-to-year).

Figure 1 demonstrates greedy sampling's implicit exploration for one random seed. As the years progress, greedy is (correctly) more biased toward sampling arms with high rewards. Nonetheless, it yields a large number of arms that are the same as a random sample would yield. This inherent exploration backs the hypothesis that the test sample is highly stochastic, leading to implicit exploration. It is worth emphasizing that in a larger population and with a larger budget, greedy's exploration may not be sufficient and more explicit exploration may be needed.

The key difference from our result and that of Bietti et al. (2018) and Bastani et al. (2021) is our additional population estimation objective. The greedy policy has a significant bias when it comes to model-based population estimation. This bias is similar – but not identical – to the bias reported in other adaptive data settings (Thrun and Schwartz, 1993; Nie et al., 2018; Shin et al., 2021; Farquhar et al., 2021). Even a 10% random sample – significantly underpowered for typical sampling-based estimation – can reduce this bias by more than 2.5× (see Table 1). Even if greedy can be optimal for a high-variance contextual bandit, it is not optimal for the optimize-and-estimate setting. ϵ-greedy achieves a compromise between variance that may be more acceptable in settings when some bias is permitted, but bias is not desirable in most public sector settings.

H.4 Using RFR significantly outperforms LDA and incorporating non-random data helps

Figure 5 shows that the cumulative return of ϵ-greedy sampling strategies using RFR-based approaches are significantly higher than LDA-based approaches. We emphasize again that the LDA model we use here is a *stylized* approximation of the current risk-based selection process and does not incorporate other policy objectives and confidential mechanisms of DIF. We note that while it serves as a *rough* baseline model, it demonstrates again that regression-based and globally non-linear models are needed for optimal performance in complex administrative settings such as this.

Fitting to *only* the random data, even with the RFR approach, reduces the ability of the model to make comparable selections and increases the variance across selection strategies. This can be seen in Figure 5, where fitting to random only leads to between \$6.2M and \$1.3M less reward cumulatively (95% CIs on effect size) and standard deviation across seeds is increased by \$700k. Since the current risk-selection approach mainly uses the NRP data, this suggests that future work on incorporating Op audits into the model training mechanism could improve overall risk-selection. This of course bears the risk of exacerbating biases and should be done with careful correction for data imbalances.

H.5 A more focused approach audits returns with larger under-reporting, which correlates with higher cumulative total positive income

A key motivator for our work is that inefficiently-allocated randomness (random samples that carry little new information over time) in audit selection will not only be suboptimal

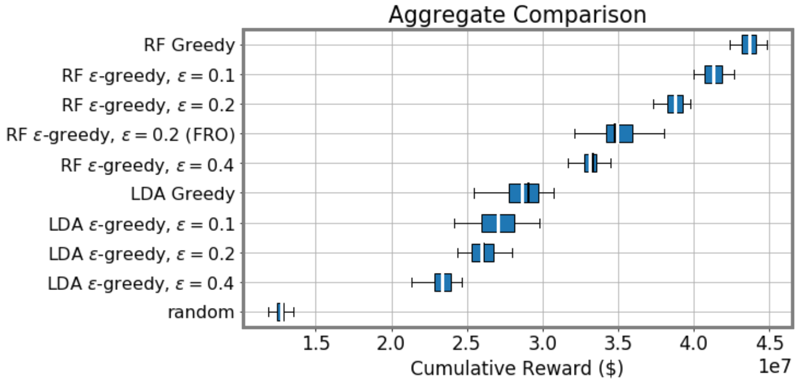
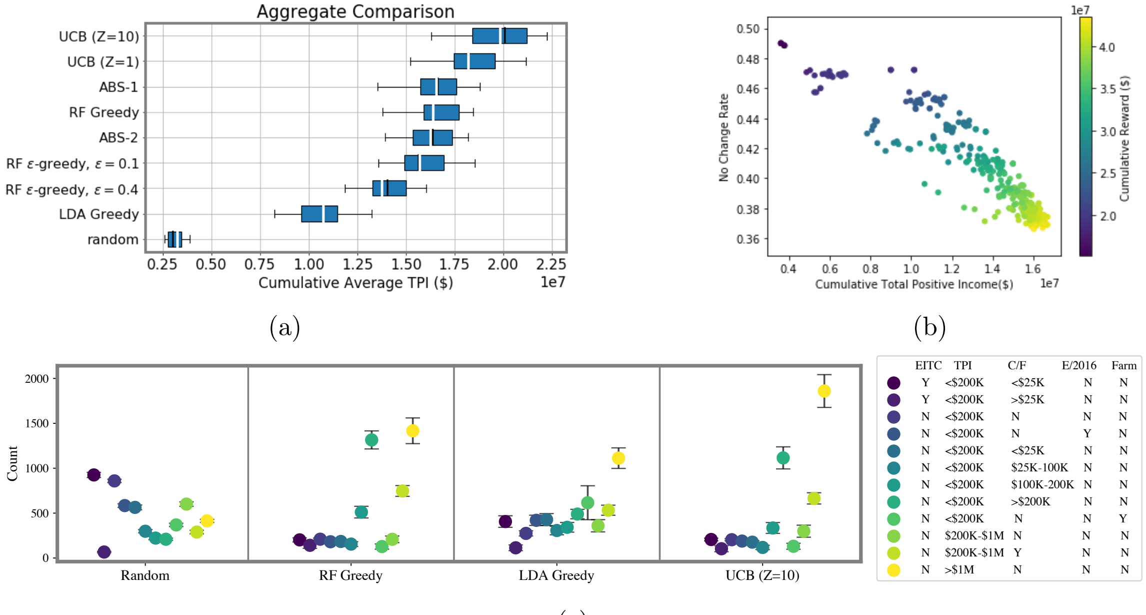


