Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits

Wenshuo Guo
University of California, Berkeley

Kumar Krishna Agrawal
University of California, Berkeley

Aditya Grover
UCLA

Vidya Muthukumar
Georgia Institute of Technology

Ashwin Pananjady
Georgia Institute of Technology

Abstract

We introduce the "inverse bandit" problem of estimating the rewards of a multi-armed bandit instance from observing the learning process of a low-regret demonstrator. Existing approaches to the related problem of inverse reinforcement learning assume the execution of an optimal policy, and thereby suffer from an identifiability issue. In contrast, we propose to leverage the demonstrator's behavior en route to optimality, and in particular, the exploration phase, for reward estimation. We begin by establishing a general information-theoretic lower bound under this paradigm that applies to any demonstrator algorithm, which characterizes a fundamental tradeoff between reward estimation and the amount of exploration of the demonstrator. Then, we develop simple and efficient reward estimators for upper-confidence-based demonstrator algorithms that attain the optimal tradeoff, showing in particular that consistent reward estimation—free of identifiability issues—is possible under our paradigm. Extensive simulations on both synthetic and semi-synthetic data corroborate our theoretical results.

1 Introduction

Reward specification plays a crucial role in building safe and reliable machine learning systems that are aligned with human values (Amodei et al., 2016). However, as pointed out in the extensive behavioral science literature, it is challenging to achieve this alignment, and hand-designed rewards are often misspecified (Anderson, 2001; Gershman and Niv, 2015; MacGlashan and Littman, 2015; Bouneffouf et al., 2017; Gershman, 2018). The paradigm of *inverse* reinforcement learning (IRL) presents a compelling workaround to explicit reward specification, and leverages the implicit optimality in expert demonstrations to infer a reward function. In particular, this paradigm places emphasis on behavioral demonstrations—that is, the demonstrator's actions themselves—as reflecting human values. Popular types of IRL include imitating the optimal policy (Ho and Ermon, 2016; Li et al., 2017b), apprenticeship learning (Abbeel and Ng, 2004), meta-learning (Finn et al., 2017) and learning the reward function (the original formulation of IRL) (Ng et al., 2000; Ramachandran and Amir, 2007; Ziebart et al., 2008; Suay et al., 2016).

Arguably the most outstanding challenge in reward-based IRL is that the reward function may not be uniquely identifiable from the agent's behavior, and infinitely many reward functions can explain the demonstrator's actions. This issue is particularly pronounced when we assume demonstrations from the *optimal* policy (Ng et al., 2000), and subsequent work in IRL has developed heuristics to regularize the space of reward functions depending on how well they explain behavior (Ziebart et al., 2008; Ramachandran and Amir, 2007). Even so, some of these approaches, including maximum-entropy IRL (Ziebart et al., 2008), still suffer from their own identifiability issues.

The central message of this paper is that the reward identifiability issue can be alleviated even in the case where we have a *single* demonstration, provided the demonstrator improves over time by exploration and then exploitation. In other words, such a demonstra-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

tor begins her trajectory facing an unknown environment, explores the environment through a sequence of actions, and eventually settles on an (approximately) optimal policy. Coincidentally, the original introduction of the IRL problem to the AI community did involve learning from this type of evolving demonstrator (Russell, 1998). Concretely, Russell (1998) frames the goal of IRL as being *"to output the reward function that the agent is optimizing...given measurements of an agent's behavior over time"*, and asks whether we can determine the reward function "by observation *during*, rather than *after*, learning". Indeed, the process of policy improvement leaks information: *when* the demonstrator ceases to use a suboptimal policy might contain useful signal about *how* suboptimal that policy is. This, in turn, provides more information about the reward function than observations solely of the optimal policy.

We make this intuition formal and provide simple, tractable and optimal reward estimators from demonstrations in the *multi-armed bandit (MAB)* setting that alleviate the aforementioned identifiability issue. Note that the identifiability issue from optimal demonstrations is particularly acute in MAB: this is because no information about the suboptimal arms' rewards is revealed from the optimal demonstration, only the fact that they are suboptimal. In addition to this conceptual motivation, studying the problem in the MAB setting also has several independent motivations. First and most directly, MAB forms the cornerstone of experiment design in several applications: two notable examples are hyperparameter selection in large-scale machine learning (Li et al., 2017a) and protocol selection for battery charging (Attia et al., 2020), where sequential experiment design is performed using popular, off-the-shelf bandit algorithms. Being able to infer the utility of various alternative options from prior experimentation holds substantial value, as we can use this inference to assess the performance of all configurations that were involved in the experiment without actually rerunning the experiment itself (which may be expensive). Second, it is also worth noting that humans frequently face MAB problems in the real world (Anderson, 2001; Bouneffouf et al., 2017). It is often desirable to make inferences about their intrinsic preferences (e.g. a latent measure of customer utility) from observing their behavior, which can, in turn, be modeled from observing past interactions with a known environment.

Our paradigm is applicable in both such cases. In contrast to learning purely from the exploitation phase in which the demonstrator pulls the optimal arm, we will use a model for the MAB algorithm—in particular, the temporal information revealed by the choices of which arms to pull over the course of the algorithm—to make inferences about the suboptimal arms.

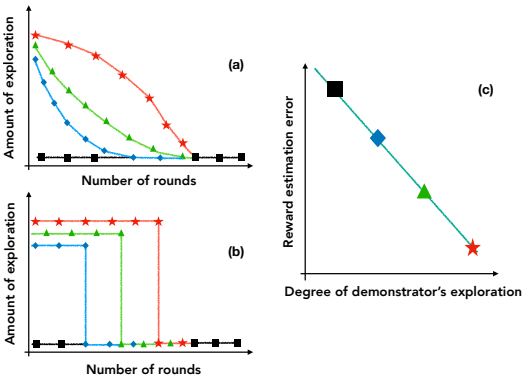


Figure 1: An illustration of exploration decreasing along the learning path for algorithms in the (a) upper-confidence-bound (UCB) family (Auer et al., 2002) (b) successive-arm-elimination (SAE) family (Even-Dar et al., 2006). (c) The tradeoff between exploration and reward estimation, signifying that algorithms that explore more are easier to estimate rewards from.

Contributions. We formally introduce the inverse bandit problem and take a fundamental approach to it, providing both information-theoretic lower bounds and provably optimal algorithms. What is notable, and perhaps surprising, from our work is that the demonstrator's sequence of choices can reveal not only the relative suboptimality of arms but also the *extent* of suboptimality, enabling consistent estimation of the reward of each arm from the behavior of a single demonstrator. To the best of our knowledge, this constitutes the first analysis of this type for inverse reward estimation, whether in bandits or RL. In more detail:

- We first derive information-theoretic lower bounds that apply to any demonstrator algorithm (Theorem 3.1), which provide a quantitative tradeoff between exploration and reward estimation. This is illustrated schematically in Figure 1(c). In the special case of two arms, these bounds show that reward estimation error is inversely proportional to the square root of the *regret* of the demonstrator's algorithm (Corollary 4.3), thereby formalizing our claim from the abstract.

- We develop simple and efficient reward estimation procedures (Procedures 1 and 2) for demonstrations based on the popular *successive-arm-elimination* (SAE) (Even-Dar et al., 2006) and *upper-confidence-bound* (UCB) (Lai and Robbins, 1985) algorithms, and prove upper bounds on the estimation error which match our lower bounds (Theorem 4.2). Both algorithms can be naturally parameterized by their

amount of in-built exploration, and are schematically represented in Figures 1(a) and 1(b), respectively. In particular, these show that for either type of demonstrator, exploration can be optimally leveraged in reward estimation, even though the exploration schedule takes different forms[1].

- Our theory is corroborated by extensive simulations and semi-synthetic experiments (e.g. on battery charging and gene expression datasets).

After discussing related work in the next subsection, we provide background on bandit algorithms and regret and formally state the inverse bandit problem in Section 2. Section 3 contains statements and discussions of our main theoretical results, and we present and discuss our experiments in Section 5. We conclude with a discussion of future work in Section 6.

Related work. *Alternative IRL paradigms.* Recent work has explored easier settings that have the scope to avoid reward identifiability issues in IRL by assuming either additional structure on the reward or access to side-information (Amin et al., 2017; Gershman, 2016; Geng et al., 2020; Ballard and McClure, 2019; Jeon et al., 2020; Fu et al., 2017). In contrast to this line of work, our reward estimation procedure recovers the exact reward function in the limit, without additional assumptions, for a natural class of low-regret demonstrations. A parallel line of work on learning from demonstrations, including imitation learning, studies alternative approaches that directly copy the demonstrators' actions without specifying a reward function (Ho and Ermon, 2016; Li et al., 2017b). While these approaches have had many successful applications, the lack of reward identification limits their use in others. For instance, planning across multiple environments—with different transition dynamics—cannot be accomplished purely by imitation learning since the optimal policy can vary significantly. On the other hand, a learned reward function can be used to transfer knowledge across environments.

Learning from "improving demonstrators". Our paradigm of learning from an exploring demonstrator is similar in spirit to a line of recent work on learning from improving demonstrators (Gao et al., 2018; Jacq et al., 2019; Wu et al., 2019; Balakrishna et al., 2020; Ramponi et al., 2020). We highlight two key differences in both the setting and theoretical scope. First, we show that reward estimation is possible from observing a *single* demonstration, and consistent estimation is obtainable as the horizon of the demonstration grows. On the other hand, related work on learning from improving demonstrators is based on estimating population-based quantities arising from RL algorithms like soft policy iteration on gradient descent, and requires a large number of demonstrations for estimation. At a lower level, our analysis proves not only consistency, but also non-asymptotic guarantees on reward estimation that are matched by information-theoretic lower bounds. On the other hand, there are no finite-sample guarantees available in the literature on learning from improving demonstrators, optimal or otherwise.

Finally, we mention that IRL in bandits has been considered by two recent papers, but the settings are motivated by social choice (Noothigattu et al., 2021) and imitation/assisted learning (Chan et al., 2019), as opposed to reward learning from a single demonstration.

2 Preliminaries

We formally define the problem of reward estimation in a multi-armed bandit (MAB) instance from a single demonstrator who uses a low-regret algorithm. We begin by setting up standard notation for the MAB problem, and by formally defining (pseudo-)regret (Lattimore and Szepesvári, 2020; Slivkins, 2019).

2.1 Multi-armed bandits (MAB) and regret minimization

Consider a K-armed bandit instance with action (arm) set $[K] := \{1, 2, \ldots, K\}$. Every time an arm $i \in [K]$ is pulled, a random reward is generated, independently of past actions, from an unknown probability distribution ν_i. We assume that the distribution ν_i is supported on the interval $[0, 1]$ for each $i \in [K]$, and denote by $\mu_i = \mathbb{E}_{X\sim\nu_i}[X]$ the expected reward of arm i. We assume throughout this paper that there is a unique best arm with highest expected reward. We let $i^* := \arg\max_{i\in[K]}\mu_i$ denote the index of the best arm, and use $\mu^* := \mu_{i^*}$ to denote its reward. We refer to the remaining arms as suboptimal arms, and define the *suboptimality gap* of arm $i \neq i^*$ to be $\Delta_i := \mu^* - \mu_i$. Owing to the uniqueness of the best arm, note that $\Delta_i > 0$ for all $i \neq i^*$.

The demonstrator takes actions on the bandit instance with the goal of maximizing her accumulated reward over a finite horizon consisting of T rounds. At each round $t \in \{1, 2, \ldots, T\}$ and based on her observations thus far, the demonstrator pulls an arm $I_t \in [K]$ and

[1] While Section 4 provides exact details of both the SAE and UCB algorithms, we provide a short high-level description here. While SAE and UCB algorithms both trade off exploration and exploitation in a similar manner, their day-to-day behavior diverges sharply. In particular, SAE-based algorithms have a marked transition between their exploration and exploitation phases. On the other hand, in UCB-based algorithms the amount of exploration reduces smoothly with time (as reflected in Figure 1(b).

receives a reward $r_t \sim \nu_{I_t}$. Define $n_{i,t}$ to be the number of times arm i has been pulled up to time t. It is also useful to define the empirical reward estimate of arm i at time t as $\bar{\mu}_{i,t} = \left(\sum_{s=1}^{t} r_s \cdot \mathbb{I}\{I_s = i\}\right)/n_{i,t}$. The performance of the demonstrator is measured by her *regret*, which quantifies the difference between the best possible reward she could accrue if she knew which arm was the best one, and the actual accumulated reward.

Definition 2.1 (Pseudo-regret (Lattimore and Szepesvári, 2020))**.** The (expected) pseudo-regret is

$$\mathbb{E}[R_T] = T\mu^* - \mathbb{E}\Big[\sum_{t=1}^{T} r_t\Big] = \sum_{i \in [K]} \Delta_i \mathbb{E}[n_{i,t}].$$

A *low-regret* demonstrator is one whose regret scales *sublinearly* in T with high probability, that is, $R_T = o(T)$ with probability that goes to 1 as $T \to \infty$.

2.2 The "inverse bandit" problem

The "inverse bandit" problem is to estimate the expected rewards $\{\mu_i\}_{i \in [K]}$ of a multi-armed bandit instance from observing only the actions of a demonstrator algorithm. Importantly, we *do not* observe the rewards accrued at each round. Consider a demonstration consisting of the sequence of actions $\{I_t\}_{t=1}^{T}$. A reward estimation *procedure* is a mapping from $\{I_t\}_{t=1}^{T} \mapsto \{\widehat{\mu}_i\}_{i \neq i^*}$, where $\widehat{\mu}_i$ denotes the mean estimate arm i. The goal of the reward estimation procedure is to minimize the expected estimation error for each arm[2] i, given by $\mathbb{E}[|\widehat{\mu}_i - \mu_i|]$. Here, the expectation is taken over the randomness of the received rewards and the sequence of actions. Furthermore, since the behavior of any natural demonstrator is invariant to constant shifts of all expected rewards, we assume that the procedure has access to the value of μ^* (but not the index i^*) to avoid trivial identifiability issues. Note that one can remove this recentering assumption and instead consider estimating the suboptimality gaps.

Note that this goal of *estimation* is significantly more challenging than simply *ranking* the arms: the latter problem is solvable by ordering the arms according to their pull counts, but does not produce cardinal reward values. Nevertheless, we will show shortly that reward estimation can indeed be performed from observing a single trajectory from a natural class of demonstrators.

[2]Our guarantees are most natural to state on the stringent arm-by-arm error metric, and yield ℓ_p guarantees.

3 Fundamental Limits on Reward Learning

To provide a concrete baseline, we first prove information-theoretic lower bounds showing a fundamental tradeoff between reward estimation and exploration, regardless of the specific reward estimation procedure and the demonstrator's learning algorithm. At a high level, the identifiability issues that arise in IRL already suggest that exploration is necessary for nontrivial reward estimation; our lower bound makes this formal. We then present some intuitive but unsuccessful attempts to achieve this lower bound.

3.1 Information-theoretic lower bound

The following theorem collects our lower bound.

Theorem 3.1. *(Proof in Appendix A) For every K-armed Bernoulli bandit instance $\mathcal{M}$ satisfying $\max_{i \in [K]} |\mu_i - 1/2| \leqslant 1/4$ and for each suboptimal arm $i \neq i^*$, the following is true. Suppose that the demonstrator employs algorithm $\mathcal{A}$, and let $\mathbb{E}[n_{i,T}^{\mathcal{A}}]$ denote the expected number of times arm i is pulled by $\mathcal{A}$ when presented the instance $\mathcal{M}$. Then there exists an instance $\mathcal{M}'$ such that for any reward estimation procedure having knowledge of μ^* and mapping $\{I_1, \ldots, I_T\} \mapsto \{\widehat{\mu}_i\}_{i \in [K], i \neq i^*}$,*

$$\max_{\widetilde{\mathcal{M}} \in \{\mathcal{M}, \mathcal{M}'\}} \mathbb{E}[|\widehat{\mu}_i - \mu_i(\widetilde{\mathcal{M}})|] \geqslant \frac{1}{16} \cdot \left(\frac{1}{\sqrt{\mathbb{E}[n_{i,T}^{\mathcal{A}}]}} \wedge 1 \right).$$

Here $\mu_i(\mathcal{M})$ denotes the i-th reward mean of the bandit instance $\mathcal{M}$.

