Entropy Regularization for Population Estimation

Ben Chugg[1], **Peter Henderson**[2], **Jacob Goldin**[3], **Daniel E. Ho**[2]

[1]Carnegie Mellon University
[2] Stanford University
[3] University of Chicago
benchugg@cmu.edu, {phend, dho}@stanford.edu, jsgoldin@uchicago.edu

Abstract

Entropy regularization is known to improve exploration in sequential decision-making problems. We show that this same mechanism can also lead to nearly unbiased and lower-variance estimates of the mean reward in the optimize-and-estimate structured bandit setting. Mean reward estimation (i.e., population estimation) tasks have recently been shown to be essential for public policy settings where legal constraints often require precise estimates of population metrics. We show that leveraging entropy and KL divergence can yield a better trade-off between reward and estimator variance than existing baselines, all while remaining nearly unbiased. These properties of entropy regularization illustrate an exciting potential for bridging the optimal exploration and estimation literatures.

Introduction

While most frameworks for online sequential decision-making focus on the objective of maximizing reward, in practice this is rarely the sole objective. Other considerations may involve budget constraints, ensuring fair treatment, or estimating various population characteristics. There has been growing recognition that these other constraints must be formally integrated into sequential decision-making frameworks, especially if such algorithms are to be used in sensitive application areas (Henderson et al. 2021). In this work, we focus on the problem of maximizing reward while simultaneously estimating the population total (equivalently, mean) in a structured bandit setting.

The most natural approach to this problem from a machine learning perspective is to use a model to predict the mean. However, this method is subject to the problem that adaptively collected data are subject to bias, which in turn biases the model estimates (Nie et al. 2018). Natural tools from survey sampling such as IPW estimators (also used commonly in off-policy evaluation (OPE)) enable mean estimation if each observation has been sampled with some known or estimable probability. However, in those settings the probabilities are given *a priori*, and can yield high-variance estimates if the sampling distribution is skewed. Here we seek to optimize over possible probability distributions in order to trade off between our expected reward and the variance of our estimator. Unfortunately, the variance formulation of such estimators is unwieldy and yields an intractable optimization problem if directly incorporated into the objective. Instead, we substitute the variance term for an information-theoretic policy regularization term such as entropy or KL divergence. Adding such terms yields analytically tractable optimization problems, while maintaining our ability to smoothly navigate reward-variance trade offs.

In reinforcement learning (RL) settings, adding policy regularization terms to the objective function has been shown to consistently improve the performance of sequential decision-making algorithms (Williams and Peng 1991; Ahmed et al. 2019; Xiao et al. 2019). Our ability to leverage these tools in a different setting enables us to take advantage of well-established theory and insights in RL, indicating an exciting convergence between our problem, survey sampling, and other areas of sequential decision-making.

In sum, our contributions are as follows: (i) We propose novel algorithms to handle the dual objective of reward maximization and mean reward estimation in the structured bandit setting; (ii) characterize the bias of the estimators in our setting, provide closed-form solutions for our proposed optimization problems, and relate entropy and KL sampling strategies to the variance of the estimators; and (iii) demonstrate the improvement of our algorithms over baselines on four datasets.

Related Work

The bandit literature is large. We focus on the most relevant related works here, and relegate a prolonged discussion of all peripheral work to Appendix A.

The closest work to ours is Henderson et al. (2022), who introduce the *optimize-and-estimate* structured bandit setting which we adopt in this paper. They introduce Adaptive Bin Sampling (ABS) which we describe below and use as our main baseline. The optimize-and-estimate setting is closest to that in Abbasi-Yadkori, Pál, and Szepesvári (2011) and Joseph et al. (2016), but extended to non-linear rewards with a required estimation objective. Crucially, in this setting, arms may change from step to step and the agent must instead rely on a per-arm context to determine which arms to pull.

A number of works have sought to yield reduced-bias (e.g., Nie et al. (2018)) or unbiased estimators (e.g., Zhang,

Janson, and Murphy (2020)) for inference with adaptively-collected data. This is because it is well known that adaptive sampling leads to bias when estimating sample means and other population characteristics (Nie et al. 2018; Shin, Ramdas, and Rinaldo 2019, 2020, 2021; Russo and Zou 2016). Attempts to remedy this bias have come in the form of differential privacy (Neel and Roth 2018), adaptive estimators (Dimakopoulou, Ren, and Zhou 2021; Chugg and Ho 2021), or re-normalization (Zhang, Janson, and Murphy 2020). Unfortunately, none of these efforts apply to our setting. They are mostly in the multi-armed and contextual bandit setting (wherein one samples the same arm multiple times) which differs from our own structured bandit setting. We emphasize that we are in the non-linear setting and thus seek to have non-parametric unbiasedness guarantees. Moreover, much previous work seeks to remedy bias *ex post*, as opposed to incorporating bias (and variance) into the objective *ex ante.*

