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

Peter Henderson[1], Ben Chugg[2], Brandon Anderson[3], Kristen Altenburger[1], Alex Turk[3], John Guyton[3], Jacob Goldin[4], Daniel E. Ho[1]

[1] Stanford University
[2] Carnegie Mellon University
[3] Internal Revenue Service
[4] University of Chicago

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 analyze optimize-and-estimate approaches to the IRS problem and propose 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 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) provides a review of this potential. 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 with these multiple objectives under a sequential decision-making 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 (see more discussion in Appendix C).[1] 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).

We highlight this framework with a case study of the Internal Revenue Service (IRS). 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). Currently, the IRS accomplishes this via two separate mechanisms: (1) a stratified random sample to estimate the tax gap; (2) a focused risk-selected sample of taxpayers to collect under-reported taxes. 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 investigate whether these separate mechanisms and objectives can be combined into one batched structured bandit algorithm, which must both maximize reward and maintain accurate population estimates. Ideally, if information is reused, the system can make strategic selections to balance the two objectives. We benchmark several sampling approaches and examine the trade-offs between them with the goal of understanding the effects of using bandit algorithms in this high-impact setting. We identify several interesting results and challenges using historical taxpayer audit data in collaboration with the IRS.

First, we introduce a novel sampling mechanism called *Adaptive Bin Sampling* (ABS) which guarantees an unbiased population estimate by employing a Horvitz-Thompson (HT) approach (Horvitz and Thompson 1952), but is comparable to other methods for cumulative reward. Its unbiasedness and comparable reward comes at the cost of additional vari-

[1] Appendices can be found at: https://arxiv.org/abs/2204.11910.

ance, though the method provides fine-grained control of this variance-reward trade-off.

Second, we compare this approach to ϵ-greedy and optimism-based approaches, where a model-based population estimate is used. We find that model-based approaches are biased absent substantial reliance on ϵ, but low in variance. Surprisingly, we find that greedy approaches perform well in terms of reward, reinforcing findings by Bietti, Agarwal, and Langford (2018) and Bastani, Bayati, and Khosravi (2021). But we find the bias from population estimates in the greedy regime to be substantial. These biases are greatly reduced even with small amounts of random exploration, but the lack of unbiasedness guarantees make them unacceptable for many public policy settings.

Third, we show that more reward-optimal approaches tend to sample high-income earners versus low-income earners. And more reward-optimal approaches tend to audit fewer tax returns that yield no change (a reward close to 0). This reinforces the importance of reducing the amount of unnecessary exploration, which would place audit burdens on compliant taxpayers. Appendix D details other ethical and societal considerations taken into account with this work.

Fourth, we show that model errors are heteroskedastic, resulting in more audits of high-income earners by optimism-based methods, but not yielding greater rewards.[2]

We demonstrate that combining random and focused audits into a single framework can more efficiently maximize revenue while retaining accuracy for estimating the tax gap. While additional research is needed in this new and challenging domain, this work demonstrates the promise of applying a bandit-like approach to the IRS setting, and optimize-and-estimate structured bandits more broadly. The results of this and related research are currently being incorporated into the continual improvement of the IRS audit selection methods.

2 Background

Related Work. The bandit literature is large. To fully engage with it, we provide an extended literature review in Appendix E , but we mention several strands of related research here. The fact that adaptively collected data leads to biased estimation (whether model-based or not) is well-known. See, e.g., Nie et al. (2018); Xu, Qin, and Liu (2013); Shin, Ramdas, and Rinaldo (2021). A number of works have sought to develop sampling strategies that combat bias. See, e.g. Dimakopoulou et al. (2017). This work has been in the multi-armed bandit (MAB) or (mostly linear) contextual bandit settings. In the MAB setting, there has also been some work which explicitly considers the trade-off between reward and model-error. See, e.g, Liu et al. (2014); Erraqabi et al. (2017). In Appendix E we provide a comparison against our setting, but crucially we have volatile arms which make our setting different and closer to the linear stochastic bandit work (a form of structured bandit) (Abbasi-Yadkori, Pál, and Szepesvári 2011; Joseph et al. 2018). However, we require non-linearity and batched selection, as well as adding the novel estimation objective to this structured bandit setting. To our knowledge, ours is the first formulation which actively incorporates bias and variance of population estimates into a batched structured bandit problem formulation. Moreover, our focus is to study this problem in a real-world public sector domain, taking on the challenges proposed by Wagstaff (2012). No work we are aware of has analyzed the IRS setting in this way.

