Upper Counterfactual Confidence Bounds: a New Optimism Principle for Contextual Bandits

Yunbei Xu* Assaf Zeevi†

Abstract

The principle of optimism in the face of uncertainty is one of the most widely used and successful ideas in multi-armed bandits and reinforcement learning. However, existing optimistic algorithms (primarily `UCB` and its variants) are often unable to deal with large context spaces. Essentially all existing well performing algorithms for general contextual bandit problems rely on weighted action allocation schemes; and theoretical guarantees for optimism-based algorithms are only known for restricted formulations. In this paper we study general contextual bandits under the realizability condition, and propose a simple generic principle to design optimistic algorithms, dubbed "Upper Counterfactual Confidence Bounds" (`UCCB`). We show that these algorithms are provably optimal and efficient in the presence of large context spaces. Key components of `UCCB` include: 1) a systematic analysis of confidence bounds in policy space rather than in action space; and 2) the potential function perspective that is used to express the power of optimism in the contextual setting. We further show how the `UCCB` principle can be extended to infinite action spaces, by constructing confidence bounds via the newly introduced notion of "counterfactual action divergence."

1 Introduction

1.1 Motivation.

Algorithms that rely on the "optimism principle" have been a major cornerstone in the study of multi-armed bandit (MAB) and reinforcement learning problems. Roughly speaking, optimistic algorithms are those that choose a deterministic action at each round, based on some optimistic estimate of future rewards. Perhaps the most representative example is the celebrated Upper Confidence Bounds (`UCB`) algorithm and its many variants. Popularity of optimistic algorithms stems from their simplicity and effectiveness: the analysis of `UCB`-type algorithms are usually more straightforward than alternative approaches, so they have become the "meta-algorithms" for more complex settings. They are also often preferable to weighted allocations among actions because of the ability to discard sub-optimal actions and achieve superior instance-dependent empirical performances.

Despite their prevalent use in traditional bandit problems, existing `UCB`-type algorithms have a glaring drawback in contextual MAB settings: their regret often scales with the cardinality of the context space. (Notable exceptions are the special "linear payoff" formulation [11] and its generalized-linear variant [24].) In particular, despite encouraging empirical observations [9], optimism-based algorithms provably achieve sub-linear regret only under restrictive distributional assumptions [15]. This motivates the main problem studied in the paper:

Is there a generic principle that ensures that optimistic algorithms are optimal and efficient for general contextual bandit problems?

*Columbia University, New York, NY; Email: `yunbei.xu@gsb.columbia.edu`.
†Columbia University, New York, NY; Email: `assaf@gsb.columbia.edu`.

Interestingly, whether computationally efficient or not, almost all existing solutions to general contextual bandits [3–5, 13, 16, 27] rely on weighted, randomized allocations among actions at each round—we refer to these as "randomized algorithms" in the paper. Moreover, there is little focus on contextual MAB with infinite actions, which we believe to be a natural setting to illustrate simplicity and universality of optimism-based algorithms. These observations motivate us to search for a new optimism principle in the presence of large context spaces.e is little focus on contextual MAB with infinite actions, which we believe to be a natural setting to illustrate simplicity and universality of optimism-based algorithms. These observations motivate us to search for a new optimism principle in the presence of large context spaces.

1.2 The contextual MAB problem

The canonical stochastic contextual bandit problem can be described as follows. Let $\mathcal{A}$ be the action set (in the initial parts of the paper, one can think of $\mathcal{A}$ as the integer set $\{1, \ldots, K\}$, which we generalize later on), and $\mathcal{X}$ be the space of contexts that supports the distribution $\mathcal{D}_{\mathcal{X}}$ (e.g., $\mathcal{X}$ can be a subset of Euclidean space). For all $x \in \mathcal{X}, a \in \mathcal{A}$, denote $\mathcal{D}_{x,a}$ a reward distribution determined by context x and action a. At each round $t = 1, \ldots, T$, the agent first observes a context x_t drawn i.i.d. according to $\mathcal{D}_{\mathcal{X}}$. She then chooses an action $a_t \in \mathcal{A}$ based on x_t and the history H_{t-1} generated by $\{x_i, a_i, r_i(x_i, a_i)\}_{i=1}^{t-1}$, and finally observes the reward $r_t(x_t, a_t)$, which is conditionally independent and distributed according to the distribution $\mathcal{D}_{x_t,a_t}$. We assume the rewards take values in the interval $[0, 1]$. An admissible contextual bandit algorithm `Alg` is a (possibly randomized) procedure that associates each realization of $\{H_{t-1}, x_t\}$ with an action a_t to employ at round t.

Previous literature on contextual MAB problems can be sorted into two categories: the realizable setting and the agnostic setting. In the realizable setting, the agent has access to a function class $\mathcal{F}$, with its members $f \in \mathcal{F}$ being mappings from $\mathcal{X} \times \mathcal{A}$ to $[0, 1]$. The following is referred to as the realizability condition [3, 11, 15, 16, 27] :

Assumption 1 (realizability). *There exists $f^* \in \mathcal{F}$ such that for all $t \geq 1$, $x \in \mathcal{X}$, $a \in \mathcal{A}$, the conditional mean reward, $\mathbb{E}[r_t(x_t, a_t)|x_t = x, a_t = a]$, is equal to $f^*(x, a)$.*

We call a mapping $\pi : \mathcal{X} \to \mathcal{A}$ from the context space $\mathcal{X}$ to the action set $\mathcal{A}$ a "policy." (Those mappings may be referred to more precisely as "deterministic stationary policies;" in this paper we often just refer to them as "policies" with slight abuse of terminology.) Let π_{f^*}, defined by $\pi_{f^*}(x) = \arg\max f^*(x, a)$, be the "ground truth" optimal policy. The cumulative (pathwise) regret of a contextual bandit algorithm `Alg` compared with the optimal policy π_{f^*} after T rounds is

$$\text{Regret}(T, \texttt{Alg}) := \sum_{t=1}^{T} (r_t(x_t, \pi_{f^*}(x_t)) - r_t(x_t, a_t)),$$

and the agent aims to minimize this cumulative regret. The agnostic setting [4, 5, 13, 23], on the other hand, does not make such realizability assumption; instead, algorithms are compared with the best policy within a given policy class. In this paper we focus on the realizable setting which lends itself more naturally to the design of optimism-based algorithms.

We present some examples of the realizable setting. The most well-studied contextual MAB problems are simple variants of the "linear payoff" model [11, 24]

$$\mathcal{F} = \{f : f(x, a) = \theta^T x_a, \theta \in \Theta\}, \quad \Theta, \mathcal{X} \subseteq \mathbb{R}^d, x = (x_a)_{a \in \{1,\ldots,K\}}. \tag{1.1}$$

One motivation towards general function classes is to encompass models of the form

$$\mathcal{F} = \{f : f(x, a) = g_a(x), \quad g_a \in \mathcal{G}, a \in \{1, \ldots, K\}\}, \tag{1.2}$$

where parameters of $g_a : \mathcal{X} \to \mathbb{R}$ can be distinct for different actions [18, 21]; it is also desirable to handle complex nonlinear models (such as neural networks) which are much more expressive than their linear counterparts.

On the computation side, we make the rather benign assumption the the agent has access to a pre-specified least square oracle over $\mathcal{F}$. Formally, after the agent inputs the historical data $\{x_i, a_i, r_i(x_i, a_i)\}_{i=1}^{t-1}$, the least square oracle outputs a solution $\widehat{f}_t \in \mathcal{F}$ that provides the best fit, namely,

$$\widehat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} (f(x_i, a_i) - r_i(x_i, a_i))^2. \tag{1.3}$$

This is the simplest optimization oracle assumed in the contextual bandit literature. We assume the least square oracle to be deterministic, for simplicity, as there may be multiple solutions to (1.3).

1.3 Introducing `UCCB`: two equivalent viewpoints

This subsection will describe the `UCCB` principle introduced in this paper from two equivalent viewpoints: 1) implicitly, it is an upper confidence bound rule in policy space; and 2) explicitly, it calculates the upper confidence bound via simulating *counterfactual* action trajectories rather than using the original action trajectory. For illustration purpose we focus on the finite-action setting where $\mathcal{A} = \{1, \ldots, K\}$; extension to infinite action spaces will be discussed later in Section 4.

Implicit strategy: maximizing upper confidence bounds in policy space. Let Π be the *policy space* that contains all deterministic stationary policies $\pi : \mathcal{X} \to \{1, \ldots, K\}$. The core idea of `UCCB` is to choose policies that maximize certain upper confidence bounds in the policy space Π. After initialization, for each round t, data $\{(x_i, a_i), r_i\}_{i=1}^{t-1}$ is sent to an offline least square oracle to compute the estimator $\widehat{f}_t \in \mathcal{F}$. Without the need to "see" x_t, the agent selects the optimistic policy $\pi_t \in \Pi$ (which is a mapping from $\mathcal{X}$ to the action set $\{1, \ldots, K\}$) such that

$$\pi_t \in \arg\max_{\pi \in \Pi} \left\{ \mathbb{E}_x[\widehat{f}_t(x, \pi(x)] + \mathbb{E}_x \left[\frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}} \right] + \frac{K\beta_t}{t} \right\}, \tag{1.4}$$

where the expectation $\mathbb{E}_x[\cdot]$ is history-independent and taken with respect to the distribution $\mathcal{D}_{\mathcal{X}}$ (over the random context x), and β_t is a parameter related to the complexity of the function class. (When there are multiple solutions to (1.4), we take π_t to be the unique solution such that for all other solutions π' to (1.4) and all $x \in \mathcal{X}$, the index of the action $\pi_t(x)$ is smaller than the index of the action $\pi'(x)$.) Then the agent observes x_t and selects the action $a_t = \pi_t(x_t)$. The right hand side of (1.4) is an upper confidence bound on the true expected reward of π, because we can prove that with high probability, for all $\pi \in \Pi$,

$$\left|\mathbb{E}_x[f^*(x, \pi(x))] - \mathbb{E}_x[\widehat{f}_t(x, \pi(x)]\right| \leq \mathbb{E}_x \left[\frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}} \right] + \frac{K\beta_t}{t}.$$

Explicit strategy: constructing confidence bounds via counterfactual actions. The distribution $\mathcal{D}_{\mathcal{X}}$ is unknown so there are both statistical and computational challenges in the optimization over policies. However, since our proposed policy optimization problem (1.4) is decomposable across contexts, there is an equivalent strategy where no explicit policy optimization is required: at round t, after observing x_t, the agent selects the optimistic *action*

$$a_t \in \arg\max_{a \in \{1, \ldots, K\}} \left\{ \widehat{f}_t(x_t, a) + \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{1}\{a = \widetilde{a}_{t,i}\}} \right\}$$

(ties broken by choosing the action with the smallest index), where $\{\widetilde{a}_{t,i}\}_{i=1}^{t-1}$ is the *counterfactual action trajectory* for context x_t, defined as realizations of all past chosen policies $\{\pi_i\}_{i=1}^{t-1}$ on the context x_t. To recover the counterfactual actions, at round t, the agent runs an inner loop to sequentially generate $\widetilde{a}_{t,1}, \ldots, \widetilde{a}_{t,t-1}$: for $i = 1, \ldots, t-1$,

$$\widetilde{a}_{t,i} \in \underset{a \in \{1,\ldots,K\}}{\arg\max} \left\{ \widehat{f}_i(x_t, a) + \frac{\beta_i}{\sum_{j=1}^{i-1} \mathbb{1}\{a = \widetilde{a}_{t,j}\}} \right\},$$

(ties are broken by choosing the action with the smallest index). Our approach is clearly quite distinct from previous variants of UCB, as we construct confidence bounds by using *simulated counterfactual actions* rather than using the *actual selected actions*.

The UCCB principle leads to provably efficient optimism-based algorithms for general function classes: their regret bounds do not scale with the cardinality of the context spaces, and the required offline least square oracle is feasible for most natural function classes.

1.4 Related literature

We review previous works in the following three areas.

Randomized solutions to general contextual bandits (with finite actions). In the agnostic contextual bandits setting, the minimax regret is $O(\sqrt{KT \log |\bar{\Pi}|})$[1] given a finite policy class $\bar{\Pi} \subset \Pi$. The earliest optimal solution to agnostic contextual bandits is the EXP4 algorithm [5] whose computation is linear in $|\bar{\Pi}|$. There are two optimal oracle-efficient randomized algorithms using the cost-sensitive classification (CSC) oracle: Randomized UCB [13] and ILOVETOCONBANDIT [4].

In this paper, we focus on the realizable contextual bandits setting. Here, the minimax regret of stochastic contextual bandits is $O(\sqrt{KT \log |\mathcal{F}|})$ for a general finite function class $\mathcal{F}$. In [3] the non-efficient algorithm Regressor Elimination was proposed to achieve optimal regret. [16] proposed the use of an online regression oracle and gave an optimal and oracle-efficient algorithm called SquareCB, however the online regression oracle is only computationally efficient for specific function classes.

The open problem of optimal realizable contextual MAB with an offline least square oracle was first solved by [27], with a randomized algorithm called FALCON. One very inspiring aspect of FALCON is that weighted allocation in policy space can be implicitly achieved by weighted allocation over actions under the realizability assumption—this implication was referred to as "bypassing the monster" in [27]. This motivates the investigation in the present paper that considers implicit optimization over policies when designing optimistic algorithms. Unlike the FALCON algorithm, our approach is predicated on computing counterfactual action trajectories.

Variants of UCB for particular contextual bandit problems. Variants of LinUCB are well-known to be regret-optimal and efficient for simple variants of (1.1). However, for general function classes, existing variants of UCB typically have their regret scaling with $|\mathcal{X}|$ [26], except under strong assumptions on the data distribution [15]. UCB has also been used as a subroutines in contextual bandits when the functions in $\mathcal{F}$ admit smoothness or Lipchitz continuity over $\mathcal{X}$ [25, 28]. These works are usually based on discretization of $\mathcal{X}$.