Finally, a large body of work has examined entropy regularization for optimal exploration in reinforcement learning (Williams and Peng 1991; Zimmert and Seldin 2021; Ahmed et al. 2019; Brekelmans et al. 2022) and bandits (Fontaine, Berthet, and Perchet 2019; Xiao et al. 2019). While the problem settings are different, we examine the fundamental question of whether entropy regularization may bring the same benefits for unbiased estimation as it brought for exploration guarantees elsewhere.

Problem Setting

Optimize-and-Estimate Structured Bandits

At time t we receive a set of N_t observations with features $X_t = \{x_{it}\} \subset \mathcal{X}$, where observation x has reward $r(x) \sim \mathcal{D}(x)$ drawn from a reward distribution conditional on the context. Each period, we can sample at most K_t observations and receive their rewards. If $K_t = 1$ we are in the *sequential* setting; if $K_t > 1$ we are in the *batched* setting. Let $S_t \subset X_t$, $|S_t| \leq K_t$ denote the set (possibly singleton) of observations selected. Like in traditional bandit problems, one of our objectives is to maximize reward (equivalently, to minimize regret). The cumulative reward at time T is

$$\textsc{Rew}(T) = \sum_{t \leq T} \sum_{x \in S_t} r(x), \tag{1}$$

where $S_t \subset X_t$ is the set of selected observations. For an event E, the function $\mathbf{1}_E$ is 1 if E occurs and 0 otherwise. Our second goal is to achieve a reliable unbiased population estimate $\textsc{Pop}(T)$ of the total of all arms in each period T:

$$\textsc{Pop}(T) = \sum_{x \in X_T} \mathbb{E}_{\mathcal{D}}[r(x)]. \tag{2}$$

Note that previous work in this setting estimated the population *mean*, instead of the total. Obviously, the difference is unimportant, and we find it cleaner to work with the total. Throughout, we will drop the parenthetical in favor of a subscript when space demands and write, e.g., r_x in lieu of $r(x)$. If the timestep T is clear from context we drop it from the notation.

We observe that observations are equivalent to arms and can be volatile from timestep to timestep, characterized only by a context and an underlying shared reward structure. This is similar to the structured bandit (Abbasi-Yadkori, Pál, and Szepesvári 2011; Joseph et al. 2016). If arms were fixed and the context singular, this would reduce to the contextual bandit. If the context were removed and arms remained fixed, this would reduce to the multi-armed bandit. As such, methods from the multi-armed bandit and contextual bandit (like Thompson sampling) do not necessarily apply to the structured bandit, and especially not the optimize-and-estimate structured bandit. We discuss this more in Appendix A.

Strategies for Population Estimation

Given a model $\hat{\varphi} : \mathcal{X} \to \mathbb{R}$ which estimates the reward $r(x)$, a natural approach to estimate $\textsc{Pop}$ is to combine the empirical population total of the selected points with the estimated total from $\hat{\varphi}$:

$$\widehat{\textsc{Pop}}_{\textsc{model}}(T) = \sum_{x \in S_T} r(x) + \sum_{x \in X_T \setminus S_T} \hat{\varphi}(x). \tag{3}$$

A separate approach is available if selection is performed according to a probability distribution $\{\pi(x) = \mathbb{P}(x \in S)\}$ over the observations. In this case we can turn to "importance sampling methods", which employ the basic idea of weighting the observations by a function of their probability. We'll focus on two popular importance sampling methods: inverse propensity weighting (IPW) and doubly-robust (DR) estimation.

Suppose that x was sampled with probability $\pi(x)$ and that we have estimates $\hat{\pi}(x)$ of $\pi(x)$ for all x (we allow for the possibility that $\hat{\pi}(x) = \pi(x)$, but we'll see the utility of allowing approximations later on). Assuming that $\hat{\pi}(x) > 0$ for all x, then the IPW estimator (Horvitz and Thompson 1952; Narain 1951) is

$$\widehat{\textsc{Pop}}_{\textsc{ipw}}(T) = \sum_{x \in X_T} \frac{r(x)}{\hat{\pi}(x)} \mathbf{1}_{x \in S_T}. \tag{4}$$

The final estimator we'll consider is the so-called doubly-robust (DR) estimator (Cassel, Särndal, and Wretman 1976; Jiang and Li 2016). This combines the model-based approach with the IPW estimator:

$$\widehat{\textsc{Pop}}_{\textsc{dr}}(T) = \sum_{x \in X_T} \left(\hat{\varphi}(x) + \frac{r(x) - \hat{\varphi}(x)}{\hat{\pi}(x)} \mathbf{1}_{x \in S_T} \right). \tag{5}$$

The IPW and DR estimators are common in off-policy evaluation (Dudík, Langford, and Li 2011; Dudík et al. 2014). The bias of all three estimators is given in Lemma 1. While the bias was previously given by Dudík et al. (2014), that was in the RL setting, which differs from ours. That said, the results do not change much.