Institutional Background. The IRS maintains two distinct categories of audit processes. National Research Program (NRP) audits enable population estimation of noncompliance 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, $\sim$15k tax returns each year for research audits (Internal Revenue Service 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, and estimate improper payments of certain tax credits. Given a 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 (Internal Revenue Service 2022), 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 DIF, 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 audits have been criticized for burdening compliant taxpayers and for failing to target areas of known non-compliance (Lawsky 2008). The current process already somewhat represents 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 its efficiency, accuracy, and fairness.

Data. The data used throughout this work is from the NRP's random sample (Andreoni, Erard, and Feinstein 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 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

[2]We note that it is possible that these stem from measurement limitations in the high income space (Guyton et al. 2021).

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 expensive nature of NRP audits, NRP sample sizes are relatively small ($\sim$15k/year) (Guyton et al. 2018). The IRS uses these audits to estimate the tax gap and average non-compliance.[3] Legal requirements for these estimates exist (Taxpayer Advocate Service 2018). The 2018 Office of Management and Budget (OMB) guidelines, for instance, state that these values should be "statistically valid" (unbiased estimates of the mean) and have "$\pm 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 an IRS policy priority.

Our NRP stratified random audit sample covers from 2006 to 2014. We use 500 covariates as inputs to the model which are a superset of those currently used for fitting the DIF model. The covariates we use include every value reported by a taxpayer on a tax return. For example, the amount reported in Box 9 of Form 1040 is Total income and would be included in these covariates. Table 5 , in the Appendix, provides 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.

Problem Formulation. We formulate the optimize-and-estimate structured bandit problem setting in the scenario where there is an extremely large, but finite, number (N_t) of arms ($a \in \mathcal{A}_t$) to select from at every round. This set of arms is the population at timestep t. The population can vary such that the set of available arms may be totally different at every step, similar to a sleeping or volatile bandit (Nika, Elahi, and Tekin 2021). In fact, it may not be possible to monitor any given arm over several timesteps.[4] To make the problem tractable, it is assumed that the reward for a given arm can be modeled by a shared 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. Assume $f \in \mathcal{F}$ is any realizable or ϵ-realizable function. Thus, as is typical of the structured bandit setting "choosing one action allows you to gain information about the rewards of other actions" (Lattimore and Szepesvári 2020, p. 301). The agent chooses a batch of K_t arms to: (1) maximize reward; (2) yield an accurate and unbiased estimate of the average reward across all arms – even those that have not been chosen (the population reward). Thus we seek to optimize a selection algorithm that chooses non-overlapping actions $(\hat{a}_1, ..., \hat{a}_K)$ according to a selection policy (ϖ) and outputs a population estimate ($\hat{\mu}_\varkappa$) according to an estimation algorithm ($\varkappa$):

$$\min_{\varpi,\varkappa} \quad \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r^*(a_k^*) - r^*(\hat{a}_k)\right] + \mathbb{V}_{\mathcal{D},\varpi}(\mu^*(t) - \hat{\mu}_\varkappa(t)) \tag{1}$$

$$\text{s.t.} \quad |\hat{\mu}_\varkappa(t) - \mu^*(t)| \to \mathcal{N}(0,\sigma) \quad \text{as} \quad K \to N, \tag{2}$$

where $\mathcal{D}$ is the underlying distribution from which all taxpayers are pulled. In our IRS setting each arm (a_t) represents a taxpayer which the policy 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 (see Appendix F for reasoning and method selection criteria).

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 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[5] from the population distribution.