Contextual bandits with infinite actions. There is far less discussion of the infinite-action contextual bandit problem with general function classes. [2] studies how to reduce realizable contextual MAB with infinite actions to an online learning oracle called knows-what-it-knows (KWIK), but this oracle is only known to exist for restricted function classes. [16] studies how to combine general function classes with a linear action model (our illustrative example (4.1) in Section 4). However, their results crucially rely on the restrictive assumption that the action set $\mathcal{A}$ is the unit ball, and they assume access to the online regression oracle which is not computationally efficient in general. Lastly, [22] studies infinite-action contextual bandits

[1] we adopt non-asymptotic big-oh notation: for functions h_1, h_2, $h_1 = O(h_2)$ if there exists constant $C > 0$ such that h_1 is dominated by Ch_2 with high probability (omitting $\log \frac{1}{\delta}$ factors); $h_1 = \tilde{O}(h_2)$ if $h_1 = O(h_2 \max\{1, \text{polylog}(h_2)\})$.

in a quite general agnostic setting. Their formulation and results are quite different from ours, and they do not provide a computationally efficient algorithm.

1.5 Organization

In Section 2 we introduce an optimal and efficient optimistic algorithm in the finite-action setting, and explain the key ideas underlying its principles. For illustrative purpose we assume the function class $\mathcal{F}$ to be finite in Section 2, and present extensions to infinite function classes in Section 3. In Section 4 we introduce a unified framework for contextual bandits with infinite action spaces, and present several interesting examples for which our work gives rise to the first efficient solutions. In Section 5 we propose an optimistic subroutine to generalize randomized algorithms to the infinite-action setting.

2 Upper counterfactual confidence bounds

Following previous works [3, 15, 16, 27], we start by assuming $\mathcal{A} = \{1, \ldots, K\}$, $|\mathcal{F}| < \infty$, and target the "gold standard" in this area—$\mathrm{Regret}(T, \texttt{Alg}) \le \tilde{O}(\sqrt{KT \log |\mathcal{F}|})$, which emphasises the logarithmic scaling in the cardinality $|\mathcal{F}|$. This is mainly for illustrative purposes, and we discuss extensions to infinite function classes in Section 3. A relatively new setting which has essentially not been explored is the infinite-action setting, which we will discuss in Section 4.

2.1 The algorithm

Algorithm 1 Upper Counterfactual Confidence Bounds (`UCCB`)

Input tuning parameters $\{\beta_t\}_{t=1}^{\infty}$.

1: **for** round $t = 1, 2, \ldots, K$ **do**
2: Choose action t.
3: **for** round $t = K+1, K+2, \ldots$ **do**
4: Compute $\widehat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} (f(x_i, a_i) - r_i(x_i, a_i))^2$ via the least square oracle.
5: Observe x_t.
6: **for** $i = K+1, K+2, \ldots, t$ **do**
7: Calculate the counterfactual action $\widetilde{a}_{t,i}$ by

$$\widetilde{a}_{t,i} \in \arg\max_{a \in \mathcal{A}} \left\{ \widehat{f}_i(x_t, a) + \frac{\beta_i}{\sum_{j=K+1}^{i-1} \mathbb{1}\{a = \widetilde{a}_{t,j}\} + 1} \right\}.$$

(ties broken by taking the action with the smallest index)

8: Take $a_t = \widetilde{a}_{t,t}$ and observe reward $r_t(x_t, a_t)$.

We present the algorithm that formalizes the high-level descriptions presented in Section 1.3, where $\{\beta_t\}_{t=1}^{\infty}$ are tuning parameters that depends on the statistical complexity of $\mathcal{F}$. With the choice $\beta_t = \sqrt{17t \log(2|\mathcal{F}|t^3/\delta)/K}$ for finite $\mathcal{F}$, the algorithm is simple and achieves $\tilde{O}(\sqrt{KT \log |\mathcal{F}|})$ regret, which is optimal up to $\log T$ factors. On the computation side, the algorithm executes no more than T^2 maximizations over actions and no more than T calls to the regression oracle.

Theorem 1 (**Regret for Algorithm 1**). *Under Assumption 1 and fixing $\delta \in (0, 1)$, set the parameter β_t in Algorithm 1 to be*

$$\beta_t = \sqrt{17t \log(2|\mathcal{F}|t^3/\delta)/K}.$$

Then with probability at least $1-\delta$, for all $T \geq 1$, the regret of Algorithm 1 after T rounds is upper bounded by

$$\text{Regret}(T, \text{Algorithm 1}) \leq 2\sqrt{17KT\log(2|\mathcal{F}|T^3/\delta)}(\log(T/K)+1) + \sqrt{2T\log(2/\delta)} + K.$$

Remark: Recall that our offline regression step can be solved by first-order algorithms and does not require any computation related to the confidence interval (i.e., maintaining a subset of $\mathcal{F}$ or inverting the Hessian). Therefore, despite having much broader applicability, Algorithm 1 is also simpler than many variants of `UCB` [11, 15, 26] from a computational perspective. The only comparable algorithm to Algorithm 1 is a randomized algorithm—`FALCON` in [27], which requires $O(T)$ maximizations over actions and $O(\log T)$ calls to the offline least square oracle. However, we believe our optimistic solution should be preferable in many practical settings as we do not require randomization and our regret bound exhibits much smaller constants.

2.2 Key ideas underlying `UCCB`

We now explain three key ideas underlying `UCCB`.

Key idea 1: building confidence bounds for policies. Previous literature typically refers to the optimism principle as choosing the optimistic action that has the largest estimate on the current context [1, 15, 26]—optimism is analyzed in the action space. In contrast, we view policies as decisions and build confidence bounds in policy space. The key step in our approach is to characterize the confidence bounds of the function estimate $\hat{f}_t$, which is the output of the least square oracle given the history H_{t-1}.

For an admissible non-randomized contextual bandit algorithm, at each round t there exists a deterministic stationary policy π_t such that the chosen action a_t is equal to $\pi_t(x_t)$ for any realization of x_t. Equivalently, the algorithm selects π_t based on H_{t-1} and chooses the action $a_t = \pi_t(x_t)$ at round t. Through this viewpoint, the following lemma is applicable to all admissible non-randomized contextual bandit algorithms:

Lemma 1 (**confidence of policies**). *Consider an admissible non-randomized contextual bandit algorithm that selects π_t based on H_{t-1} (and chooses the action $a_t = \pi_t(x_t)$) at each round t. Then $\forall\delta \in (0,1)$, with probability at least $1-\delta/2$, for all $t > K$ and all $\pi \in \Pi$, the estimation error on the expected reward of π is bounded by*

$$\left|\mathbb{E}_x[\widehat{f}_t(x,\pi(x))] - \mathbb{E}_x[f^*(x,\pi(x))]\right| \leq \sqrt{\mathbb{E}_x\left[\frac{1}{\sum_{i=1}^{t-1}\mathbb{1}\{\pi(x)=\pi_i(x)\}}\right]}\sqrt{68\log(2|\mathcal{F}|t^3/\delta)} \tag{2.1}$$

The proof of Lemma 1 may be interesting in its own right; a proof ketch will be presented in Section 2.3, and full details are deferred to Appendix A.2.

Key idea 2: the potential function perspective. The idea to establish confidence bounds in policy space is natural when one takes a potential function perspective. From the potential function perspective, the cumulative regret of an optimistic algorithm can be approximately bounded by the sum of confidence bounds at all rounds. Therefore, we would like to establish a uniform upper bound whatever the trajectory of policies is, which usually depends on the "entropy" of the policies. Although the number of policies is "large," the "entropy" of the policies is essentially bounded by $\tilde{O}(K)$ in the following manner.

Lemma 2 (**contextual potential lemma**). *Let π_t be the policy that chooses action t regardless of x for $t = 1, \ldots, K$, and from round $K+1$ up to T, its actions are given by any deterministic stationary policy. Then for all $T > K$,*

$$\sum_{t=K+1}^{T}\mathbb{E}_x\left[\frac{1}{\sum_{j=1}^{t-1}\mathbb{1}\{\pi_t(x)=\pi_j(x)\}}\right] \leq K + K\log(T/K).$$

The above lemma applies to all admissible non-randomized contextual bandit algorithms that choose each action once at the first K rounds, regardless of the order by which they are chosen. Proof of this lemma follows from the observation that for every $x \in \mathcal{X}$, the historical sum of $\mathbb{1}\{\pi_t(x) = \pi_j(x)\}$ will never exceeds a "per-context entropy" $O(K \log T)$. In short, analyzing confidence bounds in policy space helps us take expectation over the "per-context entropy," and successfully avoid the dependence on $|\mathcal{X}|$.

Key idea 3: the relaxation tricks and efficient computation. Following Lemma 1 and Lemma 2, a natural "upper confidence bound" strategy is to choose the policy that maximizes the following (unrelaxed) upper confidence bound:

$$\pi_t \in \arg\max_{\pi \in \Pi_{\mathcal{F}}} \left\{ \mathbb{E}_x[\widehat{f}_t(x, \pi(x)] + \sqrt{\mathbb{E}_x \left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}} \right]} \sqrt{68 \log(2|\mathcal{F}|t^3/\delta)} \right\},$$

where $\Pi_{\mathcal{F}}$ is the policy class defined by $\Pi_{\mathcal{F}} = \{\pi_f : \pi_f(x) \in \arg\max_{a \in \mathcal{A}} f(x, a), \forall x \in \mathcal{X}\}$, which contains π_{f^*}. While we can prove this strategy leads to optimal regret bounds, it is not directly feasible: 1) the distribution $\mathcal{D}_{\mathcal{X}}$ is unknown; and 2) the optimization over policies is computationally intractable. To solve this issue, we introduce two relaxations: we "agnostically" optimize over the full policy space Π rather than $\Pi_{\mathcal{F}}$; and we use a simple inequality to relax the confidence bound proved in Lemma 1, which we call the "square trick".

Lemma 3 (the "square trick" relaxation). *The inequality (2.1) can be further relaxed to*

$$\left|\mathbb{E}_x[\widehat{f}_t(x, \pi(x))] - \mathbb{E}_x[f^*(x, \pi(x))]\right| \leq \mathbb{E}_x \left[\frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}} \right] + \frac{K\beta_t}{t}. \tag{2.2}$$

Proof. Simply relax (2.1) by the Arithmetic Mean-Geometric Mean inequality. □

By performing the two relaxations stated above, we only need to consider the optimization problem

$$\pi_t \in \arg\max_{\pi \in \Pi} \left\{ \mathbb{E}_x[\widehat{f}(x, \pi(x)] + \mathbb{E}_x \left[\frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}} \right] \right\}. \tag{2.3}$$

This is a "per context" optimization problem, where optimality at every context implies optimality of π_t over the full policy space Π. The algorithm does not need to calculate π_t explicitly in every step. Instead, the algorithm observes x_t, and calculates all the counterfactual actions $\pi_1(x_t), \pi_2(x_t), \ldots, \pi_{t-1}(x_t)$ as if the past policies were applied at x_t. Using these counterfactual actions, the algorithm calculates a counterfactual confidence, and chooses an optimistic action a_t that maximize the upper confidence bound stated in (2.3).

The formula to calculate the counterfactual action $\pi_i(x_t)$,

$$\pi_i(x_t) \in \arg\max_{a \in \mathcal{A}} \left\{ \widehat{f}_i(x_t, a) + \frac{\beta_i}{\sum_{j=1}^{i-1} \mathbb{1}\{a = \pi_j(x_t)\}} \right\},$$

requires us to the compute the sequence $\{\widetilde{a}_{t,i}\}_{i=1}^t$ in a recursive manner: for $i = 1, \ldots, t$, compute

$$\widetilde{a}_{t,i} \in \arg\max_{a \in \mathcal{A}} \left\{ \widehat{f}_i(x_t, a) + \frac{\beta_i}{\sum_{j=K+1}^{i-1} \mathbb{1}\{a = \widetilde{a}_{t,j}\} + 1} \right\}.$$

And finally we take $a_t = \pi_t(x_t) = \widetilde{a}_{t,t}$. Therefore, we can explain the explicit steps in Algorithm 1 via the following (obvious) equivalence:

Lemma 4 (**equivalence between Algorithm 1 and implicit strategy** (2.3))**.** *After the first K initialization rounds, Algorithm 1 produce the same pathwise actions as those produced by the policies $\{\pi_t\}_{t>K}$ chosen by the upper-confidence-bound rule (2.3) and a specific tie-breaking rule (i.e., when there are multiple solutions to (2.3), taking π_t to be the unique solution such that for all other solutions π' to (2.3) and all $x \in \mathcal{X}$, the index of the action $\pi_t(x)$ is smaller than the index of the action $\pi'(x)$).*

Based on all the lemmas that we introduce in this subsection, one can prove the $\widetilde{O}(\sqrt{KT \log |\mathcal{F}|})$ regret bound for Algorithm 1 through relatively standard techniques. The full proof is deferred to Appendix A, and a sketch is provided below.

2.3 Proof sketch of Theorem 1 and Lemma 1

In this subsection we present a proof sketch of Theorem 1 (the cumulative regret of Algorithm 1) and Lemma 1 (confidence bounds in policy space, whose relaxation leads to Lemma 3).

Proof sketch of Theorem 1. From Lemma 4, we know Algorithm 1 implicitly chooses the optimistic policy π_t (i.e., solution of (2.3)) at each round t. We prove the regret bound on the event where the inequality (2.2) holds true for all $\pi \in \Pi$. From Lemma 3, the measure of this event is at least $1 - \frac{\delta}{2}$.

