Cost-Aware Near-Optimal Policy Learning

Joy He-Yueya*, Jonathan Lee*, Matthew Jörke, Emma Brunskill

Stanford University
heyueya@cs.stanford.edu, jnl@stanford.edu, joerke@stanford.edu, ebrun@cs.stanford.edu

Abstract

It is often of interest to learn a context-sensitive decision policy, such as in contextual multi-armed bandit processes. To quantify the efficiency of a machine learning algorithm for such settings, probably approximately correct (PAC) bounds, which bound the number of samples required, or cumulative regret guarantees, are typically used. However, real-world settings often have limited resources for experimentation, and decisions/interventions may differ in the amount of resources required (e.g., money or time). Therefore, it is of interest to consider how to design an experiment strategy that reduces the experimental budget needed to learn a near-optimal contextual policy. Unlike reinforcement learning or bandit approaches that embed costs into the reward function, we focus on reducing resource use in learning a near-optimal policy without resource constraints. We introduce two resource-aware algorithms for the contextual bandit setting and prove their soundness. Simulations based on real-world datasets demonstrate that our algorithms significantly reduce the resources needed to learn a near-optimal decision policy compared to previous resource-unaware methods.

Code and datasets — https://github.com/joyheyueya/cost-aware-policy-learning

Introduction

Consider designing a program to support people to attend their court date (Fishbane, Ouss, and Shah 2020; Chohlas-Wood et al. 2021). Missing a required court appearance, even for a minor offense, can sometimes lead to warrants and jail time—consequences many defendants may be unaware of. Missed appointments also waste time and money in the judicial system. Historically, if one wanted to test out interventions that were specific to each individual, one would have had to rely on mailing different interventions to different individuals. This process requires significant manual human effort and provides limited time specificity. In contrast, as the majority of people now carry a cell phone, one can now automatically send targeted intervention support (such as text messages, transit coupons, or ride-sharing fares) to specific individuals at specific times. The costs involved may vary substantially by individual and action—while a text message is very low cost, providing transportation for an individual living far from the court house is expensive.

Prior work (Chohlas-Wood et al. 2021) has modeled this as a contextual multi-armed bandit (CMAB) problem and aimed to learn high-performing policies within a short experimental period. However, such work neglects the potentially vast differences in costs incurred by different actions during the experimental period.

Indeed, considering the cost incurred to learn high-performance decision policies is relevant in many social impact settings. As a second example, we consider interventions to support voter turnout (Gerber, Green, and Larimer 2008) and learning a contextual bandit policy that could customize interventions per potential voter. There are many potential interventions that might be used to increase voter turnout, including phone calls, letters, or in person visits, which have substantially different costs. In addition, in such settings it would be of particular interest to design a data gathering experimental policy in advance to allocate samples such that after a single election a good decision policy could be learned. In line with prior work (Zanette et al. 2021), we call this static exploration, since it cannot use the observed outcomes from one context-decision to adapt and change the action taken for the next context (which would require waiting for one election for each decision). Reducing the cost of learning a near-optimal contextual policy would be highly valuable for campaigns and advocacy groups.

Motivated by such examples, we introduce a new setting of cost-aware near-optimal policy learning for contextual multi-armed bandits (CMABs). Historically, the majority of research on multi-armed bandits (MABs) and CMABs has focused on cumulative regret minimization (Lattimore and Szepesvári 2020). Cumulative regret measures policy suboptimality at each round of interaction during exploration, whereas our pure exploration setting aims to minimize policy suboptimality only at the end of exploration. There is substantial research on best arm identification in non-contextual MABs (Audibert, Bubeck, and Munos 2010; Jamieson and Nowak 2014; Kaufmann, Cappé, and Garivier 2016). Most of this work focuses on sample efficiency and does not consider cost. Another line of work considers knapsack bandits where there is a fixed budget and tries to maximize the reward obtained given that cumulative budget (Slivkins 2019). The

*These authors contributed equally.

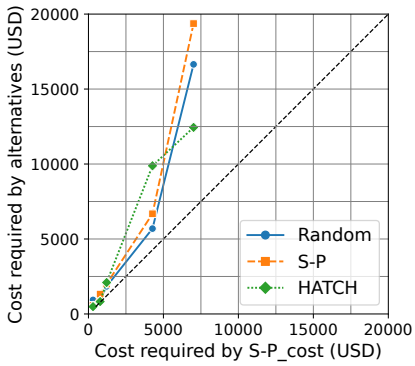
majority of this work focuses on MABs, though there is some work for the CMAB case (Wu et al. 2015), which has again focused on cumulative regret given a bound on the total resources used. Recently, there have been a few papers on pure exploration for CMABs (Zanette et al. 2021; Li et al. 2022b; Krishnamurthy et al. 2023; Pacchiano, Lee, and Brunskill 2024). These papers have not considered when actions may have heterogeneous costs. In the adaptive exploration literature (including Bayesian optimization), there has been some work on cost-aware exploration (Snoek, Larochelle, and Adams 2012; Lee et al. 2021; Belakaria et al. 2023; Astudillo et al. 2021; Paria et al. 2020), but this work has not considered the CMAB setting and has largely not provided finite sample bounds on the resulting learned optima (with the exception of Paria et al. (2020)).

Perhaps most similar to our setting is work on CMABs and Markov decision processes with constraints. Most of this work has focused on minimizing cumulative regret subject to some constraints on the policy class (Amani, Alizadeh, and Thrampoulidis 2019; Pacchiano et al. 2021), or does not consider contextual bandits (Carlsson et al. 2024). Most literature on Markov decision processes with constraints has focused on planning (Altman 2021) though there is work focusing on learning, which tends to assume that there is a fixed constraint that must be satisfied, and which constrains what policy can be optimal (Germano et al. 2023; Sun et al. 2021). In contrast, our aim is to learn a near-optimal decision policy for the contextual bandit setting while reducing the cost needed to learn this policy.