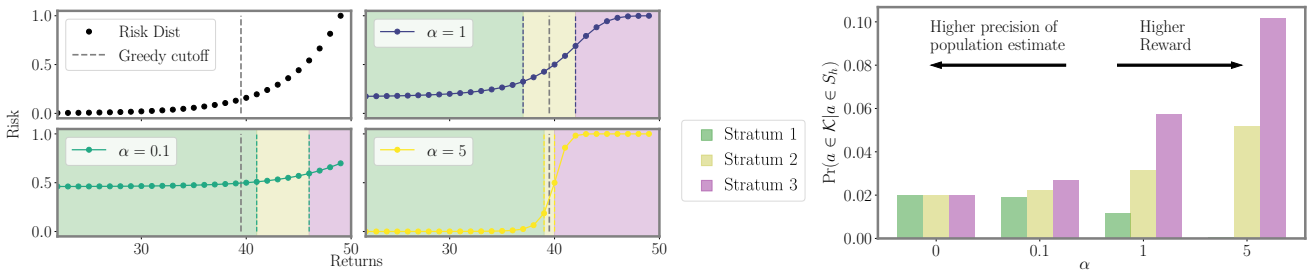
Optimism. We refer readers to Lattimore and Szepesvári (2020) for a general introduction to Upper Confidence Bound (UCB) and optimism-based methods. We import an optimism-based approach into this setting as follows. Consider a random forest with B trees $T_1, T_2, \ldots, T_B$. We form an optimistic 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, Hoos, and Leyton-Brown (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.

ABS Sampling. Adaptive Bin Sampling brings together sampling and bandit literatures to guarantee statistically unbiased population estimates, while enabling an explicit trade-off between reward and the variance of the estimate. In

[3] The IRS uses statistical adjustments to compensate naturally occurring variation in the depth of audit, and taxpayer misreporting that is difficult to find via auditing, and other NRP sampling limitations (Guyton et al. 2020; Internal Revenue Service 2019; Erard and Feinstein 2011). For the goals of this work we ignore these.

[4] Note the reason we make this assumption is because the NRP data does not track a cohort of taxpayers, but rather randomly samples. We are not guaranteed to ever see a taxpayer twice.

[5] 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 use NRP weights for population estimation. See Appendix K.



(a) Risk distribution and parameterizations

(b) Probability of sampling an individual

Figure 1: 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).

essence, ABS performs adjustable *risk-proportional* random sampling over optimized population strata. By maintaining probabilistic sampling, ABS can employ HT estimation to achieve an unbiased measurement of the population.

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 mixing function $\hat{\rho}_a$ intended to smoothly transition focus between the reward and variance objectives, but whose only requirement is that it be monotone (rank-preserving). For our empirical exploration we examine two such mixing functions, a logistic function, $\hat{\rho}_a = \frac{1}{1+\exp(-\alpha(\hat{r}_a - \kappa))}$ and 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 1 provides a visualization of this.

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:

$$\min_{S_1,\ldots,S_H : |S_h| \geq K-\zeta} \quad \sum_h \sum_{\hat{\rho} \in S_h} \|\hat{\rho} - \lambda_h\|^2, \tag{3}$$

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:

$$\pi_h = \frac{\lambda_h}{\sum_{h'} \lambda_{h'}}. \tag{4}$$

To make our selection, 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. The major benefit of ABS is that by

Algorithm 1: ABS (Logistic)

Input: α, H, ζ, 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 ζ predictions.
  $\forall_a \; \hat{\rho}_a \leftarrow (1 + \exp(-\alpha(\hat{r}_a - \kappa))^{-1}$
  Construct strata $S_1, \ldots, S_H$ by solving (3).
  Form distribution $\{\pi_h\}$ over strata via (4).
  repeat
    $h \sim (\pi_1, \ldots, \pi_H)$
    Sample arm uniformly at random from S_h.
  until $K - \zeta$ samples drawn
  Compute $\hat{\mu}_{\text{HT}}$ once true rewards are collected.
  Retrain model $\hat{f}$ on $(\cup_i^t X_i, \cup_i^t r_i)$.
end for

sampling according to the distribution π, we can employ HT estimation to eliminate bias. Indeed, if $\mathcal{K}$ is the set of arms sampled during an epoch, $\hat{\mu}_{HT}(t) = \frac{1}{\sum_a w_a} \sum_{a \in \mathcal{K}} \frac{w_a r_a}{p_a}$ is an unbiased estimate of the true population, where p_a is the probability that arm a was selected (i.e., $\Pr(a \in \mathcal{K})$) and w_a is the NRP weight. Like with other HT-based methods (Potter 1990; Alexander, Dahl, and Weidman 1997), to reduce variance we also add an option for a minimum probability of sampling a bin, which we call the trim %. See Appendix N for more details, proof of unbiased estimation, and estimator variance. See Appendix U for regret bounds.