Lemma 1 (Bias of estimators)**.** *Let* $\Delta_x = \mathbb{E}_{\mathcal{D}}[r(x)] - \hat{\varphi}(x)$, *and* $\lambda_x = \pi(x)/\hat{\pi}(x)$. *Then, at any time* T*,*

1. $|\mathbb{E}_{\mathcal{D},S}[\widehat{\textsc{Pop}}_{\textsc{model}}] - \textsc{Pop}| = \sum_x \Delta_x(\pi_x - 1)$.
2. $|\mathbb{E}_{\mathcal{D},S}[\widehat{\textsc{Pop}}_{\textsc{ipw}}] - \textsc{Pop}| = \sum_x \mathbb{E}_{\mathcal{D}}[r_x](\lambda_x - 1)$, *and*
3. $|\mathbb{E}_{\mathcal{D},S}[\widehat{\textsc{Pop}}_{\textsc{dr}}] - \textsc{Pop}| = \sum_x \Delta_x(\lambda_x - 1)$,

where all sums are over $x \in X_T$.

The proof of Lemma 1, along with all other propositions in the paper, can be found in the Appendix. Note a corollary of Lemma 1: IPW and DR are unbiased if $\hat{\pi}(x) = \pi(x)$, i.e., we sample with precise inclusion probabilities. This is computationally challenging for sufficiently large budgets, but as we discuss in Section , we employ an approximation mechanism – Pareto Sampling – that, in practice, enables unbiasedness (Figure 4).

Variance

The probabilities $\pi(x) = \mathbb{P}(x \in S)$ are called (first order) *inclusion probabilities*. The terms $\pi(x, z) = \pi_{x,z} = \mathbb{P}(x, z \in S)$ are called *second order*, or *joint* inclusion probabilities, and naturally arise in the variance of population estimators due to the covariance terms. More detail on inclusion probabilities can be found in Appendix C.

To define the variance, for an arbitrary function $\theta : \mathcal{X} \to \mathbb{R}$, let

$$A_T(\theta) = \frac{1}{2} \sum_{x,z \in X_T} \left(\frac{\theta_x}{\pi_x} - \frac{\theta_z}{\pi_z} \right)^2 (\pi_x \pi_z - \pi_{x,z}). \quad (6)$$

The variance of $\widehat{\text{POP}}_{\text{IPW}}$ and $\widehat{\text{POP}}_{\text{DR}}$ (with respect to sampling) for $\hat{\pi}(x) = \pi(x)$ can then be written as (derivation in Appendix E).

$$\mathbb{V}_\pi(\widehat{\text{POP}}_{\text{IPW}}(T)) = A_T(r), \quad (7)$$

$$\mathbb{V}_\pi(\widehat{\text{POP}}_{\text{DR}}(T)) = A_T(r - \hat{\varphi}). \quad (8)$$

That is, fixing the inclusion probabilities, the variance of the IPW estimator depends on the ratio between the true reward and the probability, while that of the DR estimator depends on the ratio between the model error $(r(x) - \hat{\varphi}(x))$ and the sampling probability. Thus, the variance of IPW estimator is zero if we sample proportionally to the true reward, while the variance of the DR estimator is zero if we sample according to the model residuals.

Methods

Optimization Objective

A natural approach to minimizing an estimator's variance while maximizing reward is to form a linear combination of the two objectives. For any set $X \subset \mathcal{X}$, consider the optimization problem

$$\sup_{\boldsymbol{\pi} \in \Pi_K(X)} \Phi_\beta(\boldsymbol{\pi}) = \mathbb{E}_\pi[\text{REW}(\pi, \hat{\varphi})] - \beta \mathbb{V}_\pi(\widehat{\text{POP}}), \quad (9)$$

which selects inclusion probabilities in order to maximize reward and minimize variance. The trade-off between the two objectives is controlled by a predetermined scalar $\beta \in \mathbb{R}_{\geq 0}$. The set of legal inclusion probabilities over which the optimization takes place is

$$\Pi_K(X) = \left\{ \pi \in (0, 1]^X : \sum_{x \in X} \pi(x) = K \right\},$$

which requires that probabilities be strictly greater than zero to ensure that $\widehat{\text{POP}}_{\text{IPW}}$ and $\widehat{\text{POP}}_{\text{DR}}$ are well-defined. Here,

$$\mathbb{E}_\pi[\text{REW}(\pi, \hat{\varphi})] = \sum_x \pi(x) \hat{\varphi}(x),$$

is the expected reward according to the model. We note that the supremum is required in (9) since $\Pi_K(X)$ is not closed.