Optimism of Algorithm 1 in policy space suggests that for all $t > K$,

$$
\begin{aligned}
\mathbb{E}_x[f^*(x, \pi_{f^*}(x)] &\leq \mathbb{E}_x[\widehat{f}_t(x, \pi_{f^*}(x))] + \mathbb{E}_x\Big[\frac{\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi_{f^*}(x) = \pi_i(x)\}}\Big] + \frac{K\beta_t}{t} \\
&\leq \arg\max_{\pi \in \Pi} \left\{ \mathbb{E}_x[\widehat{f}_t(x, \pi(x))] + \mathbb{E}_x\Big[\frac{\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi(x) = \pi_i(x)\}}\Big] \right\} + \frac{K\beta_t}{t} \\
&= \mathbb{E}_x[\widehat{f}_t(x, \pi_t(x))] + \mathbb{E}_x\Big[\frac{\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi_t(x) = \pi_i(x)\}}\Big] + \frac{K\beta_t}{t} \\
&\leq \mathbb{E}_x[f^*(x, \pi_t(x))] + \mathbb{E}_x\Big[\frac{2\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi_t(x) = \pi_i(x)\}}\Big] + \frac{2K\beta_t}{t},
\end{aligned}
$$

where the first and the last inequality are due to Lemma 3; and the second inequality due to maximization over policies. Therefore, the expected regret incurred at round t is bounded by

$$
\mathbb{E}_x[f^*(x, \pi_{f^*}(x)] - \mathbb{E}_x[f^*(x, \pi_t(x)] \leq \mathbb{E}_x\left[\frac{2\beta_t}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi_t(x) = \pi_i(x)\}}\right] + \frac{2K\beta_t}{t}. \tag{2.4}
$$

Taking the telescoping sum of (2.4) and applying the contextual potential lemma (Lemma 2), we can prove

$$
\sum_{t=1}^{T} \mathbb{E}_x[f^*(x, \pi_{f^*}(x) - f^*(x, \pi_t(x))] \leq 2\sqrt{17KT \log(2|\mathcal{F}|T^3/\delta)}(\log(T/K) + 1) + K. \tag{2.5}
$$

By Azuma's inequality and Lemma 4, with probability at least $1 - \delta/2$, we can bound the regret by

$$
\text{Regret}(T, \text{Algorithm 1}) \leq \mathbb{E}_x[f^*(x, \pi_{f^*}(x) - f^*(x, \pi_t(x))] + \sqrt{2T \log(2/\delta)}. \tag{2.6}
$$

Finally we combine (2.5) and (2.6) by a union bound to finish the proof.

Proof sketch of Lemma 1. The proof of Lemma 1 includes three key steps: characterization of the estimation error (inequality (2.7)); a counting argument (inequality (2.8)); and applying Cauchy-Schwartz inequality to (2.8). Now we describe these key steps.

The following lemma, which holds for arbitrary algorithms, characterizes the estimation errors of an arbitrary sequence of estimators.

Lemma 5 (uniform convergence over all sequences of estimators)**.** *For an arbitrary contextual bandit algorithm, $\forall \delta \in (0,1)$, with probability at least $1-\delta/2$,*

$$\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}\left[(f_t(x_i,a_i)-f^*(x_i,a_i))^2|H_{i-1}\right] \le 68\log(2|\mathcal{F}|t^3/\delta)$$
$$+2\sum_{i=1}^{t-1}(f_t(x_i,a_i)-r_i(x_i,a_i))^2-(f^*(x_i,a_i)-r_i(x_i,a_i))^2, \tag{2.7}$$

uniformly over all $t \ge 2$ and all fixed sequence $f_2, f_3, \cdots \in \mathcal{F}$.

Proof of Lemma (2.7) can be found in Appendix A.2.

Consider the contextual bandit algorithm that choose π_t based on H_{t-1} at each round t, the left hand side of (2.7) is equal to $\sum_{i=1}^{t-1}\mathbb{E}_x\left[(f(x,\pi_i(x))-f^*(x,\pi_i(x)))^2\right]$. Then by using the fact that $\forall\pi\in\Pi$, for all $x\in\mathcal{X}$,

$$\mathbb{1}\{\pi(x)=\pi_i(x)\}(f_t(x,\pi(x))-f^*(x_i,\pi(x)))^2 \le (f_t(x,\pi_i(x))-f^*(x,\pi_i(x)))^2,$$

we obtain the key inequality

$$\mathbb{E}_x\Big[\sum_{i=1}^{t-1}\mathbb{1}\{\pi(x)=\pi_i(x)\}(f_t(x,\pi(x))-f^*(x_i,\pi(x)))^2\Big] \le 68\log(2|\mathcal{F}|t^3/\delta)$$
$$+2\sum_{i=1}^{t-1}(f_t(x_i,a_i)-r_i(x_i,a_i))^2-(f^*(x_i,a_i)-r_i(x_i,a_i))^2. \tag{2.8}$$

We then apply Cauchy-Schwartz inequality to lower bound the left hand side of (2.8), and take $f_t=\widehat{f}_t$ be the least square solutions to upper bound the right hand side of (2.8).

3 Generalization to infinite $\mathcal{F}$

Extensions of our theory to "infinite" $\mathcal{F}$ with statistical complexity notions of covering number and parametric dimension are straightforward. Technically speaking, we only require some standard uniform convergence arguments to modify Lemma 5. We will first show that our results trivially generalizes to parametric $\mathcal{F}$ with suitable continuity, and then extend our results to general function classes following some more careful covering arguments.

Parametric dimension. Assume $\mathcal{F}$ is parametrized by a compact set $\Theta\subset\mathbb{R}^d$ whose diameter is bounded by Δ, and satisfies

$$|f_{\theta_1}(x,a)-f_{\theta_2}(x,a)| \le L\|\theta_1-\theta_2\|, \tag{3.1}$$

uniformly over $x\in\mathcal{X}$ and $a\in\mathcal{A}$. This case clearly covers many previous structured models (variants of the "linear payoff" formulation (1.1)).

Corollary 2 (extension to infinite $\mathcal{F}$ via parametric dimension). *Under Assumption 1 and the assumption (3.1) and fixing $\delta\in(0,1)$, set the parameter β_t in Algorithm 1 to be*

$$\beta_t=\sqrt{34t/K}\sqrt{d\log(2+\Delta Lt)+\log(2t^3/\delta)+1}.$$

Then Algorithm 1 satisfies that with probability at least $1-\delta$, for all $T\ge1$,

$$\text{Regret}(T,\text{Algorithm 1}) \le 2K\beta_T(\log(T/K)+1)+\sqrt{2T\log(2/\delta)}+K=\tilde{O}(\sqrt{KTd}).$$

Remark: While this regret bound has a worse dependence on K in the "linear payoff" formulation (1.1) compared with `SupLinUCB` in [11] (whose regret is logarithmic in K), Algorithm 1 can be applied in more general parametric settings and enjoys much lower computational demands (there is no need to invert any Hessian). While the square-root dependence on K can not be improved for general $\mathcal{F}$ (see the lower bound in [3]), we can improve this dependence for structured models by applying our results in Section 4.

Covering number formulation. Our results can be extended to general (possibly non-parametric) function classes via covering numbers and standard uniform convergence techniques. We consider formulation (3.2)—a major target of previous works on general contextual bandits [15, 16, 21]. We assume access to a general function class $\mathcal{G}$ that contains mappings from $\mathcal{X}$ to $[0, 1]$, and assume

$$\mathcal{F} = \{f : f(x, a) = g_a(x), \quad g_a \in \mathcal{G}\}. \tag{3.2}$$

Definition 1 (covering number). For a function class $\mathcal{G}$ that contains mappings from $\mathcal{X}$ to $[0, 1]$ and fixed $n \in \mathbb{Z}_+$, an empirical L_1 cover on a sequence $x_1, \dots, x_n$ at scale ε is a set $U \subseteq \mathbb{R}^n$ such that $\forall g \in \mathcal{G}, \exists u \in U, \frac{1}{n}\sum_{i=1}^n |g(x_n) - u_n| \le \varepsilon$. We define the covering number $\mathcal{N}_1(\mathcal{G}, \varepsilon, \{x_i\}_{i=1}^n)$ to be the size of the smallest such cover.

Given careful covering arguments proved in [15, 21], the following extension is straightforward:

Corollary 3 (extension to infinite $\mathcal{F}$ via covering number). *Under Assumption 1 and the assumption (3.2), given $T \ge 1$ and $\delta \in (0, 1)$, by setting all the parameters β_t in Algorithm 1 to be a fixed value*

$$\beta = \sqrt{TK} \cdot \inf_{\varepsilon > 0} \left\{ 25\varepsilon T + 80 \log \left(\frac{8KT^3 \mathbb{E}_{\{x_i\}_{i=1}^T} \mathcal{N}_1(\mathcal{G}, \varepsilon, \{x_i\}_{i=1}^T)}{\delta} \right) \right\}.$$

Then, Algorithm 1 satisfies that with probability at least $1 - \delta$,

$$\text{Regret}(T, \text{Algorithm 1}) \le 2K\beta(\log(T/K) + 1) + \sqrt{2T \log(2/\delta)} + K.$$

4 A unified framework for infinite action spaces

In this section we study infinite-action contextual bandits to illustrate the simplicity and applicability of the `UCCB` principle. In context-free settings, discussion on infinite actions can be sorted into two streams. The first stream studies variants of the linear action model. Prominent examples include linearly parametrized bandit [1, 12], and parametrized bandit with generalized linear model [14]. The second stream is based on discretization over actions and reduction to the finite-action setting (e.g., Lipchitz bandit ([20]). We focus on the first stream here, as it exhibits additional challenges of efficient exploration beyond the finite-action setting.

To focus on the core messages, we assume $\mathcal{F}$ to be finite and function in $\mathcal{F}$ take values in $[0, 1]$. We propose a generic algorithm (Algorithm 2) that achieves

$$\text{Regret}(T, \text{Algorithm 2}) \le \tilde{O}(\sqrt{\mathcal{E} \log |\mathcal{F}| T}),$$

for many models of interest. Here we call $\mathcal{E} := \mathbb{E}_x[\mathcal{E}_x]$ the "average decision entropy," where $\mathcal{E}_x$ is (informally) the complexity of the "fixed-x-model" where the context is fixed to be x. Note that unlike previous complexity measures such as "Eluder dimension" [26], the "average decision entropy" $\mathcal{E}$ does not scale with $|\mathcal{X}|$ so that this complexity measure is much more useful in the contextual settings. We will present several interesting illustrative examples, and present key ideas of our algorithm using these examples.

4.1 Illustrative models

In the context-free infinite-action bandits literature, it is well-known that $\tilde{O}(\sqrt{T})-$type regret is only possible for structured models, among which variants of linear bandits are the preponderant models. As a result, our framework mainly targets settings where all "fixed-x-model" are variants of linear bandits.

Example 1 (contextual bandit with linear action model). Given a general vector-valued function class $\mathcal{G}$ that contains mappings from $\mathcal{X}$ to $\mathbb{R}^d$, let

$$\mathcal{F} = \{f : \exists g \in \mathcal{G} \text{ s.t. } f(x, a) = g(x)^\top a, \forall x \in \mathcal{X}, \forall a \in \mathcal{A}\}. \tag{4.1}$$

We assume $\mathcal{A} \subset \mathbb{R}^d$ is an arbitrary compact set, and is available for the agent at all rounds. This formulation is a strict generalization of the finite-action realizable contextual bandit problem we studied in previous sections (it reduces to the $K-$armed setting when $\mathcal{A}$ is the set of K element vectors in $\mathbb{R}^K$). Another special case where $\mathcal{A}$ is restricted to be the unit ball is studied in Foster and Rakhlin [16], but a general solution to arbitrary compact action set is still open. Moreover, Foster and Rakhlin [16] requires online regression oracles which are not computationally efficient in general. Formulation (4.1) was also studied in [10] but the goal there was off-policy evaluation rather than regret minimization.

With knowledge on linear bandits we can prove $\mathcal{E}_x = d$ for all $x \in \mathcal{X}$. (detailed explanation is deferred to Section 4.4.1). Therefore $\mathcal{E} = d$, which is independent of the number of actions, and the order of regret is expected to be $\tilde{O}(\sqrt{d \log |\mathcal{F}| T})$.

Example 2 (contextual bandit with generalized linear action model.). Consider a broader choice of models, which contains generalized linear action models and allows a mapping φ:

$$\mathcal{F} = \left\{ f : \exists g \in \mathcal{G} \text{ s.t. } f(x,a) = \sigma_x \left(g(x)^\top \varphi(x,a) \right), \forall x \in \mathcal{X}, \forall a \in \mathcal{A} \right\}, \tag{4.2}$$

where for every $x \in \mathcal{X}$, $\sigma_x : \mathbb{R} \to [0,1]$ is a known link function that satisfies

$$\frac{\sup_a \sigma_x'(\langle g^*(x), \varphi(x,a) \rangle)}{\inf_a \sigma_x'(\langle g^*(x), \varphi(x,a) \rangle)} \leq \kappa_x;$$

and $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a known compactness-preserving mapping (e.g., continuous mappings). This model generalizes (4.1) and allows more flexibility. When we set $\varphi(x,a) = x_a$, we see that this model is significantly broader in scope than the simple "linear payoff" formulation (1.1), as $g(x)$ is a general function that depends on x rather than a fixed parameter θ.

Our analysis will show that $\mathcal{E}_x = \kappa_x^2 d$ for all $x \in \mathcal{X}$ (detailed explanation is deferred to Section 4.4.2), so that $\mathcal{E} = \mathbb{E}_x[\kappa_x^2] d$, and the order of regret is expected to be $\tilde{O}(\sqrt{E_x[\kappa_x^2] d \log |\mathcal{F}| T})$.

Example 3 (heterogeneous action set). Many real-world, customized pricing and personalized healthcare applications have a high dimensional action set $\mathcal{A}$, but the "effective dimension" of available actions after observing x is usually much smaller. To model these applications, consider the reward model

$$\mathcal{F} = \left\{ f : \exists g \in \mathcal{G} \text{ s.t. } f(x,a) = \sigma_x(g(x)^\top a), \forall x \in \mathcal{X}, \forall a \in \mathcal{A}(x) \right\}, \tag{4.3}$$

where for all $x \in \mathcal{X}$ we assume a compact action set $\mathcal{A}(x) \subset \mathcal{A}$, and assume $\mathcal{A}(x)$ is contained in a d_x-dimensional subspace. When the agent observes context x, she can only choose her action from $\mathcal{A}(x)$.