Reward Structure Models. As the data is highly non-linear and high-dimensional, we use Random Forest Regression (RFR) for our reward model. We exclude linear models from our suite of algorithms after verifying that they consistently underperform RFR (Appendix M). We do not include neural networks in this analysis as the data regime is too small. Future approaches might build on this work using pretraining methods suited for a few-shot context (Bommasani et al. 2021). We do compare to an LDA baseline (Appendix T.3).

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.

4 Evaluation Protocol

We evaluate according to three metrics: cumulative reward, percent difference of the population estimate, and the no-change rate. More details in Appendix L.

Cumulative reward (R) is simply the total reward of all arms selected by the agent across the entire time series $\mathbb{E}\Big[\Big(\sum_t^T \sum_k^K r_{a_k}\Big)\Big]$. 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 where we round down such that any reward <\$200 is considered no change $\mu_{NR} = \mathbb{E}\Big[\Big((1/T)\sum_t^T (1/K)\sum_k^K \mathbf{1}\{r_{a_k} < 200)\}\Big]\Big)$. 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.

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 in Appendix S). 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 R). 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 (diagram in Appendix O).

5 Results

We highlight several key findings with additional results and sensitivity analyses in Appendix T.

Unbiased population estimates are possible with little impact to reward. ABS sampling can achieve similar returns to the best performing methods in terms of audit selection, while yielding an unbiased population estimate (see Table 1).

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased	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	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 a cumulative reward. μ_{PE} is the average percent difference of the population estimate across seeds. σ_{PE} is the standard deviation of the percent difference across seeds. μ_{NR} is the no change rate. Extended table with hyperparameters for all displayed methods is in Appendix T, selection method in Appendix P. 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.

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 can lead to biased models (Nie et al. 2018; Neel and Roth 2018). 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 2). 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.

ABS allows fine-grained control over variance-reward trade-off. We sample a grid of hyperparameters for ABS (see Appendix P). Figure 2 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 2 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. While ϵ-greedy with the random sample only does surprisingly well, optimal ABS configurations have a better Pareto front. We can fit a function to this Pareto front and estimate the marginal value

