Bandits with Stochastic Experts: Constant Regret, Empirical Experts and Episodes

NIHAL SHARMA, The University of Texas at Austin, Austin, United States
RAJAT SEN, Google, Mountain View, United States
SOUMYA BASU, Google, Mountain View, United States
KARTHIKEYAN SHANMUGAM, Google DeepMind, Bengaluru, India
SANJAY SHAKKOTTAI, The University of Texas at Austin, Austin, United States

We study a variant of the contextual bandit problem where an agent can intervene through a set of stochastic expert policies. Given a fixed context, each expert samples actions from a fixed conditional distribution. The agent seeks to remain competitive with the "best" among the given set of experts. We propose the Divergence-based Upper Confidence Bound (D-UCB) algorithm that uses importance sampling to share information across experts and provide horizon-independent constant regret bounds that only scale linearly in the number of experts. We also provide the Empirical D-UCB (ED-UCB) algorithm that can function with only approximate knowledge of expert distributions. Further, we investigate the episodic setting where the agent interacts with an environment that changes over episodes. Each episode can have different context and reward distributions resulting in the best expert changing across episodes. We show that by bootstrapping from $O(N \log(NT^2\sqrt{E}))$ samples, ED-UCB guarantees a regret that scales as $O(E(N+1) + \frac{N\sqrt{E}}{T^2})$ for N experts over E episodes, each of length T. We finally empirically validate our findings through simulations.

CCS Concepts: • **Computing methodologies** → **Online learning settings**; **Sequential decision making**; **Learning under covariate shift**;

Additional Key Words and Phrases: Bandit problems, episodic learning, learning with experts

ACM Reference Format:
Nihal Sharma, Rajat Sen, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. 2024. Bandits with Stochastic Experts: Constant Regret, Empirical Experts, and Episodes. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 9, 3, Article 12 (August 2024), 33 pages. https://doi.org/10.1145/3680279

This work was conducted when Karthikeyan Shanmugam was at IBM Research, NY, USA.
This research was partially supported by NSF Grants 1826320, 2019844, 2107037, and 2112471, ARO grant W911NF-17-1-0359, US DOD grant H98230-18-D-0007, ONR Grant N00014-19-1-2566, and the Wireless Networking and Communications Group Industrial Affiliates Program.
Authors' Contact Information: Nihal Sharma, The University of Texas at Austin, Austin, Texas, United States; e-mail: nihal.sharma@utexas.edu; Rajat Sen, Google, Mountain View, California, United States; e-mail: rajat.sen@utexas.edu; Soumya Basu, Google, Mountain View, California, United States; e-mail: basusoumya@utexas.edu; Karthikeyan Shanmugam, Google DeepMind, Bengaluru, Karnataka, India; e-mail: karthikeyanvs@google.com; Sanjay Shakkottai, The University of Texas at Austin, Austin, Texas, United States; e-mail: shakkott@austin.utexas.edu.

ACM 2376-3639/2024/08-ART12
https://doi.org/10.1145/3680279

1 Introduction

Recommendation systems for suggesting items to users are commonplace in online services such as marketplaces, content delivery platforms, and ad placement systems. Such systems, over time, learn from user feedback and improve their recommendations. An important caveat, however, is that both the distribution of user types and their respective preferences change over time, thus inducing changes in the optimal recommendation and requiring the system to periodically "reset" its learning.

We consider systems with known change-points (a.k.a. episodes) in the distribution of user-features and preferences. Examples include seasonality in product recommendations where there are marked changes in interests based on time-of-year, or ad placements based on time of day. While a baseline strategy would be to re-learn the recommendation algorithm in each episode, it is often advantageous to share some learning across episodes. Specifically, one often has access to (potentially, a very) large number of pre-trained recommendation algorithms (a.k.a. experts), and the goal then is to quickly determine (in an online manner) which expert is best suited to a specific episode. Crucially, the expert policies are *invariant over episodes* and can be learned offline—given samples of (observed context, recommended action) pairs from the deployment of one expert in the past, one can infer the policy *approximately*. The problem is then to efficiently "transfer" this approximate knowledge to the online phase to accelerate the learning of the episode-dependent best expert.

As a specific example, we take the case of online advertising agencies, which are companies that have proprietary ad-recommendation algorithms that place ads for other product companies on newspaper websites based on past *campaigns*. In each campaign, the agencies place ads for a specific product of the client (e.g., a flagship car, gaming consoles) to maximize the click-through rate of users on the newspaper website. At any given time, the agency signs contracts for *new* campaigns with new companies. The information about product features and user profiles form the context, whose distribution changes across campaigns due to change in user traffic and updated product lineups. This could also cause shifts in user preferences. In practice, the agency already has a finite inventory of ad-recommendation models (a.k.a. experts, typically logistic models for their very low inference delays of micro-seconds that is mandated by real-time user traffic) from past campaigns. On a new campaign, online ad agencies *bid* for slots in news media outlets, depending on the profile of the user that visits their website, using these pre-learned experts (see References [28, 38]). In this setting, agencies only re-learn which experts in their inventory work best for their new campaign, possibly fine-tuning them across campaigns. Our work models this episodic setup, albeit without fine tuning of experts between campaigns.

In this article, we study the contextual bandit problem with stochastic experts in the episodic setting. In the single-episode case, we develop an Importance Sampling–based strategy that shares information across experts and provides horizon-independent regret guarantees when expert policies and context distributions are known. In the episodic case, we generalize our methods to function with approximate knowledge of these quantities.

1.1 Main Contributions

We formulate the Contextual Bandit with Stochastic Experts problem. Here, an agent interacts with an environment through a set of N experts. Each expert i is characterized by a fixed and unknown conditional distribution over actions in a set $\mathcal{V}$ given the context from $\mathcal{X}$. We also extend this to the episodic case, where this agent-environment interaction is carried out over E episodes. At the start of episode e, the context distribution $p_e(\cdot)$ as well the distribution of rewards $q_e(\cdot|v, x)$ changes and remains fixed over the length of the episode denoted by T. At each time, the agent observes the context X, chooses one of the N experts, and plays the recommended

action V to receive a reward Y. Note here that the expert policies remain invariant across **all** episodes.

The goal of the agent is to track the best expert to maximize the cumulative sum of rewards. Here, the best expert is one that generates the maximum average reward averaged over the randomness in contexts, recommendations, and rewards. In the episodic setting, the agent seeks to track the *episode-dependent* best expert, which may change across episodes due to differences in the environment. Due to the stochastic nature of experts, we can use **Importance Sampling (IS)** estimators to share reward information to leverage the information leakage.

Our main contributions are as follows:

(1) Divergence-based Upper Confidence Bound (D-UCB) Algorithm: We develop the D-UCB algorithm (Algorithm 1) for the contextual bandit with stochastic experts problem, which employs a clipped IS-based estimator to predict expert rewards. We analyze this estimator and show exponentially fast concentrations around its mean in Theorem 1. We also provide horizon-independent regret guarantees for D-UCB in Theorem 3 that scale as $O(C_1 N)$ with N experts where C_1 is a problem-dependent constant. Additionally, we extend this to the case where the expert policies are only known approximately and provide the **Empirical D-UCB (ED-UCB)** algorithm (Algorithm 3) for this setting. We also show that with well-approximated experts, using ED-UCB leads to regret performance that scales as $O(C_2 N)$ with C_2 as another constant (similar to that of the full-information setting of D-UCB) in Theorem 6.

Further, in Section 4.2, we also show that, with some mild assumptions, these regret bounds can be improved to $O(\log N)$ and present strategies to improve computational complexity of our algorithms at the cost of some regret.

(2) Theoretical Contributions: Authors in Reference [30] study the best-arm identification problem in our setting and design a successive elimination algorithm wherein the sequence of expert plays in each epoch is decided before any samples are observed. Thus, they provide concentration results for Clipped IS-based estimators in the setting where the number of samples collected under each expert is *a priori* known using Chernoff-type analyses.

In contrast, we study the cumulative regret setting in this work and develop a Upper Confidence Bound-style *randomized bandit algorithm* to choose experts at each timestep. This results in the number of samples under each expert at any time being a random quantity, which restricts the use of Chernoff bounds. We prove online concentration bounds for the Clipped IS-based estimator (Theorems 1, 14) that are valid under any *arbitrary causal policy*; that is, any policy that chooses experts based purely on observations made in the past. We achieve this by analyzing a carefully constructed martingale and show that the mean of the estimator concentrates around the true mean of the expert exponentially fast, similar to the deterministic sample setting above due to the information leakage across experts . To the best of our knowledge, these types of results have not been established, and our technique may be of independent interest.

(3) Episodic Behavior with Bootstrapping: In Section 6, we also specify the construction of the approximate experts used by ED-UCB in the case when the supports of X, V are finite in the episodic setting. We show that if the agent is bootstrapped with $O(|\mathcal{V}| \log(T) + \log(|\mathcal{X}|NT\sqrt{E}))$ samples per expert, then the use of ED-UCB over E episodes, each of length T, guarantees a regret bound of $O(E(N+1) + N\sqrt{E}/T^2)$, where the dominant term does not scale with T. These are presented in Theorem 8 and Corollary 8.1, respectively. Our regret bound lies in between those of D-UCB in the full information setting and naive optimistic bandit policies (e.g., UCB in Reference [2], KL-UCB in Reference [9]), demonstrating the merits of sharing information among experts. We also mention how our methods can be extended to continuous context spaces in Section 7.

(4) Empirical evaluation: We validate our findings empirically through simulations on the MovieLens 1M [18] and CIFAR10 [19] datasets in Section 8. We provide soft-control on the genre of the recommendation for users that are clustered according to age and show that the performance of ED-UCB is comparable to that of D-UCB and significantly better than the naive strategies for multi-armed bandits.

1.2 Related Work

Adapting to changing environments forms the basis of meta-learning [5, 34], where agents learn to perform well over new tasks that appear in phases but share underlying similarities with the tasks seen in the past. Our approach can be viewed as an instance of meta-learning for bandits, where we are presented with varying environments in each episode with similarities across episodes. Here, the objective is to act to achieve the maximum possible reward through bandit feedback, while also using the past observations (including offline data, if present). This setting is studied in Reference [4], where a finite hypothesis space maps actions to rewards with each phase having its own true hypothesis. The authors propose a UCB-based algorithm that learns the hypothesis space across phases while quickly learning the true hypothesis in each phase with the current knowledge. Similarly, linear bandits where instances have common unknown but sparse support is studied in Reference [36]. In References [11, 20], meta-learning is viewed from a Bayesian perspective, where in each phase an instance is drawn from a common meta-prior that is unknown. In particular, Reference [11] studies meta-linear bandits and provides regret guarantees for a regularized ridge regression, whereas Reference [20] uses Thompson sampling for general problems, with Bayesian regret bounds for K-armed bandits.

Collective learning in a **fixed** and contextual environment with bandit feedback, where the reward of various arms and context pairs share a latent structure, is known as Contextual Bandits (References [1, 3, 7, 13–15, 21, 32] among several others), where actions are taken with respect to a context that is revealed in each round. In various works [1, 16, 17, 32], a space of hypotheses is assumed to capture the mapping of arms and context pairs to reward, either exactly (realizable setting) or approximately (non-realizable), and bandit feedback is used to find the true hypothesis that provides the greedy optimal action, while adding enough exploration to aid learning.

In the context of online learning, **Importance Sampling (IS)** is used to transfer knowledge about random quantities under a known target distribution using samples from a known behavior distribution in the context of off-policy evaluation in reinforcement learning [26]. Clipped IS estimates are also commonly studied to reduce the variance of the estimates by introducing a controlled amount of bias [8, 12, 23, 30]. Bootstrapping from prior data has been used in References [16, 37] to warm-start the online learning process.

Meta-learning algorithms take a model-based approach, where the invariant-structure (hypothesis space in Reference [4] or meta-prior in References [11, 20]) is first learned to make the optimal decisions, while most contextual bandit algorithms are policy-based, trying to learn the optimal mapping by imposing structure on the policy space. Our approach falls in the latter category of optimizing over policies (a.k.a. experts) from a given finite set of policies. However, contrary to the commonly assumed deterministic policies, each policy in our setting is given by fixed distributions over arms conditioned on the context (learned by bootstrapping from offline data). Using the estimated experts, in each episode (where both the arm reward per context and context distributions change), we quickly learn the average rewards of the experts by collectively using samples from all the experts.

Previously, in Reference [31], we studied the non-episodic version of this problem in the full-information setting where expert policies and context distributions are known to the agent. The strategy presented therein guarantees a worst-case regret that scales logarithmically in the

time-horizon. Their analysis built upon the results of Reference [30] that studied a best-arm identification variant of the problem. In this work, we tighten the existing guarantee and prove a constant (time-independent) regret bound. Further, we provide techniques that reduce the computation complexity to be logarithmic in the number of experts per time (previously linear in the number of experts per time). We then generalize the full information setting to the case of empirical experts and unknown context distributions and also study the episodic version of the problem.

2 Problem Setting

An agent interacts with a contextual bandit environment through a set of experts $\Pi = \{\pi_1, \dots \pi_N\}$. At each time t, the agent observes a context $X_t \in \mathcal{X}$ drawn independently from a fixed but unknown distribution $p(\cdot)$. The agent then selects an expert π_{k_t} (or simply, expert k_t) that recommends an action V_t from a finite set of actions $\mathcal{V}$. The agent then receives a reward $Y_t \in [0, 1]$ distributed according to $q(\cdot|X_t, V_t)$.

Given a context X, the action recommended by expert i is sampled from the conditional distribution $\pi_i(\cdot|X)$. The choice of expert at this time t can be instructed by the set of historical observations $(X_n, k_n, V_n, Y_n)_{n<t}$ and the current context X_t. We assume that the agent is only given access to *all* conditional distributions in Π, while the context and reward distributions are *unknown*.

Regret: The objective of the agent in our contextual bandit problem is to perform competitively against the "best" expert in Π. We define $p_k(x, v, y) \triangleq p(y|v, x)\pi_k(v|x)p(x)$ as the distribution of the corresponding random variables when the expert chosen is $\pi_k \in \Pi$. The expected reward of expert k is then denoted by, $\mu_k = \mathbb{E}_k[Y]$, where $\mathbb{E}_k$ denotes expectation with respect to distribution $p_k(\cdot)$. The best expert is given by $k^* = \arg\max_{k\in[N]}\mu_k$.

The goal of the agent is to minimize the *regret* till time T, which is defined as $R(T) = \sum_{t=1}^{T} \mu^* - \mathbb{E}[\mu_{k_t}]$, where $\mu^* = \mu_{k^*}$. Note that this is analogous to the regret definition for the *deterministic expert* setting of Reference [22]. We use $\Delta_k \triangleq \mu^* - \mu_k$ as the optimality gap in terms of expected reward for expert k. Let $\mu \triangleq \{\mu_1, \dots, \mu_N\}$. We further assume that for all $i \in [N]$, $\mu_i \geq \gamma$.