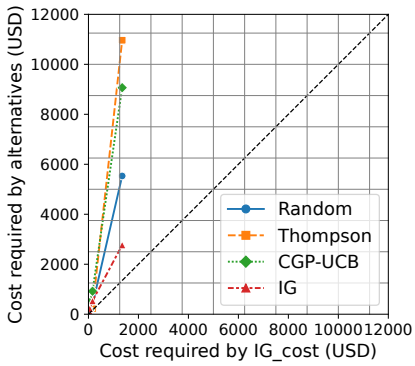
Our paper makes the following contributions:

- We introduce the setting of cost-aware near-optimal policy learning for CMABs.
- We introduce two algorithms for cost-aware near-optimal policy learning in CMABs: (1) a non-adaptive algorithm for settings where it is beneficial for the exploration policy to be static (such as the voting encouragement domain), and (2) an adaptive Bayesian algorithm for settings where we can update our policy as exploration proceeds (such as the court appearance setting).
- More significantly, we empirically show that our cost-aware algorithms require significantly fewer resources to learn a near-optimal decision policy compared to cost-unaware methods, in simulations based on real-world datasets on court appearance (Chohlas-Wood et al. 2021) and voting outcomes (Gerber, Green, and Larimer 2008). As Figure 1 illustrates, our cost-aware approaches can learn an epsilon-optimal policy with far fewer resources than existing approaches, for a range of epsilons.

We conclude the paper with a discussion of open issues.



(a) Static exploration. Voting encouragement simulation.



(b) Adaptive exploration. Court appearance support simulation.

Figure 1: Our cost-aware algorithms (S-P_cost and IG_cost) learn a near-optimal decision policy using substantially fewer resources than traditional cost-unaware algorithms. Each dot represents the mean of 50 trials, showing the cumulative exploration costs required by our method compared to alternatives at different performance levels.

Related Work

There is an extensive and growing literature on CMABs and adaptive experimental design.

Cumulative reward optimization for MABs with a budget constraint. A number of papers have considered how to adaptively pull arms to maximize cumulative reward given a constraint on the total budget used, where arms may have heterogeneous treatment effects. These are often called bandits with knapsack constraints (Badanidiyuru, Langford, and Slivkins 2014; Agrawal and Devanur 2016; Agrawal, Devanur, and Li 2016; Agrawal et al. 2016; Slivkins 2019), and most of this work focuses on the non-contextual bandit setting, with some work considering Bayesian approaches (Xia et al. 2015). Sinha et al. (2021) introduce a variant of the MAB problem where the learner is willing to tolerate a small loss from the highest reward to reduce costs. Kanarios, Zhang, and Ying (2024) consider cost-aware best-arm identification, but they focus on non-contextual settings with unknown costs. There has been less work on CMABs and most of this work has focused on theoretical analysis, and the algorithms are not always computationally tractable. For CMABs with budget constraints, Wu et al. (2015) introduce an approximate linear programming method for small discrete state spaces and provide cumulative regret bounds.

Pure exploration in CMABs. Recently, there have been several papers that focus on quickly learning a good decision policy for CMABs in the efficient pure exploration setting. Zanette et al. (2021) provide a static (non-adaptive) algorithm

with tight minimax bounds on the number of samples needed for learning a contextual policy with expected near-optimal performance in linear CMABs. Li et al. (2022b) provide an instance-optimal algorithm for PAC learning of the optimal policy within a policy class for contextual bandits, and very recent work by Krishnamurthy et al. (2023) presents an algorithm for balancing simple regret and cumulative regret minimization. None of these consider settings where actions have heterogeneous costs.

Conservative and safe bandits. Another line of work that is loosely related to our setting is conservative bandits (Wu et al. 2016), where the learner maximizes the cumulative reward while ensuring the reward of the chosen arm is above a fixed percentage of a known arm. There is also substantial research on bandits with safety constraints. Amani, Alizadeh, and Thrampoulidis (2019) consider minimizing cumulative regret with respect to the best safe actions, whereas we focus on reducing cost while learning a policy that minimizes simple regret. Pacchiano et al. (2021) consider cost but focus on minimizing cumulative regret while ensuring the deployed policy on each round satisfies a cost constraint. Similar to us, Carlsson et al. (2024) focus on learning a near-optimal policy, but they consider multi-armed bandits with no context and assume there are constraints on the arms. Additionally, there have been papers on safe data collection for evaluating a single known policy π, subject to restrictions on the exploration policy π_e being used to gather data to evaluate the value of π (Zhu and Kveton 2021; Wan, Kveton, and Song 2022).

Active learning. Our setting is loosely related to the active learning problem in machine learning, where the goal is to maximize the model accuracy while minimizing the total cost of annotating the data used to train the model. Many previous studies assume that the cost of obtaining each sample is the same, some studies consider varying costs (Settles, Craven, and Friedland 2008; Kapoor, Horvitz, and Basu 2007; Haertel et al. 2008). However, active learning is focused on supervised learning models, where the next sample is chosen and the full label is observed. In CMABs, we do not get to choose the next state—we only get to choose the action for a given state and observe its reward.

Bayesian optimization and Experimental design. Our setting also overlaps broadly with many other research areas focused on efficient data collection to learn the optima of a function (Bayesian optimization) or to gather as much information as possible about some parameters of interest (Bayesian optimal experimental design).