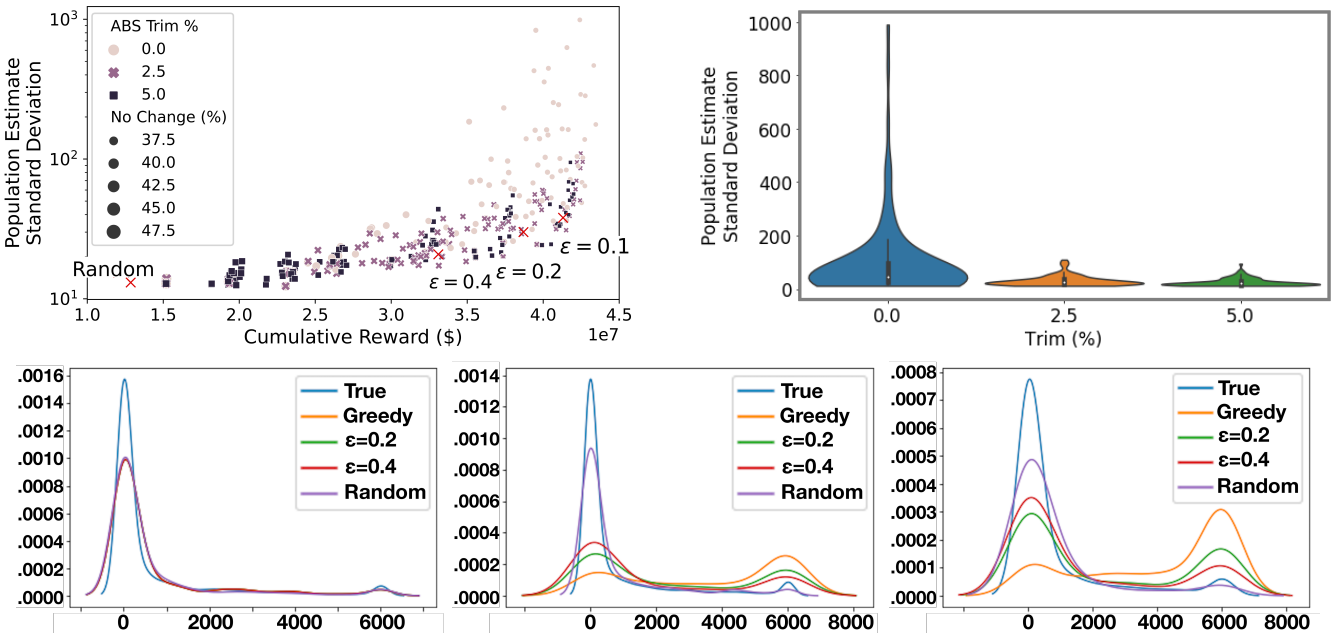


Figure 2: (Top Left) Population estimation empirical standard deviation versus reward for a grid of ABS hyperparameters. Red x's and associated labels indicate ϵ-only and fully random sample. (Top Right) Population estimation variance as a function of the trim. (Bottom) A kernel density plot of the distribution of sampled arms from 2006 (top left) to 2014 (bottom). X-axis is true reward. Y-axis is sampling distribution density.

of the reward-variance trade-off (see Appendix T.2).

Greedy is not all you need. 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, Agarwal, and Langford 2018; Kannan et al. 2018; Bastani, Bayati, and Khosravi 2021). Here, there are several intrinsic sources of exploration that may cause this result: intrinsic model error, covariate drift (see Appendix Table 5), 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 2 (bottom) 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 prior work showing greedy's surprising performance (Bietti, Agarwal, and Langford 2018; Kannan et al. 2018; Bastani, Bayati, and Khosravi 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, Ramdas, and Rinaldo 2021; Farquhar, Gal, and Rainforth 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. We also show that RFR regressors significantly outperform LDA and that incorporating non-random data helps (Appendix T.3). This is a stylized proxy of the status quo system that uses a small ϵ-only sample (NRP) for population estimates and an LDA-like algorithm (DIF) for selection.

A more focused approach audits higher cumulative total positive income. A key motivator for our work is that inefficiently-allocated randomness in audit selection will not only be suboptimal for the government, but could impose unnecessary burdens on taxpayers (Lawsky 2008; Davis-Nozemack 2012). 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 3, 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