Remark. In contextual bandits, the best arm can change with the revealed context. However, learning this context-dependent best arm is often not tractable in practice when the arm/context spaces are large. As an alternative, access to a set of functions (or experts), each mapping contexts to actions, is assumed, and the learner is now tasked with competing with the best expert in this given set [1, 3, 32, 33]. This notion of regret is especially useful when the number of experts N is much smaller than the $|\mathcal{V}|^{|\mathcal{X}|}$, where $\mathcal{V}$ is the set of arms and $\mathcal{X}$ is the set of contexts. Such experts are usually learned using offline data and in combination with domain-specific knowledge. For more details around this, we refer the reader to Chapter 18 of Reference [24].

3 Clipped Importance Sampling-based Estimator

The experts being modeled as conditional distributions over arms given contexts allows us to leverage IS to use rewards collected under one expert to estimate the rewards of all other experts. Mathematically, the expected reward of expert k can be written as

$$\mu_k = \mathbb{E}_k[Y] = \mathbb{E}_j\left[Y\frac{\pi_k(V|X)}{\pi_j(V|X)}\right]. \tag{1}$$

Note here that, after the second equality, the expectation is taken under expert j and the rewards are re-weighted appropriately. This has been termed as *information leakage* and has been leveraged before in the literature of best-arm identification [6, 23, 30].

The equation above is only valid for infinitely many samples from expert j. Further, the small values of the denominator $\pi_j(V|X)$ can introduce large variances in a standard empirical estimator of Equation (1). To work around these issues, we use a Clipped IS based estimator, similar to that introduced in Reference [30].

We first define an f-divergence metric that is crucial to the design and analysis of this estimator. We define the conditional f-divergence as follows:

Definition 1. Let $f(\cdot)$ be a non-negative convex function such that $f(1) = 0$. For two joint distributions $p_{X,Y}(x, y)$ and $q_{X,Y}(x, y)$ (and the associated conditionals), the conditional f-divergence $D_f(p_{X|Y}\|q_{X|Y})$ is given by: $D_f(p_{X|Y}\|q_{X|Y}) = \mathbb{E}_{q_{X,Y}}[f(\frac{p_{X|Y}(X|Y)}{q_{X|Y}(X|Y)})]$.

Recall that π_i is a conditional distribution of V given X. Thus, $D_f(\pi_i\|\pi_j)$ is the conditional f-divergence between the conditional distributions π_i and π_j. We now introduce the M_{ij} measure:

Definition 2. (M_{ij} measure) [30] Consider the function $f_1(x) = x\exp(x-1) - 1$. We define the following log-divergence measure: $M_{ij} = 1 + \log(1 + D_{f_1}(\pi_i\|\pi_j))$, $\forall i, j \in [N]$.

We also assume that the divergence between any two experts is upper bounded by $M < \infty$, i.e., $M \geq \max_{i,j\in[N]} M_{ij}$. This immediately implies that every expert in $[N]$ recommends *every* action in $\mathcal{V}$ with a non-zero probability for *every* context in $\mathcal{X}$.

To define our Clipped IS-based estimator, we recall that the history of observations until time t includes the set $\{X_n, k_n, V_n, Y_n\}_{n<t}$ and the context X_t. To ease notation, we will write $r_{ik_t}(t) = \frac{\pi_i(V_t|X_t)}{\pi_{k_t}(V_t|X_t)}$ to be the IS ratio between expert i and the expert k_t chosen at time t.

The estimator for the mean reward of expert i at time t is defined as:

$$\hat{\mu}_i(t) = \frac{1}{Z_i(t)} \sum_{s=1}^{t} \frac{Y_s}{M_{ik_s}} r_{ik_s}(s) \cdot \mathbb{1}\left\{ r_{ik_s}(s) \leq 2\log\left(\frac{2}{\epsilon(t)}\right) M_{ik_s} \right\}. \tag{2}$$

Here, $Z_i(t) = \sum_j N_j(t)/M_{ij}$ is the normalizing constant for expert i with $N_j(t)$ as the number of times expert j has been selected by time t. The bias-variance trade off is controlled by the adjustable term $\epsilon(t)$. Formally, we define $\epsilon(t) = Cw(\frac{\sqrt{t\log t}}{Z_i(t)})$, where C is a constant. The function $w(\cdot)$ is defined as $w(x) = y \iff y/\log(2/y) = x$.

Intuition: The clipped IS estimator is a weighted average of the samples collected under different experts, where each sample is scaled by the importance ratio as suggested by 1. At each time t, the adjustable term $\epsilon(t)$ is calculated to be used in the clipper levels. Then, the estimator is recomputed to include the re-weighted samples collected under other experts in the past that fall below the new clipper levels. Observe here that, since $\epsilon(t)$ decreases with time, the clipper levels are increasing and thus, this estimator is asymptotically consistent.

Compared to the vanilla IS estimator in Equation (1), the clipped IS estimator above drops samples with large importance sampling ratios (due to the indicator function), leading to a biased estimate of the true mean μ_i. However, as observed by authors in Reference [6], adding a controlled amount of bias to an importance sampling estimator helps its concentration behavior due to reduced variance. This variance reduction is thanks to the fact that the clipping leads to an estimate bounded within a smaller range compared to the vanilla (potentially unbounded) estimator. Additionally, the clipper-level values and the weights are dependent on the divergence terms M_{ij}'s. When the divergence M_{ij} is large, it means that the samples from expert j is not valuable for estimating the mean for expert i. We will show in Theorem 1 that this leads to exponential concentrations of the clipped IS estimate $\hat{\mu}_i(t)$ around the true mean μ_i.

4 Divergence-based Upper Confidence Bound Algorithm

Recall that the goal of the agent is to remain competitive with the best expert in the given set of experts Π. We propose an optimistic index-based algorithm motivated by the popular **Upper Confidence Bound (UCB)** algorithm for stochastic Multi-armed Bandits [2]. Our algorithm uses the Clipped IS estimators developed in the previous section to compute a high-probability upper confidence bound for the mean of each expert. At each time, the policy chooses experts greedily according to these UCB's. The **Divergence-based UCB algorithm (D-UCB)** is summarized in Algorithm 1.

ALGORITHM 1: D-UCB: Divergence-based UCB for Contextual Bandits with Stochastic Experts

1: For timestep $t = 1$, observe context X_1 and choose a random expert $\pi \in \Pi$. Play an arm drawn from the conditional distribution $\pi(V|X_1)$.
2: **for** $t = 2, \ldots, T$ **do**
3: Observe context X_t
4: Let $k_t = \text{argmax}_k \, U_k(t-1) \triangleq \hat{\mu}_k(t-1) + s_k(t-1)$.
5: Sample action V_t from the distribution $\pi_{k_t}(\cdot|X_t)$.
6: Observe the reward Y_t.
7: **end for**

Here, the confidence bonus in Algorithm 1 for the estimator $\hat{\mu}_k(t)$ at time t is chosen as $s_k(t) = \frac{3}{2}\epsilon(t)$.

4.1 Regret of D-UCB

In this section, we discuss the performance of D-UCB. To upper bound the expected regret incurred by our algorithm by time T, we require concentration guarantees for the estimators $\hat{\mu}_k(T)$. In Reference [30], the authors provide concentration bounds for these estimators assuming that the number of times an expert is played by time T is known. However, in our case, since the experts are chosen in an online fashion, the number of times an expert is chosen by time T is a random variable upper bounded by T. Since the existing analysis does not apply in this case, we prove the following Lemma using martingale concentrations in place of the Chernoff bounds used previously to account for the random nature of expert plays. This lemma provides exponentially fast concentrations for the estimator in Equation (2) around the true mean of the corresponding expert.

Theorem 1. *For any expert $j \in [N]$, the estimator $\hat{\mu}_j(t)$ defined in Equation (2) satisfies*

$$\mathbb{P}\left((1-\beta(t))\left(\mu_j - \frac{\epsilon(t)}{2}\right) \le \hat{\mu}_j(t) \le (1+\beta(t))\mu_j\right) \ge 1 - 2\exp\left(-\frac{\gamma^2\beta(t)^2 t}{128M^2(\log(2/\epsilon(t)))^2}\right),$$

when $\beta(t)$ and $\epsilon(t) < \gamma$ are fixed non-negative constants and $M \ge \max_{i,j} M_{ij}$ is the finite *upper bound on the divergence between any two experts.*

By design, in the D-UCB (Algorithm 1), the mean estimates of *all* experts are updated at each time. This departs from the stochastic bandit variant, where the mean of *only* the played arm is updated. Therefore, we expect that after sufficient time has passed, the mean estimates of all the experts are close to their true counterparts. Indeed, we show that this intuition holds and, to this end, we define a series of problem-dependent times τ_i for all $i \in [N]$ as follows:

$$\tau_1 = \min\left\{t : Cw\left(\sqrt{\log t/t}\right) \le \gamma\right\}, \; \tau_k = \min\left\{t \ge \tau_1 : \frac{t}{\log t} \ge \frac{9C^2M^2\log^2\left(6C/\Delta_k\right)}{\Delta_k^2}\right\}. \tag{3}$$

Remark 2. We note here that the times τ_i for all $i \in [N]$ defined above are ***deterministic constants*** that do not scale with time t. In particular, these are not the random number of times expert i is played, which is the notation commonly used in bandit literature.

Recall here that $\Delta_k = \mu^* - \mu_k$ is the suboptimality gap of expert k, $M = \max_{i,j\in[N]} M_{ij}$. Without loss of generality, we assume that the experts are indexed such that $0 = \Delta_1 < \Delta_2 \leq \Delta_3, \ldots \Delta_N$ and thus, $\tau_N \leq \tau_{N-1} \cdots \leq \tau_2$.

With τ_k as defined above, using the results of Theorem 1, we can establish the following statements:

(1) For all $t \geq \tau_1$, $\mathbb{P}(U_{k^*}(t) \leq \mu^*) \leq t^{-2}$.
(2) For any $k : \Delta_k > 0$, for all $t \geq \tau_k$, $\mathbb{P}(U_k(t) \geq \mu^*) \leq t^{-2}$.

Together, these two statements give us the following corollary:

Corollary 2.1. *For any suboptimal expert $k : \Delta_k > 0$, at any time $t \geq \tau_k$, $\mathbb{P}(k_t = k) \leq 2/t^2$.*

Since the above is true for each time t after τ_k, since $\lim_{T\to\infty} \sum_{t=1}^{T} 1/t^2 = \pi^2/6$, we can show that expert k is only chosen a constant number of times from then on. Extending this argument to all suboptimal experts, we can conclude that after time τ_2, the optimal expert is played all but a constant number of times. This leads to our main *constant regret* result below:

Theorem 3. *The regret incurred by Algorithm 1 by time T is upper bounded by*

$$R(T) \leq \frac{\pi^2}{3} \sum_{k=2}^{N} \Delta_k + \tau_N \Delta_N + \sum_{k=2}^{N-1} (\tau_k - \tau_{k+1}) \Delta_k.$$

Remarks. **(1) Value of C:** In the above, we use $C = 16M/\gamma$ with γ the lower bound on the reward of all experts and M the upper bound on the divergence between any pair of experts. However, our empirical evaluations show that we observe that smaller values of C also lead to constant regret.

(2) About M: We have assumed that all experts recommend all actions with non-zero probability, guaranteed by $M < \infty$. While this assumption leads to clean analysis for constant regret, it is not clear if it is necessary. For example, suppose there exists an action $v \in \mathcal{V}$ that is only ever played by only one expert in $[N]$ with a low probability under any context $x \in \mathcal{X}$. It is reasonable that the effect of this action on the mean reward of this expert is low and can be upper bounded. Hence, we might be able to recover constant regret by simply ignoring this action in the cases where the rewards obtained by this action are low. This would require a more careful regret analysis that builds on the methods used to prove Theorem 3.

(3) Comparing to Linear Bandits: The structure we impose on the experts and their interactions with the actions leads to the leakage of information across experts. Indeed, other settings, most famously that of linear bandits [25], also share similarities in that playing one action leads to non-trivial information about other actions. However, linear bandit-like formulations can not achieve the constant regret guarantees we are able to provide. This is due to the fact that, in our setting, playing one arm explicitly generates a "pseudo-sample" for *every* other arm using the clipped importance sampling estimators. When the clipper levels are large enough, this leads to us essentially operating in the full-information setting, albeit with a potentially larger variance of reward per expert. However, in a linear bandit, this property can not be guaranteed—first, there can be no information leakage in orthogonal directions and, further, rewards from one arm can not be scaled uniformly to infer rewards from another.

4.2 Reducing Computation

In each round t of the D-UCB algorithm (Algorithm 1, the agent computes mean estimates of each of the N experts using all the observations up to time $t-1$ to form the respective UCB indices. Recall that the clipper levels of the estimator defined in Equation (5) are increasing over rounds. Hence, updating these at each round is computationally expensive. To work around this cumbersome update procedure, we introduce a variant of D-UCB called D-UCB-lite in Algorithm 2. Here, at each time, the algorithm updates only a small subset of the experts using only a fraction of the collected samples. The rationale being that if we carefully choose the rate at which each arm is updated over time, then, provided enough samples, the algorithm should still be able to distinguish between the true best expert and the remainder.

ALGORITHM 2: D-UCB-lite: D-UCB with Intermittent Updates

For timestep $t = 1$, observe context X_1 and choose a random expert $\pi \in \Pi$. Play an arm drawn from the conditional distribution $\pi(V|X_1)$.
For each expert, set $B_k(1) = U_k(1)$ and initialize $N_k(t) = 1$.
for $t = 2, \ldots, T$ **do**
  Observe context X_t
  Let $k_t = \text{argmax}_k\, B_k(t-1)$.
  Sample action V_t from the distribution $\pi_{k_t}(\cdot|X_t)$, observe the reward Y_t.
  Sample a subset $S'(t)$ of experts of size $\log N - 1$ uniformly from $[N]\backslash\{k_t\}$.
  for $k \in S'(t) \cup k_t$ **do**
    if $N_k(t) \leq \lfloor t^{\frac{1}{\beta}} \rfloor$ **then**
      $B_k(t) = U_k(\lfloor t^{\frac{1}{\beta}} \rfloor)$, $N_k(t) = \lfloor t^{\frac{1}{\beta}} \rfloor$.
    else
      $B_k(t) = B_k(t-1)$.
    end if
  end for
end for

This algorithm is similar to D-UCB in all aspects other than the index update rule. In contrast, D-UCB-lite first samples a $(\log N - 1)$-sized subset $S'(t)$ of experts at time t. For each expert, it maintains a counter $N_k(t)$ that is updated to be the number of samples used to build a current index of expert k. The algorithm then updates the indices of the chosen expert k_t and those in $S'(t)$ only if they have not used $\lfloor t^{\frac{1}{\beta}} \rfloor$ samples by time t for some chosen $\beta \geq 1$. In the worst case, this algorithm updates $\log N$ experts with $\lfloor t^{\frac{1}{\beta}} \rfloor$ samples each at every timestep (rather than N experts with t samples each).