For this model we have $\mathcal{E}_x = \kappa_x^2 d_x$ (detailed explanation is deferred to Section 4.4.3) so that $\mathcal{E} = \mathbb{E}_x[\kappa_x^2 d_x]$. The salient point here is the we avoid dependence on the full dimension d. Regret therefore scales as $\tilde{O}(\sqrt{\mathbb{E}_x[\kappa_x^2 d_x] \log |\mathcal{F}| T})$.

4.2 Counterfactual action divergence

The main modification required for infinite-action settings is predicated on a central concept called "counterfactual action divergence," which generalizes the term $(\sum_{i=1}^n \mathbb{1}\{a = a_i\})^{-1}$ that was used in Algorithm 1. This new concept characterizes "how much information" is learned from action a given a sequence $\{a_i\}_{i=1}^n$, on the "fixed-x-model."

Definition 2 (counterfactual action divergence). For fixed integer n, a context x, an action a and a sequence of actions $\{a_i\}_{i=1}^n$, we say $V_x(a||\{a_i\}_{i=1}^n)$ is a proper choice of the counterfactual action divergence between a and $\{a_i\}_{i=1}^n$ evaluated at x, if

$$V_x(a||\{a_i\}_{i=1}^n) \geq \sup_{f \in \mathcal{F}} \left\{ \frac{|f(x,a) - f^*(x,a)|^2}{\sum_{i=1}^n (f(x,a_i) - f^*(x,a_i))^2} \right\}.$$

We define $V_x(a||\emptyset) = \infty$ in the case $n = 1$.

Using the definition of counterfactual action divergence, the expectation

$$\mathbb{E}_x[V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})], \tag{4.4}$$

can be used to construct an upper confidence bound on the expected reward of policy π given the past chosen policies $\{\pi\}_{i=1}^{t-1}$. Similar to the finite-action setting, the agent chooses the optimistic policy π_t that maximizes this confidence bound, and chooses $a_t = \pi_t(x_t)$ without explicitly computing π_t—this is achieved by sequentially recovering counterfactual actions, as will be illustrated in our proposed Algorithm.

Convenient choices of $V_x(a||\{a_i\}_{i=1}^n\}_{i=1}^n)$ should be taken case by case for different problems. In the following lemma, we present closed-form choices of $V_x(a||\{a_i\}_{i=1}^n\}_{i=1}^n)$ in all our illustrative examples.

Statement 1 (illustration of counterfactual action divergences). In the illustrative examples, the counterfactual action divergences are given as follows (and taken as ∞ when inverse of matrices is not well-defined):

- finite-action contextual bandit:

$$V_x(a||\{a_i(x)\}_{i=1}^n) = \frac{1}{\sum_{i=1}^n \mathbb{1}\{a = a_i\}}.$$

- linear action model (4.1):

$$V_x(a||\{a_i\}_{i=1}^n\}_{i=1}^n) = a^\top(\sum_{i=1}^n [a_i a_i^\top])^{-1} a. \tag{4.5}$$

- generalized linear action model (4.2):

$$V_x(a||\{a_i\}_{i=1}^n) = \kappa_x^2 \varphi(x,a)^\top(\sum_{i=1}^n [\varphi(x,a_i)\varphi(x,a_i)^\top])^{-1}\varphi(x,a).$$

- generalized linear action model with heterogeneous action sets (4.3):

$$V_x(a||\{a_i(x)\}_{i=1}^n) = \kappa_x^2 b_{x,a}\top(\sum_{i=1}^n [b_{x,a_i} b_{x,a_i}^\top])^{-1} b_{x,a},$$

  where $b_{x,a}$ is the coefficient vector of a with a basis $\{A_{x,1}, \ldots, A_{x,d_x}\}$ of $\mathcal{A}(x)$, i.e.,

$$a = [A_{x,1}, \ldots, A_{x,d_x}] b_{x,a}.$$

4.3 The algorithm and regret bound

Algorithm 2 is a generalization of Algorithm 1 to the infinite-action setting. It can be applied to most parametric action models that have been studied in the context-free setting, and handles heterogeneous action sets. Recall that $\mathcal{E} := \mathbb{E}[\mathcal{E}_x]$ is the average decision entropy of the problem, for which we will give the formal definition later. The "initialization oracle" and the "action maximization oracle" will also be explained shortly.

Algorithm 2 essentically provide a reduction from contextual models to the "fixed-x-models." The regret of an optimistic algorithm is usually upper bounded by the sum of confidence bounds. In our case, the sum of expectations (4.4) is decomposable over contexts, so tractability of the "fixed-x-models" suffices to make Algorithm 2 provably efficient. Formally, we require regularity conditions so that the "fixed-x-models" are solvable by the optimism principle. Motivated by the standard potential arguments used in the linear bandit literature, we make Assumption 2 below. Verification of this assumption on Examples 1-3 will be presented in the next section.

Algorithm 2 Upper Counterfactual Confidence Bound-Infinite Action (`UCCB-IA`)

Input tuning parameters $\{\beta_t\}_{t=1}^{\infty}$.

1: **for** round $t = 1, 2, \dots$ **do**

2: Compute $\widehat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{t-1} (f(x_t, a_t) - r_t(x_i, a_t))^2$ via the least square oracle.

3: Observe x_t, use the initialization oracle to obtain initializations $\{A_{x_t,i}\}_{i=1}^{d_x}$.

4: **for** $i = 1, 2, \dots, t \vee d_x$ **do**

5: Take $\widetilde{a}_{t,i} = A_{x_t,i}$.

6: **for** $i = t \wedge (d_x + 1), \dots, t$ **do**

7: Use the action maximization oracle to compute counterfactual actions:

$$\widetilde{a}_{t,i} \in \arg\max_{a \in \mathcal{A}(x_t)} \left\{ \widehat{f}_i(x_t, a) + \beta_i V_{x_t}(a || \{\widetilde{a}_{t,j}\}_{j=1}^{i-1}) \right\}.$$

8: Take $a_t = \widetilde{a}_{t,t}$ and observe reward $r_t(x_t, a_t)$.

Assumption 2 (**per-context models are solvable by optimism**). *There exists counterfactual action divergences such that the following are satisfied:*

i)for all $x \in \mathcal{X}$, there exists d_x actions $A_{x,1}, \dots, A_{x,d_x} \in \mathcal{A}(x)$ such that $V_x(a||\{A_{x,i}\}_{i=1}^{d_x}) < \infty$ for all $a \in \mathcal{A}(x)$.

ii) For all $x \in \mathcal{X}$, there exists $\mathcal{E}_x > 0$ such that for all $T \geq 1$ and all sequences $\{a_t\}_{t=1}^{T}$ that satisfy $\{a_t\}_{t=1}^{d_x \wedge T} = \{A_{x,t}\}_{t=1}^{d_x \wedge T}$, we have

$$\sum_{t=1}^{T} \left[1 \wedge V_x(a_t || \{a_j\}_{j=1}^{t-1})\right] \leq \mathcal{E}_x \mathrm{poly}(\log T)$$

for all $x \in \mathcal{X}$, where $\mathrm{poly}(\cdot)$ *is a fixed polynomial-scale function.*

Given positive values $\mathcal{E}_x$ that satisfies condition ii) in Assumption 2, we define $\mathcal{E}_x := \mathbb{E}_x[\mathcal{E}_x]$ to be (a proper choice of) the "average decision entropy" of the problem. The "average decision entropy" of a problem is not unique, and any "proper" choice of $\mathcal{E}$ leads to a rigorous regret bound of Algorithm 2.

Besides the least-square oracle, Algorithm 2 uses two other optimization oracles that are necessary in the infinite-action setting: 1) a deterministic initialization oracle which returns $\{A_{x,i}\}_{i=1}^{d_x}$ satisfying Assumption 2 after inputting $\mathcal{A}(x)$ (this is standard for Examples 1-3 using the theory of barycentric spanners, see the next subsection); and 2) a deterministic action maximization oracle whose output is a maximizer of a function over the feasible region $\mathcal{A}(x)$.

After imposing the regularity conditions proposed in Assumption 2, the regret of Algorithm 2 can be bounded as the follows.

Theorem 4 (**Regret of Algorithm 2**). *Under Assumptions 1 and 2 and fixing $\delta \in (0, 1)$, let*

$$\beta_t = \sqrt{17t \log(2|\mathcal{F}|t^3/\delta)/\mathcal{E}}.$$

Then with probability at least $1 - \delta$, for all $T \geq 1$ the regret of Algorithm 2 after T rounds is upper bounded by

$$\mathrm{Regret}(T, \text{Algorithm 2}) \leq 2\sqrt{17\mathcal{E}T \log(2|\mathcal{F}|T^3/\delta)}(\mathrm{poly}(\log T) + 1) + \sqrt{2T \log(2/\delta)} + \mathcal{E}.$$

This theorem immediately provides regret bounds for all our illustrative examples, which we will discuss in the next subsection.

Finally, we give a high-level interpretation of the average decision entropy $\mathcal{E}$: if the expectation (4.4) is the "discrete" partial gradient of a potential function, then the historical sum has the path independence property—that is, the historical sum of (4.4) can be bounded by the maximum value of a potential function, which is characterized by the average decision entropy $\mathcal{E}$. Since $\mathcal{E}$ is the average rather than the sum of the effective complexities of all "fixed-x-models," `UCCB` provides a generic solution to achieve optimal regret bounds that do not scale with $|\mathcal{X}|$.

4.4 Applications in illustrative examples

In this subsection we will carefully go through the three illustrative examples. We summarize the conclusions in the following corollary:

Corollary 5 (Theorem 4 applied to illustrative examples). *Examples 1-3 satisfy Assumptions 2 with the average decision entropy given by*

- *linear action model (4.1):* $\mathcal{E} = d$.
- *generalized linear action model (4.2):* $\mathcal{E} = \mathbb{E}_x[\kappa_x^2]d$.
- *generalized linear action model with heterogeneous action sets (4.3):* $\mathcal{E} = \mathbb{E}_x[\kappa_x^2 d_x]$.

Now we give a verification in the remaining parts of this subsection.

4.4.1 Contextual bandits with linear action model (Example 1).

We begin with contextual bandits with linear action model (4.1), with the homogeneous action set $\mathcal{A}$. For this problem, Algorithm 1 only needs to compute the initialization actions $A_1, \dots, A_d$ once, and use them during the first d rounds. This suffices to complete the required initialization for all contexts.

Based on well-known results in the linear bandit literature, it is straightforward to show that $\mathcal{E} = d$, because we can take $\mathcal{E}_x = d$ for every per-context model. The details are as follows.

As shown in Statement 1, for all $x \in \mathcal{X}$, we choose the counterfactual action divergence between any a_t and any sequence $\{a_i\}_{i=1}^{t-1}$ evaluated at x to be

$$V_x(a_t || \{a_i\}_{i=1}^{t-1}) = a_t^\top (\sum_{i=1}^{t-1} [a_i a_i^\top])^{-1} a_t.$$

Following the standard approach in the linear bandit literature (e.g., see [12]), we choose the d initialization actions $\{A_i\}_{i=1}^d$ to be the barycentric spanner of $\mathcal{A}$. A barycentric spanner is a set of d vectors, all contained in $\mathcal{A}$, such that every vector in $\mathcal{A}$ can be expressed as a linear combination of the spanner with coefficients in $[-1, 1]$. An efficient algorithm to find the barycentric spanner for an arbitrary compact set is given in [6].

The following result follows [12, Lemma 9][2], which is often referred to as the "elliptical potential lemma": let $a_i = A_i$ for $i = 1, \dots, d$, then for all $T > d$ and all trajectory $\{a_t\}_{t=d+1}^T$,

$$\sum_{t=d+1}^{T} \left[1 \wedge a_t^\top (\sum_{i=1}^{t-1} [a_i a_i^\top])^{-1} a_t \right] \le 2d \log T.$$

Therefore, we obtain for all $T \ge 1$ and all $x \in \mathcal{X}$,

$$\sum_{t=1}^{T} \left[1 \wedge V_x(a_t || \{a_i\}_{i=1}^{t-1}) \right] \le 2d \log T + d \le 3d \log T. \tag{4.6}$$

By taking $\{A_i\}_{i=1}^d$ to be a barycentric spanner of $\mathcal{A}$, setting $\mathcal{E} = d$, and taking $\mathrm{poly}(\log T) = 3 \log T$, Assumption 2 holds for problem (4.1). Despite the illustration here, we also note that our Assumption 2 is not restricted to any particular choice of initialization actions and $\mathcal{E}$: there are other ways to choose linearly independent initialization actions, giving rise to a slightly different $\mathrm{poly}(\log T)$ term in Assumption 2 (see, e.g. [1, Lemma 11]).

[2] Lemma 9 in [12] holds for an arbitrary compact set $\mathcal{A} \subset \mathbb{R}^d$, as changing the coordinate system is without the loss of generality for this lemma.

4.4.2 Contextual bandits with generalized linear action model (Example 2).

For the problem formulation (4.2), we can take $\mathbb{E}[\mathcal{E}_x] = \mathbb{E}[\kappa_x^2]d$. The details are as follows.