Unfortunately, if $\widehat{\text{POP}}$ is the IPW or the DR estimator, Equation (9) leads to a rather intractable optimization problem. We provide a more thorough discussion of the difficulties in Appendix G, but suffice it to say that the joint inclusion probabilities $\pi(x, z)$ in Equation (6) do not readily lend themselves to optimization. Indeed, for most batched sampling strategies, they do not have closed form solutions.

Model-Proportional Sampling

Despite being difficult to optimize directly, for the IPW estimator $\widehat{\text{POP}}_{\text{IPW}}$, the optimization problem (9) has an interesting property. Notice that $A_T(\theta) = 0$ in Eq. (6) if $\theta(x)/\pi(x) = C$ for some constant C for all x. Thus, if using $\widehat{\text{POP}}_{\text{IPW}}$, a natural strategy is to sample according to

$$\pi(x) = \frac{K \hat{\varphi}(x)}{\sum_z \hat{\varphi}(z)}, \quad (10)$$

if $K\hat{\varphi}(x)/\sum_z \hat{\varphi}(z) \leq 1$. We call this approach MODELPROPORTIONALSAMPLING (MPS). There are several drawbacks to this approach. The first is that it relies on the model $\hat{\varphi}$. If the model error is large, then so too will be the variance. This method also provides no way to trade-off between variance and reward; its only focus is minimizing variance. Indeed, we'll see in the results section that while the variance is indeed low (if model error is reasonable), reward is also much lower than other methods. The next section uses KL-divergence to generalize this approach, enabling us to smoothly transition between probabilities in (10) and those which place more weight on expected reward.

Entropy and KL Sampling

Due to the difficulties imposed by the variance term in objective function (9), we propose two new optimization problems which are analytically tractable while still enabling a trade-off between variance and expected reward. The first is entropy (Shannon 1948). Fix a timestep t, and consider the sequential version of the problem (i.e., $K = 1$). The *entropy* of the sample $S = S_t$ (the random variable describing which observations are sampled), is $H(S) = -\sum_{x \in X} \pi(x) \log(\pi(x))$. $H(S)$ is, roughly, a measure of how spread out the distribution π is. As $H(S)$ increases, the π resembles a uniform distribution over X; as $H(S)$ decreases π is skewed towards some subset of X. The second objective we'll consider is the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951). The KL divergence between two discrete distributions P and Q defined on the sample space Ω is

$$D_{\text{KL}}(P||Q) = \sum_{\omega \in \Omega} P(\omega) \log\left(\frac{P(\omega)}{Q(\omega)} \right).$$

Algorithm 1: Entropy-regularized Pareto Sampling

$Z \leftarrow X$
$\hat{\pi}(x) \leftarrow K \frac{e^{\hat{\varphi}(x)/\beta}}{\sum_{z \in Z} e^{\hat{\varphi}(x)/\beta}}, \forall x \in Z$ (Eq. 14 if KL)
$F \leftarrow \emptyset$
while $\exists x : \hat{\pi}(x) > 1$ **do**
  $F \leftarrow F \cup \{x : \hat{\pi}(x) > 1\}$
  $Z \leftarrow X \setminus F$
  $\hat{\pi}(x) \leftarrow \frac{K e^{\hat{\varphi}(x)/\beta}}{\sum_{z \in Z} e^{\hat{\varphi}(x)/\beta}}, \forall x \in Z$ (Eq. 14 if KL)
end while
Sample $U(x) \sim \text{unif}(0,1), \forall x \in X$
$V(x) \leftarrow U(x)(1-\hat{\pi}(x))/((1-U(x))\hat{\pi}(x))$
Relabel s.t. $V(x_1) \leq V(x_2) \leq \cdots \leq V(x_N)$
return $x_1, \ldots, x_K$

While the KL divergence has many interpretations depending on the application at hand, for our purposes we can think of it as measuring the divergence between P and Q. Minimizing the divergence as a function of P pushes P towards Q.