While Bayesian optimization and pure exploration in bandits are closely related (Srinivas et al. 2012; Krause and Ong 2011), Bayesian optimization techniques often use Gaussian processes to model complex, black-box functions, while bandit algorithms often leverage parametric structure in rewards for statistical efficiency gains. Some Bayesian optimization papers explicitly consider heterogeneous sampling cost in the acquisition function used to direct sampling (Snoek, Larochelle, and Adams 2012; Lee et al. 2021; Astudillo et al. 2021; Belakaria et al. 2023). A simple approach proposed is to move from the popular acquisition function of expected improvement, to expected improvement per unit of cost (Snoek, Larochelle, and Adams 2012). Recent work shows this can be suboptimal in the Bayesian optimization setting (Astudillo et al. 2021) and has considered unknown costs using a multi-step lookahead approach (Astudillo et al. 2021; Lee et al. 2021) with a finite fixed budget. However, this work has focused on learning the optima for generic function optimization, has not considered parametric structure, and are focused on finding the best optima given a fixed input budget. In contrast, we provide finite bounds that can be used to bound the expected simple regret of the learned contextual policy. Perhaps most similar is work by Paria et al. (2020), which uses a cost-aware version of information-directed sampling (Russo and Van Roy 2018) to guide exploration for generic Bayesian optimization. One of our algorithms is also related to information-directed sampling, but we formulate a different objective and focus on CMABs. Li et al. (2022a) introduce an algorithm for best-policy identification, but do not consider action costs or budget minimization.

Constrained Markov Decision Processes (MDPs). Altman (2021) focuses on planning with constrained MDPs; in contrast, we focus on online learning of a policy through active data collection for CMABs. There is additional work in learning in constrained MDPs (Germano et al. 2023). Such work often assumes there is a fixed constraint that must be satisfied and constrains what policy can be optimal (Sun et al. 2021). In contrast, we aim to find a near-optimal policy with no constraints and reduce the cost required to learn that policy.

Setting

We consider the stochastic contextual bandit environment where at each round $n \in [N]$, a context $s_n \in \mathcal{S}$ is sampled i.i.d. from a distribution μ. For each context s_n, a (potentially context-dependent) finite action set $\mathcal{A}_{s_n}$ is made available to the learner. The bandit instance is defined by a reward function $r : \mathcal{S} \times \mathcal{A}_s \mapsto \mathbb{R}$. Upon choosing an action $a_n \in \mathcal{A}_{s_n}$, a stochastic sample r_n with mean $r(s_n, a_n)$ is revealed to the learner. The reward model parameterization depends on the problem setting and we consider several specific settings. We assume there is a known, deterministic, and positive resource cost $c(s, a) \in \mathbb{R}^+$ for each state-action pair.

We define a decision policy π to be a mapping from contexts to actions: $\pi : \mathcal{S} \to \mathcal{A}$. Let $V(\pi)$ denote the expected reward (i.e., value) of a policy π,

$$V(\pi) := \mathbb{E}_{s\sim\mu}[r(s, \pi(s))], \tag{1}$$

where the expectation is taken over the context distribution, the stochasticity in the observed rewards, and any stochasticity in the policy π. The optimal policy $\pi^\star$ maximizes the expected reward: $\pi^\star = \max_\pi V(\pi)$.

In the pure exploration/simple regret setting, the goal is create an efficient exploration policy π_e to gather a dataset $\mathcal{D}$, such that a near-optimal policy $\widehat{\pi}(\mathcal{D})$ can be learned from the resulting dataset $\mathcal{D}$.

In prior work, efficiency has been defined by the number of samples needed to achieve a particular performance bound ϵ on the resulting policy

$$V(\pi^\star) - V(\widehat{\pi}(\mathcal{D})) \leq \epsilon \tag{2}$$

In our work, we are interested in designing resource-aware exploration algorithms, which aim to reduce the sum of costs

Algorithm 1: COST-AWARE PLANNER

1: **Input**: Contexts $\mathcal{C} = \{s_1, \ldots, s_M\}$, regularization λ
2: $\Sigma_1 = \lambda I$
3: $m = 1$
4: **for** $m = 1, 2, \ldots M$ **do**
5: **if** $\det(\Sigma_m) > 2\det(\Sigma_{\underline{m}})$ or $m = 1$ **then**
6: $\underline{m} \leftarrow m$
7: $\Sigma_{\underline{m}} \leftarrow \Sigma_m$
8: **end if**
9: Define $\pi_m : s \mapsto \text{argmax}_{a \in \mathcal{A}_s} \frac{\|\phi(s,a)\|^2_{\Sigma_{\underline{m}}^{-1}}}{c(s,a)}$
10: $\Sigma_{m+1} = \Sigma_m + \alpha\phi_m\phi_m^\top;\ \phi_m = \phi(s_m, \pi_m(s_m))$
11: **end for**
12: **return** policy mixture π_{mix} of $\{\pi_1, \ldots, \pi_M\}$

Algorithm 2: COST-AWARE SAMPLER

1: **Input**: $\pi_{mix} = \{\pi_1, \ldots, \pi_M\}$, regularization λ
2: Set $\mathcal{D}' = \emptyset$
3: **for** $n = 1, 2, \ldots N$ **do**
4: Receive context $s'_n \sim \mu$
5: Sample $m \in [M]$ uniformly at random
6: Select action $a'_n = \pi_m(s'_n)$
7: Receive feedback reward r'_n
8: Store feedback $\mathcal{D}' = \mathcal{D}' \cup \{s'_n, a'_n, r'_n\}$
9: **end for**
10: **return** dataset $\mathcal{D}'$

incurred during exploration $c(\mathcal{D}) = \sum_{(s_n,a_n)\in\mathcal{D}} c(s_n, a_n)$ relative to the ϵ-accuracy of the resulting learned policy. While provably minimizing this cost may involve complex optimization programs (similar to knapsack problems), we will shortly introduce and show that myopic cost-aware exploration strategies involve the same computational cost as prior related methods, but can offer notable improvements in the cost required to learn the same ϵ-optimal policy.

Finally, in CMABs, the primary focus has been on the realizable setting where we assume access to a statistical parameterized model that can capture the true reward function. Similarly, we assume access to a particular function class (such as a linear model) that describes the reward function. We consider both the frequentist setting where there is a single fixed but unknown parameter and a Bayesian setting in which a prior over the reward model parameters is provided. Before proceeding we briefly define our notation.