Note that in addition to applying to any reward estimation procedure, Theorem 3.1 provides a fundamental limit for any choice of demonstrator algorithm in terms of the degree of exploration in that algorithm. Its proof utilizes information-theoretic lower bounds on the demonstrator's regret (Kaufmann et al., 2016): even with the strong side information of noisy reward observations, we need sufficiently many pulls of arm i to be able to estimate its reward, since zero information is shared across arms in the MAB setting. Thus, the efficacy of any inverse procedure for estimating μ_i is fundamentally limited by $\mathbb{E}[n_{i,T}]^{-1/2}$.

3.2 Some initial observations

Theorem 3.1 constitutes a fundamental limit on reward estimation from *any* demonstrator algorithm, even if we know the algorithm beforehand. We now make some observations to help assess the types of demonstrator algorithms that allow us to match this lower bound.

The algorithm needs to satisfy instance-adaptivity. Ideally, we would aim to obtain reward estimation guarantees from *any* plausible low-regret algorithm. Unfortunately, such a general statement cannot be true (even if we are satisfied with a worse estimation error bound) as witnessed by the following simple counterexample. Suppose that the demonstrator employs the explore-then-commit algorithm (Lattimore and Szepesvári, 2020) which pulls arms randomly for $\mathcal{O}(T^{2/3})$ rounds, and then pulls the arm with the highest estimated mean reward thereafter. This algorithm achieves regret $\mathcal{O}(T^{2/3})$ for all bandit instances, and so constitutes a no-regret algorithm. However, it is easy to see (since the arm pulls provide no information about the rewards themselves) that nontrivial reward estimation is impossible from observing the actions alone. As this example shows, reward estimation is only possible when the algorithm exhibits some type of instance-dependent behavior (e.g., if the action sequence differs when the suboptimality gaps change).

Algorithm 1 Successive arm elimination (SAE) with $O(T^\alpha)$ regret (for $0 < \alpha < 1$) or $O(\log T)$ regret (for $\alpha = 0$)

Input: K arms, $\alpha \in [0, 1)$, total rounds T.
Initialize: Set SAE epoch $t_r = 1$, active set $\mathcal{S}(1) \leftarrow [K]$ and round $t = 0$.
while $|S(t_r)| > 1$ **do**
  Sample arm $i \in S(t_r)$ once and set $t \leftarrow t + 1$
  Let $\bar{\mu}_{i,t}$ be the average reward of arm i by t
  Set $C_{i,t} \stackrel{\text{def}}{=} \sqrt{\frac{2(T^\alpha - 1)}{\alpha \cdot t_r}}$.
  for each $i \in \mathcal{S}(t_r)$ and $\bar{\mu}_{i,t} \leqslant \bar{\mu}_{\max}(t) - 2C_{i,t}$ **do**
    $\mathcal{S}(t_r) \leftarrow S(t_r) \setminus \{i\}$.
  end for
  $t_r \leftarrow t_r + 1$
end while
Pull arm in $\mathcal{S}$ and set $t \leftarrow t + 1$ until $t = T$.

Does order-wise instance-optimal regret suffice? The next natural question that arises is whether it is possible to estimate the rewards from any algorithm that exhibits (order-wise) optimal instance-dependent behavior, *even when we do not know the specific details of the algorithm.* In particular, we might hope to use the number of pulls of a suboptimal arm by round T, which we denoted by $n_{i,T}$, as a sufficient statistic for our estimation procedure. For example, classic instance-dependent bounds are of the form $n_{i,T} = \Theta\left(\frac{\log T}{\Delta_i^2}\right)$, where the constant inside the $\Theta(\cdot)$ varies across arms. A possible estimator from this relation would be to construct $\widehat{\Delta}_i = C_0\sqrt{\frac{\log T}{n_{i,T}}}$ for each suboptimal arm i, for some choice of common constant C_0. While this estimator possesses the attractive property of being algorithm-agnostic, it turns out to not even be *statistically consistent* (with respect to the number of rounds T), let alone match the fundamental limit given by Theorem 3.1. In fact, an elementary analysis verifies that $|\widehat{\Delta}_i - \Delta_i| = \Theta\left(\sqrt{\frac{\log T}{n_{i,T}}}\right) = \Theta(1)$, and so the estimation error does not decay with T. At a high level, such a "naive" estimator does not effectively exploit the day-to-day structure present in a demonstrator algorithm, and consequently cannot match the lower bound in Theorem 3.1 (also see Appendix F).

Our lower bound and preliminary observations motivate a class of procedures that utilizes[3] the characteristics of structured, instance-adaptive algorithms like *successive-arm-elimination* (SAE) (Even-Dar et al., 2006) and *upper-confidence-bounds* (UCB) (Lai and Robbins, 1985) to perform reward estimation.

Algorithm 2 Upper confidence bound (UCB) with $O(T^\alpha)$ regret (for $0 < \alpha < 1$) or $O(\log T)$ regret (for $\alpha = 0$)

Input: K arms, $\alpha \in [0, 1)$, total rounds T.
Initialize: Set round $t = 1$. Set for every arm a confidence width $C_{i,0} = \infty$.
while $t < T$ **do**
  Pull arm $I_t = \arg\max_{i \in [K]} \bar{\mu}_{i,t-1} + C_{i,t-1}$ (break ties arbitrarily).
  Let $\bar{\mu}_{i,t}$ be the average reward of arm i by time t, and let $n_{i,t}$ be the number of times arm i is pulled by time t.
  Set $C_{i,t} \stackrel{\text{def}}{=} \sqrt{\frac{2(T^\alpha - 1)}{\alpha \cdot n_{i,t}}}$
  $t \leftarrow t + 1$
end while

Note: When $\alpha = 0$, we use that $\lim_{\alpha \to 0} \frac{T^\alpha - 1}{\alpha} = \log T$.

4 Optimal Reward Estimators

Two popular families of algorithms in the MAB literature are *successive-arm-elimination* (SAE) and *upper-confidence-bounds* (UCB), presented formally in Algorithms 1 and 2. While these algorithms differ in their round-by-round details, they are both based on the principle of optimism in the face of uncertainty,

[3] In addition to this conceptual motivation, assuming knowledge of the demonstrator's algorithm is reasonable, e.g., in experiment design settings where algorithms like UCB constitute the "gold standard".

whereby exploration is encouraged by constructing an "optimistic" upper-confidence-bound on the reward of an arm that is a decreasing function of the number of times that arm has been pulled thus far.

The SAE algorithm proceeds in multiple epochs; in each epoch, all active arms are pulled in a round robin fashion and their sample means are maintained. As soon as we observe that a certain arm is obviously suboptimal, we drop it from consideration and render it "inactive", or eliminated. The UCB algorithm instead intertwines exploration with exploitation.

Remark 4.1. *The use of α in Algorithms 1 and 2 is only to obtain a more general class of algorithms with a smooth variation in their regret. In particular, a higher value of α essentially inflates the confidence intervals, allowing for greater exploration. Indeed, both the SAE and UCB algorithm incur a sublinear regret of $O(T^\alpha)$ in high probability for any $\alpha \in [0,1)$ (the statement and proof of this result is in Appendix C for completeness). The smaller the value of α, the smaller the regret and—from the fundamental limits that we characterized in Theorem 3.1—the harder it is to perform reward estimation. The typical choice of $C_{i,t}$ is $\mathcal{O}\left(\sqrt{\frac{\log T}{n_{i,t}}}\right)$, which yields the instance-optimal regret guarantee $\mathcal{O}\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i^2}\right)$, is recovered by taking the limit $\alpha \to 0$. It is important to note that we obtain consistency of estimation even in this case of minimal exploration, and our main ideas are already evident here.*

4.1 Optimal reward estimation

The naive attempts from before suggest that one needs more delicate procedures in order to an optimal (or even consistent) estimator. We now present such estimators for the SAE and UCB algorithms, starting with SAE since the ideas are most intuitive when there is a clear separation between exploration and exploitation.

SAE reward estimator. Note from the description of SAE in Algorithm 1 that the transition from exploration to exploitation is particularly abrupt: for every arm i, there exists (with high probability) a round τ_i at which the condition for arm i to be eliminated is met. More formally, for a typical execution of SAE given by $\{I_1, \ldots, I_T\}$, we define this "switching round" as

$$\tau_i := \{t \geqslant 1 : I_t = i \text{ and } I_{t'} \neq i \; \forall t' > t\}. \qquad (1)$$

Procedure 1 estimates the suboptimality gap of arm i by exactly twice the width of the confidence interval at the switching round τ_i, denoted by C_{i,τ_i}. Figure 2 provides three-fold intuition for why this simple estimator is reasonable in the simplest case of 2 arms (with $i^* = 1$).

Procedure 1 SAE reward estimator

Input: Sequence of actions $\{I_1, \ldots, I_T\}$; scalar μ^*.
Set $\widehat{\imath} \in \arg\max_i n_{i,T}$
for each $i \in [K], i \neq \widehat{\imath}$: **do**
  Compute τ_i according to Eq. (1)
  $\widehat{\mu}_i \stackrel{\text{def}}{=} \mu^* - 2 \cdot C_{i,\tau_i}$.
end for
return $\widehat{\mu}_i$ for $i \in [K]$.

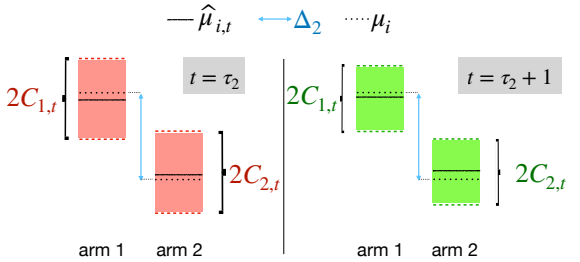


Figure 2: SAE on a 2-armed bandit instance at the rounds τ_2 and τ_2+1. In Procedure 1, we exploit the fact that on both left and right $\Delta_2 \approx C_{1,t} + C_{2,t}$.

First, at round τ_2 arm 2 is still in play; so the sum of the confidence widths $C_{1,\tau_2+1} + C_{2,\tau_2+1}$ must *upper bound* the difference in sample means $\bar{\mu}_{1,\tau_2} - \bar{\mu}_{2,\tau_2}$. This is depicted on the left hand side of Figure 2. Second, at round $\tau_2 + 1$ the condition for elimination of arm 2 must be met; so the sum of the confidence intervals $C_{1,\tau_2+1} + C_{2,\tau_2+1}$ must *lower bound* the difference in the algorithm's sample means $\bar{\mu}_{1,\tau_2+1} - \bar{\mu}_{2,\tau_2+1}$. This is depicted on the right hand side of Figure 2. Putting these together, we obtain an estimator that is close to the difference in sample means $\bar{\mu}_{1,\tau_2+1} - \bar{\mu}_{2,\tau_2+1}$ (which is in turn very close to $\bar{\mu}_{1,\tau_2} - \bar{\mu}_{2,\tau_2}$). Finally, since both arm 1 and arm 2 have been active until switching round τ_2, their confidence widths are identical. This leads to the particularly simple description of the SAE estimator in Procedure 1.

UCB reward estimator. While the details are significantly more complex for UCB, a similar idea works. In this case suboptimal arms could be pulled throughout the decision-making process, but there will still exist (with high probability) a maximal round at which arm i is pulled *and* the optimal arm is pulled at least once there-after. Let $\widehat{\imath}$ denote the index of the arm that is pulled most often in the demonstration; this is our estimate of the optimal arm. The switching round of interest is given by

$$\tau_i := \max\{t : I_t = i \text{ and } I_{t'} = \widehat{\imath} \text{ for some } t' > t\}. \quad (2)$$

Procedure 2 UCB reward estimator

Input: Sequence of actions $\{I_1, \ldots, I_T\}$; scalar μ^*.
Set $\widehat{\imath} \in \arg\max_i n_{i,T}$
for each $i \in [K], i \neq \widehat{\imath}$: **do**
  Compute τ_i according to Eq. (2)
  $\widehat{\mu}_i \stackrel{\text{def}}{=} \mu^* - (C_{i,\tau_i} - C_{\widehat{\imath},\tau_i})$.
end for
return $\widehat{\mu}_i$ for $i \in [K]$.

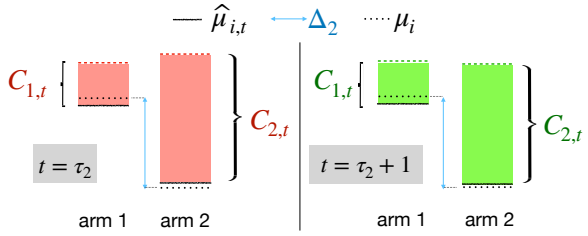


Figure 3: UCB on a 2-armed bandit instance at the rounds τ_2 and τ_2+1. In Procedure 2, we exploit the fact that on both left and right $\Delta_2 \approx C_{2,t} - C_{1,t}$.

Then, Procedure 2 directly estimates the reward of arm i by exactly the difference in confidence widths of arms i and i^* at τ_i. As illustrated in Figure 3 for the case of 2 arms, the confidence widths can be significantly different for the optimal and suboptimal arm at the switching round for the case of UCB. However, similar intuition as in the case of the SAE estimator continues to hold here; once again, arm 2 is suboptimal and we work on the high-probability event that $i^* = \widehat{\imath} = 1$. First, at round τ_2 the upper confidence bound of arm 2 must exceed that of arm 1; therefore, the difference in confidence widths must upper bound the difference in sample means. Second, at round t' the upper confidence bound of arm 1 exceeds that of arm 2; therefore, the difference in confidence widths must lower bound the difference in sample means. Putting these together, we again obtain an estimator that is close to the difference in sample means $\bar{\mu}_{1,\tau_2} - \bar{\mu}_{2,\tau_2}$.

Our main theorem makes the above intuition precise and obtains a unified characterization of the estimation error $|\widehat{\mu}_i - \mu_i|$ for each $i \in [K]$ arising from demonstrations of either SAE or UCB.

Theorem 4.2. *(Proof in Appendix D for SAE, Appendix E for UCB) Suppose*[4] $T \geqslant 64 \sum_{i \neq i^*} \frac{T^\alpha - 1}{\alpha \Delta_i^2}$, *and let* $n_{i,T}$ *denote the number of times arm* i *is pulled by either Algorithm 1 or 2. Denote the total number of arms as* K. *There is a universal positive constant* C *such that for any suboptimal arm* i, *Procedures 1 and 2 satisfy*

$$\mathbb{E}|\widehat{\mu}_i - \mu_i| \leqslant C \sqrt{\frac{\log(\mathbb{E}[n_{i,T}]\sqrt{K})}{\mathbb{E}[n_{i,T}]}}.$$

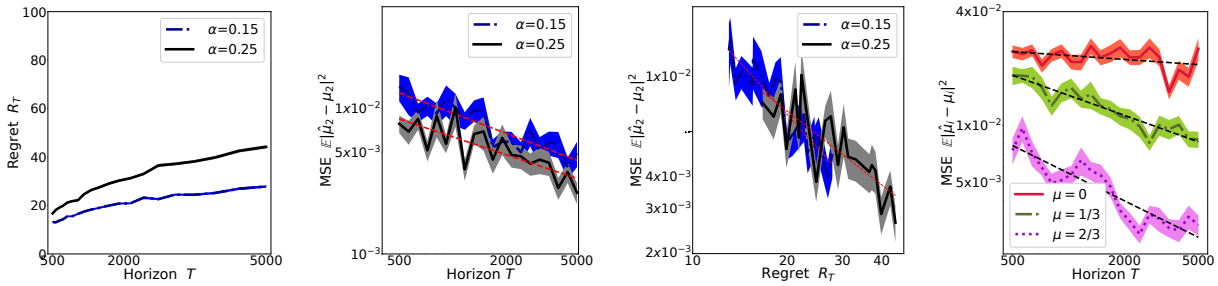
Furthermore, we have $\mathbb{E}[n_{i,T}] \geqslant c \cdot \frac{T^\alpha - 1}{\alpha \Delta_i^2}$ *for a second universal constant* $c > 0$.

Since the map $x \mapsto \log x / x$ is decreasing for large enough x, the two parts of the theorem also provide an upper bound on the estimation error purely in terms of the the tuple (T, α, Δ_i). Nevertheless, we have chosen to state it in terms of the expected number of pulls of arm i so as to bring into sharp focus the effect of exploration on reward estimation. Note that $\mathbb{E}[n_{i,T}]$ measures the degree to which the suboptimal arm i is explored; Theorem 4.2 shows that a larger value of $n_{i,T}$ will lead to a smaller error. The precise quantitative relationship is also compelling: indeed, if we had oracle access to the reward samples accrued over the course of the demonstration, simply averaging them and outputting the sample mean would achieve a rate of the order $n_{i,T}^{-1/2}$. The theorem shows that a similar rate is achievable *solely* using observations of the trajectory itself.