Set $q(x) = \hat{\varphi}(x)/\sum_z \hat{\varphi}(z)$, i.e., the MPS sampling solution. From here, define two optimization problems over $\Pi(X)$:

$$\sup_{\pi \in \Pi_K(X)} \Phi_\beta^{\text{ENT}}(\pi) = \mathbb{E}[\text{REW}(\pi, \hat{\varphi})] + \beta H(S), \quad (11)$$

$$\sup_{\pi \in \Pi_K(X)} \Phi_\beta^{\text{KL}}(\pi) = \mathbb{E}[\text{REW}(\pi, \hat{\varphi})] - \beta D_{\text{KL}}(\pi||q), \quad (12)$$

Note that the only unknowns in the above two equations are the probabilities π. The following proposition shows they have closed-form solutions.

Proposition 1. *For a set of observations $X \subset \mathcal{X}$ and $\Pi_K(X)$ as above, the solutions to optimization problems* (11), (12) *with $K = 1$ are $\{\pi^{\text{ENT}}(x)\}_{x \in X}$ and $\{\pi^{\text{KL}}(x)\}_{x \in X}$ respectively, where*

$$\pi^{\text{ENT}}(x) = \text{softmax}(\hat{\varphi}(x)/\beta), \quad (13)$$

$$\pi^{\text{KL}}(x) = \frac{\hat{\varphi}(x)\exp(\hat{\varphi}(x)/\beta)}{\sum_z \hat{\varphi}(z)\exp(\hat{\varphi}(z)/\beta)}. \quad (14)$$

A brief recap is perhaps in order. We began by attempting to write down an optimization problem to trade off between reward and variance. Optimizing over the variance term explicitly proved to be intractable however, so we substituted the variance term for an entropy-based term which indirectly controls the variance. This yields new optimization problems which are tractable, have closed-form solutions, and are more amenable to theoretical analysis.

Approximating π for $K > 1$

Equations (11) and (12) are for the sequential setting. Unfortunately, solving the naive extension to the batched setting is intractable – both analytically and computationally – because the entropy of S involves an exponentially large sum over all subsets of size K, and the calculation of the higher level inclusion probabilities. Therefore, in order to scale up the solutions in (13) to the batched setting, we begin by multiplying each $\pi(x)$ by K, i.e., $\pi^*(x) = K\text{softmax}(\hat{\varphi}(x)/\beta)$. Depending on the distribution of $\hat{\varphi}(x)$, however, this quantity might be greater than 1. In this case, we set $\pi^*(x) = 1$ (i.e., it will be sampled), and recalculate the probabilities on the subpopulation for which $K\text{softmax}(\hat{\varphi}(x)/\beta) < 1$. This is repeated until no inclusion probabilities exceed 1. This is reflected in Algorithm 1.

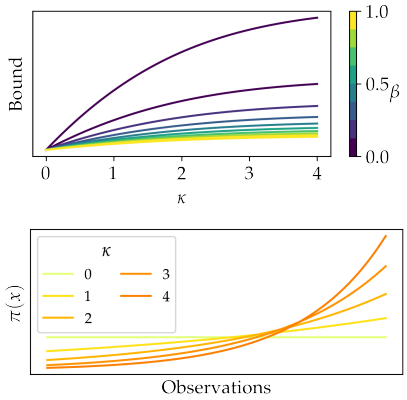

Figure 1: Illustration of Proposition 2. Top: Change in variance bound as a function of β and the shape of the inclusion probability distribution as characterized by κ, which captures the non-uniformity of $\pi(x)$. The effect of κ is shown by the bottom figure.

Once the inclusion probabilities are computed, the agent must actually sample from that distribution. This is trivial in the sequential setting, but more complicated in the batched setting. Designing a sampling scheme which respects first order inclusion probabilities precisely is difficult. Sampford sampling (Sampford 1967), for instance, can guarantee prespecified first order inclusion probabilities, but is infeasible as sample sizes become large as it is a rejective procedure. Instead, we employ Pareto Sampling (Rosén 1997). Here, given N target inclusion probabilities $\hat{\pi}(x)$, we generate N random values

$$V(x) = \frac{U(x)(1-\hat{\pi}(x))}{(1-U(x))\hat{\pi}(x)}, \quad U(x) \sim \text{unif}(0,1).$$

The K samples with the smallest values are selected. This method is fast and always yields a sample of size K. The drawback is that the method is only approximate: The true inclusion probabilities π are not precisely equal to $\hat{\pi}$. However, Rosén (2000) showed that the approximation error goes to zero as K increases: $\max_x |\pi(x)/\hat{\pi}(x) - 1| = O(\log K/\sqrt{K})$.