Notation. Unless otherwise stated, we let $\|x\|$ denote the l_2-norm of a vector $x \in \mathbb{R}^d$. For a positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, let $||x||_\Sigma = \sqrt{x^\top \Sigma x}$. For a set S we let $\Delta(S)$ denote the set of (appropriately defined) distributions over S. We use $I_d \in \mathbb{R}^{d\times d}$ to denote the d-dimensional identity matrix.

Algorithms

In the pure exploration setting, a key question is whether it is required to specify a fixed exploration policy in advance of data collection (the *static* setting) or whether it is possible to update the exploration policy during data collection in response to observed rewards (the *adaptive* setting). We present two algorithms, one for each setting, that build on prior work by introducing modifications to past algorithms to account for costs.

Static Cost-Aware Exploration

In many practical settings of interest, it is not possible to deploy a policy that is updated during exploration (Matsushima et al. 2020; Pacchiano, Lee, and Brunskill 2024; Zanette et al. 2021). Continual updates can require significant engineering overhead, and may even be infeasible when rewards are delayed or in studies with parallel treatment assignment. Consider learning a contextual bandit policy that could customize interventions per potential voter to increase voter turnout (Gerber, Green, and Larimer 2008). Many potential voters are assigned to different conditions in parallel, and voting outcomes can only be observed once per election. Here, we propose a cost-aware algorithm for static exploration in settings where different actions have different costs.

In particular, we restrict ourselves to the well-studied stochastic *linear* contextual bandits setting. We assume there is a known feature map $\phi : \mathcal{S} \times \mathcal{A}_s \mapsto \mathbb{R}^d$ and the reward model follows $r_{\theta^\star}(s, a) = \phi(s, a)^\top\theta^\star$, where $\theta^\star \in \mathbb{R}^d$ is an unknown parameter. Upon choosing an action $a \in \mathcal{A}_s$, the reward $r = r_{\theta^\star}(s, a) + \eta$ is revealed to the learner, where η is mean-zero, 1-sub-Gaussian noise. As is standard, we assume that $\|\theta^\star\| \leq 1$ and $\sup_{s,a} \|\phi(s, a)\| \leq 1$ such that $|r_{\theta^\star}(s, a)| \leq 1$. For a given parameter $\theta \in \mathbb{R}^d$, we define $\pi_\theta(s) = \text{argmax}_{a\in\mathcal{A}_s} \phi(s, a)^\top\theta$ to be a greedy policy with respect to θ. The optimal policy $\pi^\star$ is defined as $\pi_{\theta^\star}$.

Our static pure exploration setting proceeds in two phases. First, we design an exploration policy π_e to construct a dataset $\mathcal{D}' = \{(s'_n, a'_n, r'_n)\}_{n=1,\ldots,N}$. Then, using $\mathcal{D}'$, we extract the regularized least-square predictor $\hat{\theta} = \left(\Sigma'_N\right)^{-1} \sum_{i=1}^N \phi(s'_n, a'_n)r'_n$, where $\Sigma'_N = \lambda I_d + \sum_{n\in[N]} \phi(s'_n, a'_n)\phi(s'_n, a'_n)^\top$ and $\lambda > 0$. Our objective is to design π_e such that the simple regret of the greedy decision policy $\hat{\pi} = \arg\max_{a\in\mathcal{A}_s} \phi(s, a)^\top\hat{\theta}$ is minimized. It is known from prior work (Zanette et al. 2021) that to learn an ϵ-optimal policy for $\epsilon > 0$, it suffices to design the exploration policy π_e so as to minimize the maximum uncertainty $\mathbb{E}_{s\sim\mu} \max_a \|\phi(s, a)\|_{\Sigma'_N}$. Zanette et al. (2021) propose the sampler-planner algorithm (S-P), which selects the action $a_m = \arg\max_a \|\phi(s, a)\|_{\Sigma_m^{-1}}$ that maximizes the uncertainty with respect to the current covariance matrix every time a context $s_m \sim \mu$ is observed.

Building on the S-P algorithm, we propose the S-P_COST algorithm that explicitly thinks about the cost for minimizing the maximum uncertainty. S-P_COST consists of two subroutines: the *cost-aware planner* (see Alg. 1) and *cost-aware sampler* (see Alg. 2). Similar to a reward-free version of the LinUCB algorithm (Abbasi-Yadkori, Pál, and Szepesvári 2011), the planner leverages an offline set of contexts and chooses the action $a_m = \arg\max_a \frac{\|\phi(s,a)\|^2_{\Sigma_{\underline{m}}^{-1}}}{c(s,a)}$ (see line 9 of Alg. 1) every time a context $s_m \sim \mu$ is observed. Intuitively, $\frac{\|\phi(s,a)\|^2_{\Sigma_{\underline{m}}^{-1}}}{c(s,a)}$ represents the uncertainty per unit cost, and we

would like to maximize the uncertainty reduction per unit cost. We run Alg. 1 for M iterations, where M could be determined by ϵ (see Zanette et al. (2021) for details). Upon termination, the planner outputs a sequence of policies $\pi_1, \ldots, \pi_M$. The sampler then uses the average mixture policy π_{mix} to gather a dataset: for each new context, π_{mix} samples an index $m \in [M]$ uniformly at random and plays π_m. We run Alg. 2 for N iterations. The sampler's policy produces a covariance matrix close to what the planner computed with offline data, which in turn yields a bound on maximum uncertainty and thus simple regret (Zanette et al. 2021).

Adaptive Cost-Aware Exploration

Bayesian approaches are very popular in adaptive optimization and experimental design, in part because they provide a natural way to quantify information gain with respect to prior uncertainty, which can be leveraged for adaptive exploration. We introduce a simple resource-aware algorithm for pure exploration in Bayesian contextual bandits.

We here consider a more general class of reward models f, such that the observed reward when taking action a in context s is $r = f(s, a, \theta) + \eta$, where θ is the unknown parameter, η is mean-zero, 1-sub-Gaussian noise. We assume that θ is sampled from some known prior distribution.