The role of algorithmic hyperparameters. Our procedures were based on knowing not just the particular type of demonstrator algorithm being employed but also its hyperparameters (since these were used to construct the confidence intervals). It is natural to ask if the latter assumption can be relaxed. We note that even without the knowledge of the constants in the confidence widths, the same reward estimation procedures will still able to estimate the suboptimality gaps up to a scaling constant that is *common* to all arms. In particular, such a guarantee would suffice to argue statements of the form "the second arm is twice as suboptimal as the third"; such relative comparisons of the arms' rewards are often sufficient in many applications.

Technical novelty. Let us make a few comments on the technical difficulties involved in proving Theorem 4.2. Figures 2 and 3 suggest that the estimated gap closely tracks the difference in sample means $\bar{\mu}_{\widehat{\imath},\tau_i} - \bar{\mu}_{i,\tau_i}$. The first step is to make this precise: we show that in both cases, the overall estimation error is characterized, up to lower order terms, by the distance from the sample means to the true means at τ_i. The second step is to characterizing the sample-mean estimation error, and is challenging for a number of reasons: (1) the sample means both in UCB and SAE are biased even for a fixed round t due to adaptive

[4] This condition ensures, by Proposition C.1, that $\widehat{\imath} = i^*$ with high probability.



(a) (b) (c) (d)

Figure 4: Results of 100 runs of simulation experiments for the UCB algorithm. Figures (a-c) are for a two-armed bandit instance with $\mu = (1, 1/2)$ and Gaussian rewards with unit variance. Here, individual curves represent two values of $\alpha \in \{0.15, 0.25\}$. Figure (d) is a 4-armed instance with $\mu = (1, 2/3, 1/3, 0)$ and Gaussian rewards with variance 1/4. Here, individual curves represent the three suboptimal arms. Overall, these log-log plots corroborate our principal finding that better reward estimation is achievable from higher regret demonstrations; see the text for a detailed discussion.

sampling (Nie et al., 2018; Shin et al., 2019) (2) the switching round τ_i is itself random for both UCB and SAE, and (3) in the case of UCB, there is a discrepancy between the quantities n_{i,τ_i} and $n_{\widehat{i},\tau_i}$. Substantial technical effort in our proofs goes into constructing high-probability lower-bounds on n_{i,τ_i} and $n_{\widehat{i},\tau_i}$, both of the order of $\mathbb{E}[n_{i,T}]$. The lower bound on $n_{\widehat{i},\tau_i}$ appears to be the first of its kind, and does not follow even from other lower bounds on the total number of pulls of each arm derived in the literature (Syed et al., 2010). Instead, it requires a fine-grained understanding of the *day-to-day* behavior of UCB. We present a case-by-case analysis of UCB to provide these high-probability lower bounds, which may be of independent interest.

4.2 A consequence: reward estimation / regret tradeoffs for two-armed bandits

A key message of our results is that more exploration in the demonstration is both necessary and sufficient for efficient reward estimation, in an arm-by-arm sense. In the special case of a two-armed bandit problem, this tradeoff can be expressed solely in terms of the regret:

Corollary 4.3. *(Informal) Let $i^* = 1$. Procedures 1 and 2 achieve, from a demonstration of SAE or UCB with expected regret $\mathbb{E}[R_T]$, the bound $\mathbb{E}[|\widehat{\mu}_2 - \mu_2|] \lesssim \sqrt{\frac{\Delta_2}{\mathbb{E}[R_T]}}$. Conversely, any reward estimator $\widehat{\mu}_2$ from a demonstration algorithm $\mathcal{A}$ having expected regret $\mathbb{E}[R_T]$ must suffer error $\mathbb{E}|\widehat{\mu}_2 - \mu_2| \gtrsim \sqrt{\frac{\Delta_2}{\mathbb{E}[R_T]}} \wedge 1$.*

The predictions of this corollary and our other results are now verified in numerical experiments.

5 Experiments

We now implement the reward estimators in Procedures 1 and 2 on a range of synthetic bandit instances and on a physics simulator derived from a real-world application in battery charging (Attia et al., 2020; Grover et al., 2018). Further experimental results and more detailed explanations of setups in this domain and including a new domain in gene expression data can be found in Appendix F.

Simulated data. We simulate a $K = 2$ armed bandit instance with Gaussian rewards distribution $X \sim N(\mu_i, \sigma^2)$ for each arm. The arm means μ_i are bounded in the range $[0, 1]$ with $\sigma^2 = 1.0$. Our first set of experiments is based on simulations of Algorithms 1 and 2 (and the corresponding Procedures 1 and 2). The results with two arms and UCB are illustrated in Figure 4; SAE results are similar (see Appendix F).

Panel (a) of Figure 4 verifies that the regret is sublinear in T, with higher values of α incurring larger regret, as predicted by Proposition C.1. In panel (b), we plot the MSE of reward estimation ($\approx$ the square of the quantity in our theorems) from UCB against T, and observe that Procedures 1 and 2 attain smaller error when the algorithm has higher regret, i.e., for larger values of α. We also see different slopes in these plots for different values of α (as predicted by Theorem 4.2), and this motivates the question of whether a common quantity governs the scaling law across different choices of α. Panel (c) confirms that this is indeed the case: the curves collapse onto each other when we plot the MSE against regret, and the slope of the best-fit lines—as predicted by Corollary 4.3—are very close to -1.

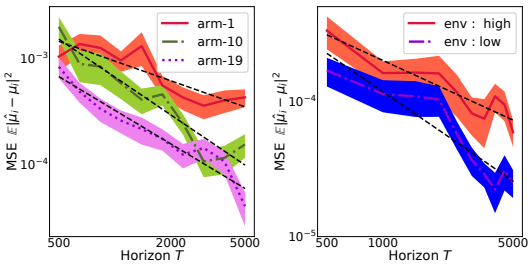


Figure 5: Results from 250 runs of estimating (normalized) battery lifetimes from a UCB experiment design procedure (a variance-adjusted version of Algorithm 2 with $\alpha = 0.25$). (a) Estimation error for a random subset of 3 arms in the "high" regime when algorithm is run on a 20-armed instance. (b) Error of estimating arm 12 in both "low" and "high" regimes with 4 protocols.

In the K-armed case with $K = 4$, panel (d) demonstrate the variation of estimation rates across arms, where arms having large gaps (or lower values of μ) are harder to estimate than those having small gaps. Once again, this corroborates the result of Theorem 4.2, where we saw that the MSE must depend near-linearly on the gap of the arm since arms with larger gaps are pulled less often.

Application: Battery charging. In many scientific domains, we are interested in studying the performance landscape of a set of configurations. For example, in battery charging, there are several electric current protocols for charging an electric battery (Attia et al., 2020). Depending on the chosen protocol and a specified temperature regime, a battery undergoes a different range of chemical reactions that eventually determine its final lifetime. Understanding relationships between charging protocols and induced battery lifetimes for different temperature regimes is crucial to designing the future generation of batteries that operate at a favorable point on this tradeoff.

Our data at hand often consists of the results of experiments that were designed to search for lifetime maximizing configurations, and we would like to estimate the landscape of lifetimes from this data. We can cast this problem as one of reward estimation from an exploring demonstration. In particular, we map every temperature regime to a bandit instance where each charging protocol is an arm and the arm's reward is given by it's expected lifetime. Given a demonstration of sequential experiments (i.e., arm pulls), our goal is to infer the lifetimes of *all* charging protocols.

We consider K distinct charging protocols from Attia et al. (2020) in two temperature regimes: *low* and *high*. These two operating regimes exhibit different ranges of expected battery lifetime: low in [901, 962] and high in [573, 1208]. We obtain lifetime distributions for each protocol by fitting a Gaussian to a mix of real-world experimental data and physical simulations (Attia et al., 2020), and perform our experiments on this semi-synthetic data using the UCB algorithm as a representative experimental design approach. The reward means are normalized to lie in the range $[0, 1]$.

Figure 5(a), plotted in the high temperature regime for $K = 20$ (see Appendix F for other regimes), shows that the estimate for each charging protocol improves as the length of the trajectory T increases. Lifetime estimation is thus possible even in cases where the number of protocols is moderately large. Next, we consider the problem of evaluating a particular charging protocol across temperature regimes with $K = 4$. In Figure 5(b), we plot the estimation error for a representative arm having similar lifetime in both temperature regimes. Here, the behavior in panels (d) of Figure 4 is observed again: since the arm-gap in the low temperature regime is lower than in the high temperature regime, the error of Procedure 2 is correspondingly reduced.

6 Discussion

We introduced and studied the inverse bandit problem of estimating rewards from observing a low-regret demonstrator. We provided information-theoretic lower bounds and simple, optimal reward estimation procedures. Our results quantify a tradeoff between exploration and reward estimation, and are corroborated by extensive synthetic and semi-synthetic experiments. While this work takes a first step towards theoretically optimal reward estimation from an exploring demonstration, many open questions remain. It is interesting to study other demonstrator algorithms, e.g., randomized algorithms, in which the reward estimation comes with new challenges. Tackling these challenges is crucial to deploying this paradigm in scenarios where humans are known to randomize their behavior (Daw et al., 2006; Schulz et al., 2015; Speekenbrink and Konstantinidis, 2015). Another interesting direction is to extend our insights to more expressive settings like contextual bandits, tabular RL, and continuous control.

Acknowledgments

We thank Krishna Acharya and Jim James for their careful reading of a draft of this paper, and for making several important suggestions. We thank Kwang-Sung Jun for sharing the gene expression data that were used for a subset of the experimental results presented in this paper. AG, VM, and AP were supported by research fellowships from the Simons Institute for the Theory of Computing when part of this work was performed. WG acknowledges support from a Google PhD Fellowship; AP acknowledges support from the National Science Foundation grant CCF-2107455.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, page 1, 2004.

Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. In *Neural Information Processing Systems*, pages 1813–1822, 2017.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Christopher Madden Anderson. *Behavioral models of strategies in multi-armed bandit problems*. PhD thesis, California Institute of Technology, 2001.

Peter M Attia, Aditya Grover, Norman Jin, Kristen A Severson, Todor M Markov, Yang-Hung Liao, Michael H Chen, Bryan Cheong, Nicholas Perkins, Zi Yang, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578(7795):397–402, 2020.

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

Ashwin Balakrishna, Brijen Thananjeyan, Jonathan Lee, Felix Li, Arsh Zahed, Joseph E Gonzalez, and Ken Goldberg. On-policy robot imitation learning from a converging supervisor. In *Conference on Robot Learning*, pages 24–41. PMLR, 2020.

Ian C Ballard and Samuel M McClure. Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, 317:37–44, 2019.

Djallel Bouneffouf, Irina Rish, and Guillermo A Cecchi. Bandit models of human behavior: Reward processing in mental disorders. In *International Conference on Artificial General Intelligence*, pages 237–248, 2017.

Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. The assistive multi-armed bandit. HRI '19, page 354–363. IEEE Press, 2019.

Nathaniel D Daw, John P O'doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, pages 1079–1105, 2006.

Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on Robot Learning*, pages 357–368, 2017.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*, 2018.

Sinong Geng, Houssam Nassif, Carlos A Manzanares, A Max Reppen, and Ronnie Sircar. Identifying reward functions using anchor actions. *arXiv preprint arXiv:2007.07443*, 2020.

Samuel J Gershman. Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71:1–6, 2016.

Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

Samuel J Gershman and Yael Niv. Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, 7(3):391–415, 2015.

Aditya Grover, Todor Markov, Peter Attia, Norman Jin, Nicolas Perkins, Bryan Cheong, Michael Chen, Zi Yang, Stephen Harris, William Chueh, et al. Best arm identification in multi-armed bandits with delayed feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 833–842. PMLR, 2018.

Linhui Hao, Akira Sakurai, Tokiko Watanabe, Ericka Sorensen, Chairul A Nidom, Michael A Newton, Paul Ahlquist, and Yoshihiro Kawaoka. Drosophila rnai screen identifies host genes important for influenza virus replication. *Nature*, 454(7206):890–893, 2008.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Neural Information Processing Systems*, pages 4572–4580, 2016.

Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. Learning from a learner. In *International Conference on Machine Learning*, pages 2990–2999, 2019.

Hong Jun Jeon, Smitha Milli, and Anca D Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *arXiv preprint arXiv:2002.04833*, 2020.

Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Artificial Intelligence and Statistics*, pages 139–148. PMLR, 2016.

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

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017a.

Yunzhu Li, Jiaming Song, and Stefano Ermon. InfoGAIL: interpretable imitation learning from visual demonstrations. In *Neural Information Processing Systems*, pages 3815–3825, 2017b.

James MacGlashan and Michael L Littman. Between imitation and intention learning. In *International Joint Conference on Artifical Intelligence*, pages 3692–3698, 2015.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, volume 1, page 2, 2000.

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269, 2018.

Ritesh Noothigattu, Tom Yan, and Ariel D. Procaccia. Inverse reinforcement learning from like-minded teachers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9197–9204, May 2021.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *International Joint Conference on Artificial Intelligence*, volume 7, pages 2586–2591, 2007.

Giorgia Ramponi, Gianluca Drappo, and Marcello Restelli. Inverse reinforcement learning from a gradient-based learner. *arXiv preprint arXiv:2007.07812*, 2020.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103, 1998.

Eric Schulz, Emmanouil Konstantinidis, and Maarten Speekenbrink. Learning and decisions in contextual multi-armed bandit tasks. In *CogSci*, pages 2122–2127, 2015.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? *arXiv preprint arXiv:1905.11397*, 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, 7(2):351–367, 2015.

Halit Bener Suay, Tim Brys, Matthew E Taylor, and Sonia Chernova. Learning from demonstration for shaping through inverse reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 429–437, 2016.

Umar Syed, Aleksandrs Slivkins, and Nina Mishra. Adapting to the shifting intent of search queries. *arXiv preprint arXiv:1007.3799*, 2010.

Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, pages 6818–6827, 2019.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438, 2008.

Supplementary Material: Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits

In the following appendices, we collect proofs of all main results, and also present some additional numerical experiments. Throughout our proofs, we suppose that T is greater than some absolute constant. We will use $c, C, c_1, C_1, \ldots$ to denote universal positive constants that may change from line to line. We also define the shorthand notation

$$\kappa_i \stackrel{\text{def}}{=} 4(T^\alpha - 1)/\alpha\Delta_i^2, \tag{3}$$

which will appear in multiple proofs and simplifies our exposition.

The appendices are organized as follows. Appendix A provides the proof of Theorem 3.1, our information-theoretic lower bound on reward estimation from a single demonstration of any algorithm. Appendix B collects preliminary lemmas for general bandit algorithms that are used as building blocks in all subsequent proofs. Appendix C provides, for completeness, proofs of high-probability regret bounds of SAE and UCB implemented with our inflated confidence widths. Appendix D provides the proof of Theorem 4.2, our upper bound on reward estimation error, from a demonstration of the SAE algorithm, and Appendix E provides the corresponding proof for the UCB case. Finally, Appendix F presents additional experimental details and results.

A Proof of Theorem 3.1

The proof of Theorem 3.1 establishes a natural link to information-theoretic lower bounds on best-arm identification. Denote the K-arm bandit instance by $\mathcal{M} = \{\text{Bern}(\mu_1), \ldots, \text{Bern}(\mu_K)\}$, and suppose without loss of generality that the arms of $\mathcal{M}$ are indexed with decreasing expected rewards, i.e. $\mu^* = \mu_1 > \mu_2 \geqslant \cdots \geqslant \mu_K$. (Note that the demonstrator's algorithm $\mathcal{A}$ does not know this indexing.) Recall that $n_{i,T}^{\mathcal{A}}$ denotes the number of times arm i is pulled by the demonstrator's algorithm $\mathcal{A}$. Further, for any t, let $\mathcal{F}_t(\mathcal{A})$ be the sigma algebra of the sequence of actions and random reward samples generated by the algorithm $\mathcal{A}$, i.e. $\mathcal{F}_t(\mathcal{A}) \stackrel{\text{def}}{=} \sigma(\{I_1, r_{I_1}, I_2, r_{I_2}, \ldots, I_t, r_{I_t}\})$ where $r_{I_t} \sim \text{Bern}(\mu_{I_t})$ denotes a random reward sample, and $\mathbb{F}(\mathcal{A}) = \{\mathcal{F}_t(\mathcal{A})\}_{t \geqslant 1}$ is a filtration.