Specifically, at round T, D-UCB incurs a computational cost that scales as $O(TN)$. This is because D-UCB needs to update the estimates of each of the N experts at the new clipper level at time T using all T collected samples. In contrast, D-UCB-lite has a computational complexity of $O(T^{\frac{1}{\beta}} \log N)$ at round T. The intermittent update strategy naturally loses performance compared to D-UCB. The following lemma specifies this loss:

Lemma 4. *With $c = N/\log N$, define $\tau_{1,\beta} = \min\{t : t - 2c\log t \geq \tau_1^{\beta}\}$ and for any $k \neq 1$, $\tau_{k,\beta} = \min\{t \geq \tau_{1,\beta} : t - 2c\log t \geq \tau_k^{\beta}\}$. The regret of D-UCB-lite in Algorithm 2 can be bounded as*

$$R(T) \leq \tau_{N,\beta}\Delta_N + \frac{\pi^2}{3}\sum_{k=2}^{N}\Delta_k + \sum_{t=1}^{T}\frac{2\Delta_N}{\lfloor t - 2c\log t \rfloor^{\frac{2}{\beta}}}.$$

Recall here that τ_i are defined in Equation (3). The final summation converges for all values of $\beta < 2$ recovering the time-independent regret guarantees as in Theorem 3, albeit with a larger numerical value.

In Appendix D, we also discuss a specific generative model under which our regret scales *logarithmically* in the number of experts N (as opposed to the linear scaling suggested by Theorem 3).

5 Empirical Experts and Unknown Context Distributions

In the previous sections, we developed the clipped importance sampling estimator that relied crucially on the knowledge of the expert distributions $\pi_i(V|X)$ for all $i \in [N]$. Further, to compute the divergence metric M_{ij}, the agent also needs to use the context distribution chosen by nature. In practice, however, it is often the case that these two distributions are not readily accessible to the policy designer.

The most natural alternative to knowing the expert policies explicitly is to estimate them empirically using offline samples and use these estimates in D-UCB to compute the mean estimates. These expert policies appear in our mean estimators in Equation (2) in the form of the IS ratios and are used to scale rewards as well as decide the state of the clippers. Therefore, to maintain our constant regret guarantee, we not only need to control for error in the estimate of $\pi_k(V|X)$, but also in the IS ratios $r_{ij}(V|X)$ for each pair of experts.

Similarly, in the case where context distributions are unknown, empirical estimation can be carried out to produce estimates of the M_{ij} divergence measures to be used in the reward estimators. Unfortunately, this is not feasible when the size of context set $\mathcal{X}$ is large. To work around this, we will only assume access to a universal lower bound on the probability of occurrence of any context.

The rest of this section will be devoted to modifying the Clipped Importance Sampling estimator in Equation (2) to use the empirical estimates of the corresponding expert policies.

5.1 Empirical Clipped IS-based Estimator

We begin by defining our empirical expert policies and translating their error to the error in the IS ratios. Recall that we write $r_{ij}(V|X) = {}^{\pi_i(V|V)}/_{\pi_j(V|X)}$ as the IS ratio between experts i and j and further abuse this by denoting $\hat{r}_{ij}(V|X)$ as the empirical IS ratio between experts i and j. That is, for empirical estimates $\hat{\pi}_k$ for all $k \in [N]$, we have $\hat{r}_{ij}(V|X) = {}^{\hat{\pi}_i(V|X)}/_{\hat{\pi}_j(V|X)}$.

Proposition 5. *Suppose ξ to be the maximum error in the empirical expert policies, i.e., $\forall i \in [N], \|\pi_i - \hat{\pi}_i\|_\infty \leq \xi$. Additionally, let $\pi_i(V|X) \geq p_V > 0$ for all i, v, x. Then, the following hold for all $(x, v) \in \mathcal{X} \times \mathcal{V}$:*

$$r_{ij}(v|x) \geq \underline{r}_{ij}(v|x) := \hat{r}_{ij}(v|x) - \frac{\xi}{p_V(p_V - \xi)},\ r_{ij}(v|x) \leq \overline{r}_{ij}(v|x) := \hat{r}_{ij}(v|x) + \frac{\xi}{p_V(p_V + \xi)}.$$

The proof of the above follows from Lemma 5.1 in Reference [10]. We note here that the assumption of a universal lower bound p_V on the probability mass $\pi_i(V|X)$ is necessary to guarantee finiteness of the IS ratio. If for any expert j this does not hold, then given samples from this expert, we would not be able to infer anything meaningful about any of the other experts. This is also true for our discussions in the previous section when the expert policies were known completely where this assumption was implicit in that $M = \max_{i,j \in [N]} M_{ij}$ was assumed to be finite.

We now introduce the Empirical Clipped Importance Sampling–based Estimator, similar to its non-empirical counterpart in Equation (2). To this end, we define the following natural lower

bound to the M_{ij} divergence measures in Definition 2.

$$\underline{D}_{f_1}(i||j) \triangleq p_X \sum_{x\in\mathcal{X}} \sum_{v\in\mathcal{V}} \left(\hat{\pi}_j(V|X) - \xi\right) f_1\left(\underline{r}_{ij}(V|X)\right), \ \underline{M}_{ij} \triangleq 1 + \log(1 + \underline{D}_{f_1}(i||j)). \tag{4}$$

Recall here that $f_1(x) = x\exp(x-1) - 1$. Additionally, $p_X > 0$ is the lower bound on the occurrence of any context, i.e., $p_X \geq \min_{x\in\mathcal{X}} p(x)$. We are now ready to define our empirical reward estimator. Suppose $(X_s, k_s, V_s, Y_s)_{s<t}$ is the historical observations up to time t and X_t is the provided context. We define our empirical clipped IS estimator for the mean of expert i at time t as

$$\tilde{Y}_i(t) = \frac{1}{Z_i(t)} \sum_{s=1}^{t} \left(\frac{Y_s}{\underline{M}_{ik_s}} \underline{r}_{ik_s}(s) \cdot \mathbb{1}\left\{ \overline{r}_{ik_s}(s) \leq 2\log\left(\frac{2}{\epsilon_i(t)}\right) \underline{M}_{ik_s} \right\} \right). \tag{5}$$

As before, we have $Z_i(t) = \sum_{s=1}^{t} 1/\underline{M}_{ik_s}$ for k_s the expert selected at time s and $\epsilon(t)$ is the term that balances bias and variance (defined using this new version of $Z_i(t)$).

The empirical IS estimator $\tilde{Y}_k(t)$ is designed to be underestimate of the true mean μ_k, on average. Additionally, its mean is also lesser than the mean of the full information IS estimator in Equation (2), making it further biased. We compensate for this by enlarging the confidence bonus ($s_k(t)$ in Section 3) by the maximum bias that $\tilde{Y}_k(t)$ can have by time t.

The UCB index of expert $i \in [N]$ at time t is set to be

$$U_i'(t) = \tilde{Y}_i(t) + e_i(t) + \frac{3}{2}\epsilon_i(t), \ e_i(t) = \max_{j\in[N],(v,x)\in\mathcal{V}\times\mathcal{X}} \left[\overline{r}_{ij} - \underline{r}_{ij} \mathbb{1}\left\{ \overline{r}_{ij} \leq 2\log\left(\tfrac{2}{\epsilon_i(t)}\right) \underline{M}_{ij} \right\} \right]. \tag{6}$$

The index $U_i'(t)$ consists of three terms: the first is the estimate we define in Equation (5). The error term $e_i(t)$ captures the maximum error in our estimate due to the inherent inaccuracies in the empirical expert policies. Finally, the last term defines the length of the upper confidence interval, which is characterized by $\epsilon_i(t)$ (which appears in the clipper level of the estimator). It can be shown that with high probability, this index $U_i'(t)$ is an over-estimate of the true mean μ_i of expert i. We summarize the ED-UCB algorithm in Algorithm 3.

ALGORITHM 3: ED-UCB: Empirical D-UCB

1: **Inputs:** Empirical experts $\hat{\pi}_i$ for $i \in [N]$ with maximum error ξ, lower bounds p_X, p_V.
2: For timestep $t = 1$, observe context X_1 and choose a random expert $\pi \in \Pi$. Play an arm drawn from the conditional distribution $\pi(V|X_1)$.
3: **for** $t = 2, \ldots, T$ **do**
4: Observe context X_t
5: Let $k_t = \text{argmax}_k\, U_k'(t-1)$ with $U_k'(t)$ as in Equation (6)
6: Observe the realization of the arm V_t and reward Y_t.
7: **end for**

5.2 Regret Bounds

The order of operations in proving regret bounds for ED-UCB are similar to that of D-UCB in 4.1. The key difference is that concentration bounds are now required on the estimator in Equation (5) instead of its D-UCB counterpart that assumes access to all expert policies. We defer the development of these bounds to the Appendix. As done previously, we then produce a sequence of times τ_i' for $i \in [N]$ as follows:

$$T_{clip} = \min\{t : \forall i, j \in [N], \overline{r}_{ij} \leq 2\log(2/\epsilon_i(t))\underline{M}_{ij}\}$$
$$\tau_1' = \min\left\{t \geq T_{clip} : C w\left(\sqrt{\log t/t}\right) \leq \gamma\right\},$$
$$\forall i : \Delta_i > 0, \tau_i' = \min\left\{t \geq \tau_1' : \frac{t}{\log t} \geq \frac{9C^2M^2\log^2\left(\frac{6C}{(\Delta_i - \gamma p_V)}\right)}{(\Delta_i - \gamma p_V)^2}\right\}. \quad (7)$$

These times can be used to a similar result to Corollary 2.1. Our final regret result follows:

Theorem 6. *Suppose the empirical estimation error ξ is such that for all experts $i \in [N]$,*

$$\|\pi_i - \hat{\pi}_i\|_\infty \leq \xi \leq 2\frac{\sqrt{1 + p_V^4\gamma^2} - 1}{p_V\gamma}.$$

Consider τ_k' for $k \in [N]$ as defined in the above. Then, for $T \geq \tau_2' > 0$, the expected cumulative regret of ED-UCB is bounded as

$$R(T) \leq \frac{\pi^2}{3}\sum_{k=2}^{N}\Delta_k + \tau_N'\Delta_N + \sum_{k=2}^{N-1}\left(\tau_k' - \tau_{k+1}'\right)\Delta_k := R(\boldsymbol{\Delta}),$$

where $\boldsymbol{\Delta} = (\Delta_2, \ldots, \Delta_N)$ is the vector of suboptimality gaps.

Remarks. **(1) Value of constants:** To analyze ED-UCB, we use $C = {}^{32M}/_{\gamma(1-p_V)}$. However, as with D-UCB, our empirical results suggest that smaller values suffice to achieve constant regret.

(2) Dependence on *N* and intermittent updates: We note here that all the discussions carried out in Section 4.2 also apply in the empirical case above. That is, ED-UCB also enjoys the improved scaling in N under the generative model for suboptimality gaps, and the computational complexity can be improved via the intermittent update strategy.

(3) Updating empirical expert policies online: We assume that the empirical expert policies remain unchanged during the learning process. However, in practice, it is sensible to improve these estimates with the online collected data. Indeed, this can be done, but it makes analyzing the online regret more complex due to evolving expert policy estimates: The form of Theorem 3 depends on a universal error bound for all empirical expert estimates at a level ξ. By updating experts online based on collected samples, it is unclear how this error evolves with time, as the rewards are random. We note that our regret result above is powerful, as it implies that updating expert policy estimates online is *not necessary to achieve constant regret.*

(4) Misspecifying ξ: The legitimacy of the upper bound on expert policy estimation errors ξ is imperative to guarantee the constant regret of ED-UCB. If this value is inaccurate, i.e., $\exists i \in [N] : \|\pi_i - \hat{\pi}_i\|_\infty > \xi$, then we can not guarantee that the UCB index of this expert is strictly below the true mean of the best expert (or above, if this was the best expert). Owing to this, we can no longer guarantee constant regret.

6 Extending to Episodes

In this section, we study the episodic setting. Here, the agent is to act on the environment for a total of E episodes, each with T timesteps. In each episode, the distribution of contexts $p_e(x)$ as well as the reward distribution $q_e(y|v, x)$ is held fixed but can change across episodes. Since the set of experts that the agent accesses remain fixed across episodes, the respective policies can be learned offline and reused in each episode to guarantee constant regret via Theorem 6. We recall our advertising example, where the agency learns the experts in its inventory before deployment and reduces the task of placing advertisements in each campaign to one of selecting the best expert from its roster. In most cases, learning these policies while the recommendations are generating

rewards can be prohibitive due to the size of the context set $\mathcal{X}$ and the action set $\mathcal{V}$. However, in practice (including our advertising example), this learning can be carried out offline by repeatedly querying these experts for recommendations, thus decoupling this process with that of reward accumulation.

Based on this observation, we suggest the use of basic sampling to build empirical expert policies that are bootstrapped into ED-UCB at the start of each episode. This choice helps us to implicitly handle the changes in the reward distributions in each episode. The reward distribution q_e only affects the generated rewards Y_s and the average rewards $\mu_{i,e}$, which affects the suboptimality vector Δ_e. The estimation procedure in Equation (5) adapts to the former, while the latter simply causes the constant upper bound to vary with episodes. Thus, in what follows, we study the bootstrapping process and provide a corollary for the case when the expert policies are learned online. We also abuse notation here by using subscripts of the episode index wherever necessary.

6.1 Bootstrapping with Offline Sampling and Online Sampling

To use empirical estimates in each episode, the agent must learn **all** expert policies sufficiently well, characterized by the maximum error ξ. To this end, we assume that the agent is allowed to sample from a context distribution $p_{prior}(x)$ such that for any $x \in \mathcal{X}$, $p_{prior}(x) \geq p_X > 0$. In the context of our practical advertising example, this can also be thought of as trying to estimate the expert policies from a large pool of data collected from previous ad campaigns. Bootstrapping and warm-starting methods are common in contextual bandit algorithms; for examples, see References [16, 37], among others.

To achieve constant regret per episode, Theorem 6 suggests the use of $\xi = 2\frac{\sqrt{1-p_V^4\gamma^2}-1}{p_V\gamma}$. Using Chernoff bounds, for a fixed expert, under a fixed context, collecting $n = \frac{2|\mathcal{V}|\log(2T)}{\xi^2}$ would imply a maximum error of ξ with probability at least $1 - T^{-1}$ [35]. However, since the sampling procedure is probabilistic through $p_{prior}(x)$, we have the following lemma:

Lemma 7. *Define* $A = \frac{2n}{p_X} + \frac{\log(|\mathcal{X}|NT\sqrt{E})}{2p_X^2}$. *Let* $\mathcal{E}$ *be the event that after sampling each expert* $i \in [N]$ *A times, we acquire at least n samples under each context* $x \in \mathcal{X}$. *Then,* $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{1}{T\sqrt{E}}$.