Let $\mathcal{F}_n = \{s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_{n-1}, a_{n-1}, r_{n-1}\}$ be the sequence of states observed, actions taken, and rewards observed up to the current time point. Define $\mathbb{E}_n[X] \coloneqq \mathbb{E}[X|\mathcal{F}_n]$. Recall that the entropy of a probability distribution P_x is defined as $H(P_x) = -\sum_x P(x) \log P(x)$. Given a history $\mathcal{F}_n$, a prior $p(\theta)$, and an observed state s_n, we can define a posterior probability distribution over the optimal action $a^\star = \text{argmax}_{a \in \mathcal{A}_{s_n}} f(s_n, a, \theta)$ in state s_n:

$$\alpha_n(s_n, a) = P(a^\star = a | s_n, \mathcal{F}_n). \quad (3)$$

The information gain $g_n(a')$ of selecting a particular action a' in state s_n is defined as the expected reduction in entropy over the optimal action for state s_n after taking action a':

$$g_n(a') = \mathbb{E}_n \left[H(\alpha_n(s_n, \cdot)) - H(\alpha_{n+1}(s_n, \cdot)) \right]. \quad (4)$$

A common approach in Bayesian optimization that can also be easily applied in the pure exploration simple multi-armed bandit setting is to select actions to maximize the information gain. Russo and Van Roy (2018) introduced information-directed sampling for Bayesian cumulative regret minimization in bandits and extend the above by considering the ratio of the expected regret to the information gain.

In our setting, we are instead interested in considering the information gain in relation to resources spent. To do so, we define our exploration policy as one that maximizes the relative information gain per unit cost:

$$\pi_e(s_n) = \arg\max_{a'} \frac{g_n(a')}{c(s_n, a')}. \quad (5)$$

Our objective is very similar to work in Bayesian optimization that uses expected improvement per unit of cost an acquisition function (Snoek, Larochelle, and Adams 2012), though that work did not consider MABs or the contextual setting, nor provided finite sample analysis.

Algorithm 3: IG_cost

1: **Input**: K, r, q
2: Set $\mathcal{D}'_0 = \emptyset$
3: **for** $n = 1, 2, \ldots N$ **do**
4: Receive context $s'_n \sim \mu$
5: Draw $\theta^1, \ldots, \theta^K$ from the posterior $p(\theta|\mathcal{F}_n)$
6: $\hat{\Theta}_a \leftarrow \{m | a = \arg\max_{a'} \sum_y q_{\theta^m, s'_n, a'}(y) r(y)\}$
7: $\hat{p}(a^*) \leftarrow |\hat{\Theta}_{a^*}|/K \qquad \forall a^*$
8: $\hat{p}_a(y) \leftarrow \sum_m q_{\theta^m, s'_n, a}(y)/K \qquad \forall y$
9: $\hat{p}_a(a^*, y) \leftarrow \sum_{m \in \hat{\Theta}_{a^*}} q_{\theta^m, s'_n, a}(y)/K \qquad \forall a^*, y$
10: $\vec{g}_a \leftarrow \sum_{a^*, y} \hat{p}_a(a^*, y) \log \frac{\hat{p}_a(a^*, y)}{\hat{p}(a^*)\hat{p}_a(y)} \qquad \forall a \in \mathcal{A}_{s'_n}$
11: Select action $a'_n = \arg\max_{a \in \mathcal{A}_{s'_n}} \frac{\vec{g}_a}{c(s'_n, a)}$
12: Receive feedback reward r'_n
13: Store feedback $\mathcal{D}'_n = \mathcal{D}'_{n-1} \cup \{s'_n, a'_n, r'_n\}$
14: **end for**
15: **return** dataset $\mathcal{D}'_N$

In general, computing the information gain is computationally challenging due to intractable posteriors. Prior work often considers approximations, and we draw from (Russo and Van Roy 2018)'s algorithm for a sample-based approximation to the information gain and present a cost-aware information gathering algorithm for contextual bandits (IG_cost) in Alg. 3. This uses a sample-based approximation to Equation 5 and we restrict our attention to settings with discrete reward outcomes. We let $y(s, a) \in \mathcal{Y}$ denote the outcome of choosing action a in context s, where $\mathcal{Y}$ is a discrete set.

Alg. 3 takes as input K, r, and q. K is the number of samples drawn independently from the posterior $p(\theta|\mathcal{F}_n)$ and $r : \mathcal{Y} \mapsto \mathbb{R}$ is a reward function mapping outcomes to scalar rewards. We let $q_{\theta,s,a}(y) = P(y(s, a) = y|\theta)$ be the probability, conditioned on θ, of observing y when action a is selected in context s. Line 6 computes the optimal action for each value of θ. Line 7 computes the probability that each action is optimal. Line 8 computes the marginal distribution over the particular rewards, and line 9 computes the joint probability distribution of the optimal action and a particular reward outcome. These quantities are used to compute the information gain (see a derivation in "Definition of Information Gain in Alg. 3" in Appendix), which is then scaled by the inverse of the cost.

Experiments

Voting Encouragement Simulation

We evaluate S-P_cost on a voting dataset focused on the August 2006 primary election[1], collected by Gerber, Green, and Larimer (2008) to study the effect of social pressure on voter turnout. The researchers designed one control and 4 treatment actions that involved sending different types of letters to selected individuals before the 2006 Michigan primary election. The actions are:

- Nothing (control): No letter sent.

[1]Data available at https://github.com/gsbDBI/ExperimentData/tree/master/Social

- Civic: A letter stating "Do your civic duty."
- Hawthorne: A letter stating "You are being studied."
- Self-History: A letter showing the voter's and their household's past voting records, with a promise to send a follow-up letter after the election showing updated records.
- Neighbors-History: A letter showing the voting records of the individual, their household, and their neighbors, with a notification that a follow-up letter would be sent after the election to update everyone, thereby making the individual's participation public among the neighbors.