Corresponding to some suboptimal arm $i \neq 1$, we construct another bandit instance $\mathcal{M}' = \{\text{Bern}(\mu_1'), \ldots, \text{Bern}(\mu_K'))$ with μ_j' defined as follows. Let $\mu_j' = \mu_j$ for each $j \neq i$, and set

$$\mu_i' = \begin{cases} \mu_i + \varepsilon & \text{if } \mu_i \leqslant 1/2, \\ \mu_i - \varepsilon & \text{otherwise.} \end{cases}$$

for some scalar $\varepsilon \in (0, 1/4)$ that we will subsequently specify. Because $\mu_i \in [0, 1]$, we have $\mu_i' \in [1/4, 3/4]$ for all $i \in [K]$.

We now reduce the reward estimation problem to one of binary testing via the classic Le-Cam approach. Suppose one of instance $\mathcal{M}$ or $\mathcal{M}'$ is chosen uniformly at random, and we observe sequence $\xi_T = \{I_1, I_2, \cdots, I_T\}$ generated by algorithm $\mathcal{A}$. Let ξ_T^0 denote this random sequence under the bandit instance $\mathcal{M}$, and denote by ξ_T^1 the random sequence observed under the bandit instance $\mathcal{M}'$. We denote the distributions of ξ_T^0 and ξ_T^1 by ν_T^0 and ν_T^1, respectively. We use $\mathbb{E}_0[\cdot]$ to denote expectations under the bandit instance $\mathcal{M}$, and $\mathbb{E}_1[\cdot]$ to denote expectations under the bandit instance $\mathcal{M}'$. Analogously, we use $\mathbb{P}_0(\cdot)$ to denote $\mathbb{E}_0[\mathbb{I}(\cdot)]$, and $\mathbb{P}_1(\cdot)$ to denote $\mathbb{E}_1[\mathbb{I}(\cdot)]$.

Now suppose the reward estimation procedure has knowledge of $\mu_1 = \mu_1' = \mu^*$, and must estimate the sequence of reward means $\{\mu_i\}_{i \in [K]}$. Since the error of estimation is lower bounded by the error of testing between the instances $\mathcal{M}$ and $\mathcal{M}'$, we have

$$\begin{aligned}\max\{\mathbb{E}_0[|\widehat{\mu}_i - \mu_i|], \mathbb{E}_1[|\widehat{\mu}_i - \mu_i|]\} &\geqslant \frac{\varepsilon}{2}\left(1 - \|\nu_T^0 - \nu_T^1\|_{\mathrm{TV}}\right) \\ &\geqslant \frac{\varepsilon}{2}\left(1 - \sup_{\mathcal{E}\in\mathcal{F}_T(\mathcal{A})} |\mathbb{P}_0(\mathcal{E}) - \mathbb{P}_1(\mathcal{E})|\right),\end{aligned} \tag{4}$$

where the last step follows from the definition of the total variation (TV) distance and the fact that the action sequence is in the filtration.

We now apply Kaufmann et al. (2016, Lemma 1) to obtain

$$\sup_{\mathcal{E}\in\mathcal{F}_T(\mathcal{A})} |\mathbb{P}(\mathcal{E}) - \mathbb{P}_1(\mathcal{E})| \leqslant \sqrt{\frac{\mathbb{E}[n_{i,T}] \cdot \mathrm{KL}(\mathrm{Bern}(\mu_i), \mathrm{Bern}(\mu_i'))}{2}}$$

where $\mathrm{KL}(\cdot,\cdot)$ denotes the Kullback-Leibler (KL) divergence between two distributions. Then, we have

$$\begin{aligned}\mathrm{KL}(\mathrm{Bern}(\mu_i), \mathrm{Bern}(\mu_i')) &= (\mu_i' + \varepsilon)\log\left(\frac{\mu_i' + \varepsilon}{\mu_i'}\right) + (1 - \mu_i' - \varepsilon)\log\left(\frac{1 - \mu_i' - \varepsilon}{1 - \mu_i'}\right) \\ &\leqslant \left(\frac{1}{\mu_i'} + \frac{1}{1 - \mu_i'}\right) \cdot \varepsilon^2 \leqslant \frac{16}{3}\varepsilon^2\end{aligned}$$

where the first inequality follows from applying $\log(1 + x) \leqslant x$, and the second inequality follows from the fact that $\mu_i' \in [1/4, 3/4]$.

Therefore, we have

$$\sup_{\mathcal{E}\in\mathcal{F}_T(\mathcal{A})} |\mathbb{P}(\mathcal{E}) - \mathbb{P}_1(\mathcal{E})| \leqslant \sqrt{\frac{8}{3} \cdot \varepsilon^2 \cdot \mathbb{E}_0[n_{i,T}^{\mathcal{A}}]}.$$

Combining the above with Eq (4), we have

$$\max\{\mathbb{E}_0[|\widehat{\mu}_i - \mu_i|], \mathbb{E}_1[|\widehat{\mu}_i - \mu_i|]\} \geqslant \frac{\varepsilon}{2}\left(1 - \varepsilon\sqrt{\frac{8}{3}\mathbb{E}[n_{i,T}^{\mathcal{A}}]}\right) \geqslant \frac{\varepsilon}{2}\left(1 - \varepsilon\sqrt{\frac{8}{3}(\mathbb{E}[n_{i,T}^{\mathcal{A}}] \vee 3/2)}\right)$$

Picking $\varepsilon = \sqrt{3}/\{4\sqrt{2(\mathbb{E}[n_{i,T}^{\mathcal{A}}] \vee 3/2)}\} < 1/4$ to maximize the right hand side of the above equation, we have

$$\max\{\mathbb{E}_0[|\widehat{\mu}_i - \mu_i|], \mathbb{E}_1[|\widehat{\mu}_i - \mu_i|]\} \geqslant \frac{\sqrt{3}}{16\sqrt{2}} \cdot \left(\frac{1}{\sqrt{\mathbb{E}[n_{i,T}^{\mathcal{A}}]}} \wedge \frac{1}{\sqrt{3/2}}\right) \geqslant \frac{1}{16} \cdot \left(\frac{1}{\sqrt{\mathbb{E}[n_{i,T}^{\mathcal{A}}]}} \wedge 1\right).$$

This completes the proof. ■

Remark A.1. *Note from the proof that an identical lower bound applies even if the procedure has access to the random reward samples themselves, in addition to the demonstrator's action sequence.*

B Preliminary lemmas for general bandit algorithms

We first present a convenient interpretation of the multi-armed bandit instance using the notion of "reward tapes" (Slivkins, 2019, Chapter 1). We consider a reward tape of length T for each arm $i \in \mathcal{A}$, each cell of which contains a random reward sample from that arm. In particular, cell j on the tape corresponding to arm i contains the reward sample $X_{i,j} \sim \nu_i$ (recall that ν_i denotes the reward distribution of arm i). Each time arm i is pulled, we move one cell forward on its reward tape, and obtain a reward from the new cell. Note that $n_{i,T}$ simply denotes the number of cells we have gone through on the reward tape of arm i by round T, and we trivially have $n_{i,T} \leqslant T$. Corresponding to the n^{th} cell of the reward tape, we define confidence width $C(n) \stackrel{\mathrm{def}}{=} \sqrt{\frac{2(T^\alpha - 1)}{\alpha n}}$.

The reward tape construction applies to a generic adaptive sampling algorithm (including both the SAE and UCB algorithms), and simplifies the construction of certain critical events concerning the concentration of sample means of arms around their true means. We start by stating and proving a basic lemma, which essentially follows from Hoeffding's inequality.

Lemma B.1. *Denote by $\bar{\mu}_i(n)$ the sample mean of arm i obtained by moving n cells along the reward tape. Then, for any $n \geqslant 1$, we have*

$$|\bar{\mu}_i(n) - \mu_i| < \sqrt{\frac{\log(2/\delta)}{2n}},$$

with probability at least $1 - \delta$.

Proof. By construction of the reward tape, the j-th pull of arm i generates the random reward $X_{i,j} \sim \nu_i$. This random variable is bounded in the range $[0, 1]$ and has expectation $\mathbb{E}[X_{i,j}] = \mu_i$. The sample mean is given by $\bar{\mu}_i(n) = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}$. Applying Hoeffding's inequality yields

$$\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}.$$

Setting $\varepsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$, we obtain

$$\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| < \sqrt{\frac{\log(2/\delta)}{2n}}\right) \geqslant 1 - \delta,$$

which completes the proof. ■

The following series of events will be used as building blocks in all of our proofs.

Definition B.2. We define the following events that ensure concentration of the sample means of arms obtained along the reward tape around their true means.

1. "Anytime" concentration events:

$$\mathcal{E}_0^{(i)} \stackrel{\text{def}}{=} \{|\bar{\mu}_i(n) - \mu_i| \leqslant C(n) \text{ for all } n = 1, \ldots, T\}, \tag{5}$$

   corresponding to each suboptimal arm $i \in [K]$. These events will be used to prove sub-linear regret guarantees for the SAE and UCB algorithms.

2. "Small-sample" concentration events

$$\mathcal{E}_1^{(i)} \stackrel{\text{def}}{=} \{\sqrt{n}|\bar{\mu}_i(n) - \mu_i| \leqslant \sqrt{\log(8\kappa_i)} \text{ for all } n = 1, \ldots, 8\kappa_i\}, \tag{6}$$

   corresponding to each suboptimal arm $i \in [K]\backslash i^*$. These events will be used to provide an eventual guarantee on estimation error of rewards of suboptimal arms. With a slight abuse of notation, we also define the event

$$\mathcal{E}_1^{(j,i)} \stackrel{\text{def}}{=} \{\sqrt{n}|\bar{\mu}_j(n) - \mu_j| \leqslant \sqrt{\log(8\kappa_i\sqrt{K})} \text{ for all } n = 1, \ldots, 8\kappa_i\}, \tag{7}$$

   where j is the index of an arm that remains active during the first $8\kappa_i$ rounds.

3. Tighter concentration events

$$\mathcal{E}_2^{(i)} \stackrel{\text{def}}{=} \left\{|\bar{\mu}_i(n) - \mu_i| \leqslant \sqrt{\frac{3}{4}}C(n) \text{ for all } n = 1, \ldots, T\right\} \text{ and} \tag{8a}$$

$$\mathcal{E}_3^{(i)} \stackrel{\text{def}}{=} \left\{|\bar{\mu}_i(n) - \mu_i| \leqslant \frac{C(n)}{\sqrt{2}} \text{ for all } n = 1, \ldots, \frac{\kappa_i}{32}\right\}, \tag{8b}$$

   corresponding to each arm $i \in [K]$. These events will be used to ensure that suboptimal arms are pulled sufficiently often to guarantee low error in estimation of their rewards.

4. "Large-sample" concentration events

$$\mathcal{E}_4^{(i^*,i)} \stackrel{\text{def}}{=} \begin{cases} |\bar{\mu}_{i^*}(n) - \mu_{i^*}| \leqslant \sqrt{\frac{2\log\kappa_i}{c\kappa_i}} \text{ for all } n \in \{c\kappa_i, \ldots, \kappa_i^2\} \text{ and} \\ |\bar{\mu}_{i^*}(n) - \mu_{i^*}| \leqslant \sqrt{\frac{\log T}{\kappa_i^2}} \text{ for all } n \in \{\kappa_i^2 + 1, \ldots, T\}, \end{cases} \tag{9}$$

   defined for the optimal arm i^* with reference to a suboptimal arm $i \in [K] \setminus i^*$. These events will be used in the case of the UCB algorithm to ensure high-probability lower bounds on the random variable n_{i^*,τ_i}.

The following lemma shows that each of these events occurs with high probability.

Lemma B.3. *For each $i \in [K]$, the following results hold:*

- *Event $\mathcal{E}_0^{(i)}$ occurs with probability at least $1 - 2/T^3$.*
- *Event $\mathcal{E}_1^{(i)}$ occurs with probability at least $1 - 1/4\kappa_i$.*
- *Event $\mathcal{E}_1^{(j,i)}$ occurs with probability at least $1 - 1/4\kappa_i K$.*
- *Event $\mathcal{E}_2^{(i)}$ holds with probability at least $1 - 2/T^2$.*
- *Event $\mathcal{E}_3^{(i)}$ holds with probability at least $1 - 1/16\kappa_i$.*
- *Event $\mathcal{E}_4^{(i^*,i)}$ holds with probability at least $1 - 2/T - c/\kappa_i$.*

Proof. The proof of Lemma B.3 proceeds by repeatedly applying the basic using the basic Lemma B.1 for different choices of δ and union bounding over varying ranges of n. We prove each claim separately. **Proof for event $\mathcal{E}_0^{(i)}$:** For each $i \in [K]$ and a fixed $n \geqslant 1$, we have

$$\begin{aligned}\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant C(n)\right) &\leqslant \mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant \sqrt{\frac{2\log T}{n}}\right) \\ &\leqslant \frac{2}{T^4},\end{aligned}$$

where the first inequality follows because $C(n) \geqslant \sqrt{2\log T/n}$, and the second inequality follows by applying Lemma B.1 with the choice $\delta = 2/T^4$. Taking a union bound over $n = 1, \ldots, T$ yields

$$\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant C(n) \text{ for some } n = 1, \ldots, T\right) \leqslant \frac{2}{T^3}.$$

This shows that the event $\mathcal{E}_0^{(i)}$ holds with probability at least $1 - 2/T^3$, completing the proof.

Proof for event $\mathcal{E}_1^{(i)}$: For each $i \in [K]$ and each $n = 1, \ldots, 8\kappa_i$, we apply Lemma B.1 with $\delta = 2/64\kappa_i^2$. Then, we take a union bound over all $n = 1, \ldots, 8\kappa_i$ to obtain

$$\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| > \sqrt{\frac{\log(8\kappa_i)}{n}} \text{ for some } n = 1, \ldots, 8\kappa_i\right) \leqslant 8\kappa_i \cdot \frac{2}{64\kappa_i^2} = \frac{1}{4\kappa_i}.$$

This completes the proof.

Proof for event $\mathcal{E}_1^{(j,i)}$: Applying Lemma B.1 with the choice $\delta = 2/64\kappa_i^2 K$ yields

$$\mathbb{P}\left(|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \geqslant \sqrt{\frac{\log 8\kappa_i\sqrt{K}}{n}}\right) \leqslant \frac{2}{64\kappa_i^2 K},$$

for each fixed n, and taking a union bound over n in the desired range completes the proof.

Proof for event $\mathcal{E}_2^{(i)}$: For each $i \in [K]$ and a fixed $n \geqslant 1$, we have

$$\begin{aligned}\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant \sqrt{\frac{3}{4}}C(n)\right) &\leqslant \mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant \sqrt{\frac{3\log T}{2n}}\right) \\ &\leqslant \frac{2}{T^3},\end{aligned}$$

where the first inequality follows because $C(n) \geqslant \sqrt{2\log T/n}$, and the second inequality follows by applying Lemma B.1 with the choice $\delta = 2/T^3$. Taking a union bound over $n = 1, \ldots, T$ yields

$$\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant C(n) \text{ for some } n = 1, \ldots, T\right) \leqslant \frac{2}{T^2}.$$

This shows that the event $\mathcal{E}_2^{(i)}$ holds with probability at least $1 - 2/T^2$, completing the proof.

Proof for event $\mathcal{E}_3^{(i)}$: For each $i \in [K]$ and a fixed $n \in \{1, \ldots, \kappa_i/32\}$, we have

$$\begin{aligned}\mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant \frac{C(n)}{\sqrt{2}}\right) &\leqslant \mathbb{P}\left(|\bar{\mu}_i(n) - \mu_i| \geqslant \sqrt{\frac{\log \kappa_i}{n}}\right) \\ &\leqslant \frac{2}{\kappa_i^2},\end{aligned}$$

where the first inequality follows because $C(n)/\sqrt{2} \geqslant \sqrt{\log \kappa_i/n}$ for the specified range of n, and the second inequality follows by applying Lemma B.1 with the choice $\delta = 2/\kappa_i^2$. Taking a union bound over the specified range of n completes the proof.