ALGORITHM 4: Meta-algorithm: ED-UCB for Episodic Bandits with Bootstrapping

1: **Inputs:** Sampling oracles for true agent policies $\pi_i(\cdot|\cdot)$, parameters p_X, p_V, γ, E, T.
2: **Bootstrapping:** Play each expert A times to build approximate experts $\hat{\pi}_i(\cdot|x)$.
3: **Episodic Interaction:**
4: **for** $e = 1, 2, \ldots E$ **do**
5: Play a fresh instance of ED-UCB (Algorithm 3) with parameters $\hat{\pi}_i, p_V, p_X, \gamma$ for T steps.
6: **end for**

This lemma specifies that under the event $\mathcal{E}$, the maximum error in the empirical expert policies is bounded by ξ with probability at least $1 - T^{-1}$. The agent then instantiates ED-UCB at the start of each episode. This process is summarized in Algorithm 4. A straightforward application of the Law of Total Probability leads to the following theorem:

Theorem 8. *The regret of the agent in Algorithm 4 is bounded as*

$$R(T, E) \leq \sqrt{E} + E + \left(\sum_{e=1}^{T} R(\Delta_e)\right)\left(1 + \frac{1}{T^2\sqrt{E}}\right) = O\left(E(N+1) + \frac{N\sqrt{E}}{T^2}\right).$$

Here, $R(\Delta_e)$ is defined as the regret bound defined by Theorem 6 for episode e.

This result extends to the online setting naturally. In this case, the agent spends the first AN timesteps collecting samples and builds the empirical estimates of the experts. After this time, the agent continues as if it were bootstrapped. Since these expert policies do not change with episodes, the agent only incurs this additional AN regret once. We summarize this in the following corollary:

Corollary 8.1. *The online estimation of the estimation oracles adds an additional regret of AN to that in Theorem 8. The total regret of the online process can be bounded as*

$$R(T,E) = O\left(N\log(NT^2\sqrt{E}) + E(N+1) + \frac{N\sqrt{E}}{T^2}\right).$$

Remark. As observed at the end of Section 5.2, using online samples to improve estimates of expert policies in practice could lead to improved regret performance. Even if we were to improve these estimates at the end of each episode, the improvements would depend on the trajectory of observations that are random. Thus, for our analysis, we assume that empirical estimates are not updated after bootstrapping.

7 Discussions

How useful is information leakage: The most natural alternative to considering the information leakage across the experts is to treat each of them independently. In this case, the problem reduces to a standard multi-armed bandit problem, where each expert is treated as an arm. In this case, the optimistic UCB strategies (as in References [2, 9]) explore each suboptimal expert $\log T$ times by time T, thus resulting in an overall regret that scales as $O(N\log T)$. In contrast, leveraging the information leakage through Importance Sampling as in D-UCB does not necessitate any exploration beyond the time τ_1(analogously τ_1' in the empirical formulation) and thus suffers regret that does not scale in the horizon. This carries on to the episodic setting, where naive UCB-type algorithms (that reset after each episode) incur a regret of the order $O(EN\log T)$, but without the need for any bootstrapping.

Lack of lower bound p_V: Our assumption of the knowledge of p_V allows us to develop IS estimates that are finite, leading to constant regret. If there exist a sub-optimal expert that does not satisfy this bound, then it would have to be explored at a logarithmic rate, since there is no information leakage with respect to this expert. However, samples could still be shared among the other experts through the use of D-UCB, making it a stronger baseline than naive policies. The question of optimality in this case remains open.

Infinite context spaces: The context distribution $p_e(x)$ has only been used in the quantities $\underline{M}_{ij}$ and the upper bound M (Equation (4)). For continuous contexts, the results in Section 5 can be extended by assuming knowledge of upper and lower bounds on the density function $p_e(x)$ and swapping out the summation for an integral in Equation (4). This can further be extended to the episodic case by assuming access to approximate oracles for expert policies.

Stochastic episode lengths and unknown change-points: Our analysis extends to the setting where the number of episodes and the length of each episode are random quantities with upper bounds E, T, respectively. In the case where the end of the episode is not communicated to the agent, ED-UCB can potentially be slow to adapt to the modified environment, which leads to linear regret. Thus, tracking the best expert in unknown non-stationary environments is an important avenue for future work.

Lower bounds: The use of Importance Sampling in our estimators implies that samples collected under one expert also serve as samples (up to scaling and clipping) for all other experts. Intuitively, this implies that after some initial exploration, playing the best expert at each time

should provide the agent enough evidence to discredit other experts. Therefore, in the single-episode setting, it is reasonable to expect a time-independent constant lower bound; the upper bounds of ED-UCB and D-UCB are in agreement of this intuition. We note however, that an algorithm that learns the best arm per-context can incur negative linear regret with respect to our defined best expert. Thus, producing lower bounds on regret in our setting and under stronger notions of the best expert as well the number of bootstrapping samples necessary in the episodic case serve as two important avenues of future work.

8 Experiments

We now present numerical experiments to validate our results above. We build two sets of semi-synthetic experiments using the CIFAR-10 [19] and MovieLens 1M [18] datasets; we describe them below. We compare our D-UCB and ED-UCB algorithms with the naive UCB [2] and KL-UCB [9] methods that do not leverage the information leakage. For all our experiments, we set the value of the constant $C = 0.02$ for both D-UCB and ED-UCB. We use the best values of constants for the naive algorithms as suggested by Reference [24].

An Image Classification Setup: Using the CIFAR-10 dataset, we build 5 classifiers. Each of these is trained over data from 9 of the 10 labels, with a different label being omitted per classifier. We use these to build 5 experts: Given an input image, with probability 0.8, the expert recommends the class suggested by its associated classifier or recommends a uniformly random class with probability 0.2. This gives us experts that output each of the 10 classes with probability at least 0.02 for any image. Further, we use classes to form contexts as follows: For an image from class c, the context is provided as "`possibly an image of class` c". We also sub-sample a set of 1,000 test images (100 per class), called the "test set."

Mapping this back to our episodic bandit setup, we have $|\mathcal{X}| = 10$ contexts, $N = 5$ experts, and $|\mathcal{V}| = 10$ arms (or recommended classes). In the online phase, in each episode, we sample a context from the corresponding context distribution $X_t \sim p_e(x)$ and an image from the test set with this (possible) class uniformly at random. We choose an expert according to each of our evaluated algorithms and provide the recommended label as its output. The expert then observes a binary reward: 1 if the output was the true label of the image and 0 otherwise. We set $p_X = 0.05$ and use $p_V = 0.02$ to generate the necessary number of samples to form the empirical expert policies for ED-UCB. The results are averaged over 300 independent runs and presented in Figure 1.

A Movie Recommendation Setup: We use the MovieLens 1M dataset [18] with 1 million ratings of approximately 3,900 movies by 6,000 users to construct a semi-synthetic bandit instance. First, we complete the reward matrix (scaled down to $(0, 1)$) using the SoftImpute algorithm of Reference [27] included in the `fancyimpute` package [29]. We filter the number of movies to 618 using the completed matrix by eliminating ones that are mostly rated 0. Then, we cluster these movies based on seven genres, namely: `Action, Children, Comedy, Drama, Horror, Romance, Thriller` and users based on ages between 0–17, 18–24, 25–49, 49+. At this stage, the average reward of all *(age,genre)* pairs are close to each other. To induce some diversity, we boost the rewards of the following *(age,genre)* pairs by 0.008: (0–17,`Children`), (18–24,`Horror`), (18–24,`Thriller`), (25–49,`Action`), and (25–49,`Drama`).

Experts are then randomly generated over the set of genres randomly with $p_V = 0.002$. Given an age group (context) and genre (expert-recommended action), a movie of the selected genre is picked uniformly, and the reward is obtained from the completed reward matrix. We build empirical expert policies using Theorem 8 with $p_V = 0.2$ and a prior distribution satisfying $p_X = 0.05$ for 5 episodes of length 10^6 each. The averaged results of 100 independent runs are presented in Figure 2.

In both Figures 1 and 2, our Importance Sampling–based policies show large improvements in regret over the naive baselines. In the movie recommendation case of Figure 2, the mismatch in

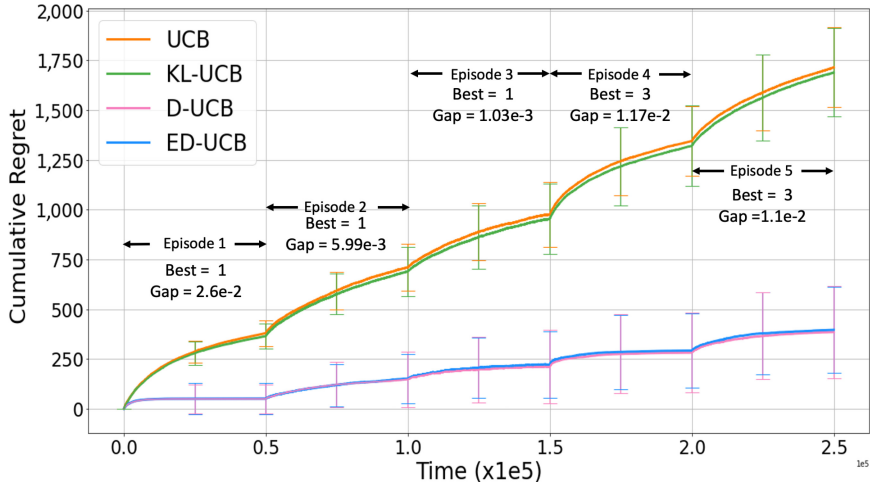


Fig. 1. Experiments on the CIFAR-10 dataset: The experiment consists of 5 episodes with 5×10^5 steps each. Plots are averaged over 300 independent runs, error bars indicate one standard deviation. Indices of the best expert and the minimum suboptimality gaps are presented.

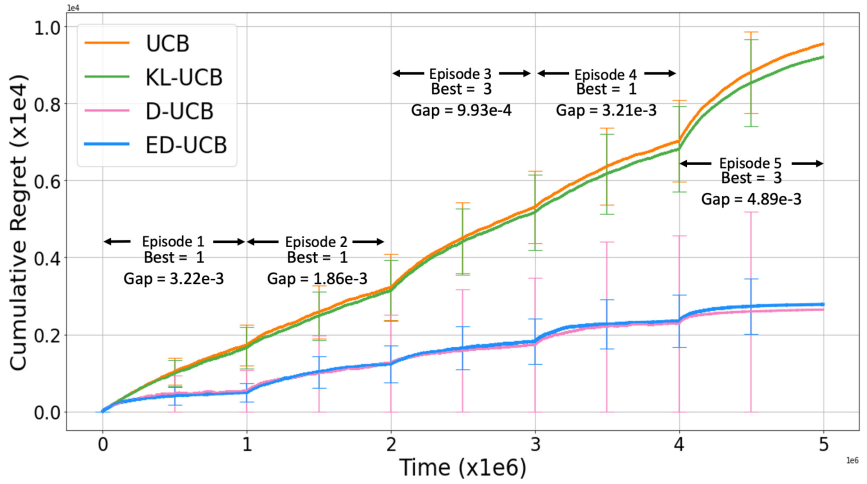


Fig. 2. Experiments on the MovieLens 1M dataset: The experiment consists of 5 episodes with 10^6 steps each. Plots are averaged over 100 independent runs, error bars indicate one standard deviation. Indices of the best expert and the minimum suboptimality gaps are presented.

p_V in the sampling process causes the empirical estimates being formed with fewer samples than theoretically recommended. However, we empirically observe that ED-UCB still heavily outperforms the naive bandit policies and is comparable in regret to D-UCB. Some additional empirical results can be found in Appendix F. These present some cases where the assumptions we make about the environment do not hold.

Appendices

The Appendix is structured as follows: We first discuss the proof of our regret results for ED-UCB (Algorithm 3) in Section 5 at length. The results for D-UCB (Algorithm 1) in Section 4.1 will then

follow simply from these with similar arguments; we provide short proofs of these results. Next, we provide the proofs for our scaling and computational improvements discussed in Section 4.2. Finally, we prove the episodic results from Section 6.

A Useful Concentrations

We begin by proving regret guarantees for ED-UCB in Algorithm 3, which will immediately imply all our results for D-UCB in Algorithm 1. We begin with some basic concentrations. First is a result about L_1 deviations of empirical probability distributions.

Lemma 9. *Let p be a probability vector with S points of support. Let $\hat{p} \sim \frac{1}{n}Multinomial(n, p)$ be an empirical estimate of p using n i.i.d. draws. Then, for any $S \geq 2$ and $\delta \in (0, 1)$, it holds that*

$$\mathbb{P}\left(\|p - \hat{p}\|_\infty \geq \sqrt{\frac{2S\log\left(\frac{2}{\delta}\right)}{n}}\right) \leq \mathbb{P}\left(\|p - \hat{p}\|_1 \geq \sqrt{\frac{2S\log\left(\frac{2}{\delta}\right)}{n}}\right) \leq \delta.$$

Proof. The result follows from that of Reference [35] as $\|x\|_\infty \leq \|x\|_1$ for any vector x. □

Next is the proof of Proposition 5, which provides confidence bounds for ratios of random variables. We restate it here for convenience:

Proposition (Restatement of Proposition 5). *Suppose ξ to be the maximum error in the empirical expert policies, i.e., $\forall i \in [N], \|\pi_i - \hat{\pi}_i\|_\infty \leq \xi$. Additionally, let $\pi_i(V|X) \geq p_V > 0$ for all i, v, x. Then, the following hold for all $(x, v) \in \mathcal{X} \times \mathcal{V}$:*

$$r_{ij}(v|x) \geq \underline{r}_{ij}(v|x) := \hat{r}_{ij}(v|x) - \frac{\xi}{p_V(p_V - \xi)},\ r_{ij}(v|x) \leq \overline{r}_{ij}(v|x) := \hat{r}_{ij}(v|x) + \frac{\xi}{p_V(p_V + \xi)}.$$

Proof. The proof follows the result of Lemma 5.1 in [10]. To ease notation, we fix an arbitrary $s \in \mathcal{S}$ and denote $\mu_k = \pi_k(s)$, $X_k = \hat{\pi}_k(s)$ for $k \in \{i, j\}$. Under the event that $\|\mu_k - X_k\|_\infty \leq \xi$ for $k \in \{i, j\}$, we have

$$\begin{aligned}\frac{X_i}{X_j} &\geq \frac{\mu_i - \xi}{\mu_j + \xi} = \frac{\mu_i}{\mu_j} - \frac{\xi}{\mu_j + \xi}\left(1 + \frac{\mu_i}{\mu_j}\right)\\ &\geq \frac{\mu_i}{\mu_j} - \frac{\xi}{c + \xi}\left(1 + \frac{1 - c}{c}\right)\\ &= \frac{\mu_i}{\mu_j} - \frac{\xi}{c(c + \xi)}.\end{aligned}$$

The upper bound is proved similarly. Since the choice of $s \in \mathcal{S}$ was arbitrary and the event $\|\mu_k - X_k\|_\infty \leq \xi$ holds with probability at least $1 - \delta$, the result follows. □

B Proofs of Results in Section 5.2

B.1 The Simpler Case of Two Experts

We begin by considering two experts, i, j. We are given access to t samples from expert j and seek to estimate the mean of expert i using the approximate policies $\hat{\pi}_i, \hat{\pi}_j$ with maximum error ξ. The arguments in this section closely follow the analysis of the two-armed estimator in Reference [30]. We note here that we work with values of ξ as in Theorem 6; i.e.,

$$\xi \leq 2\frac{\sqrt{1 + p_V^4\gamma^2} - 1}{p_V\gamma} \iff \frac{\xi}{p_V}\left(\frac{1}{(p_V - \xi)} + \frac{1}{(p_V + \xi)}\right) \leq \frac{\gamma p_V}{2}.$$

We refer the reader to Proposition 5 for the definitions of $\underline{r}_{ij}(v|x), \overline{r}_{ij}(v|x)$. Using the proposition, we also have

$$\underline{r}_{ij}(v|x) \leq \hat{r}_{ij}(v|x) \leq \overline{r}_{ij}(v|x).$$

For an arbitrarily chosen $\epsilon \in (0,1)$, we write

$$\eta' = \min\left\{a : \mathbb{P}_i\left(\underline{r}_{ij} > a\right) \leq \frac{\epsilon}{2}\right\}, \tag{8}$$

$$\eta = \min\left\{a : \mathbb{P}_i\left(r_{ij} > a\right) \leq \frac{\epsilon}{2}\right\}. \tag{9}$$

One can easily check the following claim using basic properties of indicator functions:

Claim 1. *With η, η' as above, for any $\epsilon \in (0,1)$, we have that $\eta' \leq \eta$. Further, $\mathbb{1}\{r_{ij} \leq \eta\} \geq \mathbb{1}\{\overline{r}_{ij} \leq \eta'\}$.*

Our estimator for μ_i based on t samples from expert j is then defined as

$$\tilde{Y}_i(j,t) = \frac{1}{t}\sum_{s=1}^{t} Y_s \underline{r}_{ij}(s) \mathbb{1}\left\{\overline{r}_{ij}(s) \leq \eta'\right\}. \tag{10}$$

We recall the following result on the full information IS estimator (Lemma 1 in Reference [30]):

Lemma 10. *With η as in Equation (8) and the full information IS estimator as*

$$\hat{Y}_i^{\eta}(j,t) := \frac{1}{t}\sum_{s=1}^{t} Y_s r_{ij}(s) \mathbb{1}\left\{r_{ij}(s) \leq \eta\right\}. \tag{11}$$

Then, for all $t \geq 1$, it holds that $\mathbb{E}_j[\hat{Y}_i^{\eta}(j,t)] \leq \mu_i \leq \mathbb{E}_j[\hat{Y}_i^{\eta}(j,t)] + \frac{\epsilon}{2}$.