The data collection policy sampled actions with probability $\frac{5}{9}$ for the control action and $\frac{4}{9}$ for other actions. The reward is a binary indicator of voter participation in the 2006 primary election. For simplicity, we assume the Civic and Hawthorne actions cost $\$0.5$ for each individual, roughly aligned with the 2006 postage rate and printing costs. The Self-History action costs $\$2.5$ due to the extra effort in accessing detailed voting records, and the Neighbors-History action is the most expensive at $\$10$, requiring extensive data. Since our algorithm requires $c(s, a) > 0, \forall s, a$, we set the control action cost at a nominal $\$0.1$.

The dataset consists of $180,002$ entries, each representing a voter from a unique household across the state of Michigan. The dataset includes ten voter characteristics, such as age and gender, which are used as the context features (see details in "Experiment Details for Voting Dataset" in Appendix).

Each action in our experiment is encoded using a 6-dimensional feature vector. The first dimension identifies if the action is the control (1 for control, 0 otherwise). The second dimension denotes whether the action includes mailing a letter (1 for the four treatment actions, 0 for the control). The third through sixth dimensions respectively indicate whether the action is Civic, Hawthorne, involves checking neighbors' voting history (Neighbors-History), or requires access to household voting history (applies to Neighbors-History and Self-History), with a 1 for yes and a 0 for no. We concatenate these action features with context features about each voter and use a linear model to predict the reward based on the combined features[2].

We partition the dataset by city, using 90% for training and the remaining 10% for testing. We use rejection sampling for simulating data collection and evaluation since only the outcomes of the chosen actions were observed in the real data: For each data point, we accept it if the action selected by the policy is the same as the action selected in the historical data and reject it otherwise.

We compare our S-P_cost against three baselines: (1) a random exploration algorithm (Random) that chooses actions uniformly, (2) the cost-unaware sampler-planner algorithm (S-P) proposed in Zanette et al. (2021), and (3) a variant of S-P_cost inspired by HATCH (Yang et al. 2020), which adaptively allocates a fixed exploration budget based on the remaining resources, action costs, and estimated uncertainty reduction. Unlike our S-P_cost, HATCH requires a fixed constraint on the number of samples. To evaluate its performance, we run HATCH with various constraints (e.g., $M = N = 10000, \ldots, 50000$). For S-P and S-P_cost, the planner is first run on the training set repeatedly until it has processed M data points ($M = N = 50000$). All algorithms are then used to collect a dataset of size N. After collecting each sample during exploration, we calculate the value of the resulting greedy decision policy for each algorithm on the test set of 53505 data points. The policy value is computed as the average reward on the test set. Additionally, we estimate the optimal value that could be achieved by a supervised learning oracle that has access to all state-action pairs in the training set, which represents an approximate upper bound. All algorithms use $\lambda = 1$, and the planner uses $\alpha = 1$, as these work well empirically (Zanette et al. 2021). We run the experiment for 50 trials using random seeds 1-50.

Figure 1a shows cumulative exploration costs required by our S-P_cost versus alternatives to achieve various performance gaps ϵ (i.e., the difference between the value of the optimal policy and the learned policy). We averaged costs within 5 equally spaced ϵ intervals, and each dot represents a particular epsilon interval. Our S-P_cost learns an ϵ-optimal policy at a lower exploration cost than S-P. S-P_cost tends to choose the cheaper actions to get information about most of the coordinates, while the cost-unaware S-P often selects the most informative yet expensive actions, such as Neighbors-History, without considering cost-efficiency. Figure 2a displays these costs plotted against the performance gap ϵ. We find similar results across various nominal values (i.e., the small cost assigned to actions that might otherwise be considered free), as shown in Figure 3 in the appendix.

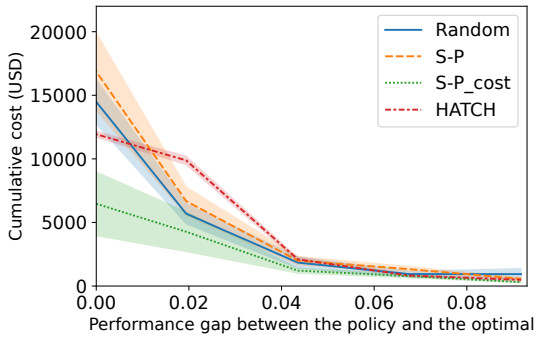
Interestingly, Random outperforms S-P in terms of cost. This highlights the importance of considering costs directly, as even sample-efficient algorithms may not necessarily lower the real experimental costs compared to random data collection.

Court Appearance Simulation

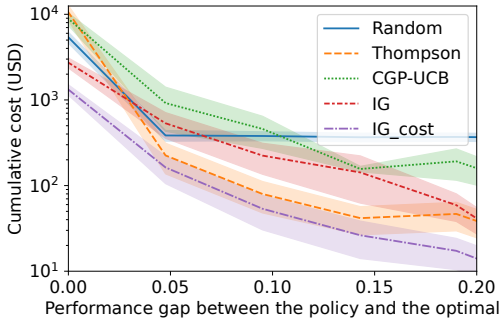
We evaluate IG_cost using a semi-synthetic court appearance simulator from Chohlas-Wood et al. (2021), which is grounded in real case data from the Santa Clara County Public Defender Office. In this setting, a policymaker seeks to help individuals attend their court dates by providing government-sponsored transportation assistance. Individuals can receive one of three mutually exclusive interventions a: rideshare assistance, a transit voucher, or no transportation assistance. The round-trip rides cost $\$5$ for every mile between an individual's home address and the courthouse and back. The transit voucher costs $\$7.5$. Since our algorithm requires $c(s, a) > 0, \forall s, a$, we assume that the no transportation assistance intervention has a cost of $\$0.1$, which is negligible compared to the other interventions. The simulator considers the binary outcome $y \in \{0, 1\}$ that indicates whether a client appeared at their court date. The simulator uses a logistic reward model, where $\mathbb{P}(r_{\theta^\star}(s, a) = 1) = \text{logit}^{-1}(\phi(s, a)^\top \theta^\star)$ for some unknown $\theta^\star \in \mathbb{R}^d$. The reward is independent across draws.