Proof for event $\mathcal{E}_4^{(i^*,i)}$: First, consider the case where $c\kappa_i \leqslant n \leqslant \kappa_i^2$. In this case, we have

$$\begin{aligned}\mathbb{P}\left(|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \geqslant \sqrt{\frac{2\log \kappa_i}{c\kappa_i}}\right) &\leqslant \mathbb{P}\left(|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \geqslant \sqrt{\frac{2\log \kappa_i}{n}}\right) \\ &\leqslant \frac{2}{\kappa_i^4},\end{aligned}$$

where the first inequality follows because $n \geqslant c\kappa_i$, and the second inequality follows by applying Lemma B.1 with the choice $\delta = 2/\kappa_i^4$. Second, consider the case where $\kappa_i^2 < n \leqslant T$. In this case, we have

$$\begin{aligned}\mathbb{P}\left(|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \geqslant \sqrt{\frac{\log T}{\kappa_i^2}}\right) &\leqslant \mathbb{P}\left(|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \geqslant \sqrt{\frac{\log T}{n}}\right) \\ &\leqslant \frac{2}{T^2},\end{aligned}$$

where the first inequality follows because we are in the case $n > \kappa_i^2$, and the second inequality follows by applying Lemma B.1 with the choice $\delta = 2/T^2$. Taking a union bound over $n = c\kappa_i, \ldots, T$ completes the proof. ■

We will work on combinations of these events to prove Theorem 4.2 for the case of the SAE algorithm (Appendix D) and the case of the UCB algorithm (Appendix E).

C Sub-linear regret guarantees for UCB and SAE

For completeness, we provide a proof for Proposition C.1, which bounds the regret of the UCB and SAE algorithms.

Proposition C.1. *Recall that $\Delta_i = \mu^* - \mu_i$. For any $T > K$, Algorithm 1 and Algorithm 2 both incur regret $R_T \leqslant \sum_{i \neq i^*} \frac{32(T^\alpha - 1)}{\alpha \Delta_i}$ with probability at least $1 - \frac{4K}{T^3}$.*

For clarity, we prove it separately for the SAE and UCB algorithms in Sections C.1 and C.2, respectively. These proofs follow from straightforward modifications to classical results (Even-Dar et al., 2006; Lattimore and Szepesvári, 2020), and readers familiar with regret analysis are advised to skip to Sections D and E for novel analyses of our reward estimation procedures.

C.1 Proof of Proposition C.1 with SAE algorithm

We begin with a useful lemma whose proof is provided at the end of the subsection (see Section C.1.1). Recall the definition of the scalar κ_i from Equation (3).

Lemma C.2. *For the SAE algorithm, we have*

$$n_{i,T} \leqslant 8\kappa_i,$$

simultaneously for all suboptimal arms i with probability at least $1 - 2K/T^3$. Furthermore, on the same event, the optimal arm i^* is never eliminated.

With this lemma in hand, the proof of Proposition C.1 for the SAE algorithm follows immediately.

Proof of Proposition C.1, SAE. By the definition of pseudo-regret, we have

$$\begin{aligned} R_T &= T\mu^* - \sum_{t=1}^{T} r_{t,I_t} \\ &= \sum_{i\in[K]} \Delta_i n_{i,T} \\ &\leqslant \sum_{i\in[K]} 8\kappa_i\Delta_i \overset{\text{def}}{=} \frac{32(T^\alpha - 1)}{\alpha\Delta_i}, \end{aligned}$$

where the final inequality holds with probability at least $1 - 2K/T^3$ by applying Lemma C.2. This completes the proof. ■

C.1.1 Proof of Lemma C.2

Throughout this proof, we work on the event $\cap_{i=1}^K \mathcal{E}_0^{(i)}$, which we showed in Lemma B.3 holds with probability at least $1 - 2K/T^3$. Recall the definition of κ_i from Equation (3). It is easy to verify that $C(8\kappa_i) \leqslant \Delta_i/4$. Consequently, we have

$$\bar{\mu}_i(8\kappa_i) + C(8\kappa_i) \leqslant \mu_i + 2C(8\kappa_i) \leqslant \mu_i + \frac{\Delta_i}{2}. \tag{10}$$

Similarly, for the optimal arm i^* we have

$$\bar{\mu}_{i^*}(8\kappa_i) - C(8\kappa_i) \geqslant \mu_{i^*} - 2C(8\kappa_i) \geqslant \mu_{i^*} - \frac{\Delta_i}{2}. \tag{11}$$

On the other hand, we have $\bar{\mu}_i(n) - C(n) \leqslant \mu_i$ and $\bar{\mu}_{i^*}(n) + C(n) \geqslant \mu_{i^*}$ for every $n = 1, \ldots, T$ and every $i \in [K]$. Because $\mu_{i^*} > \mu_i$, this yields

$$2C(n) \geqslant \mu_{i^*} - \bar{\mu}_{i^*}(n) + \bar{\mu}_i(n) - \mu_i > \bar{\mu}_i(n) - \bar{\mu}_{i^*}(n), \tag{12}$$

for every $n = 1, \ldots, T$ and every $i \in [K] \setminus i^*$. Equation (12) guarantees that arm i^* remains active throughout, as claimed.

To complete the proof, we show that each arm $i \neq i^*$ is eliminated by the time we arrive at epoch $8\kappa_i$. Denote by $\bar{t}(s)$ the (random) last round of epoch s. If arm i has already been eliminated in an epoch preceding epoch $8\kappa_i$, we are done. Otherwise, since arm i^* is always active, combining Equations (10) and (11) gives us

$$\begin{aligned} 2C_{i,\bar{t}(8\kappa_i)} &\leqslant \mu_i + \frac{\Delta_i}{2} - \mu_{i^*} + \frac{\Delta_i}{2} + \bar{\mu}_{i^*,\bar{t}(8\kappa_i)} - \bar{\mu}_{i,\bar{t}(8\kappa_i)} \\ &= \bar{\mu}_{i^*,\bar{t}(8\kappa_i)} - \bar{\mu}_{i,\bar{t}(8\kappa_i)} \\ &\leqslant \bar{\mu}_{\max}(\bar{t}(8\kappa_i)) - \bar{\mu}_{i,\bar{t}(8\kappa_i)}. \end{aligned}$$

In summary, the condition for arm i to be eliminated is met by epoch at most $8\kappa_i$, directly implying that $n_{i,T} \leqslant 8\kappa_i$. This completes the proof. ■

C.2 Proof of Proposition C.1 for UCB

The structure of this proof is identical to the SAE case. Recall the definition of the scalar κ_i from Equation (3).

Lemma C.3. *For the UCB algorithm, we have*

$$n_{i,T} \leqslant 8\kappa_i$$

for a suboptimal arm i with probability at least $1 - 4/T^3$.

As with the SAE case, the proof of Proposition C.1 follows immediately from this lemma. The steps are exactly identical and we omit them for brevity. We conclude this section by proving Lemma C.3.

C.2.1 Proof of Lemma C.3

Throughout this proof, we work on the event $\mathcal{E}_0^{(i)} \cap \mathcal{E}_0^{(i^*)}$, which we showed in Lemma B.3 holds with probability at least $1 - 4/T^3$. Since $C(n) \leqslant \Delta_i/4$ for all $n \geqslant 8\kappa_i$, we have

$$\bar{\mu}_i(n) + C(n) \leqslant \mu_i + 2C(n) \leqslant \mu_i + \frac{\Delta_i}{2} = \mu_{i^*} - \frac{\Delta_i}{2} \tag{13}$$

for all $n \geqslant 8\kappa_i$.

On the other hand, for the optimal arm i^* we have

$$\bar{\mu}_{i^*}(n) + C(n) \geqslant \mu_{i^*} \tag{14}$$

for all $n = 1, \ldots, T$. Denote by $\bar{t}(8\kappa_i)$ the (random) earliest round after which arm i was pulled for the $8\kappa_i$-th time. If no such round exists, then we are done. Otherwise, combining Equations (13) and (14) gives us

$$\bar{\mu}_{i^*,t} + C_{i^*,t} > \bar{\mu}_{i,t} + C_{i,t}$$

for all $t \geqslant \bar{t}(8\kappa_i)$. Thus, we have shown that the upper-confidence bound of arm i^* dominates the upper confidence bound of arm i for all rounds $t \geqslant \bar{t}(8\kappa_i)$, implying that arm i is never pulled thereafter. This directly gives us $n_{i,T} = n_{i,\bar{t}(8\kappa_i)} \leqslant 8\kappa_i$, which completes the proof for each suboptimal arm i. ■

D Proof of Theorem 4.2 for SAE

In this section, we provide the proof of Theorem 4.2 for the case of the SAE algorithm. Recall that we need to bound the estimation error $|\widehat{\mu}_i - \mu_i|$, and recall the notation $\kappa_i \stackrel{\text{def}}{=} 4(T^\alpha - 1)/\alpha\Delta_i^2$ from Equation (3). This proof will follow as a series of deterministic statements working on the high-probability event

$$\mathcal{E}_{\mathsf{SAE}} \stackrel{\text{def}}{=} \bigcap_{i=1}^{K} \left(\mathcal{E}_0^{(i)} \cap \mathcal{E}_2^{(i)}\right) \cap \mathcal{E}_1^{(i)} \cap \left(\cap_{j\in[K]} \mathcal{E}_1^{(j,i)}\right). \tag{15}$$

Lemma B.3 together with an application of the union bound ensures that the event $\mathcal{E}_{\mathsf{SAE}}$ holds with probability at least $1 - 2K/T^3 - 1/2\kappa_i - 2K/T^2$.

First, we claim that on the event $\mathcal{E}_{\mathsf{SAE}}$ and under our assumption that $T \geqslant 32\sum_{i\neq i^*}\kappa_i$, we have $\widehat{\imath} = i^*$. In order to see this, note that on the event $\cap_{i=1}^K \mathcal{E}_0^{(i)}$ we may apply the statement of Lemma C.2 to conclude that $n_{i,T} \leqslant 8\kappa_i$ simultaneously for all suboptimal arms $i \neq i^*$. Consequently, we have

$$n_{i^*,T} = T - \sum_{i\neq i^*} n_{i,T} \geqslant 8\sum_{i\neq i^*}\kappa_i > \max_{i\neq i^*} n_{i,T}.$$

Next, we consider the estimation error $|\widehat{\mu}_i - \mu_i|$ for any suboptimal arm i. Recall that τ_i is the round at which suboptimal arm $i \in \mathcal{A}$ is eliminated (Equation (1)). Since we identified the optimal arm, i.e. $\widehat{\imath} = i^*$, we have $\tau_i < T$. Further, since arm i is eliminated at round τ_i, we have

$$2C_{i,\tau_i} \leqslant \bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i}. \tag{16}$$

On the other hand, denote by τ_i' the *penultimate* round on which arm i is pulled. Since arm i is still active during this round, we have

$$2C_{i,\tau_i'} > \bar{\mu}_{\max}(\tau_i') - \bar{\mu}_{i,\tau_i'}. \tag{17}$$

Note that $n_{i,\tau_i'} = n_{i,\tau_i} - 1 = n_{i,T} - 1$. Therefore, we have

$$\begin{aligned} 2C_{i,\tau_i} = 2\sqrt{\frac{T^\alpha - 1}{\alpha n_{i,T}}} &\geqslant 2\sqrt{\frac{T^\alpha - 1}{\alpha(n_{i,T}-1)}} - 4\sqrt{\frac{T^\alpha - 1}{\alpha}} \cdot (n_{i,T}-1)^{-3/2} \\ &= 2C_{i,\tau_i'} - 4\sqrt{\frac{T^\alpha - 1}{\alpha}} \cdot (n_{i,T}-1)^{-3/2} \\ &> \bar{\mu}_{\max}(\tau_i') - \bar{\mu}_{i,\tau_i'} - 4\sqrt{\frac{T^\alpha - 1}{\alpha}} \cdot (n_{i,T}-1)^{-3/2}. \end{aligned}$$

Above, the first inequality follows from the fact that $\frac{1}{\sqrt{x}} - \frac{1}{\sqrt{x+1}} \leqslant 2x^{-3/2}$ for any $x \geqslant 1$, and the second inequality is a direct substitution of Equation (17).

Furthermore, we obtain

$$\bar{\mu}_{\max}(\tau_i') - \bar{\mu}_{i,\tau_i'} \geqslant \bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i} - \frac{2}{n_{i,T}},$$

as a consequence of the rewards being bounded between $[0,1]$. Therefore, we have

$$(\bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i}) - 2C_{i,\tau_i} \leqslant \frac{2}{n_{i,T}} + 4\sqrt{\frac{T^\alpha - 1}{\alpha}} \cdot (n_{i,T}-1)^{-3/2}. \tag{18}$$

Proceeding now to the error term of interest, we have

$$\begin{aligned} |\widehat{\mu}_i - \mu_i| &= |2C_{i,\tau_i} - (\mu^* - \mu_i)| \\ &\leqslant |2C_{i,\tau_i} - (\bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i})| + |\bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i} - (\mu^* - \mu_i)| \\ &\leqslant \frac{2}{n_{i,T}} + 4\sqrt{\frac{T^\alpha - 1}{\alpha}} \cdot (n_{i,T}-1)^{-3/2} + |\bar{\mu}_{\max}(\tau_i) - \mu^*| + |\bar{\mu}_{i,\tau_i} - \mu_i| \\ &\leqslant \frac{2}{n_{i,T}} + 2\sqrt{\kappa_i} \cdot (n_{i,T}-1)^{-3/2} + |\bar{\mu}_{\max}(\tau_i) - \mu^*| + |\bar{\mu}_{i,\tau_i} - \mu_i|, \end{aligned} \tag{19}$$

where the first inequality follows from triangle inequality and rearranging terms, and the second inequality follows from Equation (18) and noting that $|2C_{i,\tau_i} - (\bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i})| = (\bar{\mu}_{\max}(\tau_i) - \bar{\mu}_{i,\tau_i}) - 2C_{i,\tau_i}$ as a consequence of Equation (16).

It remains to bound the sample-mean deviations $|\bar{\mu}_{\max}(\tau_i) - \mu^*|$ and $|\bar{\mu}_{i,\tau_i} - \mu_i|$. Towards that end, we require two technical lemmas, stated below and proved at the end of this section. The first lemma bounds the deviation of the sample mean for each arm.

Lemma D.1. *Fix a suboptimal arm i. On the event $\mathcal{E}_{\mathsf{SAE}}$, there are universal positive constants C, C_1, C_2 such that for any arm $j \in [K]$ that remains active at round τ_i, we have*

$$|\bar{\mu}_{j,\tau_i} - \mu_j| < C\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}}.$$

The second lemma provides a high-probability lower bound on n_{i,τ_i}, which is significantly more intricate than the typical lower bound on $\mathbb{E}[n_{i,\tau_i}]$.

Lemma D.2. *There is a universal constant $c > 0$ such that on the event $\mathcal{E}_{\mathsf{SAE}}$, we have $n_{i,T} \geqslant c\kappa_i$.*

Having stated these technical lemmas, we now use them to complete the proof of Theorem 4.2 for SAE. First, Lemma D.1 applied to the case $j = i$ directly gives us

$$|\bar{\mu}_{i,\tau_i} - \mu_i| < C\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}}.$$

Next, we again use Lemma D.1 to bound $|\bar{\mu}_{\max}(\tau_i) - \mu^*|$. We denote by $i_{\max} \in [K]$ the arm index such that $\bar{\mu}_{i_{\max}} = \bar{\mu}_{\max}(\tau_i)$. On one hand, we have

$$\begin{aligned}\bar{\mu}_{\max}(\tau_i) - \mu^* &\leqslant \bar{\mu}_{i_{\max}} - \mu_{i_{\max}} \\ &< C\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}}\end{aligned}$$

where the last step follows from Lemma D.1. On the other hand, we have

$$\begin{aligned}\mu^* - \bar{\mu}_{\max}(\tau_i) &\leqslant \mu^* - \bar{\mu}_{i^*} \\ &< C\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}}\end{aligned}$$

where, again, the last step follows from Lemma D.1.