As shown in Statement 1, given $x \in \mathcal{X}$, we choose the counterfactual action divergence between any a_t and any sequence $\{a_i\}_{i=1}^{t-1}$ evaluated at x to be

$$V_x(a_t||\{a_i\}_{i=1}^{t-1}) = \kappa_x^2 \varphi(x, a_t)^\top (\sum_{i=1}^{t-1}[\varphi(x, a_i)\varphi(x, a_i)^\top])^{-1}\varphi(x, a_t).$$

Given $x \in \mathcal{X}$, we take $\{A_{x,i}\}_{i=1}^{d}$ such that $\{\varphi(x, A_{x,i})\}_{i=1}^{d}$ consists of a barycentric spanner of $\{\varphi(x, a) : a \in \mathcal{A}\}$[3]. Note that a different basis $\{A_{x,i}\}_{i=1}^{d}$ should be computed for each x. From our previous result (4.6) and the fact $\kappa_x \geq 1$, for all $T \geq 1$ and all sequences $\{a_i\}_{i=1}^{T}$ that satisfy $\{a_i\}_{i=1}^{d_x \wedge T} = \{A_{x,i}\}_{i=1}^{d_x \wedge T}$,

$$\sum_{t=1}^{T} \left[1 \wedge V_x(a_t||\{a_i\}_{i=1}^{t-1})\right] = \sum_{t=1}^{T} \left[1 \wedge \kappa_x^2 \varphi(x, a)^\top (\sum_{i=1}^{t-1}[\varphi(x, a_i)\varphi(x, a_i)^\top])^{-1}\varphi(x, a)\right] \leq \kappa_x^2 3d \log T.$$

By taking $\mathcal{E}_x = \kappa_x^2 d$, and $\text{poly}(\log T) = 3 \log T$, Assumption 2 holds with $\mathcal{E} = \mathbb{E}_x[\kappa_x^2]d$.

4.4.3 Contextual bandits with heterogeneous action set (Example 3)

We consider the problem formulation (4.3) where the action set $\mathcal{A}(x)$ is heterogeneous for different $x \in \mathcal{X}$. Note that $\mathcal{A}(x)$ is a compact set contained in a d_x-dimensional subspace. Given $x \in \mathcal{X}$, we choose $\{A_{x,i}\}_{i=1}^{d_x}$ as the barycentric spanner of $\mathcal{A}(x)$ and take $a_i = A_{x,i}$ for $i = 1, \ldots, d_x$. As stated in Statement 1, given $x \in \mathcal{X}$, the counterfactual action divergence between a_t and $\{a_i\}_{i=1}^{t-1}$ evaluated at x is

$$V_x(a_t||\{a_i\}_{i=1}^{t-1})) = \kappa_x^2 b_{x,a_t}^\top (\sum_{i=1}^{t-1} b_{x,a_t} b_{x,a_t}^\top)^{-1} b_{x,a_t},$$

where b_{x,a_t} is the coefficient vector of a_t with respect to the basis $\{A_{x,i}\}_{i=1}^{d_x}$. From our previous result (4.6) and the fact $\kappa_x \geq 1$, for all $T \geq 1$ and all sequences $\{a_i\}_{i=1}^{T}$ that satisfy $\{a_i\}_{i=1}^{d_x \wedge T} = \{A_{x,i}\}_{i=1}^{d_x \wedge T}$,

$$\sum_{t=1}^{T} 1 \wedge \left[V_x(a_t||\{a_i\}_{i=1}^{t-1})\right] = \sum_{t=1}^{T} \left[1 \wedge \kappa_x^2 b_{x,a_t}^\top (\sum_{i=1}^{t-1}[b_{x,a_i} b_{x,a_i}^\top])^{-1} b_{x,a_t}\right] \leq \kappa_x^2 3 d_x \log T.$$

By taking $\mathcal{E}_x = \kappa_x^2 d_x$, and $\text{poly}(\log T) = 3 \log T$, we verify Assumption 2 with $\mathcal{E} = \mathbb{E}[\kappa_x^2]d_x$. We note that under the heterogeneous formulation, Algorithm 2 needs to compute a different basis for each $\mathcal{A}(x)$, and the computation of counterfactual action divergence also requires a coefficient decomposition for each $x \in \mathcal{X}$.

One significant advantage of Algorithm 2 is that the regret does not rely on the full dimension d—this means that we can increase feature context as long as we can control the average decision entropy $\mathbb{E}[\kappa_x^2 d_x]$.

5 Using "optimistic subroutines" to generalize randomized algorithms

What is the connection between our proposed optimistic algorithms and existing randomized algorithms? In this section, we show that by combining the idea of counterfactual confidence bounds and a non-trivial "optimistic subroutine," we can also generalize an existing randomized algorithm to the infinite-action setting. However, the analysis and implementation of the resulting randomized algorithm is much more complex than

[3] in formulation (4.2) we have asked φ to preserve compactness with respect to a (e.g. the continuous ones), so such barycentric spanner must exists.

the optimistic algorithm we introduced before. Through this extension, we see the simplicity and importance of the optimism principle for complex settings like infinite-action spaces.

The first least-square-oracle-efficient randomized algorithm in the general realizable contextual bandits, `FALCON` from [27], is restricted to the finite-action setting. `FALCON` performs implicit optimization in policy space, but the allocation of policies reduces to a closed-form weighted allocation rule for actions (this design principle also influences the design of `UCCB`). We find that it becomes more crucial to exploit the counterfactual confidence bounds in the infinite-action setting: the optimization of weighted allocation rules no longer has closed-form solutions, and we need to design a technical "optimistic subroutine" to find feasible weighted allocations.

As the required subroutine is a bit complex, we focus on the linear action model (4.1) stated in Example 1 for simplicity. Extensions to more complex models follow similar ideas, and the structure of the proposed algorithm remain mostly unchanged. We assume a deterministic initialization oracle that outputs a barycentric spanner of the compact set $\mathcal{A}$ (e.g. the algorithm in [6]), and an action maximization oracle that outputs the maximizer of the input function over $\mathcal{A}$. The following algorithm extends `FALCON` to the linear action model (4.1), where the step 6 is a novel optimization problem to find the "right" weighted allocation over actions. Here the "$a^\top(\mathbb{E}_{\widetilde{a}\sim p_t}[\widetilde{a}\widetilde{a}^\top])^{-1}a$" term in (5.2) is a continuous analogue to the counterfactual action divergence (4.5).

Algorithm 3 a generalized version of `FALCON` for linear action model (4.1)

Input epoch schedule $\{\tau_m\}_{m=1}^\infty$, $\tau_0 = 0$, tuning parameters $\{\beta_m\}_{m=1}^\infty$, an arbitrary function $\widehat{f}_1 \in \mathcal{F}$.

1: **for** epoch $m = 1, 2, \ldots$ **do**
2: Compute $\widehat{f}_m = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^{\tau_m - 1}(f(x_t, a_t) - r_t(x_t, a_t))^2$ via the least square oracle when $m \geq 1$.
3: **for** round $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**
4: Observe context x_t.
5: Use the action maximization oracle to compute $\widehat{a}_t \in \max_{a\in\mathcal{A}} \widehat{f}_m(x_t, a)$.
6: Run the algorithm `OptimisticSubroutine`$(\mathcal{A}, \widehat{a}_t, \widehat{f}_m(x_t, \cdot), \beta_m)$ to find a distribution p_t over $\mathcal{A}$ such that,

$$\mathbb{E}_{a\sim p_t}[\widehat{f}_m(x_t, \widehat{a}_t) - \widehat{f}_m(x_t, a)] \leq 2\beta_m d, \tag{5.1}$$

$$\forall a \in \mathcal{A}, \quad \widehat{f}_m(x_t, a) + \beta_m a^\top(\mathbb{E}_{\widetilde{a}\sim p_t}[\widetilde{a}\widetilde{a}^\top])^{-1}a \leq \widehat{f}_m(x_t, \widehat{a}_t) + 2\beta_m d. \tag{5.2}$$

7: Sample $a_t \sim p_t$ and observe reward $r_t(a_t)$.

Algorithm 3 runs in an epoch schedule and only calls the least square oracle at the pre-specified rounds $\tau_1, \tau_2, \ldots$. We take $\tau_m = 2^m$ for all $m \geq 1$ to simplify the statement of the theorem, though other choices of the epoch schedule are also possible [27].

Theorem 6 (Regret of Algorithm 3). *Consider the problem formulation (4.1) stated in Example 1, under Assumption 1. Take the epoch schedule $\tau_m = 2^m$ for $m \geq 1$. Let*

$$\beta_m = 30\sqrt{\log(|\mathcal{F}|\tau_{m-1}/\delta)/(2d\tau_{m-1})}$$

for $m = 2, \ldots,$ and $\beta_1 = 1$. Then with probability at least $1-\delta$, for all $T \geq 1$, the regret of Algorithm 3 after T rounds is upper bounded by

$$\text{Regret}(T, \text{Algorithm 3}) \leq 608.5\sqrt{2dT\log(|\mathcal{F}|T/\delta)} + 2\sqrt{2T\log(2/\delta)} + 2.$$

Theorem 6 can be obtained by modifying the regret analysis of the original `FALCON` algorithm. (We refer the readers to [27] for the background and intuition of the original `FALCON` algorithm, especially the "Observation 2" in that paper.) However, the key challenge is to provide an efficient algorithm to find a weighted allocation rule that satisfy both (5.1) and (5.2) in Algorithm 3.

We provide Algorithm 4 as a subroutine to achieve this. The core idea of this algorithm is to use a coordinate descent procedure to compute a sparse distribution over actions, which is motivated by the

Algorithm 4 `OptimisticSubroutine`$(\widehat{a}, \beta, \mathcal{A}, \widehat{h})$

input action set $\mathcal{A}$, greedy action $\widehat{a} \in \mathcal{A}$, function $\widehat{h} : \mathcal{A} \to [0,1]$, parameter $\beta > 0$.

Obtain a barycentric spanner $\{A_i\}_{i=1}^d$ of $\mathcal{A}$ via the initialization oracle.

Set $q_0 = \sum_{i=1}^d \frac{1}{d} \mathbb{1}_{A_i}$.

for $t = 1, 2, \ldots$ **do**

  Set

$$q_{t-\frac{1}{2}} = \min\{\frac{2d}{2d + \mathbb{E}_{a\sim q}[(\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta]}, 1\} \cdot q_{t-1}. \tag{5.3}$$

  Use the action maximization oracle to compute

$$a_t = \arg\max_{a\in\mathcal{A}} \left\{ \widehat{h}(a) + \beta a^\top (\mathbb{E}_{\widetilde{a}\sim q_{t-\frac{1}{2}}}[\widetilde{a}\widetilde{a}^\top])^{-1} a \right\}. \tag{5.4}$$

  if $\widehat{h}(a_t) + \beta a_t^\top (\mathbb{E}_{a\sim q_{t-\frac{1}{2}}}[aa^\top])^{-1} a_t > \widehat{h}(\widehat{a}) + 2\beta d$, **then**

    Run the coordinate descent step

$$q_t = q_{t-\frac{1}{2}} + \frac{-2a_t^\top (\mathbb{E}_{a\sim q}[aa^\top])^{-1} a_t + 2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta}{(a_t^\top (\mathbb{E}_{a\sim q}[aa^\top])^{-1} a_t)^2} \mathbb{1}_{a_t}. \tag{5.5}$$

  else

    Let $q_t = q_{t-\frac{1}{2}}$, halt and output

$$q_t + (1 - \int_{\mathcal{A}} q_t(a) \mathrm{d}a) \mathbb{1}_{\widehat{a}}.$$

optimization procedure used in Agarwal et. al. [4]—however, we extend their idea from the finite-action setting to the linear action model, which requires further matrix analysis and may be interesting in its own right. We call this algorithm `OptimisticSubroutine` as the algorithm is built upon the optimistic step (5.4), where the "$a^\top(\mathbb{E}_{\widetilde{a}\sim q_{t-\frac{1}{2}}}[\widetilde{a}\widetilde{a}^\top])^{-1}a$" term is a continuous analogue to the counterfactual action divergence (4.5) in the linear action model.

Proposition 1 (optimization through `SubOpt`). *At each round within epoch m, Algorithm 4 outputs a probability distribution that satisfies (5.1) and (5.2) within at most $\lceil \frac{4}{\beta_m} + 8d(\log d + 1)\rceil$ iterations.*

According to this proposition, the optimistic subroutine outputs an efficient solution that satisfies the requirements (5.1) and (5.2) within finite number of iterations at every rounds. One advantage of Algorithm 3 is that it requires only $O(\log T)$ calls to the least-square oracle. However, the design and analysis of the optimistic subroutine becomes challenging in the infinite-action setting, especially for complex problem formulations. On the other hand, Algorithm 2 exhibits much cleaner structure and a principled analysis that covers many problem formulations of interest.

6 Conclusion and future directions

In this paper we propose `UCCB`, a simple generic principle to design optimistic algorithms in the presence of large context spaces. Key ideas underlying `UCCB` include: 1) confidence bounds in policy space rather than in action space; and 2) the potential function perspective that explains the power of optimism in the contextual setting. We present the first optimal and efficient optimistic algorithm for realizable contextual bandits with general function classes. Besides the traditional finite-action setting, we also discuss the infinite-action setting and provide the first solutions to many interesting models of practical interest.

Moving forward, there are many interesting future directions that may leverage the ideas presented in this work. The principle of optimism in the face of uncertainty plays an essential role in reinforcement learning. Currently the majority of existing provably efficient algorithms are developed for the "tabular" case, and their regret scales with the cardinality of the state space. However, empirical reinforcement learning problems typically have a large state space and rely on function approximation [19]. Motivated by this challenge, a natural next step is to adapt the `UCCB` principle to reinforcement learning problems with large state space. This paper can be viewed as an initial step towards this goal, as the contextual MAB problem is a special case of episodic reinforcement learning where the episode length is equal to one. Within the scope of bandit problems, `UCB`-type algorithms are often the "meta-algorithms" for many complex formulations when there is no contextual information. Since `UCCB` improves over `UCB`-type algorithms in several fundamental contextual settings, this work may be a building block to combine contextual information and function approximation with more complex formulations such as Gaussian process optimization [29], bandits with long-term constraints [7], and bandits in non-stationary environments [17]. We leave these directions to future work.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Jacob Abernethy, Kareem Amin, Michael Kearns, and Moez Draief. Large-scale bandit problems and kwik learning. In *International Conference on Machine Learning*, pages 588–596, 2013.