We now compare our estimator in Equation (10) to the full information estimator in Equation (11). Due to Claim 1, it holds trivially that $\tilde{Y}_i(j,t) \leq \hat{Y}_i(j,t)$. Consider the following chain:

$$\begin{aligned}
\hat{Y}_i(j,t) &= \hat{Y}_i(j,t) + \tilde{Y}_i(j,t) - \tilde{Y}_i(j,t) \\
&\leq \tilde{Y}_i(j,t) + \frac{1}{t}\sum_{s=1}^{t} Y_s\left(\overline{r}_{ij}(s) - \underline{r}_{ij}(s)\mathbb{1}\{\overline{r}_{ij}(s) \leq \eta'\}\right) \\
&\leq \tilde{Y}_i(j,t) + \frac{1}{t}\sum_{s=1}^{t}\left(\overline{r}_{ij}(s) - \underline{r}_{ij}(s)\mathbb{1}\{\overline{r}_{ij}(s) \leq \eta'\}\right) \\
&\leq \tilde{Y}_i(j,t) + e_{ij}.
\end{aligned}$$

Here, the second inequality holds, since $e_{ij} := \max_{v,x} \overline{r}_{ij} - \underline{r}_{ij}\mathbb{1}\{\overline{r}_{ij} \leq \eta'\} \geq 0$ and $Y_s \leq 1$. As a result of the arguments above, we have the following Lemma:

Lemma 11. *With $\hat{Y}_i(j,t)$ and $\tilde{Y}_i(j,t)$ as defined in Equations (11) and (10), respectively, and $e_{ij} := \max_{v,x} \overline{r}_{ij} - \underline{r}_{ij}\mathbb{1}\{\overline{r}_{ij} \leq \eta'\}$, it holds that*

$$\mathbb{E}_j[\tilde{Y}_i(j,t)] \leq \mathbb{E}_j[\hat{Y}_i(j,t)] \leq \mathbb{E}_j[\tilde{Y}_i(j,t)] + e_{ij}.$$

Together with Lemma 10, we have the following corollary:

Corollary 11.1. *The mean of the estimator $\tilde{Y}_i(j,t)$ according to the distribution of arm j satisfies*

$$\mathbb{E}_j[\tilde{Y}_i(j,t)] \in \left[\mu_i - \frac{\epsilon}{2} - e_{ij}, \mu_i\right]. \tag{12}$$

B.1.1 Simpler Clipper Levels. In this section, we move from the abstract clipper levels in Equation (8) to those based on divergences as in the estimator used in ED-UCB. To this end, using Markov's inequality, we have $\mathbb{P}_i(\underline{r}_{ij} > a) \leq \exp(-a)\mathbb{E}_i[\exp(\underline{r}_{ij})]$. Suppose that the RHS of the above is upper bounded by $\frac{\epsilon}{2}$. That is, $\exp(a) \geq \frac{2}{\epsilon}\mathbb{E}_i[\underline{(r}_{ij})]$. Consider the following chain:

$$\mathbb{E}_i[\exp(\underline{r}_{ij})] = e + e\left(\mathbb{E}_j[r_{ij}\exp(\underline{r}_{ij} - 1) - 1]\right) \geq e + e\left(\mathbb{E}_j[\underline{r}_{ij}\exp(\underline{r}_{ij} - 1) - 1]\right) = e + e\underline{D}_{f_1}(\pi_i||\pi_j).$$

Therefore, the RHS of the Markov inequality is upper bounded by $\frac{\epsilon}{2}$ if $\exp(a) \geq \frac{2}{\epsilon}(e + e\underline{D}_{f_1}(\pi_i||\pi_j))$ or, equivalently, if $a \geq \log(\frac{2}{\epsilon}) + \underline{M}_{ij}$. However, by definition, we must have that $\eta' \leq \log(\frac{2}{\epsilon}) + \underline{M}_{ij} \leq 2\log(\frac{2}{\epsilon})\underline{M}_{ij}$.

Therefore, we redefine our original estimator in Equation (10) to use this new clipper level $\alpha_i(j,\epsilon) := 2\log(\frac{2}{\epsilon})\underline{M}_{ij}$. We have

$$\tilde{Y}_i(j,t) := \frac{1}{t}\sum_{s=1}^{t} Y_s \underline{r}_{ij}(s)\mathbb{1}\{\overline{r}_{ij}(s) \leq \alpha_i(j,\epsilon)\}. \tag{13}$$

We also restate Lemma 11 for convenience:

Lemma 12. *For $\tilde{Y}_i(j,t)$ defined as above, and $e_{ij} = \max_{v,x} \overline{r}_{ij} - \underline{r}_{ij}\mathbb{1}\{\overline{r}_{ij} \leq \alpha_i(j,\epsilon)\}$,*

$$\mathbb{E}_j[\tilde{Y}_i(j,t)] \in \left[\mu_i - \frac{\epsilon}{2} - e_{ij}, \mu_i\right]. \tag{14}$$

B.2 Estimator Concentrations

The Clipped IS-based estimator with empirical policies is defined in Equation (5). To provide concentrations for this estimator, we define the following filtration by time t: $\mathcal{F}_t = \sigma(\{k_s, X_s, V_s, Y_s\}_{s=1}^{t-1}, k_t)$. Note that the filtration at time t contains information about all the observations up to time $t-1$ as well as the choice of the arm at time t. We now define the following martingale that will be used to analyze the estimator:

$$A_0 := 0, A_s := \sum_{l=1}^{s} \frac{L_i(l)}{\underline{M}_{ik_l}} - \sum_{l=1}^{s} \mathbb{E}\left[\frac{L_i(l)}{\underline{M}_{ik_l}}\middle|\mathcal{F}_{l-1}\right],$$

where, to ease notation, we write $L_i(l) = Y_l \underline{r}_{ik_l}(l)\mathbb{1}\{\overline{r}_{ik_l}(l) \leq \alpha_i(k_l,l)\}$. Since $\mathcal{F}_{l-1}$ contains knowledge of k_l, the second term in A_s can be further simplified using $\mu_i(l) = \mathbb{E}[L_i(l)|\mathcal{F}_{l-1}]$ as

$$A_s = \sum_{l=1}^{s} \frac{L_i(l)}{\underline{M}_{ik_l}} - \sum_{l=1}^{s} \frac{\mu_i(l)}{\underline{M}_{ik_l}}.$$

Remark 13. Using Lemma 12, we can write that $\mu_i(l) \in [\mu_i - \frac{\epsilon}{2} - e_i(t), \mu_i]$, where $e_i(t) := \max_j e_{ij}$.

It is easy to see that $|A_s - A_{s-1}| \leq 4\log(\frac{2}{\epsilon})$. Therefore, using the Azuma-Hoeffding inequality for martingales with bounded differences, we can write:

$$\mathbb{P}\left(\left|\sum_{l=1}^{t} \frac{L_i(l)}{\underline{M}_{ik_l}} - \sum_{l=1}^{t} \frac{\mu_i(l)}{\underline{M}_{ik_l}}\right| \geq \beta\right) \leq 2\exp\left(-\frac{\beta^2}{32t\log^2\left(\frac{2}{\epsilon}\right)}\right).$$

We are now ready to state and prove our concentration result.

Theorem 14. *The estimator $\tilde{Y}_i(t)$, as in Equation (5), satisfies*

$$\mathbb{P}\left(\tilde{Y}_i(t) \notin \left[(1-\beta)\left(\mu_i - \frac{\epsilon_i(t)}{2} - e_i(t)\right), (1+\beta)\mu_i\right]\right) \leq 2\exp\left(-\frac{t\beta^2(\gamma - 2e_i(t))^2}{128M^2\log^2\left(\frac{2}{\epsilon_i(t)}\right)}\right)$$

for t such that $\epsilon_i(t) \leq \gamma$, $\beta > 0$ and $M = (1-p_X)|\mathcal{X}||\mathcal{V}|f_1(p_V/1-p_V)$.

Proof. **Upper Tail:** For any $\beta(t) > 0$, we have the following chain:

$$\begin{aligned}(1+\beta(t))\left(\max_{l\in[t]}\mu_i(l)\right)\sum_{l=1}^{t}\frac{1}{\underline{M}_{ik_l}} &\geq \sum_{l=1}^{t}\frac{\mu_i(l)}{\underline{M}_{ik_l}} + \frac{t}{M}\times\beta(t)\times\max_{l\in[t]}\mu_i(l)\\ &\geq \sum_{l=1}^{t}\frac{\mu_i(l)}{\underline{M}_{ik_l}} + \frac{\beta(t)t}{M}\left(\frac{\gamma}{2} - e_i(t)\right).\end{aligned}$$

The final inequality uses Remark 13 with $\epsilon(t) \leq \gamma \leq \mu_i$ for all $i \in \mathcal{A}$. Thus, we have that

$$\begin{aligned}\mathbb{P}\left(\tilde{Y}_i(t) \geq (1+\beta(t))(\max_{l\in[t]}\mu_i(l))\right) &\leq \mathbb{P}\left(\sum_{l=1}^{t}\frac{L_i(l)}{\underline{M}_{ik_l}} \geq \sum_{l=1}^{t}\frac{\mu_i(l)}{\underline{M}_{ik_l}} + \frac{\beta(t)t}{M}\left(\frac{\gamma}{2} - e_i(t)\right)\right)\\ &\leq \exp\left(-\frac{t\beta^2(t)\left(\frac{\gamma}{2} - e_i(t)\right)^2}{32M^2\log^2\left(\frac{2}{\epsilon(t)}\right)}\right).\end{aligned}$$

(2) Lower Tail: Again, for any $\beta(t) > 0$,

$$(1-\beta(t))\left(\min_{l\in[t]}\mu_i(l)\right)\sum_{l=1}^{t}\frac{1}{\underline{M}_{ik_l}} \leq \sum_{l=1}^{t}\frac{\mu_i(l)}{\underline{M}_{ik_l}} - \frac{\beta(t)t}{M}\left(\frac{\gamma}{2} - e_i(t)\right).$$

Therefore, we have that

$$\begin{aligned}\mathbb{P}\left(\tilde{Y}_i(t) \leq (1-\beta(t))(\min_{l\in[t]}\mu_i(l))\right) &\leq \mathbb{P}\left(\sum_{l=1}^{t}\frac{L_i(l)}{\underline{M}_{ik_l}} \leq \sum_{l=1}^{t}\frac{\mu_i(l)}{\underline{M}_{ik_l}} - \frac{\beta(t)t}{M}\left(\frac{\gamma}{2} - e_i(t)\right)\right)\\ &\leq \exp\left(-\frac{t\beta^2(t)\left(\frac{\gamma}{2} - e_i(t)\right)^2}{32M^2\log^2\left(\frac{2}{\epsilon(t)}\right)}\right).\end{aligned}$$

Combining the two tails above and using Remark 13, we have

$$\mathbb{P}\left(\tilde{Y}_i(t) \notin \left[(1-\beta(t))\left(\mu_i - \frac{\epsilon(t)}{2} - e_i(t)\right), (1+\beta(t))\mu_i\right]\right) \leq 2\exp\left(-\frac{t\beta^2(t)\left(\frac{\gamma}{2} - e_i(t)\right)^2}{32M^2\log^2\left(\frac{2}{\epsilon(t)}\right)}\right).$$

□

B.3 Per-expert Concentrations

In this section, we provide concentration results for optimal and sub-optimal experts separately. We begin with the following claim:

Claim 2. *For all $t \geq T_{clip}$ and all $k \in [N]$, $e_k(t) = \frac{\gamma p_V}{2}$.*

Proof. Fix $i, j \in [N]$ and $(v, x) \in \mathcal{V} \times \mathcal{X}$, arbitrary. Consider at some $t \geq T_{clip}$

$$\begin{aligned}
&\overline{r}_{ij}(v|x) - \underline{r}_{ij}(v|x)\mathbb{1}\left\{\overline{r}_{ij}(v|x) \leq \alpha_{ij}(t)\right\} \\
&= \overline{r}_{ij}(v|x) - \underline{r}_{ij}(v|x) \text{ (By definition of } T_{clip}) \\
&= \frac{\xi}{p_V}\left(\frac{1}{(p_V - \xi)} + \frac{1}{(p_V + \xi)}\right) \leq \frac{\gamma p_V}{2}.
\end{aligned}$$

Since this holds for the arbitrary choice of i, j, v, x, it must hold that $e_k(t) \leq \frac{\gamma p_V}{2}$. □

The following lemmas provide concentrations for the best expert and suboptimal experts:

Lemma 15. *Let $\beta'_{k^*}(t) = Cw(\sqrt{\frac{\log t}{t}})$ and τ'_1 be as in Equation (7). Then, for all $t \geq \tau'_1$, it holds that $\mathbb{P}(U'_{k^*}(t) \leq \mu^*) \leq \frac{1}{t^2}$.*