Proceeding from equation (19) and applying the inequalities established above, we have we have

$$|\widehat{\mu}_i - \mu_i| \leqslant \frac{C}{\kappa_i} + C'\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}}$$

on the event $\mathcal{E}_{\mathsf{SAE}}$. Taking an expectation to include the complement of $\mathcal{E}_{\mathsf{SAE}}$, we have

$$\mathbb{E}|\widehat{\mu}_i - \mu_i| \leqslant C\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}} + \frac{C_1}{\kappa_i} + \frac{C_2K}{T^2} \leqslant C''\sqrt{\frac{\log(\kappa_i\sqrt{K})}{\kappa_i}}.$$

We have used that $T^2/K \gtrsim \sqrt{\kappa_i}$ in stating the second inequality.

To complete the proof of the theorem, note that the proof of Proposition C.1 (see Lemma C.2) yields $\mathbb{E}[n_{i,T}] \leqslant c'\kappa_i$ for some positive constant $c' > 0$. Since the map $x \mapsto \log x/x$ is decreasing, we obtain

$$\mathbb{E}|\widehat{\mu}_i - \mu_i| \leqslant C''\sqrt{\frac{\log(\mathbb{E}[n_{i,T}\sqrt{K})]}{\mathbb{E}[n_{i,T}]}}$$

for some adjusted constant C''. This completes the proof of the first part of the theorem. The second part follows directly from Lemma D.2. ■

D.1 Proof of Lemma D.1

As detailed at the beginning of Appendix D, working on the event $\mathcal{E}_{\mathsf{SAE}}$ guarantees that $n_{i,T} \leqslant 8\kappa_i$ for each suboptimal arm i and that arm i^* remains active throughout. In addition, because event $\mathcal{E}_{\mathsf{SAE}}$ holds, we have that $\mathcal{E}_1^{(i)} \cap (\cup_{j\in\mathcal{S}_i}\mathcal{E}_1^{(j,i)})$ (defined in Equations (6) and (7), and $\mathcal{S}_i$ denotes the set of arms that remain active in the first $8\kappa_i$ rounds) holds. This gives us

$$\sup_{1\leqslant n\leqslant 8\kappa_i} \sqrt{n}|\bar{\mu}_k(n) - \mu_k| \leqslant \sqrt{\log(8\kappa_i\sqrt{K})} \tag{20}$$

where k can denote either the optimal arm i^* or any suboptimal arm j that remains active until round τ_i. Finally, note by the definition of τ_i that for all arms $j \in [K]$ that are active, we have $n_{j,\tau_i} = n_{i,\tau_i} = n_{i,T} \leqslant 8\kappa_i$, which ensures that n_{j,τ_i} lies in the required range $\{1,\ldots,8\kappa_i\}$ to apply Equation (20). Moreover, by Lemma D.2 (which also holds on event $\mathcal{E}_{\mathsf{SAE}}$) we have $n_{i,\tau_i} \geqslant c\kappa_i$ for some constant $c > 0$. Combining this with Equation (20) yields

$$|\bar{\mu}_{j,\tau_i} - \mu_j| < \sqrt{\frac{\log(8\kappa_i\sqrt{K})}{c\kappa_i}}$$

for all active arms $j \in [K]$. This completes the proof. ■

D.2 Proof of Lemma D.2

Because the event $\mathcal{E}_{\mathsf{SAE}}$ holds, we have that $\cap_{i=1}^{K}\mathcal{E}_2^{(i)}$ (defined in Equation (8a)) holds, which guarantees that

$$|\bar{\mu}_i(n) - \mu_i| \leqslant \sqrt{\frac{3}{4}}C(n) \text{ for all } n = 1, \ldots, T \text{ and all } i \in [K]. \tag{21}$$

For only this proof, we let $c \stackrel{\text{def}}{=} (2-\sqrt{3})^2$ for brevity. It is easy to verify that $\Delta_i = c \cdot C\left(\frac{c^2\kappa_i}{2}\right)$. By the triangle inequality, we have

$$\begin{aligned} |\bar{\mu}_{i'}(n) - \bar{\mu}_i(n)| &\leqslant |\bar{\mu}_{i'}(n) - \mu_{i'}| + \Delta_i + |\bar{\mu}_i(n) - \mu_i| \\ &\leqslant 2C(n) \end{aligned} \tag{22}$$

for all $n \leqslant \frac{c^2\kappa_i}{2}$ and every $i \neq i'$. In other words, provided the number of pulls of arm i does not exceed $\frac{c^2\kappa_i}{2}$, the condition for elimination is not met. Arm i thus stays active for at least $\frac{c^2\kappa_i}{2}$ epochs, establishing the desired lemma. ■

E Proof of Theorem 4.2 for UCB

In this section, we provide the proof of Theorem 4.2 for the more complex case of the UCB algorithm. The structure of the proof resembles the SAE case, but the steps themselves are significantly more involved. Recall that we need to bound the estimation error $|\widehat{\mu}_i - \mu_i|$, and recall the notation $\kappa_i \stackrel{\text{def}}{=} 4(T^\alpha - 1)/\alpha\Delta_i^2$. As before, the proof will follow as a series of deterministic statements working on the high-probability event

$$\mathcal{E}_{\mathsf{UCB}} \stackrel{\text{def}}{=} \left(\bigcap_{i=1}^{K}\mathcal{E}_0^{(i)}\right) \cap \mathcal{E}_1^{(i)} \cap \mathcal{E}_1^{(i^*,i)} \cap \mathcal{E}_2^{(i)} \cap \mathcal{E}_3^{(i)} \cap \mathcal{E}_3^{(i^*)} \cap \mathcal{E}_4^{(i^*,i)} \tag{23}$$

From Lemma B.3 and the union bound, we have that the event $\mathcal{E}_{\mathsf{UCB}}$ holds with probability at least $1 - C_1/\kappa_i - C_2/T$ for universal constants $C_1, C_2 > 0$.

First, we note that on the event $\mathcal{E}_{\mathsf{UCB}}$ and under our assumption of $T \geqslant 32\sum_{i\neq i^*}\kappa_i$, we have $\widehat{\imath} = i^*$ via an argument that is identical to the SAE case (provided at the beginning of Appendix D). For any suboptimal arm i, let $\overline{\tau}_i$ denote the first round after τ_i in which the best arm is pulled, noting that such a round always exists by the definition of τ_i.

Since arm i is pulled at round τ_i and arm i^* is pulled at round $\overline{\tau}_i$, the respective upper confidence relations yield the bounds

$$\begin{aligned} \bar{\mu}_{i,\tau_i} + C_{i,\tau_i} - (\bar{\mu}_{i^*,\tau_i} + C_{i^*,\tau_i}) &\geqslant 0 \quad \text{and} \\ \bar{\mu}_{i,\overline{\tau}_i} + C_{i,\overline{\tau}_i} - (\bar{\mu}_{i^*,\overline{\tau}_i} + C_{i^*,\overline{\tau}_i}) &\leqslant 0. \end{aligned}$$

Combining the above two equations, rearranging terms, and applying the triangle inequality, we obtain

$$\begin{aligned} |C_{i,\tau_i} - C_{i^*,\tau_i} - (\bar{\mu}_{i^*,\tau_i} - \bar{\mu}_{i,\tau_i})| &\leqslant |C_{i,\tau_i} - C_{i,\overline{\tau}_i}| + |C_{i^*,\tau_i} - C_{i^*,\overline{\tau}_i}| + |\bar{\mu}_{i^*,\tau_i} - \bar{\mu}_{i^*,\overline{\tau}_i}| + |\bar{\mu}_{i,\tau_i} - \bar{\mu}_{i,\overline{\tau}_i}| \\ &= |C_{i,\tau_i} - C_{i,\overline{\tau}_i}| + |C_{i^*,\tau_i} - C_{i^*,\overline{\tau}_i}| + |\bar{\mu}_{i,\tau_i} - \bar{\mu}_{i,\overline{\tau}_i}|, \end{aligned} \tag{24}$$

where the last equality follows because arm i^* is not pulled between round τ_i and $\overline{\tau}_i$. Thus, we have $|\bar{\mu}_{i^*,\tau_i} - \bar{\mu}_{i^*,\overline{\tau}_i}| = 0$. Proceeding now to the error term of interest, we have

$$\begin{aligned} |\widehat{\mu}_i - \mu_i| &= |C_{i,\tau_i} - C_{i^*,\tau_i} - (\mu^* - \mu_i)| \\ &\leqslant |C_{i,\tau_i} - C_{i^*,\tau_i} - (\bar{\mu}_{i^*,\tau_i} - \bar{\mu}_{i,\tau_i})| + |\bar{\mu}_{i^*,\tau_i} - \mu^*| + |\bar{\mu}_{i,\tau_i} - \mu_i| \\ &\leqslant |C_{i,\tau_i} - C_{i,\overline{\tau}_i}| + |C_{i^*,\tau_i} - C_{i^*,\overline{\tau}_i}| + |\bar{\mu}_{i,\tau_i} - \bar{\mu}_{i,\overline{\tau}_i}| \\ &\quad + |\bar{\mu}_{i^*,\tau_i} - \mu^*| + |\bar{\mu}_{i,\tau_i} - \mu_i|, \end{aligned}$$

where the second inequality follows from equation (24). We bound each of the above terms separately. First, because arm i^* is not pulled between rounds τ_i and $\overline{\tau}_i$, we have $n_{i^*,\tau_i} = n_{i^*,\overline{\tau}_i}$, and consequently, its confidence interval stays the same, with

$$|C_{i^*,\tau_i} - C_{i^*,\overline{\tau}_i}| = 0. \tag{25}$$

On the other hand, the confidence interval of arm i changes, but not by much. We have

$$\begin{aligned} |C_{i,\tau_i} - C_{i,\overline{\tau}_i}| &= \sqrt{\frac{2(T^\alpha - 1)}{\alpha}} \left(\frac{1}{\sqrt{n_{i,\tau_i}}} - \frac{1}{\sqrt{n_{i,\overline{\tau}_i}}} \right) \\ &= \sqrt{\frac{2(T^\alpha - 1)}{\alpha}} \left(\frac{1}{\sqrt{n_{i,\tau_i}}} - \frac{1}{\sqrt{n_{i,\tau_i} + 1}} \right) \\ &\leqslant \sqrt{\frac{2(T^\alpha - 1)}{\alpha}} \cdot n_{i,\tau_i}^{-3/2}. \end{aligned} \tag{26}$$

where the inequality uses the fact that $\sqrt{j+1} - \sqrt{j} \leqslant j^{-1/2}$ for any integer $j \geqslant 1$. Next, using our reward tape notation, note that

$$\begin{aligned} |\bar{\mu}_{i,\overline{\tau}_i} - \bar{\mu}_{i,\tau_i}| &\stackrel{\text{def}}{=} \left| \frac{\sum_{j=1}^{n_{i,\overline{\tau}_i}} X_{i,j}}{n_{i,\overline{\tau}_i}} - \frac{\sum_{j=1}^{n_{i,\tau_i}} X_{i,j}}{n_{i,\tau_i}} \right| \\ &= \left| \frac{\bar{\mu}_{i,\tau_i} \cdot n_{i,\tau_i} + X_{n_{i,\overline{\tau}_i}}}{n_{i,\tau_i} + 1} - \bar{\mu}_{i,\tau_i} \right| \\ &= \left| \frac{X_{n_{i,\overline{\tau}_i}} - \bar{\mu}_{i,\tau_i}}{n_{i,\tau_i} + 1} \right| \leqslant \frac{1}{n_{i,\tau_i}}, \end{aligned} \tag{27}$$

where the last inequality follows from the fact that both $X_{i,j}$ and $\bar{\mu}_{i,\tau}$ are bounded between 0 and 1.

It remains to bound the sample-mean deviations $|\bar{\mu}_{i^*,\tau_i} - \mu^*|$ and $|\bar{\mu}_{i,\tau_i} - \mu_i|$. Towards that end, we require three technical lemmas, stated below and proved at the end of this section. The first lemma bounds the deviation of the sample mean for arm i in terms of the number of times it has been pulled.

Lemma E.1. *On the event $\mathcal{E}_{\mathsf{UCB}}$, we have*

$$\sup_{1 \leqslant t \leqslant T} \sqrt{n_{i,t}} |\bar{\mu}_{i,t} - \mu_i| < \sqrt{\log 8\kappa_i} \tag{28}$$

for any suboptimal arm i.

The second lemma provides a high-probability lower bound on n_{i,τ_i}, which is significantly more intricate than the typical lower bound on $\mathbb{E}[n_{i,\tau_i}]$.

Lemma E.2. *On the event $\mathcal{E}_{\mathsf{UCB}}$, there is an absolute constant $c > 0$ such that we have $n_{i,\tau_i} \geqslant c\kappa_i$ for any suboptimal arm i.*

Our third and final lemma bounds the deviation of the sample mean for arm i^*.

Lemma E.3. *On the event $\mathcal{E}_{\mathsf{UCB}}$, there is an absolute constant $C > 0$ such that*

$$|\bar{\mu}_{i^*,\tau_i} - \mu^*| < C \sqrt{\frac{\log \kappa_i}{\kappa_i}}. \tag{29}$$

Having stated these technical lemmas, let us now complete the proof of Theorem 4.2 for UCB, operating throughout on the event $\mathcal{E}_{\mathsf{UCB}}$. Applying Lemma E.1 yields

$$\sup_{1 \leqslant t \leqslant T} \sqrt{n_{i,t}} |\bar{\mu}_{i,t} - \mu_i| < \sqrt{\log 8\kappa_i},$$

from which we obtain

$$|\bar{\mu}_{i,\tau_i} - \mu_i| < \sqrt{\frac{\log \kappa_i}{n_{i,\tau_i}}}. \tag{30}$$

In addition, Lemma E.3 yields the bound

$$|\bar{\mu}_{i^*,\tau_i} - \mu_{i^*}| < C \sqrt{\frac{\log \kappa_i}{\kappa_i}}. \tag{31}$$

Putting together equations (25), (26), (27), (30) and (31), the following sequence of bounds holds (where constants change from line-to-line, but are always absolute):

$$|\widehat{\mu}_i - \mu_i| \leqslant \frac{1}{n_{i,\tau_i}} + \sqrt{\frac{2(T^\alpha - 1)}{\alpha}} \cdot n_{i,\tau_i}^{-3/2} + \sqrt{\frac{\log \kappa_i}{n_{i,\tau_i}}} + C\sqrt{\frac{\log \kappa_i}{\kappa_i}}$$
$$\leqslant \frac{1}{n_{i,\tau_i}} + C\sqrt{\kappa_i} \cdot n_{i,\tau_i}^{-3/2} + \sqrt{\frac{\log \kappa_i}{n_{i,\tau_i}}} + C\sqrt{\frac{\log \kappa_i}{\kappa_i}}.$$

Here, the second inequality holds by definition of κ_i. Finally, Lemma E.2 provides a lower bound on n_{i,τ_i}, which gives us

$$|\widehat{\mu}_i - \mu_i| \leqslant C\sqrt{\frac{\log \kappa_i}{\kappa_i}}$$

on the event $\mathcal{E}_{\mathsf{UCB}}$. Recall that the event $\mathcal{E}_{\mathsf{UCB}}$ holds with probability greater than $1 - {}^{C_1}/_{\kappa_i} - {}^{C_2}/_{T}$ for absolute constants C_1, C_2. Substituting the value of κ_i and reasoning exactly as in the SAE case about the complementary event, we obtain the desired upper bound on the expected error. ■

E.1 Proof of Lemma E.1

As detailed at the beginning of Appendix E, working on the event $\mathcal{E}_{\mathsf{UCB}}$ guarantees that $n_{i,T} \leqslant 8\kappa_i$ for each suboptimal arm i (see Lemma C.3). Moreover, since event $\mathcal{E}_{\mathsf{UCB}}$ holds, we have that event $\mathcal{E}_1^{(i)}$ (defined in Equation (6)) holds, which guarantees that

$$\sqrt{n} \cdot |\bar{\mu}_i(n) - \mu_i| \leqslant \sqrt{\log 8\kappa_i} \quad \text{for all } n = 1, \ldots, 8\kappa_i.$$

Thus, it follows directly that

$$\sup_{1 \leqslant t \leqslant T} \sqrt{n_{i,t}} |\bar{\mu}_{i,t} - \mu_i| \leqslant \sup_{1 \leqslant n \leqslant 8\kappa_i} \sqrt{n} |\bar{\mu}_i(n) - \mu| \leqslant \sqrt{\log 8\kappa_i},$$

which completes the proof. ■

E.2 Proof of Lemma E.2

Since event $\mathcal{E}_{\mathsf{UCB}}$ holds, we have that events $\cap_{i=1}^K \mathcal{E}_0^{(i)}$ (defined in Equation (5)) and $\mathcal{E}_2^{(i)}$ (defined in Equation (8a)) hold. Our first observation is that we have $n_{i,\tau_i} \geqslant n_{i,3T/4}$ on the event $\mathcal{E}_{\mathsf{UCB}}$. To see this, we define $\widetilde{\tau}_i$ to be the *last* time that arm i was pulled before round $3T/4$, and make a series of observations:

1. $\widetilde{\tau}_i$ *always exists and is well-defined*, since an examination of Algorithm 2 reveals that all arms will be pulled at least once in the first K rounds, including any suboptimal arm i.
2. On the event $\cap_{i=1}^K \mathcal{E}_0^{(i)}$, we have $n_{i,T} \leqslant 8\kappa_i$ for all $i \neq i^*$. Since $T \geqslant 32 \sum_{i \neq i^*} \kappa_i$, this implies that the optimal arm i^* will be pulled at least $3T/4$ times, and hence, at least once between rounds ${}^{3T}/_4$ and T. Thus, *the optimal arm i^* is pulled at least once after $\widetilde{\tau}_i$.*
3. By definition, $n_{i,\widetilde{\tau}_i} = n_{i,3T/4}$.