The resulting dataset consists of $12,636$ example cases. Each data point is a 7-dimensional feature vector associated with the true appearance probability of the individual, the observed binary outcome, and the cost of the intervention if provided each of the three interventions. The simulator is designed in such a way that the type of assistance that is

[2]We also tested a logistic model and observed a similar fit.



(a) Static exploration. Voting encouragement simulation.



(b) Adaptive exploration. Court appearance support simulation.

Figure 2: The x-axis shows the performance gap ϵ, which is the difference between the value of the optimal policy and the learned decision policy. The y-axis shows the total cost of the experiment. Each line represents the mean of 50 trials, with error bars indicating the standard error.

best for each individual varies across the population. The goal is to learn which intervention maximizes the appearance probability for each individual.

We compare our IG_cost against four baselines: (1) a random exploration algorithm (Random) that always chooses actions uniformly, (2) Thompson sampling (Thompson) (Russo et al. 2018), (3) Contextual Gaussian Process Upper Confidence Bound (CGP-UCB) (Krause and Ong 2011), and (4) the cost-unaware algorithm that only considers the information gain (IG). Thompson maintains a posterior over the parameters of the reward model. Every time a context $s \sim \mu$ is observed, Thompson samples θ from this posterior and selects the action $a = \arg\max_a r_\theta(s, a)$. We compare against Thompson because prior work (Chohlas-Wood et al. 2021) has shown that Thompson is good at simple regret minimization even though it optimizes for cumulative regret. CGP-UCB is a Bayesian optimization approach to pure exploration that relies on an underlying Gaussian process (GP) that takes as input a policy's parameters and is trained to predict policy value. CGP-UCB proceeds by selecting a candidate policy that maximizes the upper confidence bound over policy value as predicted by the GP model. We also introduce IG, which selects action $a = \arg\max \vec{g}_a$ that maximizes the information gain instead of the relative information gain per cost (see line 11 of Alg. 3). Following Chohlas-Wood et al. (2021), we use non-informative priors, and we use the `sim` function in `arm` (Gelman 2011) to do posterior sampling (see details in "Court Appearance Experiment" in Appendix).

We run the experiment for 50 trials using random seeds 1-50. In each trial, we randomly sample 1500 data points as the training data and test data, respectively. Following Chohlas-Wood et al. (2021), we start each trial with a randomly selected warm-up group of 4 people. We run all algorithms on the training data such that they observe the same contexts but may take different actions. After each training observation, we calculate the value of the resulting greedy decision policies, which is the appearance rate under the policy on test data. As a baseline, we calculate the value of the optimal policy given access to the true reward model.

We find that IG_cost can learn an ϵ-optimal decision policy using substantially fewer exploration resources than the other cost-unaware algorithms. Figure 1b shows the cumulative exploration costs required by our IG_cost compared to alternatives at different performance levels. We averaged costs within 10 equally spaced ϵ intervals. Figure 2b shows cumulative exploration costs plotted over the performance gap ϵ. Similar to the voting encouragement setting, Random outperforms two of the three more sample-efficient alternatives in terms of cost, further demonstrating the advantages of directly considering costs.

Theoretical Guarantees

One might worry that our modifications could break the established theoretical properties of S-P and IG. However, we show that our cost-aware algorithms can provably recover ϵ-optimal policies in certain settings, provided the ratio between the maximum costs and minimum costs is bounded. We prove this in "Proofs of Theoretical Results" in the appendix, but our primary contribution is introducing this setting and the empirical results for important societal applications.

Discussion and Conclusion

Our approach for Bayesian pure exploration for contextual bandits maximizes the information gain per unit cost for the current context but does not consider the potential benefit for all contexts. Recent work by Hao, Lattimore, and Qin (2022) has demonstrated that such conditional strategies may be outperformed by context-aware strategies in cumulative regret minimization (without costs), which presents an interesting direction for further investigation. A very interesting open question is whether it is important to consider longer horizons when designing sampling strategies to learn near-optimal policies given heterogeneous resource costs.

Our work has highlighted the substantial potential impact of considering costs when learning a near-optimal decision policy. While our algorithms are not guaranteed to find the minimal cumulative cost, we view our primary contribution as providing a new important problem of how to reduce costs required to learn a good final decision policy and two practical algorithms that demonstrated substantial gains in simulations based on real-world datasets.

Acknowledgments

We would like to thank Alex Chohlas-Wood, Ethan Prihar, and Ge Gao for their feedback. This work was supported in part by the Junglee Corporation Stanford Graduate Fellowship and NSF grant #2112926.

References

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

Agrawal, S.; Avadhanula, V.; Goyal, V.; and Zeevi, A. 2016. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, 599–600.

Agrawal, S.; and Devanur, N. R. 2016. Linear contextual bandits with knapsacks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3458–3467.

Agrawal, S.; Devanur, N. R.; and Li, L. 2016. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, 4–18. PMLR.

Altman, E. 2021. *Constrained Markov decision processes*. Routledge.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32.

Astudillo, R.; Jiang, D.; Balandat, M.; Bakshy, E.; and Frazier, P. 2021. Multi-step budgeted bayesian optimization with unknown evaluation costs. *Advances in Neural Information Processing Systems*, 34: 20197–20209.

Audibert, J.-Y.; Bubeck, S.; and Munos, R. 2010. Best arm identification in multi-armed bandits. In *COLT*, 41–53.

Badanidiyuru, A.; Langford, J.; and Slivkins, A. 2014. Resourceful contextual bandits. In *Conference on Learning Theory*, 1109–1134. PMLR.