[3] Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26, 2012.

[4] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

[5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[6] Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

[7] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

[8] Peter L Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. 2008.

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

[10] Victor Chernozhukov, Mert Demirer, Greg Lewis, and Vasilis Syrgkanis. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems*, pages 15039–15049, 2019.

[11] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[12] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[13] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

[14] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[15] Dylan J Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E Schapire. Practical contextual bandits with regression oracles. *arXiv preprint arXiv:1803.01088*, 2018.

[16] Dylan J Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*, 2020.

[17] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[18] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[19] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

[20] Robert D Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.

[21] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20(65):1–50, 2019.

[22] Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *arXiv preprint arXiv:1902.01520*, 2019.

[23] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[24] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org, 2017.

[25] Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

[26] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

[27] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *arXiv preprint arXiv:2003.12699*, 2020.

[28] Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

[29] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[30] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

A Proofs for the finite-action setting

A.1 Proof of Theorem 1.

We prove the theorem on the clean event stated in Lemma 3, whose measure is at least $1-\delta/2$. For all $t > K$,

$$\begin{aligned}
\mathbb{E}_x[f^*(x,\pi_{f^*}(x)] &\le \mathbb{E}_x[\widehat{f}_t(x,\pi_{f^*}(x))] + \mathbb{E}_x\Big[\frac{\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi_{f^*}(x)=\pi_i(x)\}}\Big] + \frac{K\beta_t}{t} \\
&\le \operatorname*{arg\,max}_{\pi\in\Pi}\Big\{\mathbb{E}_x[\widehat{f}_t(x,\pi(x))] + \mathbb{E}_x\Big[\frac{\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi(x)=\pi_i(x)\}}\Big]\Big\} + \frac{K\beta_t}{t} \\
&= \mathbb{E}_x[\widehat{f}_t(x,\pi_t(x))] + \mathbb{E}_x\Big[\frac{\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi_t(x)=\pi_i(x)\}}\Big] + \frac{K\beta_t}{t} \\
&\le \mathbb{E}_x[f^*(x,\pi_t(x))] + \mathbb{E}_x\Big[\frac{2\beta_t}{\sum_{i=1}^t \mathbb{1}\{\pi_t(x)=\pi_i(x)\}}\Big] + \frac{2K\beta_t}{t},
\end{aligned} \tag{A.1}$$

where the first and the last inequality are due to Lemma 3; the second inequality due to maximization over policies.

Therefore, we have the following:

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[f^*(x_t, \pi_{f^*}(x_t)) - f^*(x_t, a_t)|H_{t-1}] &= \sum_{t=1}^{T} \big(\mathbb{E}_x[f^*(x, \pi_{f^*}(x))] - \mathbb{E}_x[f^*(x, \pi_t(x))]\big) \\
&\le \sum_{t=K+1}^{T} \mathbb{E}_x\Big[\frac{2\beta_t}{\sum_{i=1}^{t} \mathbb{1}\{\pi_t(x) = \pi_i(x)\}}\Big] + \sum_{t=K+1}^{T} \frac{2K\beta_t}{t} + K \\
&\le 2\beta_T \sum_{t=K+1}^{T} \mathbb{E}_x\Big[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi_t(x) = \pi_i(x)\}}\Big] + 2\sqrt{17KT\log(|\mathcal{F}|T^3/\delta)} + K \\
&\le 2\sqrt{17KT\log(2|\mathcal{F}|T^3/\delta)}(\log(T/K) + 1) + K, \qquad \text{(A.2)}
\end{aligned}
$$

where the first line uses the equivalence proved in Lemma 4; the second line is due to (A.1); the third line is due to $\beta_t \le \beta_T$ and $\sum_{K+1}^{T} 1/\sqrt{t} \le \sqrt{T}$; and the last line is due to the contextual potential lemma (Lemma 2).

By Azuma's inequality, with probability at least $1 - \delta/2$, we can bound the regret by

$$
\text{Regret}(T, \text{Algorithm } 1) \le \sum_{t=1}^{T} \mathbb{E}[f^*(x_t, \pi_{f^*}(x_t)) - f^*(x_t, a_t)|H_{t-1}] + \sqrt{2T\log(2/\delta)}. \qquad \text{(A.3)}
$$

Therefore, by a union bound and inequalities (A.2) (A.3), with probability at least $1 - \delta$, the regret of Algorithm 1 after T rounds is upper bounded by

$$
\text{Regret}(T, \text{Algorithm } 1) \le 2\sqrt{17KT\log(2|\mathcal{F}|T^3/\delta)}(\log(T/K) + 1) + \sqrt{2T\log(2/\delta)} + K.
$$

□

A.2 Analysis on the confidence

The main goal of this subsection is to prove Lemma 1. For a fixed f, we denote $Y_{f,i} = (f(x_i, a_i) - r_i(x_i, a_i))^2 - (f^*(x_i, a_i) - r_i(x_i, a_i))^2$, $i = 1, 2, \ldots$.

A.2.1 Proof of Lemma 1.

For a fixed $f \in \mathcal{F}$, when conditioned on Υ_{i-1}, we have

$$
\begin{aligned}
\mathbb{E}_{x_i, a_i}\left[(f(x_i, a_i) - f^*(x_i, a_i))^2 | H_{i-1}\right] &= \mathbb{E}_{x_i}\left[(f(x_i, \pi_i(x_i)) - f^*(x_i, \pi_i(x_i)))^2 | H_{i-1}\right] \\
&= \mathbb{E}_{x}\left[(f(x, \pi_i(x)) - f^*(x, \pi_i(x)))^2 \middle| H_{i-1}\right] \\
&= \mathbb{E}_{x}\left[(f(x, \pi_i(x)) - f^*(x, \pi_i(x)))^2\right],
\end{aligned}
$$

where the first equation is because $a_i = \pi_i(x_i)$ and the fact that π_i is completely determined by H_{t-1}; the second equation is because the independence between x_i and H_{i-1}; and the third inequality is because $(f(x_i, \pi_i(x)) - f^*(x_i, \pi_i(x_i)))^2$ depends on H_{i-1} only through π_i.

Therefore,

$$
\sum_{i=1}^{t-1} \mathbb{E}_{x_i, a_i}\left[(f(x_i, a_i) - f^*(x_i, a_i))^2 | H_{i-1}\right] = \sum_{i=1}^{t-1} \mathbb{E}_{x}\left[(f(x, \pi_i(x)) - f^*(x_i, \pi_i(x)))^2\right].
$$

Applying Lemma 5, we know that $\forall \delta \in (0, 1)$, with probability at least $1 - \delta/2$,

$$
\sum_{i=1}^{t-1} \mathbb{E}_{x}\left[(f_t(x, \pi_i(x)) - f^*(x_i, \pi_i(x)))^2\right] \le 68\log(2|\mathcal{F}|t^3/\delta) + 2\sum_{i=1}^{t-1} Y_{f_t, i}, \qquad \text{(A.4)}
$$

uniformly over all $t \geq K$ and all fixed sequence $f_K, f_{K+1}, \cdots \in \mathcal{F}$.

Therefore, $\forall \pi \in \Pi$,

$$\begin{aligned}
&\mathbb{E}_x\Big[\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}(f_t(x, \pi(x)) - f^*(x, \pi(x)))^2\Big] \\
&= \mathbb{E}_x\Big[\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}(f_t(x, \pi_i(x)) - f^*(x, \pi_i(x)))^2\Big] \\
&= \sum_{i=1}^{t-1} \mathbb{E}_x[\mathbb{1}\{\pi(x) = \pi_i(x)\}(f_t(x, \pi_i(x)) - f^*(x, \pi_i(x)))^2] \\
&\leq \sum_{i=1}^{t-1} \mathbb{E}_x[(f_t(x, \pi_i(x)) - f^*(x, \pi_i(x)))^2] \\
&\leq 68 \log(2|\mathcal{F}|t^3/\delta) + 2\sum_{i=1}^{t-1} Y_{f_t,i},
\end{aligned} \tag{A.5}$$

where the first inequalities are due to $\mathbb{1}\{\pi(x) = \pi_i(x)\} \leq 1$ and the second inequality is (A.4).

Since Algorithm 1 pick all actions exactly once during the first K rounds, $t > K$ will ensure $\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi(x)\} \geq 1, \forall x \in \mathcal{X}$.

From Cauchy-Schwarz's inequality, $\forall t > K$, $\forall \pi \in \Pi$,

$$\begin{aligned}
&|\mathbb{E}_x[f_t(x, \pi(x)) - f^*(x, \pi(x))]| \\
&\leq \sqrt{\mathbb{E}_x\big[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}}\big]}\sqrt{\mathbb{E}_x\Big[\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}(f_t(x, \pi(x)) - f^*(x_i, \pi(x)))^2\Big]}.
\end{aligned}$$

Combine the above inequality with (A.5), we prove

$$\begin{aligned}
&|\mathbb{E}_x[f_t(x, \pi(x)) - f^*(x, \pi(x))]| \\
&\leq \sqrt{\mathbb{E}_x\big[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}}\big]}\sqrt{68 \log(2|\mathcal{F}|t^3/\delta) + 2\sum_{i=1}^{t-1} Y_{f_t,i}}.
\end{aligned}$$

Taking $f_t = \widehat{f}_t$ in the above inequality, and use the fact $\sum_{i=1}^{t-1} Y_{\widehat{f}_t,i} \leq 0$ (as the least square solution $\widehat{f}_t$ minimizes $\sum_{i=1}^{t-1}(f(x_i, a_i) - r_i(x_i, a_i))^2$), we obtain: with probability at least $1 - \delta/2$, $\forall t > K$, $\forall \pi \in \Pi$,

$$\begin{aligned}
&\big|\mathbb{E}_x[\widehat{f}_t(x, \pi(x))] - \mathbb{E}_x[f^*(x, \pi(x))]\big| \\
&\leq \sqrt{\mathbb{E}_x\big[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}}\big]}\sqrt{68 \log(2|\mathcal{F}|t^3/\delta)} \\
&\leq \sqrt{\mathbb{E}_x\big[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x) = \pi_i(x)\}}\big]}\sqrt{68 \log(2|\mathcal{F}|t^3/\delta)}
\end{aligned}$$

□

A.2.2 Proof of Lemma 5

We now prove Lemma 5 and the supporting lemmas required to prove Lemma 5.

Proof of Lemma 5. Fix a $\delta \in (0, 1)$. Take $\delta_t = \delta/2t^3$, and apply a union bound to Lemma 6 with all $t \geq 2$. From

$$\sum_{t=1}^{\infty} \delta_t \log_2(t-1) \leq \sum_{t=2}^{\infty} \delta/2t^2 \leq \delta/2,$$

we know that with probability at least $1 - \delta/2$,

$$\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}\left[(f_t(x_i,a_i) - f^*(x_i,a_i))^2 | H_{i-1}\right] \leq 68 \log(2|\mathcal{F}|t^3/\delta) + 2\sum_{i=1}^{t-1} Y_{f_t,i},$$

uniformly over all $t \geq 2$ and all fixed sequence $f_2, f_3, \dots \in \mathcal{F}$. $\square$

Lemma 6 (**uniform convergence over** $\mathcal{F}$)**.** *For a fixed $t \geq 2$ and a fixed $\delta_t \in (0, 1/e^2)$, with probability at least $1 - \log_2(t-1)\delta_t$, we have*

$$\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}\left[(f(x_i,a_i) - f^*(x_i,a_i))^2 | H_{i-1}\right] \leq 68 \log(|\mathcal{F}|/\delta_t) + 2\sum_{i=1}^{t-1} Y_{f,i}, \tag{A.6}$$

uniformly over all $f \in \mathcal{F}$.

Proof of Lemma 6. We have $|Y_{f,i}| \leq 1, \forall i$. From Lemma 7, for $\delta_t/|\mathcal{F}| \leq \delta_t < 1/e^2$, with probability at least $1 - \log_2(t-1)\delta_t/|\mathcal{F}|$,

$$\sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}] - \sum_{i=1}^{t-1} Y_{f,i} \leq 4\sqrt{\sum_{i=1}^{t-1} \mathbf{Var}[Y_{f,i}|H_{i-1}] \log(|\mathcal{F}|/\delta_t)} + 2\log(|\mathcal{F}|/\delta_t).$$

Applying union bound to all $f \in \mathcal{F}$, we obtain that with probability at least $1 - \log_2(t-1)\delta_t \geq 1 - \log_2 t\delta_t$,

$$\sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}] - \sum_{i=1}^{t-1} Y_{f,i} \leq 4\sqrt{\sum_{i=1}^{t-1} \mathbf{Var}[Y_{f,i}|H_{i-1}] \log(|\mathcal{F}|/\delta_t)} + 2\log(|\mathcal{F}|/\delta_t), \quad \forall f \in \mathcal{F}.$$

From Lemma 8 we have $\mathbf{Var}[Y_{f,i}|H_i] \leq 4\mathbb{E}[Y_{f,i}|H_i]$. Therefore

$$\begin{aligned}\sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}] &\leq 4\sqrt{\sum_{i=1}^{t-1} \mathbf{Var}[Y_{f,i}|H_{i-1}] \log(|\mathcal{F}|/\delta_t)} + 2\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1} Y_{f,i} \\ &\leq 8\sqrt{\sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}] \log(|\mathcal{F}|/\delta_t)} + 2\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1} Y_{f,i}, \quad \forall f \in \mathcal{F}.\end{aligned}$$

This implies $\forall f \in \mathcal{F}$,

$$\left(\sqrt{\sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}]} - 4\sqrt{\log(|\mathcal{F}|/\delta_t)}\right)^2 \leq 18\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1} Y_{f,i},$$

which further implies $\forall f \in \mathcal{F}$,

$$\sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}] \leq 68\log(|\mathcal{F}|/\delta_t) + 2\sum_{i=1}^{t-1} Y_{f,i}.$$

From Lemma 8, we have

$$\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}\left[f(x_i,a_i) - f^*(x_i,a_i))^2|H_{i-1}\right] = \sum_{i=1}^{t-1} \mathbb{E}[Y_{f,i}|H_{i-1}] \leq 68\log(|\mathcal{F}|/\delta_t) + 2\sum_{i=1}^{t-1} Y_{f,i}.$$

This finish the proof to Lemma 6. □

The following two lemmas are used in the proof of Lemma 6.