The first two observations (italicized) are also satisfied for the round τ_i, except that it is the maximal round for which these observations hold. Thus, we have $n_{i,\tau_i} \geqslant n_{i,\widetilde{\tau}_i} = n_{i,3T/4}$, implying that it suffices to lower bound $n_{i,3T/4}$.

Consider the reward tapes for arms i^* and i, indexed by $n = 1, \ldots, {}^{3T}/_4$. Recall that the confidence interval in reward-tape notation is given by $C(n) \stackrel{\text{def}}{=} \sqrt{\frac{2(T^\alpha - 1)}{\alpha n}}$. First, note that for $n \geqslant 8\kappa_i$ we have $C(n) \leqslant {}^{\Delta_i}/_4$. Thus, we have

$$\bar{\mu}_{i^*}(n) + C(n) \leqslant \mu_{i^*} + 2C(n) \leqslant \mu_{i^*} + \frac{\Delta_i}{2} \text{ for all } n \geqslant 8\kappa_i \tag{32}$$

where the first inequality holds on event $\mathcal{E}_0^{(i^*,i)}$. On the other hand, event $\mathcal{E}_2^{(i)}$ gives us

$$\bar{\mu}_i(n) \geqslant \mu_i - \sqrt{\frac{3}{4}} \cdot C(n) \text{ for all } n \geqslant 1. \tag{33}$$

Finally, Lemma C.3 guarantees (see also its proof) that under event $\mathcal{E}_0^{(i)} \cap \mathcal{E}_0^{(i^*)}$, we have $n_{i,\tau_i} \leqslant 8\kappa_i$.

We now use these statements to prove the lemma. We consider indices n, n' for reward tapes corresponding to arms i^* and i respectively. Provided that $n \geqslant 8\kappa_i$ and $n' \leqslant \kappa_i/32$, we obtain

$$\begin{aligned}
\bar{\mu}_i(n') + C(n') &> \mu_i - \sqrt{\frac{3}{4}} \cdot C(n') + C(n') \\
&= \mu_i + C(n') \left(1 - \sqrt{\frac{3}{4}}\right) \\
&\geqslant \mu_{i^*} + \frac{\Delta_i}{2} \\
&\geqslant \bar{\mu}_{i^*}(n) + C(n),
\end{aligned}$$

where the first and third inequality follow from Equations (33) and (32) respectively, and the second inequality follows from the constraint on n'. Ultimately, we obtain

$$\bar{\mu}_i(n') + C_i(n') \geqslant \bar{\mu}_{i^*}(n) + C_{i^*}(n) \text{ as long as } n \geqslant 8\kappa_i \text{ and } n' \leqslant \frac{\kappa_i}{32}. \tag{34}$$

In essence, Equation (34) describes a sufficient condition for arm i^* *not* to be picked, i.e. the reward tape for arm i^* has been run for greater than $8\kappa_i$ cells and the reward tape for arm i has been run for at most $\kappa_i/32$ steps. At a high level, our proof strategy is as follows: on the event $\cap_{i=1}^K \mathcal{E}_0^{(i)}$, arm i^* has to be picked sufficiently often at "regular intervals". For this to be possible, arm i needs to be pulled a minimal number of times to ensure that the condition in Equation (34) is *not* satisfied.

We expand on this proof intuition below. Consider the round $T/2$ and note from Lemma C.3 that the optimal arm i^* needs to be pulled at least once between rounds $T/2$ and $3T/4$. This requires

$$\bar{\mu}_{i^*}(n_{i^*,t}) + C_{i^*,t} \geqslant \bar{\mu}_i(n_{i,t}) + C_{i,t} \text{ for some } t \in \left[\frac{T}{2}, \frac{3T}{4}\right]. \tag{35}$$

We now split the proof into two cases.

Case $n_{i,T/2} \geqslant \kappa_i/32$: In this case, we have $n_{i,3T/4} \geqslant n_{i,T/2} \geqslant \kappa_i/32$ and we are done.

Case $n_{i,T/2} < \kappa_i/32$: We provide a proof-by-contradiction for this case. Suppose that $n_{i,3T/4} < \kappa_i/32$. By Lemma C.3, arm i^* has to be pulled at least $8\kappa_i$ times within the horizon $T/2$, i.e. we have $n_{i^*,t} \geqslant 8\kappa_i$ for all $t \in [T/2, 3T/4]$. Thus, if we had $n_{i,3T/4} < \kappa_i/32$, the condition in Equation (34) would be satisfied for all $t \in [T/2, 3T/4]$, implying that arm i^* can *never* be picked in this interval. This contradicts our statement that arm i^* has to be pulled at least once in this interval, and shows that we require $n_{i,3T/4} \geqslant \kappa_i/32$ in this case. This completes the proof. ■

E.3 Proof of Lemma E.3

Since event $\mathcal{E}_{\mathsf{UCB}}$ holds, we have that the event $\mathcal{E}_4^{(i^*,i)}$ (defined in Equation (9)) holds. We begin with the following claim, which we return to prove momentarily.

Claim E.4. *Under the event $\mathcal{E}_{\mathsf{UCB}}$, there exists a universal positive constant c such that*

$$n_{i^*,\tau_i} \geqslant c\kappa_i. \tag{36}$$

Taking this claim as given, we split the proof of the lemma into two cases:

Case 1: $c\kappa_i \leqslant n_{i^*,\tau_i} \leqslant \kappa_i^2$. Here, the first case under the event $\mathcal{E}_4^{(i^*,i)}$ directly yields

$$|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \leqslant \sqrt{\frac{2\log \kappa_i}{c\kappa_i}}.$$

Case 2: $n_{i^*,\tau_i} > \kappa_i^2$. Here, the second case under the event $\mathcal{E}_4^{(i^*,i)}$ directly yields

$$|\bar{\mu}_{i^*}(n) - \mu_{i^*}| \leqslant \sqrt{\frac{\log T}{\kappa_i^2}}.$$

To complete the proof for this case, note that

$$\kappa_i = \frac{4(T^\alpha - 1)}{\alpha \Delta_i^2} \geqslant \frac{4\log T}{\Delta_i^2} \geqslant 4\log T,$$

which gives us $|\bar{\mu}_{i^*}(n_{i^*,\tau_i}) - \mu_{i^*}| \leqslant \sqrt{1/4\kappa_i}$. It only remains to establish Claim E.4, which we do below.

Proof of Claim E.4: Since event $\mathcal{E}_{\mathsf{UCB}}$ holds, we have that events $\mathcal{E}_3^{(i)}$ and $\mathcal{E}_3^{(i^*)}$ (defined in Equation (8b)) hold and the statement of Lemma E.2 holds. Then, we have:

$$\bar{\mu}_{i^*}(n) + C(n) \geqslant \mu_{i^*} + \left(1 - \frac{1}{\sqrt{2}}\right) C(n) \quad \text{for all } n \leqslant \frac{\kappa_i}{32} \tag{37a}$$

$$\bar{\mu}_i(n) + C(n) \leqslant \mu_i + \left(1 + \frac{1}{\sqrt{2}}\right) C(n) \quad \text{for all } n \leqslant \frac{\kappa_i}{32}, \text{ and} \tag{37b}$$

$$n_{i,\tau_i} \geqslant \frac{\kappa_i}{32}. \tag{37c}$$

Let us define γ_i as the $\kappa_i/32$-th time that arm i is pulled, and $\underline{\gamma_i}$ as the $(\kappa_i/32 - 1)$-th time that arm i is pulled. We will prove the lemma for the explicit choice $c = 1/968$. We now have two cases.

Case 1: $n_{i^*,\gamma_i - 1} \geqslant c\kappa_i$. In this case, the claim follows immediately, since on event $\mathcal{E}_3''$, we have $n_{i^*,\tau_i} \geqslant n_{i^*,\gamma_i}$.

Case 2: $n_{i^*,\gamma_i - 1} < c\kappa_i$. As a consequence of the above inequalities, we have:

$$\begin{aligned}
\bar{\mu}_{i^*}(n_{i^*,\gamma_i - 1}) + C(n_{i^*,\gamma_i - 1}) &\geqslant \mu_{i^*} + \left(1 - \frac{1}{\sqrt{2}}\right) C(n_{i^*,\gamma_i - 1}) \\
&\geqslant \mu_{i^*} + \left(1 - \frac{1}{\sqrt{2}}\right) C(c\kappa_i) \\
&= \Delta_i + \mu_i + \left(1 - \frac{1}{\sqrt{2}}\right) C(c\kappa_i) \\
&\geqslant \Delta_i + \bar{\mu}_i\left(\frac{\kappa_i}{32} - 1\right) - \frac{1}{\sqrt{2}} C\left(\frac{\kappa_i}{32} - 1\right) + \left(1 - \frac{1}{\sqrt{2}}\right) C(c\kappa_i),
\end{aligned}$$

where the first inequality follows from Equation (37a), the second inequality is because $C(\cdot)$ is decreasing in its argument, and the third inequality follows from Equation (37b). By definition, $n_{i,\underline{\gamma_i}} = \kappa_i/32 - 1$, and by Lemma E.2 arm i must be pulled at least one more time. Furthermore, since $c = 1/968$ we have

$$\begin{aligned}
-\frac{1}{\sqrt{2}} C\left(\frac{\kappa_i}{32} - 1\right) + \left(1 - \frac{1}{\sqrt{2}}\right) C(c\kappa_i) &= -\frac{1}{\sqrt{2}} C\left(\frac{\kappa_i}{32} - 1\right) + 5.5\left(1 - \frac{1}{\sqrt{2}}\right) C\left(\frac{\kappa_i}{32}\right) \\
&\geqslant 3.2\Delta_i \\
&> C\left(\frac{\kappa_i}{32} - 1\right) - \Delta_i,
\end{aligned}$$

where the final two steps follow from the relations

$$4\Delta_i \leqslant C(\kappa_i/32) \leqslant C(\kappa_i/32 - 1) \leqslant 4.2\Delta_i.$$

Putting together the pieces yields

$$\bar{\mu}_{i^*}(n_{i^*,\gamma_i-1}) + C(n_{i^*,\gamma_i-1}) > \bar{\mu}_i(n_{i,\gamma_i-1}) + C(n_{i,\gamma_i-1}) = \bar{\mu}_i(n_{i,\underline{\gamma}_i}) + C(n_{i,\underline{\gamma}_i}).$$

But by definition, arm i is pulled at round γ_i, and so we have the desired contradiction. Consequently, we must have $n_{i^*,\gamma_i} > c\kappa_i$. The claim then follows from the observation that $n_{i,\tau_i} \geqslant n_{i,\gamma_i}$ (by Eq (37c)). ■

F Additional Experimental Details and Results

In this section, we provide additional details on the experiments with simulated and battery charging data in Section 5, as well as further experimental results.

F.1 Simulation results with SAE

First, we present the simulation results with the SAE algorithm in Figure 6.

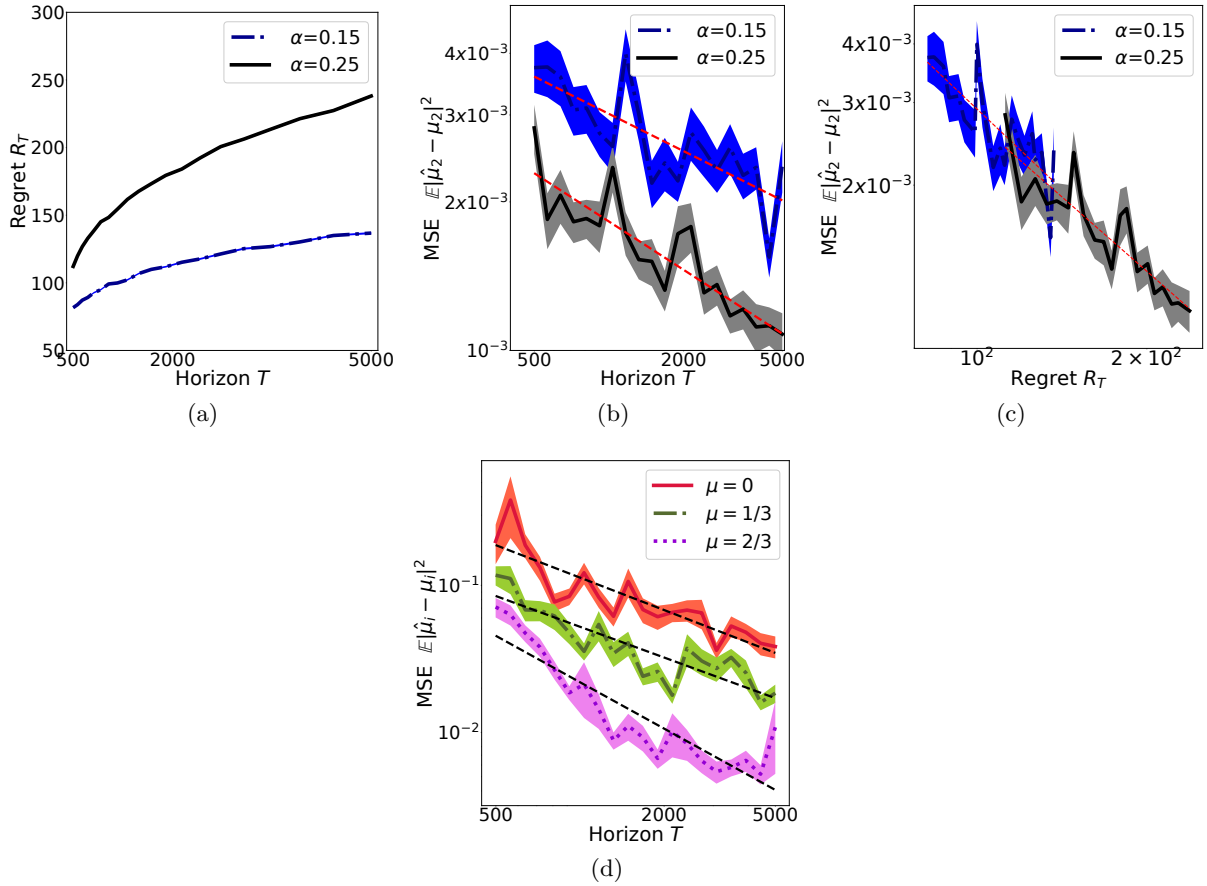


Figure 6: Results of 250 runs of simulation experiments for the SAE algorithm. Figures (a-c) are for a two-armed bandit instance with $\mu = (1, 1/2)$ and Gaussian rewards with unit variance. Here, individual curves represent two values of $\alpha \in \{0.15, 0.25\}$. Figure (d) is a 4-armed instance with $\mu = (1, 2/3, 1/3, 0)$ and Gaussian rewards with variance $1/4$. Here, individual curves represent the three suboptimal arms. Overall, these log-log plots corroborate our principal finding that better reward estimation is achievable from higher regret demonstrations; see the text for a detailed discussion.