Moreover, in practice the method works extremely well. See Figure 5 in Appendix H for an illustration of its accuracy even at the relatively low budget of $K = 20$. Thus, as Rosén (2000) discusses, in practice any bias introduced by this approximation is empirically low. And this can be further reduced by calculating exact inclusion probabilities through numerical means at the cost of time.

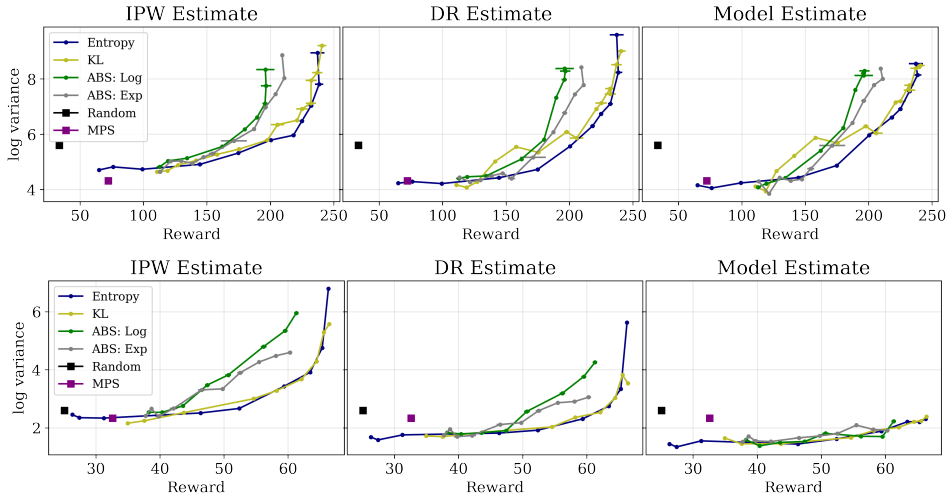

Figure 2: Reward-variance curves for all three estimators on the ACS dataset (top) and the AllState dataset (bottom). We used a budget of 1000 for both. Experiments were repeated at least 50 times to obtain errorbars (95% CI) on reward. ABS: Log and ABS: Exp correspond to ABS with logistic and exponential smoothing, respectively. The scale of the y-axis is held constant for each dataset to keep variance in perspective across estimators. For reference, an omniscient oracle achieves a reward of 330 on the ACS data and 98 on the AlState data.

Variance Bounds

Throughout this section, we fix a time T and condition on the previous observations and model choices, in addition to the population draw from $\mathcal{D}$. The only randomness stems from the sampling itself. Moreover, we focus on the set of observations which will not be sampled with certainty. This is captured by the following assumption.

Assumption 1. For all $x \in \mathcal{X}$, $K\pi^{\text{ENT}}(x) \leq 1$ and $K\pi^{\text{KL}}(x) \leq 1$ where π^{ENT} and π^{KL} are as in Equation (13).

We further assume that $\hat{\pi}(x) = \pi(x)$, i.e., we sample precisely according to π^{KL} and π^{ENT}. We provide upper bounds on the variance which do not involve the joint inclusion probabilities. This is useful, because they can be calculated *a priori* using only the model predictions. For a given period T and set of observations X_T, let $\hat{\varphi}_{\min} = \min_{x \in X_T} \hat{\varphi}(x)$.

Proposition 2. *Let $g_x = \exp(\hat{\varphi}(x)/\beta - \hat{\varphi}_{\min}/\beta)$ the β-weighted gap between the model prediction for x and the minimum prediction in exponential space. Define*

$$C_1 = \frac{1}{K} \sum_{x,z} g_x g_z - \sum_x g_x^2.$$

Then Entropy Sampling obeys $\mathbb{V}(\widehat{\text{POP}}_{\text{IPW}}(T)) \leq C_1$ and $\mathbb{V}(\widehat{\text{POP}}_{\text{DR}}(T)) \leq 2C_1$.

Figure 1 demonstrates the bound for various values of β and shapes of the inclusion probability distribution $\pi(x)$. We find that the value of β has a much greater effect on the bound than the shape of the distribution. The following proposition gives the equivalent bound for KL sampling.

Proposition 3. *Let g_x be as in Proposition 2, and define*

$$C_2 = \frac{1}{K} \sum_{x,z} \frac{\hat{\varphi}_x \hat{\varphi}_z}{\hat{\varphi}_{\min}^2} g_x g_z - \sum_x \frac{\hat{\varphi}_x^2}{\hat{\varphi}_{\min}^2} g_x^2.$$

Then, for KL Sampling, $\mathbb{V}(\widehat{\text{POP}}_{\text{IPW}}(T)) \leq C_2$ and $\mathbb{V}(\widehat{\text{POP}}_{\text{DR}}(T)) \leq 2C_2$.