Belakaria, S.; Doppa, J. R.; Fusi, N.; and Sheth, R. 2023. Bayesian optimization over iterative learners with structured responses: A budget-aware planning approach. In *International Conference on Artificial Intelligence and Statistics*, 9076–9093. PMLR.

Carlsson, E.; Basu, D.; Johansson, F.; and Dubhashi, D. 2024. Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, 334–342. PMLR.

Chohlas-Wood, A.; Coots, M.; Zhu, H.; Brunskill, E.; and Goel, S. 2021. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792*.

Fishbane, A.; Ouss, A.; and Shah, A. K. 2020. Behavioral nudges reduce failure to appear for court. *Science*, 370(6517): eabb6591.

Gelman, A. 2011. arm: Data analysis using regression and multilevel/hierarchical models. *http://cran. r-project. org/web/packages/arm*.

Gerber, A. S.; Green, D. P.; and Larimer, C. W. 2008. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1): 33–48.

Germano, J.; Stradi, F. E.; Genalti, G.; Castiglioni, M.; Marchesi, A.; and Gatti, N. 2023. A best-of-both-worlds algorithm for constrained mdps with long-term constraints. *arXiv preprint arXiv:2304.14326*.

Haertel, R. A.; Seppi, K. D.; Ringger, E. K.; and Carroll, J. L. 2008. Return on investment for active learning. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 72. Citeseer.

Hao, B.; Lattimore, T.; and Qin, C. 2022. Contextual information-directed sampling. In *International Conference on Machine Learning*, 8446–8464. PMLR.

Jamieson, K.; and Nowak, R. 2014. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, 1–6. IEEE.

Kanarios, K.; Zhang, Q.; and Ying, L. 2024. Cost Aware Best Arm Identification. *arXiv preprint arXiv:2402.16710*.

Kapoor, A.; Horvitz, E.; and Basu, S. 2007. Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning. In *IJCAI*, volume 7, 877–882.

Kaufmann, E.; Cappé, O.; and Garivier, A. 2016. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17: 1–42.

Krause, A.; and Ong, C. 2011. Contextual gaussian process bandit optimization. *Advances in neural information processing systems*, 24.

Krishnamurthy, S. K.; Zhan, R.; Athey, S.; and Brunskill, E. 2023. Proportional Response: Contextual Bandits for Simple and Cumulative Regret Minimization. *arXiv preprint arXiv:2307.02108*.

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

Lee, E. H.; Eriksson, D.; Perrone, V.; and Seeger, M. 2021. A nonmyopic approach to cost-constrained Bayesian optimization. In *Uncertainty in Artificial Intelligence*, 568–577. PMLR.

Li, X.; Mehta, V.; Kirschner, J.; Char, I.; Neiswanger, W.; Schneider, J.; Krause, A.; and Bogunovic, I. 2022a. Near-optimal policy identification in active reinforcement learning. *arXiv preprint arXiv:2212.09510*.

Li, Z.; Ratliff, L.; Jamieson, K. G.; Jain, L.; et al. 2022b. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35: 37590–37603.

Matsushima, T.; Furuta, H.; Matsuo, Y.; Nachum, O.; and Gu, S. 2020. Deployment-efficient reinforcement learning via model-based offline optimization. *arXiv preprint arXiv:2006.03647*.

Pacchiano, A.; Ghavamzadeh, M.; Bartlett, P.; and Jiang, H. 2021. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, 2827–2835. PMLR.

Pacchiano, A.; Lee, J.; and Brunskill, E. 2024. Experiment planning with function approximation. *Advances in Neural Information Processing Systems*, 36.

Paria, B.; Neiswanger, W.; Ghods, R.; Schneider, J.; and Póczos, B. 2020. Cost-aware bayesian optimization via information directed sampling. In *Adaptive Experimental Design and Active Learning in the Real World Workshop at ICML*.

Russo, D.; and Van Roy, B. 2018. Learning to optimize via information-directed sampling. *Operations Research*, 66(1): 230–252.

Russo, D. J.; Van Roy, B.; Kazerouni, A.; Osband, I.; Wen, Z.; et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1): 1–96.

Settles, B.; Craven, M.; and Friedland, L. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Sinha, D.; Sankararaman, K. A.; Kazerouni, A.; and Avadhanula, V. 2021. Multi-armed bandits with cost subsidy. In *International Conference on Artificial Intelligence and Statistics*, 3016–3024. PMLR.

Slivkins, A. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning®*, 12(1-2): 1–286.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. W. 2012. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5): 3250–3265.

Sun, H.; Xu, Z.; Fang, M.; Peng, Z.; Guo, J.; Dai, B.; and Zhou, B. 2021. Safe exploration by solving early terminated mdp. *arXiv preprint arXiv:2107.04200*.

Wan, R.; Kveton, B.; and Song, R. 2022. Safe exploration for efficient policy evaluation and comparison. In *International Conference on Machine Learning*, 22491–22511. PMLR.

Wu, H.; Srikant, R.; Liu, X.; and Jiang, C. 2015. Algorithms with Logarithmic or Sublinear Regret for Constrained Contextual Bandits. *Advances in Neural Information Processing Systems*, 28: 433–441.

Wu, Y.; Shariff, R.; Lattimore, T.; and Szepesvári, C. 2016. Conservative bandits. In *International Conference on Machine Learning*, 1254–1262. PMLR.

Xia, Y.; Li, H.; Qin, T.; Yu, N.; and Liu, T.-Y. 2015. Thompson Sampling for Budgeted Multi-Armed Bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 3960–3966. AAAI Press. ISBN 9781577357384.

Yang, M.; Li, Q.; Qin, Z.; and Ye, J. 2020. Hierarchical adaptive contextual bandits for resource constraint based recommendation. In *Proceedings of the web conference 2020*, 292–302.

Zanette, A.; Dong, K.; Lee, J. N.; and Brunskill, E. 2021. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34: 22720–22731.

Zhu, R.; and Kveton, B. 2021. Safe data collection for offline and online policy learning. *arXiv preprint arXiv:2111.04835*.