F.2 Simulating multi-armed bandits

We design a simple simulator for K-arm multi-bandit instances. In all our experiments, we assume Gaussian rewards for each arm, i.e $r_i \sim N(\mu_i, \sigma^2)$. Note that we fixed the variance σ^2 across all arms. The code for

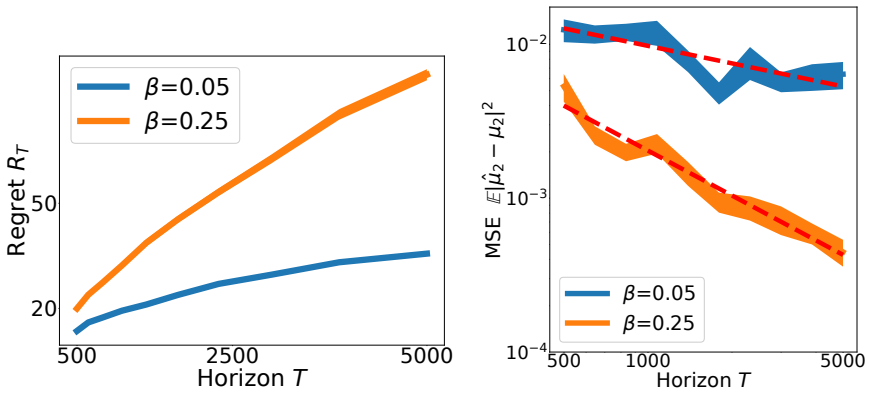
reproducing the results will be shared publicly on publication.

Algorithm implementation: Algorithms 1 and 2 provide $O(T^\alpha)$ regret for any $\alpha \in (0, 1)$. Using our simulator, we collect demonstrations for different $\alpha \in \{0.15, 0.25\}$.

For the two-armed bandit instance, we let $\mu_1 = 1$ and $\mu_2 = 0.5$, i.e with a fixed $\Delta = 0.5$. For the K-arm instances, we choose the means μ_i to be linearly spaced between $[0, 1]$, (e.g for K=4, $\mu = (1, 2/3, 1/3, 0)$) with fixed variance $\sigma^2 = 0.25$ across all arms. We report results averaged over 100 independent demonstrations. We evaluate the estimators in Procedures 1 and 2 for different time-horizons T, evenly spaced in log space $\in [500, 5000]$.

Mean-squared error vs regret: In Corollary 4.3, we characterized the relationship between error in estimating rewards, and regret of the demonstrator's algorithm. Recall that for different values of T, the regret of both our upper-confidence-bound algorithms grows as $O(T^\alpha)$. To study relationship between mean-squared error (MSE) and regret, we fix T and collect multiple demonstrations for instance with a fixed gap Δ. The mean regret R_T and corresponding standard error are computed by averaging across these demonstrations. Similarly, we estimate the gap using Procedure 2 and 1, measuring MSE as average of the squared error for $T \in [500, 5000]$.

F.3 Dependence on Δ



(a) (b)

Figure 7: For the two-arm case, we construct MAB instances of varying difficulty by choosing the value of the suboptimality gap Δ_T for horizon T as $\Delta_T = 1/T^\beta$. In figure (a), we verify that the regret increases for higher β (i.e for fixed T, suboptimality gap reduces with higher β). Figure (b) empirically supports our predictions; for fixed β, estimation error decreases with T.

In this section, we explore the role of the suboptimality gap Δ in reward estimation for the case $K = 2$. Corollary 4.3 predicts that decreasing Δ will make reward estimation easier because it increases the regret. To investigate whether this happens empirically, we make Δ_T decay smoothly with increasing time-horizon T, following the power law $\Delta_T = 1/T^\beta$, for $\beta \in (0, 0.5)$. We expect the following behavior:

1. for fixed β, the reward estimation error decays with increasing horizon T.
2. for fixed horizon T, the rate of decay of reward estimation error increases with β.

We consider two cases: $\beta \in \{0.05, 0.5\}$. Figure 7 shows that the regret indeed increases with a decrease in suboptimality gap (higher T). We observe that the reward estimation error decreases with T for both values of β. Moreover, the estimation error decays faster for larger values of β.

F.4 Comparisons with the naive estimator

We provide further comparisons between the proposed estimator and the naive estimator as described in Section 4. We evaluated the naive estimator with different values of $C_0 \in \{0.2, 0.75, 1.0, 1.5\}$ in Fig. 8. The experiment

setting is similar to Fig 4 with two-arm stochastic bandits and a UCB demonstrator, where the mean rewards of the two arms are $\mu_1 = 1.0, \mu_2 = 0.5$ with standard deviation $\sigma = 1.0$. For the baseline plots in Fig 8 (b)-(e), the horizon T is linearly spaced in $[500, 1000]$. All the results are averaged over 50 runs and plotted with the standard errors.

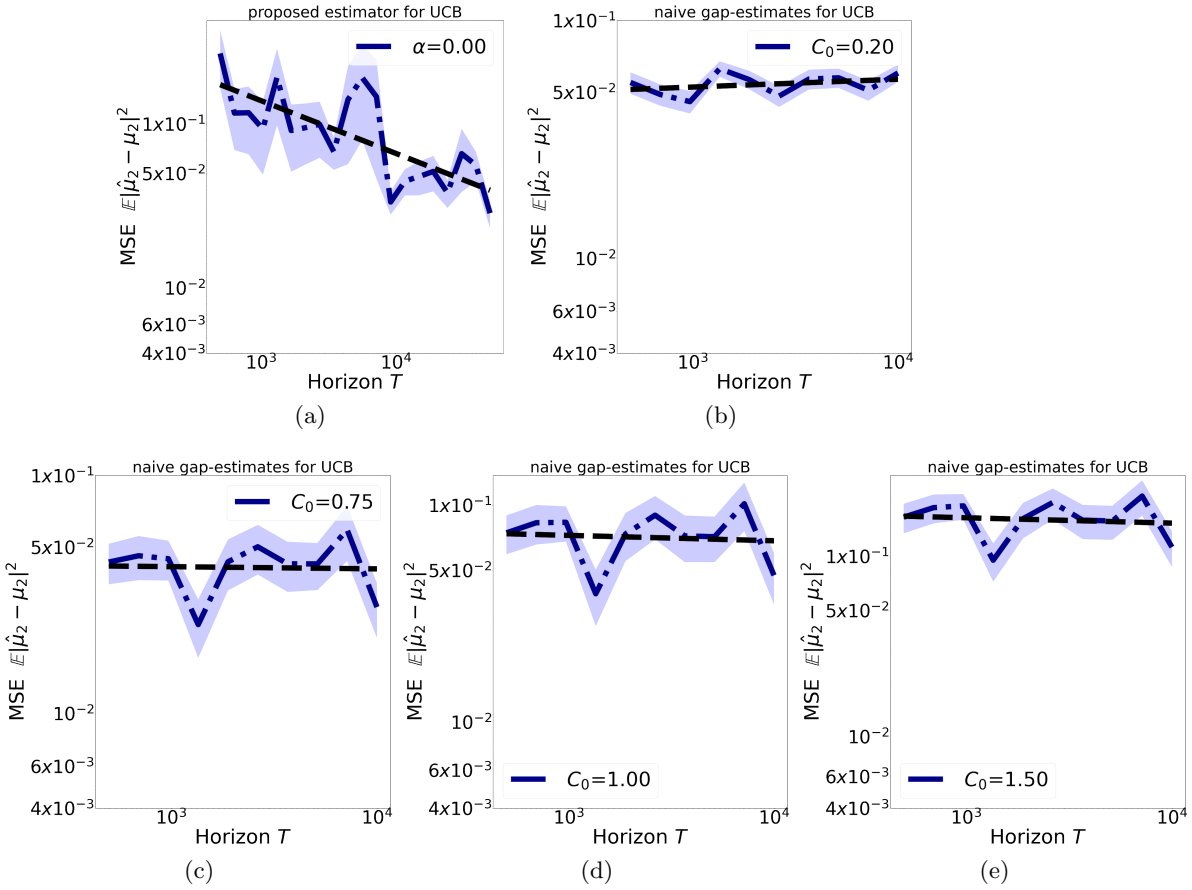


Figure 8: Comparing our proposed estimator with a simple estimator $\widehat{\Delta} = C_0\sqrt{\frac{\log T}{n_a}}$ for $C_0 \in \{0.2, 0.75, 1.0, 1.5\}$.

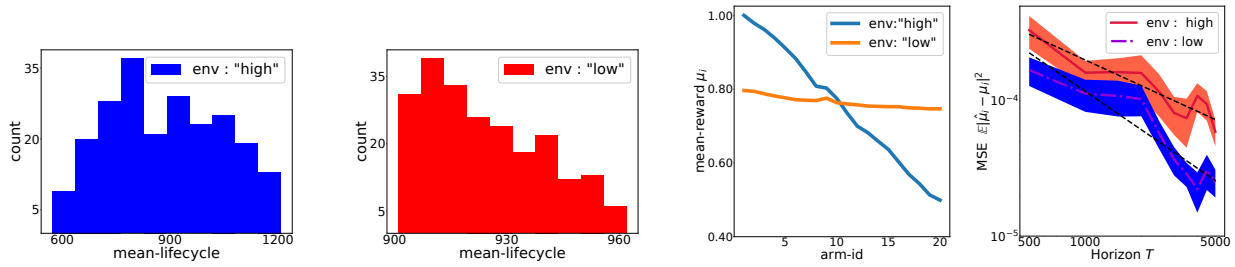
F.5 Battery charging

Dataset: The original dataset (Attia et al., 2020) provides battery life-cycles for 224 protocols, in different temperature regimes. The problem of identifying the optimal protocol (with highest mean lifetime) is cast as a MAB problem with 224 arms, where the three regimes are different instances. The distribution of reward means μ_i varies significantly across the regimes (see Figure 9). In our experiments, we compare the "low" and "high" temperature regimes. We subsample 20 arms which are representative of the distribution. In particular, we generate a histogram of rewards for the "high" setting with $n = 20$ bins, and pick an arm randomly from each bin. We fix this subset of 20 arms for all our experiments in this section. Unless mentioned, we evaluate the estimators with number of independent runs (N) as $N = 100$.

Normalization: The lifecycle of batteries across the regimes is in the range [573, 1208], with empirical standard-deviation (for the Gaussian prior) given by $\sigma = 164$. We normalize the distribution parameters such that $\mu_i \in [0, 1]$ for all arms $i \in [K]$. The normalization constant is fixed to be maximum of the life-cycles across all environments, i.e., $\mu_{max} = 1208$. This preprocessing provides instances with $\mu_i \in (0.474, 1]$ and $\sigma^2 = 0.018$.

Adjusting for variance: The estimators in Procedures 1 and 2 are defined under the assumption that $\sigma = 1$. For non-unit σ, we extend the procedures to their *variance-adjusted* versions, scaling the confidence interval by

5σ, i.e $C_{i,t} = 5\sigma\sqrt{\frac{T^\alpha - 1}{\alpha n_{i,t}}}$, while still using the same estimators.



(a) (b) (c) (d)

Figure 9: Reward distributions in the "low" and "high" temperature regimes vary significantly. Figure (d) represents the error of estimating arm 12 in both "low" and "high" regimes with 4 protocols.

Additional results: While the distributions of mean lifetime vary between the high and low temperature regimes, there are protocols that enjoy similar performance across both regimes. For instance, Figure 9(a) shows that arm 8 has normalized rewards of 0.802 and 0.775 in the high and low regimes respectively. In Figure 9(d), we demonstrate that it is easier to estimate the mean lifetime of this arm in the low regime. We run the same experiment for large subset of arms in the dataset. In this setting we take $\alpha = 0.001$ to get a low-regret demonstrator, and fix T$\in \{25000, 45000, 70000\}$. In Figure 10, we verify pictorially that the reward estimation error reduces uniformly across all arms as we increase the demonstration horizon (see the caption for a detailed explanation).

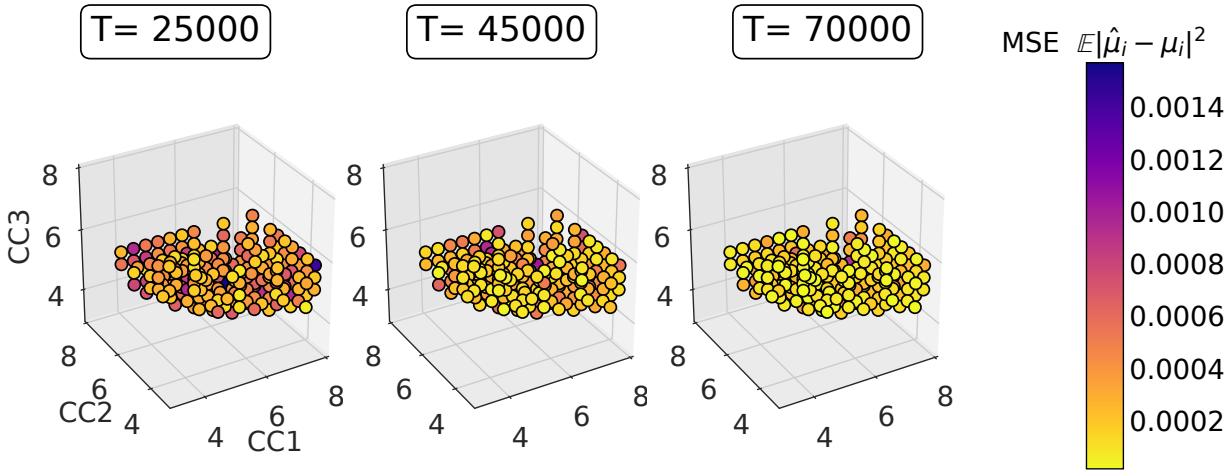


Figure 10: Each charging protocol in the battery lifecycle dataset is defined by three independent variables (CC1, CC2, CC3). These parameters correspond to constant-currents applied to the battery in a specified range (0–20%, 20–40% and 40–60%, respectively). Here, each point in the plot corresponds to one such protocol and the color profile represents mean-squared-error in estimating the average lifetime. As we increase the time-horizon T of the demonstration, our estimates improve uniformly across protocols.

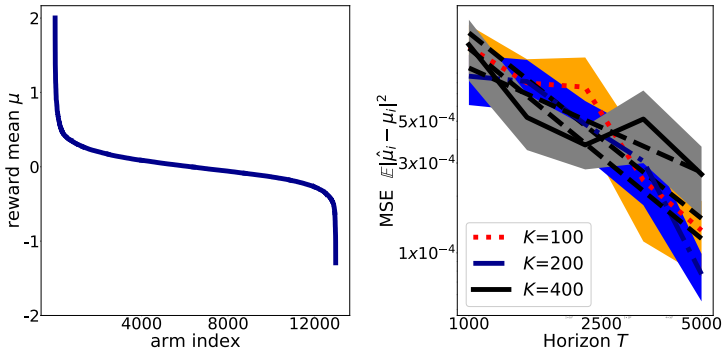
F.6 Gene expression

Dataset: Identifying the top genes responsible for virus replication could provide information about potential targets for antiviral therapy in the host. In one such study, Hao et al. (2008) investigate 13K genes in *drosophila* in the context of influenza, by adding fluorescence virus to single-gene knock-down cell strains. Measuring the

fluorescence level, the authors estimate the importance of the corresponding gene in replication; where lower fluorescence indicates that the knock-down gene encourages replication. This problem of identifying top-k genes under noisy measurements has previously been studied under the best-arm identification setting (Jun et al., 2016).

Normalization: Following the original dataset, we model rewards for arm i to follow $N(\mu_i, 0.1)$. As indicated by Figure 11 (a), the reward means μ_i lie in the range $(-1.3, 2.01)$. We normalize the reward means to be within the range $\mu_i \in [0,\ 1]$ by centering and scaling. Accordingly, the variance per arm is normalized to 0.0092. In summary, we have $r_i \sim N(\mu_i, 0.0092)$.

Results: Our goal is to estimate the mean reward μ_i of each knock-down gene from a single demonstration with uniform error guarantees. We subsample K $\in \{100, 200, 400\}$ arms from a dataset of 12979 arms, and evaluate our estimator on each of the resulting instances. While sampling the subset with K arms, we ensure that arm 12979 is present across all instances, and track the error in estimating its mean reward across different instances. In Figure 11(b), we demonstrate that our estimator works well across all values of K. Figure 11(b) also shows that the estimation error depends minimally on K as predicted by our theory.



(a) (b)

Figure 11: (a) depicts mean-reward per-arm for all the 12979 arms before normalization. (b) We track the reward estimation error of arm 12979 (this arm is added to all instances) as a function of T for $K \in \{100, 200, 400\}$.