Experiments

Experimental results, datasets, and code can be found at https://github.com/bchugg/ent-reg-pop-est.

Datasets

We run experiments on four publicly available datasets: The Current Population Survey (CPS), the American Community Survey (ACS), a voter turnout dataset, and data on AllState severity claims. These four were chosen because they each correspond to a real-world optimize-and-estimate setting. For instance, the CPS dataset is closely related to the tax-gap estimation done by the IRS each year (Henderson et al. 2022). The AllState population estimation task corresponds to an insurance company which must estimate the average cost of claims across the population. More detail on each dataset and further justification for their selection can be found in Appendix B.

Baselines

We compare entropy sampling to both MPS (described above), and also to simple random sampling (SRS). SRS calculates a population estimate based on the sampled rewards, i.e., it does not use a model. We also use Adaptive Bin Sampling (ABS), introduced by Henderson et al. (2022). A full overview of ABS is given in Appendix F. While the authors of ABS use only the IPW estimator in their experiments, we test ABS with all three estimators: IPW, DR, and Model-based.

Experimental Protocol

For each dataset and method, observations for the first period are selected uniformly at random to provide a initial training set for the model. Because we're interested primarily in the performance of the sampling methods themselves, we hold the model constant across sampling algorithms and datasets. Due to the relatively small budget sizes and some evidence that tree-based methods outperform neural networks on tabular data (Grinsztajn, Oyallon, and Varoquaux 2022), we use random forest regressors. We perform a randomized grid search on a small holdout set to determine a suitable set of hyperparameters for each dataset (see Appendix I for more details). Throughout our experiments, we keep the budget between approximately 5-10% of the dataset size in each period, i.e., $K_t \in [0.05, 0.1]X_t$ (depending on the dataset).

Results

The full suite of experimental results as well as further discussion can be found in Appendix J. Here, we distill the results into four main takeaways. We find that the results are consistent across datasets and, as such, we do not show figures for each one. To avoid selection bias, all figures in this section are plotted using the results from the final period. One broad takeaway worth mentioning concerns the choice of β. While the optimal value will depend on the practitioner's goals and the data itself, we find that values in $[0.05, 0.1]$ are prudent choices.

KL and entropy sampling improve the reward-variance tradeoff. As exemplified by Figure 2, both KL and Entropy sampling (mostly) improve over ABS in terms of the reward-variance trade off, especially for IPW and DR estimators. In fact, both are pareto improvements over ABS except occasionally in the low-reward region. For model estimation, KL and Entropy are never worse than ABS, and sometimes better. These trends hold across all datasets. Interestingly, Entropy and KL Sampling perform comparably, except perhaps for the fact that Entropy Sampling tends to ride the variance-reward curve more smoothly. We conjecture that this is due to form of π^{KL} (Equation 12), which has added multiplicative terms dependent on the model outcomes as compared to the softmax.

DR is more biased than IPW but less biased than model estimation. As testified by Lemma 1, all three estimators exhibit some form of bias if the model is mis-specified or the inclusion probabilities are only approximate. Figure 3 demonstrates the bias of entropy sampling on the ACS data, which is representative of the result across all datasets. Reliably, model estimation is the most biased, IPW the least, and DR between them. This is because DR is affected by model error (given inexact inclusion probabilities), while IPW is not.

Model estimation has unreliable variance. On the All-State data, the variance of the model estimate hardly changes as a function of reward (Figure 2). For ACS data meanwhile, the model estimates exhibit a similar reward-variance trade off to IPW and DR. This can be explained by model fit, which is significantly worse on ACS than on CPS. Given a sufficiently well-fit model, the variance of the model estimate should be low, as the inclusion probabilities play no part in the estimator. The trade off is that model-estimates are more biased than either the IPW or DR estimator.

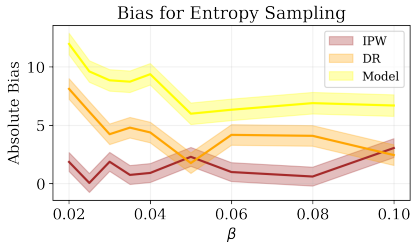

Figure 3: Bias (with 95% CI bands) plotted across values of β for Entropy Sampling on the CPS dataset.

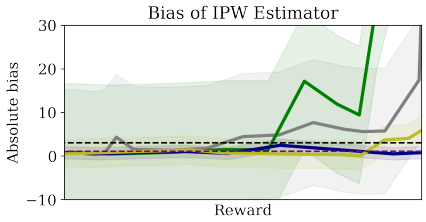

Figure 4: Absolute bias (with 95% CI bands) of IPW estimator on CPS data as a function of reward. The x-axis is clipped to the largest reward range covered by all algorithms.