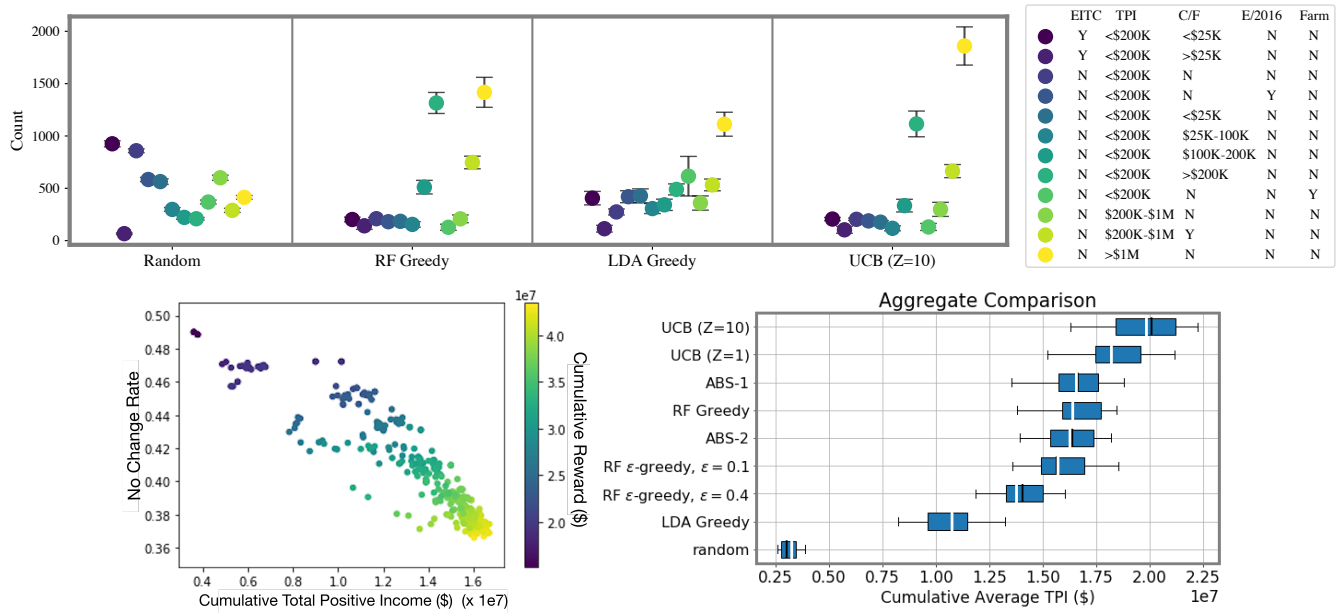


Figure 3: (Top) The distribution of audit classes for several approaches. Markers follow the same order as the legend from left to right. EITC Stands for "Earned Income Tax Credit" and "TPI" for Total Positive Income. (Bottom Left) Distribution of TPI for ABS hyperparameter settings. (Bottom Right) Distribution of TPI for various methods.

Greedy, effect size 95% CI matched by seed). We also 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 3 (top) 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).

Errors are heteroskedastic, causing difficulties in using model-based optimism methods. Surprisingly, our optimism-based approach audits tax returns with higher TPI more often ($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) NRP 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.

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. We empirically investigate the use of structured bandits in the IRS setting and show that ABS conforms to IRS specifications (unbiased estimation) and enables parties to explicitly trade off population estimation variance and reward maximization. This framework could help address longstanding concerns in the real-world setting of IRS detection of tax evasion. It could shift audits toward tax returns with larger understatements (correlating with more total positive income) 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.

Acknowledgements

We would like to thank Emily Black, Jason DeBacker, Hadi Elzayn, Tom Hertz, Andrew Johns, Dan Jurafsky, Mansheej Paul, Ahmad Qadri, Evelyn Smith, and Ben Swartz 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. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors and do not necessarily reflect the views or the official positions of the U.S. Department of the Treasury or the Internal Revenue Service. Any taxpayer data used in this research was kept in a secured Treasury or IRS data repository, and all results have been reviewed to ensure no confidential information is disclosed.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24.

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

Andreoni, J.; Erard, B.; and Feinstein, J. 1998. Tax compliance. *Journal of Economic Literature*, 36(2): 818–860.

Bastani, H.; Bayati, M.; and Khosravi, K. 2021. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3): 1329–1349.

Beer, S.; Kasper, M.; Kirchler, E.; and Erard, B. 2015. Audit Impact Study. Technical report, National Taxpayer Advocate.

Belnap, A.; Hoopes, J. L.; Maydew, E. L.; and Turk, A. 2020. Real effects of tax audits: Evidence from firms randomly selected for IRS examination. *SSRN*.

Bietti, A.; Agarwal, A.; and Langford, J. 2018. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

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

Congressional Budget Office. 2020. Trends in the Internal Revenue Service's Funding and Enforcement. https://www.cbo.gov/publication/56467#_idTextAnchor002. Accessed: 2022-09-01.

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

DeBacker, J.; Heim, B. T.; Tran, A.; and Yuskavage, A. 2018. The effects of IRS audits on EITC claimants. *National Tax Journal*, 71(3): 451–484.

Dimakopoulou, M.; Zhou, Z.; Athey, S.; and Imbens, G. 2017. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

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

Erraqabi, A.; Lazaric, A.; Valko, M.; Brunskill, E.; and Liu, Y.-E. 2017. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, 709–717. PMLR.