Lemma 7 (**Freeman's inequality, [8]**). *Suppose $Z_1, Z_2, \ldots, Z_t$ is a martingale difference sequence with $|Z_i| \leq b$ for all $i = 1, \ldots, t$. Then for any $\delta < 1/e^2$, with probability at least $1 - (\log_2 t)\delta$,*

$$\sum_{i=1}^{t} Z_i \leq 4\sqrt{\sum_{i=1}^{t} \mathbf{Var}[Z_i|Z_1, \ldots, Z_{i-1}] \log(1/\delta)} + 2b\log(1/\delta).$$

Lemma 8 (**Lemma 4.2 in [3]**). *Fix a function $f \in \mathcal{F}$. Suppose we sample x from the data distribution $\mathcal{D}_\mathcal{X}$, and $r(x,a)$ from $\mathcal{D}_{x,a}$. Define the random variable*

$$Y = (f(x,a) - r(x,a))^2 - (f^*(x,a) - r(x,a))^2.$$

Then we have

$$\mathbb{E}_{x,r,a}[Y] = \mathbb{E}_{x,a}[(f(x,a) - f^*(x,a))^2],$$
$$\mathbf{Var}_{x,r,a}[Y] \leq 4\mathbb{E}_{x,r,a}[Y].$$

A.3 Proof of Lemma 2

Proof of Lemma 2. For any fixed $x \in \mathcal{X}$, we have

$$\sum_{t=K+1}^{T} \frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\{\pi_t(x) = \pi_j(x)\}} \leq \sum_{a\in\mathcal{A}} \sum_{i=1}^{\sum_{t=1}^{T} \mathbb{1}\{\pi_t(x)=a\}} \frac{1}{i}$$
$$\leq \sum_{a\in\mathcal{A}} (1 + \log(\sum_{t=1}^{T} \mathbb{1}\{\pi_t(x) = a\})) \leq K + K\log(T/K),$$

where the last inequality is due to Jensen's inequality. By taking expectation on both sides of the above inequality, we prove the lemma. □

B Proofs for the extensions to infinite function classes

B.1 Proof of Corollary 2

From the well-known result on the covering of $d-$dimensional balls [30], the covering number of a $d-$dimensional ball with radius $\frac{\Delta}{2}$ and discretization error $\frac{1}{Lt}$ is bounded by $(1 + \Delta Lt)^d$, so there exists a set V_t of size no more than $(1 + \Delta Lt)^d + 1 \leq (2 + \Delta Lt)^d$ that contains θ^* and satisfies

$$\forall \theta \in \Theta\ \exists v \in V_t \text{ s.t. } \|\theta - v\| \leq \frac{1}{Lt}.$$

We see $\log|V_t| \le d\log(2+\Delta L t)$. $\forall f_\theta \in \mathcal{F}, x \in \mathcal{X}, a \in \mathcal{A}$, take v to be the closest point to θ in V_t, we have

$$\begin{aligned}(f_\theta(x,a) - f^*(x,a))^2 &= (f_\theta(x,a) - f_v(x,a) + f_v(x,a) - f_{\theta^*}(x,a))^2 \\ &\le 2(f_\theta(x,a) - f_v(x,a))^2 + 2(f_v(x,a) - f_{\theta^*}(x,a))^2 \\ &\le 2L^2\|\theta - v\|^2 + 2(f_v(x,a) - f_{\theta^*}(x,a))^2 \\ &\le \frac{2}{t^2} + 2(f_v(x,a) - f_{\theta^*}(x,a))^2.\end{aligned}$$

Sine V_t is a finite function class, we can prove a slight modification of Lemma 6, with the result (A.6) becomes

$$\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}[(f(x_i,a_i) - f^*(x_i,a_i))^2 | H_{i-1}] \le 136\log(|V_t|/\delta_t) + 2 + 4\sum_{i=1}^{t-1} Y_{f,i}.$$

Following the same path in the proof of Lemma 5, we can prove a slight modification of Lemma 5: with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}\left[(f_t(x_i,a_i) - f^*(x_i,a_i))^2 | H_{i-1}\right] &\le 136\Big(d\log(2+\Delta L t) + \log\frac{2t^3}{\delta}\Big) + 2 + 4\sum_{i=1}^{t-1} Y_{f_t,i} \\ &\le 136\Big(d\log(2+\Delta L t) + \log\frac{2t^3}{\delta} + 1\Big) + 4\sum_{i=1}^{t-1} Y_{f_t,i},\end{aligned}$$

uniformly over all $t \ge 2$ and all fixed sequence $f_2, f_3, \cdots \in \mathcal{F}$.

By setting the parameter β_t to be

$$\beta_t = \sqrt{34t/K}\sqrt{d\log(2+\Delta L t) + \log(2t^3/\delta) + 1}$$

in Algorithm 1, we can prove a slight modification of Theorem 1, with the result being

$$\text{Regret}(T, \text{Algorithm 1}) \le 2K\beta_T(\log(T/K) + 1) + \sqrt{2T\log(2/\delta)} + K.$$

B.2 Proof of Corollary 3

We introduce the following Lemma adopt from [15]:

Lemma 9 (a consequence of Lemma 4 in [15]). $\forall \delta \in (0,1)$, *with probability at least* $1-\delta$,

$$\sum_{i=\tau_1}^{\tau_2} \mathbb{E}_{x_i,a_i}[(f(x_i,a_i) - f^*(x_i,a_i))^2 | H_{i-1}] \le 2\sum_{i=\tau_1}^{\tau_2} Y_{f,i} + \\ K \cdot \inf_{\varepsilon > 0}\left\{100\varepsilon T + 320\log\left(\frac{4K\mathbb{E}_{\{x_i\}_{i=1}^T}\mathcal{N}_1(\mathcal{G}, \varepsilon, \{x_i\}_{i=1}^T)}{\delta}\right)\right\}.$$

for all $1 \le \tau_1 \le \tau_2 \le T$ *and* $g \in \mathcal{G}$.

We then prove a slight modification of Lemma 5, with the result (2.7) becomes

$$\sum_{i=1}^{t-1} \mathbb{E}_{x_i,a_i}\left[(f_t(x_i,a_i) - f^*(x_i,a_i))^2 | H_{i-1}\right] \le 2\sum_{i=1}^{t-1} Y_{f_t,i} + \\ K \cdot \inf_{\varepsilon > 0}\left\{100\varepsilon T + 320\log\left(\frac{8KT^3\mathbb{E}_{\{x_i\}_{i=1}^T}\mathcal{N}_1(\mathcal{G}, \varepsilon, \{x_i\}_{i=1}^T)}{\delta}\right)\right\},$$

uniformly over all $t \geq 2$ and all fixed sequence $f_2, f_3, \cdots \in \mathcal{F}$.

By setting the parameter β_t in Algorithm 1 to be the fixed value

$$\beta = \sqrt{TK} \cdot \inf_{\varepsilon > 0} \left\{ 25\varepsilon T + 80 \log \left(\frac{8KT^3 \mathbb{E}_{\{x_i\}_{i=1}^T} \mathcal{N}_1(\mathcal{G}, \varepsilon, \{x_i\}_{i=1}^T)}{\delta} \right) \right\},$$

we can prove a slight modification of Theorem 1, with the result being

$$\text{Regret}(T, \text{Algorithm 1}) \leq 2K\beta(\log(T/K) + 1) + \sqrt{2T \log(2/\delta)} + K.$$

C Proofs for the infinite-action setting

Proof of Theorem 4. We prove the theorem on the clean event stated in Lemma 10, whose measure is at least $1 - \delta/2$. For all $t \geq 2$,

$$\begin{aligned}
\mathbb{E}_x[f^*(x, \pi_{f^*}(x)] &\leq \mathbb{E}_x[1 \wedge \beta_t V_x(\pi_{f^*}(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + \mathcal{E}\beta_t/t \\
&\leq \mathbb{E}_x[\widehat{f}_t(x, \pi_t(x))] + \mathbb{E}_x[1 \wedge \beta_t V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + \mathcal{E}\beta_t/t \\
&\leq \mathbb{E}_x[f^*(x, \pi_t(x))] + 2\mathbb{E}_x[1 \wedge \beta_t V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + 2\mathcal{E}\beta_t/t.
\end{aligned}$$

where the first and the last inequalities are due to Lemma 10; the second inequality is due to the definition of π_t in Lemma 11. The above argument implies that for all $t \geq 2$

$$\mathbb{E}_x[f^*(x, \pi_{f^*}(x) - f^*(x, \pi_t(x)] \leq 2\mathbb{E}_x[1 \wedge \beta_t V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + 2\mathcal{E}\beta_t/t. \tag{C.1}$$

When $t = 1$, inequality (C.1) trivially holds true, because $V_x(\pi_1(x)||\emptyset) = \infty$ by definition. So inequality (C.1) holds true for all $t \geq 1$.

When $T \leq \mathcal{E}$, we can bound the regret by $\mathcal{E}$. We now give the regret bound for the case $T > \mathcal{E}$. We have the following:

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}[f^*(x_t, \pi_{f^*}(x_t)) - f^*(x_t, a_t)|H_{t-1}] &= \sum_{t=1}^T \left(\mathbb{E}_x[f^*(x, \pi_{f^*}(x))] - \mathbb{E}_x[f^*(x, \pi_t(x))]\right) \\
&\leq \sum_{t=1}^T 2\mathbb{E}_x[1 \wedge \beta_t V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + \sum_{t=1}^T 2\mathcal{E}\beta_t/t \\
&\leq 2\sum_{t=1}^T \mathbb{E}_x[1 \wedge \beta_t V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + 2\sqrt{17\mathcal{E}T \log(2|\mathcal{F}|T^3/\delta)} \\
&\leq 2\sum_{t=1}^T \mathbb{E}_x[\beta_T \wedge \beta_T V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + 2\sqrt{17\mathcal{E}T \log(2|\mathcal{F}|T^3/\delta)} \\
&= 2\beta_T \mathbb{E}_x[\sum_{t=1}^T 1 \wedge V_x(\pi_t(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + 2\sqrt{17\mathcal{E}T \log(2|\mathcal{F}|T^3/\delta)} \\
&\leq 2\beta_T \mathcal{E}\text{poly}(\log T) + 2\sqrt{17\mathcal{E}T \log(2|\mathcal{F}|T^3/\delta)} \\
&= 2\sqrt{17\mathcal{E}T \log(2|\mathcal{F}|T^3/\delta)}(\text{poly}(\log T) + 1),
\end{aligned} \tag{C.2}$$

where the first line uses the equivalence proved in Lemma 11; the second line is due to (C.1); the third line is due to $\sum_{t=1}^T 1/\sqrt{t} \leq \sqrt{T}$; the fourth line is due to $\beta_T > \beta_t$ and $\beta_T > 1$ when $T > \mathcal{E}$; and the sixth line is due to the condition II in Assumption 2. By Azuma's inequality, with probability at least $1 - \delta/2$, we can bound the regret by

$$\text{Regret}(T, \text{Algorithm 2}) \leq \sum_{t=1}^T \mathbb{E}[f^*(x_t, \pi_{f^*}(x_t)) - f^*(x_t, a_t)|H_{t-1}] + \sqrt{2T \log(2/\delta)}. \tag{C.3}$$

Therefore, by a union bound and inequalities (C.2) (C.3), with probability at least $1-\delta$, the regret of Algorithm 1 after T rounds is upper bounded by

$$\text{Regret}(T, \text{Algorithm 2}) \leq 2\sqrt{17\mathcal{E}T\log(2|\mathcal{F}|T^3/\delta)}(\text{poly}(\log T)+1) + \sqrt{2T\log(2/\delta)}.$$

Combine the case $T \leq \mathcal{E}$ and $T > \mathcal{E}$ we finish the proof. □

Lemma 10 (counterfactual confidence bound). *Consider a non-randomized contextual bandit algorithm that selects π_t based on H_{t-1} and chooses the action $a_t = \pi_t(x_t)$ at all rounds t. Then $\forall \delta \in (0,1)$, with probability at least $1-\delta/2$, we have*

$$\left|\mathbb{E}_x[\widehat{f}_t(x,\pi(x)] - \mathbb{E}_x[f^*(x,\pi(x))]\right| \leq \mathbb{E}[1 \wedge \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + \mathcal{E}\beta_t/t.$$

uniformly over all $\pi \in \Pi$ and all $t \geq 2$.

Proof of Lemma 10. For a fixed f, we denote $Y_{f,i} = (f(x_i,a_i) - r_i(x_i,a_i))^2 - (f^*(x_i,a_i) - r_i(x_i,a_i))^2$, $i = 1,2,\dots$. From Lemma 5, $\forall\delta \in (0,1)$, with probability at least $1-\delta/2$, we have

$$\sum_{i=1}^{t-1}\mathbb{E}_{x_i,a_i}\left[(f_t(x_i,a_i) - f^*(x_i,a_i))^2|H_{i-1}\right] \leq 68\log(2|\mathcal{F}|t^3/\delta) + \sum_{i=1}^{t-1}Y_{f_t,i}, \tag{C.4}$$

uniformly over all $t \geq 2$ and all fixed sequence $f_2, f_3, \dots f_{t+1}, \dots \in \mathcal{F}$.