Entropy and KL Sampling maintain low bias with IPW. Figure 4 illustrates the absolute bias of the IPW estimator across sampling strategies as a function of reward. Entropy and KL sampling both maintain low bias, even into the high-reward region where ABS begins to falter. This is a function of the reward-variance trade off of Figure 2: For the same reward, ABS has higher variance, making bias more apparent unless averaged over a sufficient number of model runs. We note that the variance of ABS is much higher than Entropy and KL at the same reward level, consistent with Figure 2. Figure 4 also testifies to the efficiency of the Pareto approximation. Indeed, the bias of entropy and KL rivals that of random and MPS.

MPS is a pareto improvement over random sampling. MPS tends to have low variance but also low reward (though higher reward than random sampling, Figure 2). Surprisingly, it also has low bias while maintaining low variance, demonstrating the possibility of model-based approaches in pure population estimation tasks.

Conclusion

We introduced two algorithms in the optimize-and-estimate structured bandit setting. Unlike previous work, our algorithms provide explicit sampling probabilities, thus making the approaches more amenable to theoretical analysis. The two algorithms improve upon the reward-variance trade off of current baselines, in addition to maintaining minimal bias of the population estimate.

Ackowledgements

We thank Dilip Arumugam, Brandon Anderson, Jia Wan for helpful feedback and discussions, and Vaden Masrani for shortening an otherwise overlong title. PH is supported by the Open Philanthropy AI Fellowship. This work was conducted while BC was at Stanford University.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Ahmed, Z.; Le Roux, N.; Norouzi, M.; and Schuurmans, D. 2019. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, 151–160. PMLR.

Brekelmans, R.; Genewein, T.; Grau-Moya, J.; Delétang, G.; Kunesch, M.; Legg, S.; and Ortega, P. 2022. Your Policy Regularizer is Secretly an Adversary. *arXiv preprint arXiv:2203.12592*.

Cassel, C. M.; Särndal, C. E.; and Wretman, J. H. 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615–620.

Chugg, B.; and Ho, D. E. 2021. Reconciling Risk Allocation and Prevalence Estimation in Public Health Using Batched Bandits. *arXiv preprint arXiv:2110.13306*.

Dimakopoulou, M.; Ren, Z.; and Zhou, Z. 2021. Online Multi-Armed Bandits with Adaptive Inference. *Advances in Neural Information Processing Systems*, 34.

Dudík, M.; Erhan, D.; Langford, J.; and Li, L. 2014. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4): 485–511.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Fontaine, X.; Berthet, Q.; and Perchet, V. 2019. Regularized contextual bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2144–2153. PMLR.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*.

Henderson, P.; Chugg, B.; Anderson, B.; Altenburger, K.; Turk, A.; Guyton, J.; Goldin, J.; and Ho, D. E. 2022. Integrating Reward Maximization and Population Estimation: Sequential Decision-Making for Internal Revenue Service Audit Selection. *arXiv preprint arXiv:2204.11910*.

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.

Jiang, N.; and Li, L. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 652–661. PMLR.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2016. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Narain, R. 1951. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3(2): 169–175.

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.

Rosén, B. 1997. On sampling with probability proportional to size. *Journal of statistical planning and inference*, 62(2): 159–191.

Rosén, B. 2000. On inclusion probabilities for order πps sampling. *Journal of statistical planning and inference*, 90(1): 117–143.

Russo, D.; and Zou, J. 2016. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, 1232–1240. PMLR.

Sampford, M. R. 1967. On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54(3-4): 499–513.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Shin, J.; Ramdas, A.; and Rinaldo, A. 2019. Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems*, 32.

Shin, J.; Ramdas, A.; and Rinaldo, A. 2020. On conditional versus marginal bias in multi-armed bandits. In *International Conference on Machine Learning*, 8852–8861. PMLR.

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.

Williams, R. J.; and Peng, J. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3): 241–268.

Xiao, C.; Huang, R.; Mei, J.; Schuurmans, D.; and Müller, M. 2019. Maximum entropy monte-carlo planning. *Advances in Neural Information Processing Systems*, 32.

Zhang, K.; Janson, L.; and Murphy, S. 2020. Inference for batched bandits. *Advances in neural information processing systems*, 33: 9818–9829.

Zimmert, J.; and Seldin, Y. 2021. Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits. *J. Mach. Learn. Res.*, 22(28): 1–49.