Figure 5: A comparison of cumulative reward earned between LDA-based approaches and RFR-based approaches. FRO indicates that the RFR was fit to the random sample only.



(a)

(b)

(c)

Figure 6: (a) The cumulative average total positive income (TPI) of all batches of taxpayers audited by each algorithm. (b) Across all ABS hyperparameters sampled, which allows fine-grained control over the population estimate variance and reward trade-off, the relationship between reward, no change rate, and cumulative average TPI. (c) The distribution of audit classes for several approaches. Top ABS selections are similar to RF Greedy. More details and more information on labels for (c) found in full version.

for the government, but could impose unnecessary burdens on taxpayers (Lawsky, 2008; Davis-Nozemack, 2012). In particular, an issue that has received increasing attention by policymakers and commentators in recent years concerns the fair allocation of audits by income (Kiel, 2019; Internal Revenue Service, 2021; Treasury, 2021). Although we do not take a normative position on the precise contours of a fair distribution of audits, we examine how alternative models shape the income distribution of audited taxpayers.

As shown in Figure 6a, we find that as methods become more optimal we see an increase in the total positive income (TPI) of the individuals selected for audit (RF Greedy selects between \$1.8M and \$9.4M more cumulative TPI than LDA Greedy, effect size 95% CI matched by seed). In Figure 6b we show the distribution of ABS hyperparameter settings we sampled. As the settings are more likely to increase reward and decrease no change rates, the cumulative TPI increases. This indicates that taxpayers with lower TPI are less likely to be audited as models are more likely to sample in the higher range of the risk distribution. We confirm this in Figure 6c which shows the distribution of activity classes sampled by different approaches. These classes are used as strata in the NRP sample. The UCB and RF Greedy approaches are more likely to audit taxpayers with more than \$1M in TPI (with UCB sampling this class significantly more, likely due to heteroskedasticity). More optimal approaches also significantly sample those with <\$200K in TPI, but more than \$200K reported on their Schedule C or F tax return forms (used to report business and farm income, respectively).

H.6 Errors are heteroskedastic, causing difficulties in using model-based optimism methods

We also find that, surprisingly, our optimism-based approach audits taxpayers with higher TPI more (\$1.2M to \$5.8M million cumulative TPI more than RF Greedy) despite yielding similar returns as the greedy approach. We believe this is because adjustments and model errors are heteroskedastic. Though TPI is correlated with the adjustment amount (Pearson $r = 0.49, p < 10^{-5}$), all errors across model fits were heteroskedastic according to a Breusch–Pagan test ($p < 10^{-5}$). A potential source of large uncertainty estimates in the high income range could be because: (1) there are fewer datapoints in that part of the feature space; (2) audits may not give an accurate picture of misreporting at the higher part of the income space, resulting in larger variance and uncertainty (Guyton et al., 2021); or (3) additional features are needed to improve precision in part of the state space. This makes it difficult to use some optimism-based approaches since there is a confound between aleatoric and epistemic uncertainty. As a result, optimism-based approaches audit higher income individuals more often, but do not necessarily achieve higher returns. This poses another interesting challenge for future research.