Use the fact that π_i is completely determined by H_{i-1} and independent with x_i, we obtain:

$$\sum_{i=1}^{t-1}\mathbb{E}_x\left[(\widehat{f}_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2\right] \leq 68\log(2|\mathcal{F}|t^3/\delta) + \sum_{i=1}^{t-1}Y_{f_t,i} = 4\mathcal{E}\beta_t^2/t + \sum_{i=1}^{t-1}Y_{f_t,i}, \tag{C.5}$$

uniformly over all $t \geq 2$.

From the definition of counterfactual action divergence, we know $\forall x \in \mathcal{X}$,

$$|f_t(x,\pi(x)) - f^*(x,\pi(x))| \leq \sqrt{V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})}\sqrt{\sum_{i=1}^{t-1}(f_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2}.$$

Applying the AM-GM inequality to the above inequality, we obtain

$$|f_t(x,\pi(x)) - f^*(x,\pi(x))| \leq \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1}) + \frac{1}{4\beta_t}\sum_{i=1}^{t-1}(f_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2.$$

Since $\mathcal{F}$ is bounded by $[0,1]$, we further obtain

$$\begin{aligned}
|f_t(x,\pi(x)) - f^*(x,\pi(x))| &\leq \max\left\{\beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1}) + \frac{1}{4\beta_t}\sum_{i=1}^{t-1}(f_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2, 1\right\} \\
&\leq \max\left\{1, \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})\right\} + \frac{1}{4\beta_t}\sum_{i=1}^{t-1}(f_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2 \\
&= 1 \wedge \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1}) + \frac{1}{4\beta_t}\sum_{i=1}^{t-1}(\widehat{f}_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2
\end{aligned} \tag{C.6}$$

By taking expectation on both side of (C.6) and using (C.5), we obtain that with probability at least $1-\delta/2$,

$$
\begin{aligned}
\left|\mathbb{E}_x[f_t(x,\pi(x)) - f^*(x,\pi(x))]\right| &\le \mathbb{E}_x|f_t(x,\pi(x)) - f^*(x,\pi(x))| \\
\le \mathbb{E}_x[1\wedge \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})] &+ \frac{1}{4\beta_t}\mathbb{E}_x[\sum_{i=1}^{t-1}(f_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2] \\
= \mathbb{E}_x[1\wedge \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})] &+ \frac{1}{4\beta_t}\sum_{i=1}^{t-1}\mathbb{E}[(f_t(x,\pi_i(x)) - f^*(x,\pi_i(x)))^2] \\
&\le \mathbb{E}_x[1\wedge \beta_t V_x(\pi(x)||\{\pi_i(x)\}_{i=1}^{t-1})] + \mathcal{E}\beta_t/t + \frac{1}{4\beta_t}\sum_{i=1}^{t-1}Y_{f_t,i}.
\end{aligned}
$$

uniformly over all $\pi\in\Pi$, all $t\ge 2$ and all fixed sequence $f_2, f_3, \ldots, \in \mathcal{F}$. Here the first inequality is due to the triangle inequality; the second inequality is due to (C.6); the last inequality is due to (C.5).

By taking $f_t = \widehat{f}_t$ the the least square solution that minimizes $\sum_{i=1}^{t-1}(f(x_i,a_i) - r_i(x_i,a_i))^2$, we have $Y_{f_t,i}\le 0$ and finish the proof. □

Lemma 11. *Consider an algorithm that choose policy π_t by*

$$
\pi_t(x) = \begin{cases} A_{x,t} & \text{if } t\le d_x, \\ \arg\max\limits_{\mathcal{A}(x)} \left\{\widehat{f}_t(x,\cdot) + \beta_t V_{x_t}(\cdot||\{\pi(x)\}_{j=1}^{t-1})\right\} & \text{if } t > d_x. \end{cases}
$$

($A_{x,t}$ is determined the initialization oracle and the input $\mathcal{A}(x)$; the "argmax" problem when $t>d_x$ is computed via the action maximization oracle.) Then this algorithm produces the same actions as those produced by Algorithm 2.

Proof. The proof to this lemma is straightforward. □

D Proofs for the "optimistic subroutine" in Section 5

In this subsection we prove Proposition 1. Our proof is motivated by Agarwal et. al. [4, Lemma 6, Lemma 7].

We call q an improper distribution if: 1) $\int_{\mathcal{A}} q(a)\mathrm{d}a$ is within $(0,1]$ but not necessarily equal to one; and 2) $\mathbb{E}_{a\sim q}[aa^\top]$ is invertible. We define the improper expectation $\mathbb{E}_{a\sim q}[W(a)]$ for any random variable $W:\mathcal{A}\to\mathbb{R}$ by the integral $\int_{a\in\mathcal{A}} W(a)q(a)\mathrm{d}a$.

We aim to minimize the potential function

$$
\Phi(q) := -2\log(\det(\mathbb{E}_{a\sim q}[b_a {b_a}^\top])) + \mathbb{E}_{a\sim q}[2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta], \tag{D.1}
$$

where b_a is the coefficient vector of a when the basis is the barycentric spanner $\{A_i\}_{i=1}^d$. We prove that after each iteration, either Algorithm 4 outputs a desired distribution that satisfies both (5.1) and (5.2), or

$$
\Phi(q_t) \le \Phi(q_{t-1}) - \frac{1}{4}. \tag{D.2}
$$

Since Φ function is bounded the algorithm must halt within finite iterations. (D.2) is a consequence of the following two lemmas:

Lemma 12. *When* $\mathbb{E}_{a\sim q(a)}[2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta] \geq 2d$*, the objective* $\Phi(q)$ *will not increase if we multiply* q *by* $\frac{2d}{2d+\mathbb{E}_{a\sim q}[2d+(\widehat{h}(\widehat{a})-\widehat{h}(a))/\beta]}$*. That is, after step* (5.3) *in Algorithm 4, we always have*

$$\Phi(q_{t-\frac{1}{2}}) \leq \Phi(q_{t-1}).$$

Lemma 13. *If Algorithm 4 does not halt at round* t*, then after the coordinate descent step* (5.5) *in Algorithm 4, we always have*

$$\Phi(q_t) \leq \Phi(q_{t-\frac{1}{2}}) - \frac{1}{4}.$$

Now we present the proof of Proposition 1, as well as proofs of Lemma 12 and Lemma 13.

Proof of Proposition 1. From Lemma 12 and Lemma 13 we know that if the algorithm does not halt at round t, then $\Phi(q_t) \leq \Phi(q_{t-\frac{1}{2}}) - \frac{1}{4}$. Assume Algorithm 4 does not halt after t rounds. Then we have

$$\begin{aligned}
2d\log d + 2d + \frac{1}{\beta} &\geq 2d\log d + \mathbb{E}_{a\sim q_0}[2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta] = \Phi(q_0) \\
\geq \Phi(q_t) + \frac{t}{4} &= -2\log(\det(\mathbb{E}_{a\sim q_t}[b_a b_a^\top])) + \mathbb{E}_{a\sim q_t}[2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta] + \frac{t}{4} \\
&\geq -2\log(\det(\mathbb{E}_{a\sim q_t}[b_a b_a^\top])) + \frac{t}{4}. \qquad \text{(D.3)}
\end{aligned}$$

where the first inequality is due to $\int_A q_0(a)\mathrm{d}a = 1$; the first equation is due to $-2\log(\det(\mathbb{E}_{a\sim q_0}[b_a b_a^\top])) = -2\log(\det(\frac{1}{d}I)) = 2d\log d$; and the last inequality is due to $\mathbb{E}_{a\sim q_t}[2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta] \geq 0$.

Since the initialization actions consist of a barycentric spanner of $\mathcal{A}$, all coordinates of b_a is within $[-1,1]$, $\forall a \in \mathcal{A}$. Clearly $\|b_a\| \leq \sqrt{d}$ for all $a \in \mathcal{A}$. We know that q_t is a improper distribution with at most $d+t$ non-zero supports, so we assume $q_t = \sum_{i=1}^{d+t} q_t(A_i)\mathbb{1}_{A_i}$.

$$\begin{aligned}
\det(\mathbb{E}_{a\sim q_t}[b_a b_a^\top]) &= \det(\sum_{i=1}^{d+t} q_t(A_i) b_{A_i} b_{A_i}^\top) \\
\leq (\frac{\mathrm{tr}(\sum_{i=1}^{d+t} q_t(A_i) b_{A_i} b_{A_i}^\top)}{d})^d &= (\frac{\sum_{i=1}^{d+t} q_t(A_i)\mathrm{tr}(b_{A_i} b_{A_i}^\top)}{d})^d \\
&\leq (\frac{\sum_{i=1}^{d+t} q_t(A_i)\|b_{A_i}\|^2}{d})^d \leq 1,
\end{aligned}$$

where the first inequality is due to the AM-GM inequality; the last inequality is due to $\|b_a\| \leq \sqrt{d}$ for all $a \in \mathcal{A}$ and $\sum_{i=1}^{d+t} q_t(A_i) = \int_{\mathcal{A}} q_t(a)da \leq 1$. As a result, we obtain $\log(\det(\mathbb{E}_{a\sim q_t}[b_a b_a^\top])) \leq 0$. Combine this result with (D.3), we obtain

$$t \leq 8d(\log d + 1) + \frac{4}{\beta}.$$

So Algorithm 4 must halt within at most $\lceil \frac{4}{\beta_m} + 8d(\log d + 1) \rceil$ iterations. When it halts, it is straightforward to verify that the output distribution is proper and satisfies both (5.1) and (5.2). □

Proof of Lemma 12. Denote $w(a) = (\widehat{h}(\widehat{a}) - \widehat{h}(a))/\beta$. Given an arbitrary improper distribution q, we view $\Phi(c \cdot q)$ as a function on the scaling factor c. By the chain rule, we can compute the derivative of this function with respect to c,

$$\partial_c \Phi(c\cdot q) = \int_{\mathcal{A}} \Big[\partial_{cq(a)}\Phi(c\cdot q)\Big]\Big(\partial_c cq(a)\Big)\mathrm{d}a$$
$$= \int_{\mathcal{A}} \Big[-2a^\top(\mathbb{E}_{\widetilde{a}\sim cq}[\widetilde{a}\widetilde{a}^\top])^{-1}a + 2d + w(a)\Big]q(a)\mathrm{d}a$$
$$= \frac{2}{c}\mathbb{E}_{a\sim q}[a^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a] + 2d + \mathbb{E}_{a\sim q}[w(a)],$$

where the second equation use the fact that the partial gradient of $\log(\det(\mathbb{E}_{a\sim cq}[b_a b_a^\top]))$ with respect to the coordinate $cq(a)$ is $b_a^\top(\mathbb{E}_{\widetilde{a}\sim cq}[b_{\widetilde{a}}b_{\widetilde{a}}^\top])^{-1}b_a = a^\top(\mathbb{E}_{\widetilde{a}\sim cq}[\widetilde{a}\widetilde{a}^\top])^{-1}a$.

By the "trace trick", we have

$$\begin{aligned}\mathbb{E}_{a\sim q}[a^T(\mathbb{E}_{a\sim q}[aa^T])^{-1}a] &= \mathrm{tr}(\mathbb{E}_{a\sim q}[a^T(\mathbb{E}_{a\sim q}[aa^T])^{-1}a])\\ &= \mathbb{E}_{a\sim q}[\mathrm{tr}(a^T(\mathbb{E}_{a\sim q}[aa^T])^{-1}a)]\\ &= \mathbb{E}_{a\sim q}[\mathrm{tr}(aa^T(\mathbb{E}_{a\sim q}[aa^T])^{-1})]\\ &= \mathrm{tr}(\mathbb{E}_{a\sim q}aa^T(\mathbb{E}_{a\sim q}[aa^T])^{-1})) = d.\end{aligned}$$

So we have

$$\partial_c\Phi(c\cdot q) = -\frac{2d}{c} + 2d + \mathbb{E}_{a\sim q}[w(a)].$$

This means that for all $c \in [\frac{2d}{2d+\mathbb{E}_{a\sim q}[w(a)]}, 1]$, $\partial_c\Phi(c\cdot q) \geq 0$.

Proof of Lemma 13. We define

$$\Delta_t = \frac{-2a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t + 2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta}{(a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t)^2},$$

then the coordinate descent step (5.5) is $q_t = q_{t-\frac{1}{2}} + \Delta_t \mathbb{1}_{a_t}$.

$$\begin{aligned}\Phi(q_{t-\frac{1}{2}}) - \Phi(q_t) &= 2\log(\frac{\det(\mathbb{E}_{a\sim q^+}[b_a b_a{}^\top])}{\det(\mathbb{E}_{a\sim q}[b_a b_a{}^\top])} - \Delta_t\big(2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta\big)\\ &= 2\log(1 + \Delta_t a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t) - \Delta_t\big(2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta\big) \qquad \text{(D.4)}\\ &\geq 2a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t\Delta_t - (a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t)^2\Delta_t^2 - \Delta_t\big(2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta\big) \qquad \text{(D.5)}\\ &= \frac{\Big(-2a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t + 2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta\Big)^2}{4(a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t)^2}, \qquad \text{(D.6)}\end{aligned}$$

where (D.4) is due to the matrix determinant lemma as well as $a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t = b_{a_t}^\top(\mathbb{E}_{a\sim q}[b_a b_a{}^\top])^{-1}b_{a_t}$; (D.5) is due to the inequality $\log(1+w) \geq w - \frac{w^2}{2}$ for all $w \geq 0$; (D.6) is due to the definition of Δ_t which maximize the quadratic function in (D.5).

As the algorithm does not halt, we have $a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t \geq 2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta$, so

$$\begin{aligned}&|-2a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t + 2d + (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta|\\ &= 2a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t - 2d - (\widehat{h}(\widehat{a}) - \widehat{h}(a_t))/\beta\\ &\geq a_t^\top(\mathbb{E}_{a\sim q}[aa^\top])^{-1}a_t.\end{aligned}$$

Combine this inequality with (D.6) we obtain

$$\Phi(q_{t-\frac{1}{2}}) - \Phi(q_t) \geq \frac{1}{4}.$$

□