Proof. We have that for $t \geq \tau'_1$,

$$\begin{aligned}
\mathbb{P}(U'_{k^*} \leq \mu^*) &\leq \mathbb{P}\left(\tilde{Y}_{k^*}(t) \leq \mu^* - \frac{3}{2}\beta'_{k^*}(t) - e_{k^*}\right) \\
&\leq \mathbb{P}\left(\tilde{Y}_{k^*}(t) \leq \mu^* - \mu^*\beta'_{k^*}(t) - (1 - \beta'_{k^*}(t))\left(\frac{\beta'_{k^*}(t)}{2} + e_{k^*}\right)\right) \\
&= \mathbb{P}\left(\tilde{Y}_{k^*}(t) \leq (1 - \beta'_{k^*}(t))\left(\mu^* - \frac{\beta'_{k^*}(t)}{2} - e_{k^*}\right)\right).
\end{aligned}$$

In the above, the first inequality follows, since $(Z^*_k(t) \leq t, w(x)$ increasing $\Rightarrow \beta_{k^*}(t) \geq \beta'_{k^*}(t))$ and the second is due to $(\beta'_{k^*}(t) \leq \gamma \leq \mu^* \leq 1$ for $t \geq \tau'_1)$. Applying Theorem 14 with $\epsilon(t) = \beta(t) = \beta'_{k^*}(t)$ (this choice is valid for $\epsilon(t)$ due to the definition of τ'_1) and using Claim 2, we have

$$\mathbb{P}(U'_{k^*}(t) \leq \mu^*) \leq \exp\left(-\frac{\beta'_{k^*}(t)^2 t\gamma^2 (1 - p_V)^2}{128M^2 \log^2\left(\frac{2}{\beta'_{k^*}(t)}\right)}\right).$$

Consider the exponent:

$$\frac{\beta'_{k^*}(t)^2 t\gamma^2(1 - p_V)^2}{128M^2 \log^2\left(\frac{2}{\beta'_{k^*}(t)}\right)^2} = \frac{t\gamma^2(1 - p_V)^2}{128M^2} C^2 \left(\frac{w\left(\sqrt{\log t/t}\right)}{\log\left(2/Cw\left(\sqrt{\frac{\log t}{t}}\right)\right)}\right)^2 \geq 2\log t.$$

The final inequality holds for the choice of $C = \frac{32M}{\gamma(1-p_V)}$ and because $\frac{w(x)}{\log(2/aw(x))} \geq x$ for $a \geq 1$, as is the case with C. Substituting this exponent completes the proof. □

Lemma 16. *Let τ'_k be as in Equation (7). Then, for all $t \geq \tau'_k$ and $k \neq k^*$, $\mathbb{P}(U'_k(t) > \mu^*) \leq \frac{1}{t^2}$.*

Proof. Note that $Z_k(t) \geq \frac{t}{M}$. Since $w(x)$ is increasing in x, using $t \geq \tau'_k$, we can write

$$\begin{aligned}
\frac{3}{2}\beta_k(t) + e_k(t) &\leq \frac{3C}{2} w\left(\frac{M\sqrt{t\log t}}{t}\right) + \frac{\gamma p_V}{2} \\
&\leq \frac{3C}{2} w\left(\frac{(\Delta_k - \gamma p_V)/3C}{\log\left(\frac{2}{(\Delta_k - \gamma p_V)/3C}\right)}\right) + \frac{\gamma p_V}{2} \\
&= \frac{3C}{2} \cdot \frac{\Delta_k - \gamma p_V}{3C} + \frac{\gamma p_V}{2} = \frac{\Delta_k}{2}.
\end{aligned}$$

Here, the penultimate equality follows from the definition of the w function. Therefore, we can write

$$\begin{aligned}\mathbb{P}\left(U'_k(t) > \mu^*\right) &= \mathbb{P}\left(\tilde{Y}_k(t) > \mu^* - \frac{3\beta_k(t)}{2} - e_k(t)\right)\\ &\leq \mathbb{P}\left(\tilde{Y}_k(t) > \mu^* - \frac{\Delta_k}{2}\right)\\ &\leq \mathbb{P}\left(\tilde{Y}_k(t) > \mu_k + \frac{\Delta_k}{2}\right)\\ &\leq \mathbb{P}\left(\tilde{Y}_k(t) > \mu_k\left(1 + \frac{\Delta_k}{2}\right)\right)\\ &\leq \mathbb{P}\left(\tilde{Y}_k(t) > \mu_k\left(1 + \beta'_k(t)\right)\right).\end{aligned}$$

The penultimate inequality uses $\mu_k \leq 1$ and the final one follows, since $\beta'_k(t) \leq \beta_k(t) \leq \frac{3}{2}\beta(t)_k + e_k(t) \leq \frac{\Delta_k}{2}$. Thus, we can now apply Theorem 14 with $\beta(t) = \epsilon(t) = \beta'_k(t)$ and follow the exponent bounding arguments as in Lemma 15 to obtain the required result. □

The two lemmas above lead to the following useful corollary:

Corollary 16.1. *For all suboptimal experts $k \neq k^*$ and $t \geq \tau'_k$, we have that*

$$\mathbb{P}(k_t = k) \leq \frac{2}{t^2}.$$

Proof. Consider any suboptimal expert $k : \Delta_k > 0$ and time $t \geq \tau'_k$. We bound the probability that it is played as follows:

$$\begin{aligned}\mathbb{P}(k_t = k) &= \mathbb{P}\left(k_t = k, U'_{k^*}(t) \geq \mu^*\right) + \mathbb{P}\left(k_t = k, U'_{k^*}(t) < \mu^*\right)\\ &= \mathbb{P}\left(k_t = k | U'_{k^*}(t) \geq \mu^*\right) \cdot \mathbb{P}(U'_{k^*}(t) \geq \mu^*) + \mathbb{P}\left(k_t = k | U'_{k^*}(t) < \mu^*\right) \cdot \mathbb{P}(U'_{k^*}(t) < \mu^*)\\ &\leq \mathbb{P}\left(k_t = k | U'_{k^*}(t) \geq \mu^*\right) + \mathbb{P}\left(U'_{k^*}(t) < \mu^*\right)\\ &\leq \mathbb{P}(U'_k(t) > \mu^*) + \mathbb{P}(U'_{k^*}(t) < \mu^*)\\ &\leq \frac{2}{t^2}.\end{aligned}$$

Here, the first inequality uses the fact that probabilities are bounded by 1, then, we use $\{k_t = k | U'_{k^*}(t) \geq \mu^*\} \subseteq \{U'_k(t) > \mu^*\}$. Finally, we use the results of Lemmas 15, 16 above, since $t \geq \tau'_k \geq \tau'_1$. □

B.4 Proof of Theorem 6

Proof. Using Corollary 16.1, we can bound the regret using the following chain:

$$\begin{aligned}\mathbb{E}[R(T)] &= \sum_{t=1}^{T}\sum_{k=2}^{N}\Delta_k\mathbb{P}(k_t = k)\\ &\leq \sum_{t=1}^{\tau'_N - 1}\Delta_N + \sum_{t=\tau'_N}^{T}\Delta_N\mathbb{P}(k_t = N) + \sum_{t=\tau'_N}^{T}\sum_{k=2}^{N-1}\Delta_k\mathbb{P}(k_t = k)\end{aligned}$$

$$\leq \tau'_N \Delta_N + \Delta_N \frac{\pi^2}{3} + \sum_{t=\tau'_N}^{\tau'_{N-1}-1} \Delta_{N-1} + \sum_{t=\tau'_{N-1}}^{T} \Delta_{N-1} \mathbb{P}(k_t = N-1) + \sum_{t=\tau'_{N-1}}^{T} \sum_{k=2}^{N-2} \Delta_k \mathbb{P}(k_t = k)$$

$$\leq \Delta_N \left(\tau'_N + \frac{\pi^2}{3}\right) + \Delta_{N-1}\left((\tau'_{N-1} - \tau'_N) + \frac{\pi^2}{3}\right) + \cdots + \Delta_2 \left((\tau'_2 - \tau'_3) + \frac{\pi^2}{3}\right)$$

$$= \frac{\pi^2}{3} \sum_{k=2}^{N} \Delta_k + \tau'_N \Delta_N + \sum_{k=2}^{N-1} \left(\tau'_k - \tau'_{k+1}\right).$$

□

C Proofs of Results in Section 4.1

All our results for ED-UCB can be used to derive results for D-UCB by setting $\xi = 0$ and replacing the empirical IS ratios and divergences with their true values. In particular, Theorem 14 can be modified to prove Theorem 1. Then, with τ_i defined as in Equation (3), we arrive at high probability error bounds for the UCB indices of the best and suboptimal arms separately (analogous to Lemmas 15, 16). These are then used to conclude Corollary 2.1. The result of Theorem 3 then follows from arguments similar to those in Theorem 6. We present short versions of these results below.

C.1 Estimator Concentrations

We begin with the following counterpart to Remark 13:

Lemma 17. *For all times l and all experts $j \in [N]$, we have*

$$\mu_j - \frac{\epsilon(t)}{2} \leq \mu_j(l) \leq \mu_j, \tag{15}$$

where $\mu_j(l)$ is defined in the proof of Theorem 1.

Proof. We first note that under the filtration $\mathcal{F}_{l-1}$, $M_{jk(l)}$ is a constant and $k(l)$ is fixed. Therefore, following the notation in Theorem 1, we have the following chain:

$$\begin{aligned}
\mu_j(l) &= \mathbb{E}\left[L_j(l)|\mathcal{F}_{l-1}\right] \\
&= \mathbb{E}_{k(l)}\left[Y_l \frac{\pi_j(V_{k(l)}(l)|X_{k(l)}(l))}{\pi_{k(l)}(V_{k(l)}(l)|X_{k(l)}(l))}\right] \\
&\quad - \mathbb{E}_{k(l)}\left[Y_l \frac{\pi_j(V_{k(l)}(l)|X_{k(l)}(l))}{\pi_{k(l)}(V_{k(l)}(l)|X_{k(l)}(l))} \times \mathbb{1}\left\{r_l > \alpha_l\right\}\right] \\
&\overset{(i)}{\geq} \mu_j - \mathbb{P}_j\left(\frac{\pi_j(V_{k(l)}(l)|X_{k(l)}(l))}{\pi_{k(l)}(V_{k(l)}(l)|X_{k(l)}(l))} > 2\log(2/\epsilon(t))M_{jk(l)}\right) \\
&\overset{(ii)}{\geq} \mu_j - \frac{\epsilon(t)}{2}.
\end{aligned}$$

Here, (i) follows from the fact that $Y \in [0,1]$ and (ii) follows from Lemma 2 in Reference [30]. □

Now, we are ready to prove Theorem 1.

Proof of Theorem 1. We will reuse notation from Theorem 14 for convenience. Let $\mathcal{F}_s$ be the filtration formed by the observation until time s *and* the expert chosen at time $s+1$. We define the martingale $\{A_s\}$ with $A_0 = 0$ and

$$A_s = \sum_{l=1}^{s} \frac{L_j(l)}{M_{jk(l)}} - \sum_{l=1}^{s} \mathbb{E}\left[\frac{L_j(l)}{M_{jk(l)}}\middle|\mathcal{F}_{l-1}\right],$$

where, to ease notation, we define

$$r_l := \frac{\pi_j(V_{k(l)}(l)|X_{k(l)}(l))}{\pi_{k(l)}(V_{k(l)}(l)|X_{k(l)}(l))},\ \alpha_l := 2\log(2/\epsilon(t))M_{jk(l)},\ L_j(l) := Y_l \times r_l \times \mathbb{1}\{r_l \le \alpha_l\}.$$

With $\mu_j(l) := \mathbb{E}[L_j(l)|F_{l-1}]$ and $B_t = \sum_{l=1}^{t} \frac{L_j(l)}{M_{jk(l)}}$, we rewrite A_s as $A_s = B_s - \sum_{l=1}^{s} \frac{\mu_j(l)}{M_{jk(l)}}$. Since $|A_s - A_{s-1}| \le 4\log(2/\epsilon(t))$, the Azuma-Hoeffding inequality implies that

$$\mathbb{P}\left(\left|B_t - \sum_{l=1}^{t} \frac{\mu_j(l)}{M_{jk(l)}}\right| \ge \chi\right) \le 2\exp\left(-\frac{\chi^2}{32t(\log(2/\epsilon(t)))^2}\right). \tag{16}$$

(1) Upper tail We have that

$$\mathbb{P}\left(B_t \ge \sum_{l=1}^{t} \frac{\mu_j(l)}{M_{jk(l)}} + \chi\right) \le \exp\left(-\frac{\chi^2}{32t(\log(2/\epsilon(t)))^2}\right)$$

Consider the following chain:

$$(1+\beta(t))\left(\max_{l\in[t]} \mu_j(l)\right)\sum_{l=1}^{t}\frac{1}{M_{jk(l)}} \ge \sum_{l=1}^{t}\frac{\mu_j(l)}{M_{jk(l)}} + \frac{\beta(t)t}{M}\max_{l\in[t]}\mu_j(l) \ge \sum_{l=1}^{t}\frac{\mu_j(l)}{M_{jk(l)}} + \frac{\gamma\beta(t)t}{2M},$$

where the final inequality comes about as a consequence of $\epsilon(t) < \gamma$ and $\mu_j(l) \ge \mu_j - \frac{\epsilon(t)}{2} \ge \gamma - \frac{\epsilon(t)}{2}$. The latter is proved in Lemma 17. We also use the fact that $M := \max_{j,k} M_{jk}$. Thus, we have

$$\begin{aligned}\mathbb{P}\left(B_t \ge (1+\beta(t))\left(\max_l \mu_j(l)\right)\sum_{l=1}^{t}\frac{1}{M_{jk(l)}}\right) &\le \mathbb{P}\left(B_t \ge \sum_{l=1}^{t}\frac{\mu_j(l)}{M_{jk(l)}} + \frac{\gamma\beta(t)t}{2M}\right)\\ &\le \exp\left(-\frac{\gamma^2\beta(t)^2 t}{128M^2(\log(2/\epsilon(t)))^2}\right).\end{aligned}$$

This implies that

$$\mathbb{P}\left(\hat{\mu}_k(t) \ge (1+\beta(t))\left(\max_l \mu_j(l)\right)\right) \le \exp\left(-\frac{\gamma^2\beta(t)^2 t}{128M^2(\log(2/\epsilon(t)))^2}\right). \tag{17}$$

(2) Lower Tail We have that

$$\mathbb{P}\left(B_t \le \sum_{l=1}^{t}\frac{\mu_j(l)}{M_{jk(l)}} - \chi\right) \le \exp\left(-\frac{\chi^2}{32t(\log(2/\epsilon(t)))^2}\right).$$

Using similar arguments as above, we can write

$$\begin{aligned}\mathbb{P}\left(B_t \le (1-\beta(t))\left(\min_l \mu_j(l)\right)\sum_{l=1}^{t}\frac{1}{M_{jk(l)}}\right) &\le \mathbb{P}\left(B_t \le \sum_{l=1}^{t}\frac{\mu_j(l)}{M_{jk(l)}} - \frac{\gamma\beta(t)t}{2M}\right)\\ &\le \exp\left(-\frac{\gamma^2\beta(t)^2 t}{128M^2(\log(2/\epsilon(t)))^2}\right).\end{aligned}$$

This implies that

$$\mathbb{P}\left(\hat{\mu}_k(t) \le (1-\beta(t))\left(\min_l \mu_j(l)\right)\right) \le \exp\left(-\frac{\gamma^2\beta(t)^2 t}{128M^2(\log(2/\epsilon(t)))^2}\right). \tag{18}$$

Since the choice of $j \in [N]$ was arbitrary, combining Equations (17), (18) above with Equation (15) from the lemma above, we have the result. □

C.2 Per-expert Concentrations

We begin with the best expert and show that it is overestimated with high probability.