Esteban, J.; McRoberts, R. E.; Fernández-Landa, A.; Tomé, J. L.; and Næsset, E. 2019. Estimating forest volume and biomass and their changes using random forests and remotely sensed data. *Remote Sensing*, 11(16): 1944.

Farquhar, S.; Gal, Y.; and Rainforth, T. 2021. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*.

Government Accountability Office. 2002. New Compliance Research Effort Is on Track, but Important Work Remains. https://www.gao.gov/assets/gao-02-769.pdf. United States General Accounting Office: Report to the Committee on Finance, U.S. Senate. Online; Accessed Jan 10, 2022.

Government Accountability Office. 2003. IRS Is Implementing the National Research Program as Planned. https://www.gao.gov/assets/gao-03-614.pdf. United States General Accounting Office: Report to the Committee on Finance, U.S. Senate. Online; Accessed Jan 10, 2022.

Guyton, J.; Langetieg, P.; Reck, D.; Risch, M.; and Zucman, G. 2020. 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.

Guyton, J.; Langetieg, P.; Reck, D.; Risch, M.; and Zucman, G. 2021. Tax Evasion at the Top of the Income Distribution: Theory and Evidence. Technical report, National Bureau of Economic Research.

Guyton, J.; Leibel, K.; Manoli, D. S.; Patel, A.; Payne, M.; and Schafer, B. 2018. The effects of EITC correspondence audits on low-income earners. Technical report, National Bureau of Economic Research.

Henderson, P.; Chugg, B.; Anderson, B.; and Ho, D. E. 2021. Beyond Ads: Sequential Decision-Making Algorithms in Law and Public Policy. *arXiv preprint arXiv:2112.06833*.

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

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

Internal Revenue Service. 2016. Federal tax compliance research: Tax gap estimates for tax years 2008–2010. https://www.irs.gov/pub/irs-soi/p1415.pdf. Accessed: 2022-09-01.

Internal Revenue Service. 2019. Federal tax compliance research: Tax gap estimates for tax years 2011–2013. https://www.irs.gov/pub/irs-prior/p1415--2019.pdf. Accessed: 2022-09-01.

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

Internal Revenue Service. 2022. https://www.irs.gov/irm/part4/irm_04-001-002. Accessed: 2022-09-01.

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

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2018. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 158–163.

Kannan, S.; Morgenstern, J. H.; Roth, A.; Waggoner, B.; and Wu, Z. S. 2018. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in Neural Information Processing Systems*, 31.

Kiel, P. 2019. 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. Online; Accessed Jan 10, 2022.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

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

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

Liu, Y.-E.; Mandel, T.; Brunskill, E.; and Popovic, Z. 2014. Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits. In *EDM*, 161–168.

Lowe, V. L. 1976. 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. *Subcommittee on Oversight House Committee on Ways Hearing Transcripts*.

Marr, C.; and Murray, C. 2016. IRS funding cuts compromise taxpayer service and weaken enforcement. http://www.cbpp.org/sites/default/files/atoms/files/6-25-14tax.pdf. Accessed: 2022-09-01.

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

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

Nika, A.; Elahi, S.; and Tekin, C. 2021. Contextual Combinatorial Volatile Bandits via Gaussian Processes. *arXiv preprint arXiv:2110.02248*.

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

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

Potter, F. J. 1990. 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.

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

Shin, J.; Ramdas, A.; and Rinaldo, A. 2021. On the Bias, Risk, and Consistency of Sample Means in Multi-armed Bandits. *SIAM Journal on Mathematics of Data Science*, 3(4): 1278–1300.

Taxpayer Advocate Service. 2018. 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 Annual Report to Congress — Volume One. Online; Accessed Jan 10, 2022.

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

Treasury, U. 2021. The American Families Plan Tax Compliance Agenda. *Department of Treasury, Washington, DC*.

Wagstaff, K. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656*.

Xu, M.; Qin, T.; and Liu, T.-Y. 2013. Estimation bias in multi-armed bandit algorithms for search advertising. *Advances in Neural Information Processing Systems*, 26.