Lemma 18. *Let* $\tau_1 = \min\{t : \beta'(t) := Cw(\frac{\sqrt{c_1 t \log t}}{t}) \le \gamma\}$ *be as in Equation (3). Then, for all* $t \ge \tau_1$, *the index of the best arm formed using the Clipped Estimator satisfies*

$$\mathbb{P}\left(U_{k^*}(t) > \mu^*\right) \ge 1 - \frac{1}{t^2}.$$

Proof. Since $Z_k(t) \le t$ and $w(x)$ is increasing, we have that $\beta(t) \ge \beta'(t)$. Now, we have the following chain:

$$\begin{aligned}
\mathbb{P}\left(U_{k^*}(t) \le \mu^*\right) &= \mathbb{P}\left(\hat{\mu}_{k^*}(t) \le \mu^* - \frac{3}{2}Cw\left(\frac{\sqrt{c_1 t \log t}}{Z_k(t)}\right)\right) \\
&\overset{(i)}{\le} \mathbb{P}\left(\hat{\mu}_{k^*}(t) \le \mu^* - \mu^* Cw\left(\frac{\sqrt{c_1 t \log t}}{t}\right) - \frac{1}{2}Cw\left(\frac{\sqrt{c_1 t \log t}}{t}\right)\right) \\
&\overset{(ii)}{\le} \mathbb{P}\left(\hat{\mu}_{k^*}(t) \le \mu^* - \mu^*\beta'(t) - (1-\beta'(t))\frac{1}{2}\beta'(t)\right) \\
&\le \mathbb{P}\left(\hat{\mu}_{k^*}(t) \le (1-\beta'(t))\left(\mu^* - \frac{\beta'(t)}{2}\right)\right).
\end{aligned}$$

Here, (i) uses the observation above and that $\mu^* \le 1$. (ii) follows from the fact that $\beta'(t) \le 1$. Now, we use Theorem 1 with $\beta(t)$ set to be sample-path independent $\beta'(t)$ to write

$$\mathbb{P}\left(U_{k^*}(t) \le \mu^*\right) \le \exp\left(-\frac{\gamma^2 \beta'(t)^2 t}{128M^2(\log(2/\beta'(t)))^2}\right).$$

Similar to the arguments in Lemma 15, with $C = \frac{16M}{\gamma}$, we can upper bound the exponent by $-2\log t$. This leads to the result. □

We now bound the probability of overestimating a suboptimal expert.

Lemma 19. *Let* τ_1 *be as in Lemma 18 and* τ_k *as in Equation (3). Then, for any* $k \ne k^*$, *for any* $t \ge \tau_k$, *we have that*

$$\mathbb{P}\left(U_k(t) < \mu^*\right) \ge 1 - \frac{1}{t^2}.$$

Proof. We have that $Z_k(t) \ge \frac{t}{M}$ and $t \ge \frac{9C^2M^2 \log T \log^2(6C/\Delta_k)}{\Delta_k^2}$. Additionally, $w(x)$ is increasing in x. Therefore, we can write that

$$\begin{aligned}
\frac{3C}{2}w\left(\frac{\sqrt{c_1 t \log t}}{Z_k(t)}\right) &\le \frac{3C}{2}w\left(\frac{M\sqrt{c_1 t \log t}}{t}\right) \\
&\le \frac{3C}{2}w\left(\frac{M\sqrt{c_1 t \log T}}{t}\right) \\
&\le \frac{3C}{2}w\left(\frac{M\sqrt{c_1 \log T}}{\frac{3CM\sqrt{c_1 \log T}\log(6C/\Delta_k)}{\Delta_k}}\right)
\end{aligned}$$

$$= \frac{3C}{2} w\left(\frac{\Delta_k/3C}{\log\left(\frac{2}{\Delta_k/3C}\right)} \right)$$
$$= \frac{\Delta_k}{2},$$

where the last equality follows from the fact that $w(x) = y \iff \frac{y}{\log(2/y)} = x$. Using the argument above along with the fact that $\Delta_k = \mu^* - \mu_k$ and $\mu_k < \mu^* \leq 1$, we have the following chain:

$$\begin{aligned}
\mathbb{P}\left(U_k(t) > \mu^*\right) &= \mathbb{P}\left(\hat{\mu}_k(t) > \mu^* - \frac{3\beta(t)}{2}\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_k(t) > \mu^* - \frac{\Delta_k}{2}\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_k(t) > \mu_k + \frac{\Delta_k}{2}\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_k(t) > \mu_k \left(1 + \frac{\Delta_k}{2}\right)\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(\hat{\mu}_k(t) > \mu_k \left(1 + \frac{3C}{2} w\left(\frac{M\sqrt{t\log t}}{t}\right)\right)\right) \\
&\leq \exp\left(-\frac{\gamma^2 \left(3C/2\right)^2 t}{128M^2} \times \left(\frac{w\left(M\sqrt{\log t/t}\right)}{\log\left(2/Cw\left(\sqrt{\log t/t}\right)\right)}\right)^2\right). \qquad (19)
\end{aligned}$$

In (i), we have used the fact that $\Delta_k/2 \geq \frac{3C}{2} w(\frac{M\sqrt{t\log t}}{t})$ from the chain just before. Here, the final inequality applies Theorem 1 (bounds for the upper tail error) with $\chi(t) = \frac{3C}{2} w(\frac{M\sqrt{t\log t}}{t})$ and $\epsilon(t) = \beta'(t)$ defined in Lemma 18. Upper bounding the exponent by $-2\log t$ using arguments in Lemma 15 gives us the result. □

These lead to Corollary 2.1.

Proof of Corollary 2.1. The proof follows the same chain of reasoning as Corollary 16.1 with all the empirical quantities replaced by their full-information counterparts. We have

$$\begin{aligned}
\mathbb{P}(k_t = k) &= \mathbb{P}\left(k_t = k, U_{k^*}(t) \geq \mu^*\right) + \mathbb{P}\left(k_t = k, U_{k^*}(t) < \mu^*\right) \\
&\leq \mathbb{P}\left(k_t = k | U_{k^*}(t) \geq \mu^*\right) + \mathbb{P}\left(U_{k^*}(t) < \mu^*\right) \\
&\leq \mathbb{P}(U_k(t) > \mu^*) + \mathbb{P}(U_{k^*}(t) < \mu^*) \\
&\leq \frac{2}{t^2}.
\end{aligned}$$

Again, we use that $\{k_t = k | U_{k^*}(t) \geq \mu^*\} \subseteq \{U_k(t) \geq \mu^*\}$ in the second inequality. The result follows by applying Lemmas 18, 19. □

C.3 Proof of Theorem 3

Proof. Using Corollary 2.1, we can bound the regret using the following chain:

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{t=1}^{T}\sum_{k=2}^{N}\Delta_k\mathbb{P}(k_t=k)\\
&\le \sum_{t=1}^{\tau_N-1}\Delta_N + \sum_{t=\tau_N}^{T}\Delta_N\mathbb{P}(k_t=N) + \sum_{t=\tau_N}^{T}\sum_{k=2}^{N-1}\Delta_k\mathbb{P}(k_t=k)\\
&\le \tau_N\Delta_N + \Delta_N\frac{\pi^2}{3} + \sum_{t=\tau_N}^{\tau_{N-1}-1}\Delta_{N-1} + \sum_{t=\tau_{N-1}}^{T}\Delta_{N-1}\mathbb{P}(k_t=N-1) + \sum_{t=\tau_{N-1}}^{T}\sum_{k=2}^{N-2}\Delta_k\mathbb{P}(k_t=k)\\
&\le \Delta_N\left(\tau_N+\frac{\pi^2}{3}\right) + \Delta_{N-1}\left((\tau_{N-1}-\tau_N)+\frac{\pi^2}{3}\right) + \cdots + \Delta_2\left((\tau_2-\tau_3)+\frac{\pi^2}{3}\right)\\
&= \frac{\pi^2}{3}\sum_{k=2}^{N}\Delta_k + \tau_N\Delta_N + \sum_{k=2}^{N-1}(\tau_k-\tau_{k+1})\,.
\end{aligned}$$

□

D Proofs of Results in Section 4.2 and Tighter Regret Bounds

We prove our improved computational complexity result:

Proof of Lemma 4. Let $S(t) = S'(t)\cup k_t$ for $S'(t)$ defined as in Algorithm 2. Define $E_k(t) = \{\exists t' \in [t-\tau, t] : k \in S(t')\}$ as the event that expert k has been updated at least once in the last τ timesteps from t, for some τ to be chosen. We have that

$$\begin{aligned}
\mathbb{P}(E_k^C(t)) &= \mathbb{P}(\nexists t' \in [t-\tau,t] : k \in S(t')) \le \mathbb{P}(\nexists t' \in [t-\tau,t] : k \in S'(t))\\
&\le \left(1-\frac{\log N}{N-1}\right)^{\tau} \le \left(1-\frac{\log N}{N}\right)^{\tau}.
\end{aligned}$$

Recall that in vanilla D-UCB, we have that for $t \ge \tau_1$, $\mathbb{P}(U_k^*(t) \le \mu^*) \le t^{-2}$(analog to Lemma 15). In case of D-UCB-lite, we can write for $t \ge \tau_{1,\beta}$,

$$\begin{aligned}
\mathbb{P}(B_k^*(t) \le \mu^*) &= \mathbb{P}(B_k^*(t) \le \mu^*, E_{k^*}(t))\\
&\quad + \mathbb{P}(B_k^*(t) \le \mu^*, E_{k^*}^C(t))\\
&\le \mathbb{P}(B_k^*(t) \le \mu^* | E_{k^*}(t)) + \mathbb{P}(E_{k^*}^C(t))\\
&= \mathbb{P}(U_{k^*}(\lfloor t'^{\frac{1}{\beta}}\rfloor) \le \mu^*) + \mathbb{P}(E_{k^*}^C(t))\\
&\text{(for some } t' \in [t-\tau, t])\\
&\le \frac{1}{t'^{\frac{2}{\beta}}} + \left(1-\frac{\log N}{N}\right)^{\tau}\\
&\le \frac{1}{(t-\tau)^{\frac{2}{\beta}}} + \left(1-\frac{\log N}{N}\right)^{\tau}\\
&= \frac{1}{t^2} + \frac{1}{(t-2c\log t)^{\frac{2}{\beta}}},
\end{aligned}$$

where the final inequality holds for the choice of $\tau = 2c\log t$. Similarly, for any suboptimal arm k, after time $\tau_{k,\beta}$, we can write $\mathbb{P}(B_k(t) > \mu^*) \le t^{-2} + (t-2c\log t)^{-\frac{2}{\beta}}$.

The regret result follows using these two statements and the regret decomposition in the proof of our regret theorem in Appendix B.4. □

The result of Theorem 3 implies that average regret of D-UCB is bounded by a problem-dependent constant (for any $i \in [N]$, τ_i is a constant). However, there are two things to note here: First, the upper bound expression is linear in the number of experts N and, second, the algorithm updates each of these N experts once per iteration. In this section, we further examine these observations.

Scaling with N: We first provide the regret improvement result in the case where the suboptimality gaps are drawn from a generative model. Specifically, consider a generative model where $\Delta_3 \leq \cdots \leq \Delta_N$ are the order statistics of $N-2$ random variables drawn i.i.d. uniform over the interval $[\Delta_2, 1]$. Let p_Δ denote the measure over these Δ's.

Corollary 19.1. *Under the generative model p_Δ above, the regret of Algorithm 1 can be bounded as* $\mathbb{E}_{p_\Delta}[R(T)] = O(\frac{M^4 \log N \log^2(1/\Delta_{(2)})}{\Delta_{(2)}})$.

Proof of Corollary 19.1. The analog to Lemma 16 in the full information setting guarantees that for arm k, $\mathbb{P}(U_k(t) \geq \mu^*) \leq t^{-2}$ for $t > \frac{\alpha C^2 M^2 \log^2(\frac{6C}{\Delta_2})}{\Delta_k^2}$ for an α large enough. Note here Δ_k in the log term has been replaced with a smaller Δ_2. The regret can now be decomposed as

$$\begin{aligned} R(T) &\leq \frac{\pi^2}{3}\sum_{k=2}^{N}\Delta_k + \frac{\alpha C^2 M^2 \log^2\left(\frac{6C}{\Delta_2}\right)}{\Delta_N} + \sum_{k=2}^{N-1}\frac{\alpha C^2 M^2 \log^2\left(\frac{6C}{\Delta_2}\right)}{\Delta_k}\left(1 - \frac{\Delta_k^2}{\Delta_{k+1}^2}\right) \\ &\leq O\left(\frac{M^4 \log^2\left(\frac{1}{\Delta_2}\right)}{\Delta_2}\left(1 + \sum_{k=2}^{N-1} 1 - \frac{\Delta_k^2}{\Delta_{k+1}^2}\right)\right). \end{aligned}$$

It only remains to prove that under the considered generative model, the inner expression is $O(\log N)$. We start by observing that by Jensen's inequality, we get

$$1 - \mathbb{E}_{p_\Delta}\left[\frac{\Delta_k^2}{\Delta_{k+1}^2}\right] \leq 1 - \mathbb{E}_{p_\Delta}\left[\frac{\Delta_k}{\Delta_{k+1}}\right]^2.$$

Let $X = \Delta_k, Y = \Delta_{k+1}$ for some $k \geq 3$. Then, we have that the joint distribution of X, Y under the generative model to be

$$f(x,y) = \frac{(N-1)!}{(k-1)!(N-k-3)!}\left(\frac{x-\Delta_2}{1-\Delta_2}\right)^{k-1}\left(1 - \frac{y-\Delta_2}{1-\Delta_2}\right)^{N-3-k} \cdot \frac{1}{(1-\Delta_2)^2}.$$

Thus, we have

$$\begin{aligned} \mathbb{E}[X/Y] &= \int_{y=\Delta_2}^{1}\int_{x=\Delta_2}^{y}\left(\frac{x}{y}\frac{(N-1)!}{(k-1)!(N-k-3)!}\left(\frac{x-\Delta_2}{1-\Delta_2}\right)^{k-1} \cdot \left(1 - \frac{y-\Delta_2}{1-\Delta_2}\right)^{N-3-k} \cdot \frac{1}{(1-\Delta_2)^2}dxdy\right) \\ &= \int_{b=0}^{1}\int_{a=0}^{b}\left(\frac{(1-\Delta_2)a + \Delta_2}{(1-\Delta_2)b + \Delta_2} \cdot \frac{(N-1)!}{(k-1)!(N-k-3)!} \cdot a^{k-1}b^{N-3-k}dxdy\right) \\ &\leq \int_{b=0}^{1}\int_{a=0}^{b}\left(\frac{a}{b} \cdot \frac{(N-1)!}{(k-1)!(N-k-3)!}a^{k-1}b^{N-3-k}dxdy\right) \\ &= \frac{k}{k+1}. \end{aligned}$$

Therefore, we have

$$\mathbb{E}_{p_\Delta}\left[\sum_{k=2}^{N-1} 1 - \frac{\Delta_k^2}{\Delta_{k+1}^2}\right] \leq 1 + \sum_{k=3}^{N-1} 1 - \frac{k^2}{(k+1)^2} = 1 + \sum_{k=3}^{N-1} \frac{2k+1}{(k+1)^2} \leq 1 + \sum_{k=3}^{N-1} \frac{2}{(k+1)} \leq 1 + 2\log N.$$

This completes the proof. □

Note that this corollary shows that under the defined generative model for the gaps in the means of experts, the dependence on N can be improved from linear to *logarithmic*.

E Proofs for Results in Section 6

We begin by proving that with A, n as defined, with high probability if each expert is played A times, each context $x \in \mathcal{X}$ is seen n times.

Proof of Lemma 7. We fix a context x and expert i arbitrary. Then, we have that

$$\begin{aligned}\mathbb{P}(x \text{ seen} < n \text{ times after } A \text{ pulls of } i) \leq \mathbb{P}(Z \leq n) &\leq \exp\left(-2A\left(p(x) - \frac{n}{A}\right)^2\right) \\ &\leq \exp\left(-2A\left(p_X^2 - 2\frac{np_X}{A}\right)\right) = \frac{1}{|\mathcal{X}|NT\sqrt{E}}.\end{aligned}$$

In the above, $Z \sim Binomial(A, p(x))$ and the final inequality uses the definition of A. We say expert i is incomplete if there exists a context $x \in \mathcal{X}$ s.t. x has been seen $< n$ times after A pulls of arm i. Then, a series of union bounds gives us the result.

$$\begin{aligned}\mathbb{P}(\mathcal{E}^C) = \mathbb{P}(\exists i \in [N] \text{ incomplete}) &\leq \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \mathbb{P}(x \text{ seen} < n \text{ times in } A \text{ pulls of arm } i) \\ &\leq \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|NT\sqrt{E}} = \frac{1}{T\sqrt{E}}\end{aligned}$$

□

Next, we provide the final regret result for the agent bootstrapped with A samples per expert.

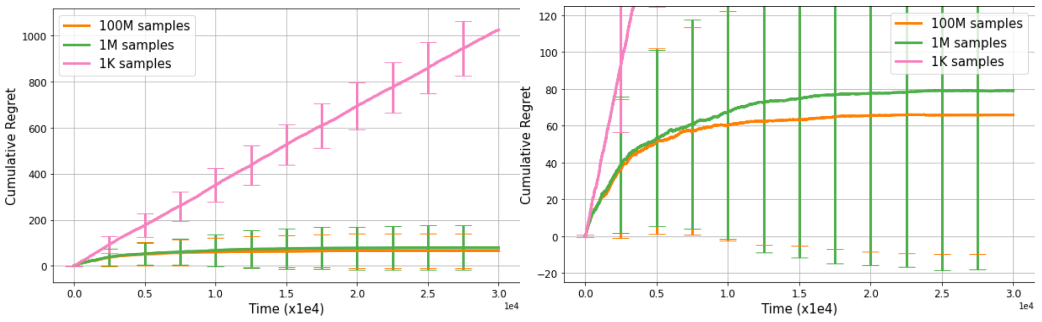
Proof of Theorem 8. Lemma 7 gives us that with probability at least $1 - \frac{1}{T\sqrt{E}}$, each expert has at least n samples before the agent interacts with the environment. Under this event, using Lemma 9, the agent can build empirical experts with maximum error ξ w.p. at least $1 - T^{-1}$. Therefore, conditioning on the event $\mathcal{E}$ and then on the event that the approximate experts are accurate, we have the following chain:

$$\begin{aligned}R(E,T) = \sum_{e \in [E]} R_e(T) &\leq \frac{1}{T\sqrt{E}} \times ET + \left(1 - \frac{1}{T\sqrt{E}}\right)\left(\frac{1}{T} \times ET + \left(1 - \frac{1}{T}\right)\sum_{e=1}^{E} R(\Delta_e)\right) \\ &\leq \sqrt{E} + E + \left(\sum_{e=1}^{T} R(\Delta_e)\right)\left(1 + \frac{1}{T^2\sqrt{E}}\right).\end{aligned}$$

□

F Additional Empirical Evaluations

In this section, we will provide a few additional experimental results that shed light on some of the key assumptions we use to develop D-UCB and ED-UCB. First, we will see how the precision of the empirical policy estimates affects the regret of ED-UCB and then move on to the assumption of bounded divergence between experts.



(a) Full-size figure (b) Zoomed in

Fig. 3. Precision of empirical estimates on regret of ED-UCB: The experiment consists of one episode of 3×10^4 steps. The legend indicates the number of samples used to form the empirical expert policies used by ED-UCB in Algorithm 3. Plots are averaged over 300 independent runs. The results suggest that using estimates with higher precision leads to lower regret.

F.1 Precision of Empirical Estimates

For the results shown in Figure 3, we consider the same setting as our image classification setup in Section 8. Instead of using Theorem 8 to instruct the number of samples to be used to form the empirical expert policies, we set a fixed budget of $100M$, $1M$, and $1K$ samples. Then, we split these samples across the different contexts using a fixed context distribution. We reuse this same distribution in the online evaluation. We then spawn three instances of ED-UCB with ξ being set at the value recommended by 6, which is inaccurate for all three sets of empirical estimates. All other parameters are kept unchanged. As is to be expected, the results show that there is an inverse relationship between the number of samples used to compute the empirical policies (or equivalently, their precision) and the regret they achieve.

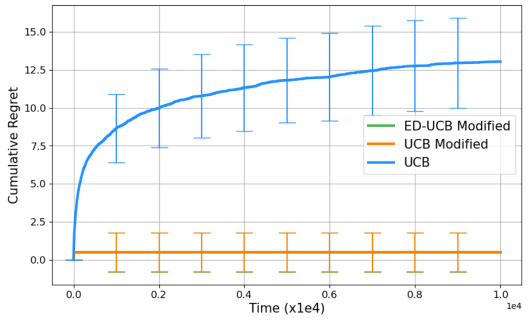
However, we note the following: We believe that our recommendation for the number of samples to use in Lemma 7 is loose. The constant regret achieved using 1M samples in Figure 3 provides potential evidence of this. However, we stress that the results displayed in the figure are only a random sample path of choosing a random context distribution, partitioning the samples randomly across contexts, and sampling random rewards according to this distribution. That is, the figure by itself is not proof that using lesser samples than recommended can guarantee constant regret.

F.2 Infinite Divergence

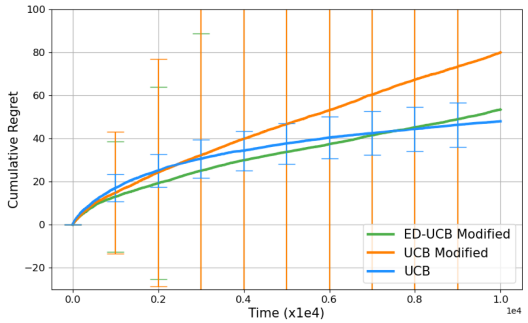
Here, we consider the case of unbounded divergence where our D-UCB and ED-UCB algorithms are not suitable. In this case, the M_ij measure is infinite for some tuple of experts i, j. This occurs when there exists an arm that is never recommended by expert j. To fit this, we create modified versions of the two algorithms above. Specifically, we replace $D_{f_1}(\pi_i \| \pi_j), M_{ij}$ in D-UCB and $\underline{D}_{f_1}(i \| j), \underline{M}_{ij}$ in ED-UCB with

$$D'_{f_1}(\pi_i \| \pi_j) = \sum_{x \in \mathcal{X}} p(x) \sum_{v \in \mathcal{V}: \pi_j(v|x) > 0} f\left(\frac{\pi_i(v|x)}{\pi_j(v|x)}\right) \pi_j(v|x), \qquad M'_{ij} = 1 + \log(1 + D'_{f_1}(\pi_i \| \pi_j))$$

$$\underline{D}'_{f_1}(i \| j) = p_X \sum_{x \in \mathcal{X}} \sum_{v \in \mathcal{V}: \hat{\pi}_j(v|x) > 0} \left(\hat{\pi}_j(V|X) - \xi\right) f_1\left(\underline{r}_{ij}(V|X)\right), \quad \underline{M}'_{ij} = 1 + \log(1 + \underline{D}'_{f_1}(i \| j)).$$



(a) Regret under Reward Setting 1



(b) Regret under Reward Setting 2

Fig. 4. Unbounded divergence with modified D-UCB and ED-UCB: The experiments consist of one episode of 3×10^3 steps. Plots are averaged over 250 independent runs. We consider modified versions of our proposed algorithms and the toy environments with setups detailed in Section F.2, where the maximal divergence between a pair of experts is unbounded. In the first setting, the mean reward and the probability of picking the problematic arm are low. Therefore, it does not affect the regret much and thus we only suffer constant regret as before. However, in the second setting, the low-probability problematic arm has high mean reward, thus leading to logarithmic exploration much like vanilla UCB.

The above quantities are finite surrogates for the potentially infinite original divergence measures. They are constructed by ignoring the terms from the problematic arm. We note here that these modified versions are only sensible when the zero-probability arm is chosen with low probability. This modification makes it so the Clipped importance sampling estimates are always well defined. Further, samples of some arm $v \in \mathcal{V}$ that is only picked by expert $k \in [N]$ are only used to update the estimates of this expert and have no impact on any other estimates. This also leads to unequal number of samples being used to compute the estimates of different experts.

For the experiments, we create toy environments with two contexts X_1, X_2 and three actions A_1, A_2, A_3. We consider two experts operating on these environments with expert policies and reward settings summarized in Tables 1 and 2.

As Expert 2 never picks arm A_3, the upper bound on the divergence is no longer finite. For context distribution, we pick X_1 with probability 0.55 and X_2 with probability 0.45. We set $p_X = 0.1, p_V = 0.05$ and note that the value of p_V is inaccurate (the true value is 0). In Figure 4, we present the results of our experiments on both reward settings. Under Setting 1, the rewards from A_3 as well as the chance of Expert 1 picking it are small. Thus, it does not affect the mean rewards of the experts much. This leads to our modified methods achieving constant regret. However, in Setting 2, this arm is associated with high rewards, leading to the high uncertainty. In this case,

Table 1. Expert Policies

Expert 1	A_1	A_2	A_3		Expert 2	A_1	A_2	A_3
X_1	0.8	0.1	0.1		X_1	0.2	0.8	0.0
X_2	0.2	0.8	0.1		X_2	0.8	0.2	0.0

Table 2. Reward Distributions

Reward Setting 1	A_1	A_2	A_3		Reward Setting 2	A_1	A_2	A_3
X_1	0.9	0.1	0.1		X_1	0.1	0.1	0.9
X_2	0.1	0.9	0.1		X_2	0.1	0.1	0.9

the performance of the modified D-UCB algorithm is comparable to that of vanilla UCB, while modified ED-UCB is slightly worse. We note that ED-UCB is subpar in this case, as it works off of inaccurate information about the environment (specifically, its value of p_V).

Acknowledgements

We sincerely thank all the reviewers and editors for the feedback that helped improve the quality of the article.

References

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

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47, 2 (2002), 235–256.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32, 1 (2002), 48–77.

[4] Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. 2013. Sequential transfer in multi-armed bandit with finite set of models. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2220–2228.

[5] Jonathan Baxter. 1998. Theoretical models of learning to learn. In *Learning to Learn*. Springer, 71–94.

[6] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *J. Mach. Learn. Res.* 14, 1 (2013), 3207–3260.

[7] Sébastien Bubeck and Nicolo Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721* (2012).

[8] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. 2013. Bandits with heavy tail. *IEEE Trans. Inf. Theor.* 59, 11 (2013), 7711–7717.

[9] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. 2013. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* (2013), 1516–1541.

[10] Semih Cayci, Atilla Eryilmaz, and Rayadurgam Srikant. 2019. Learning to control renewal processes with bandit feedback. *Proc. ACM Measur. Anal. Comput. Syst.* 3, 2 (2019), 1–32.

[11] Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. 2020. Meta-learning with stochastic linear bandits. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1360–1370.

[12] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research* 14, 101 (2013), 3207–3260. Retrieved from http://jmlr.org/papers/v14/bottou13a.html

[13] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 208–214.

[14] Aniket Anand Deshmukh, Urun Dogan, and Clayton Scott. 2017. Multi-task learning for contextual bandits. *arXiv preprint arXiv:1705.08618* (2017).

[15] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. 2011. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369* (2011).
[16] Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. 2018. Practical contextual bandits with regression oracles. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1539–1548.
[17] Dylan Foster and Alexander Rakhlin. 2020. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3199–3210.
[18] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (2015), 1–19.
[19] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009). https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[20] Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. 2021. Meta-Thompson sampling. *arXiv preprint arXiv:2102.06129* (2021).
[21] John Langford and Tong Zhang. 2007. Epoch-greedy algorithm for multi-armed bandits with side information. *Advan. Neural Inf. Process. Syst.* 20 (2007), 1.
[22] John Langford and Tong Zhang. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'08)*. 817–824.
[23] Finnian Lattimore, Tor Lattimore, and Mark D. Reid. 2016. Causal bandits: Learning good interventions via causal inference. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'16)*. 1181–1189.
[24] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press.
[25] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. 661–670.
[26] Ashique Rupam Mahmood, Hado Van Hasselt, and Richard S. Sutton. 2014. Weighted importance sampling for off-policy learning with linear function approximation. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'14)*. 3014–3022.
[27] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* 11, Aug. (2010), 2287–2322.
[28] Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. 2014. Machine learning for targeted display advertising: Transfer learning in action. *Mach. Learn.* 95, 1 (2014), 103–127.
[29] Alex Rubinsteyn and Sergey Feldman. 2016. fancyimpute: An Imputation Library for Python. Retrieved from https://github.com/iskandr/fancyimpute
[30] Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. 2017. Identifying best interventions through online importance sampling. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3057–3066.
[31] Rajat Sen, Karthikeyan Shanmugam, and Sanjay Shakkottai. 2018. Contextual bandits with stochastic experts. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 852–861.
[32] David Simchi-Levi and Yunzong Xu. 2022. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research* 47, 3 (2022), 1904–1931.
[33] Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. 2016. Efficient algorithms for adversarial contextual learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2159–2168.
[34] Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to Learn*. Springer, 181–209.
[35] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. 2003. *Inequalities for the L1 Deviation of the Empirical Distribution*. Technical Report. *Hewlett-Packard Labs, Tech. Rep.* http://shiftleft.com/mirrors/www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf
[36] Jiaqi Yang, Wei Hu, Jason D. Lee, and Simon S. Du. 2020. Provable benefits of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531* (2020).
[37] Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N. Negahban. 2019. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301* (2019).
[38] Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1077–1086.

Received 6 November 2023; revised 2 July 2024; accepted 3 July 2024