Leveraging Demonstrations to Improve Online Learning: Quality Matters

Botao Hao [1]   Rahul Jain [2]   Tor Lattimore [1]   Benjamin Van Roy [1]   Zheng Wen [1]

Abstract

We investigate the extent to which offline demonstration data can improve online learning. It is natural to expect some improvement, but *the question is how, and by how much?* We show that the degree of improvement must depend on the *quality* of the demonstration data. To generate portable insights, we focus on Thompson sampling (TS) applied to a multi-armed bandit as a prototypical online learning algorithm and model. The demonstration data is generated by an expert with a given *competence* level, a notion we introduce. We propose an informed TS algorithm that utilizes the demonstration data in a coherent way through Bayes' rule and derive a prior-dependent Bayesian regret bound. This offers insight into how pretraining can greatly improve online performance and how the degree of improvement increases with the expert's competence level. We also develop a practical, approximate informed TS algorithm through Bayesian bootstrapping and show substantial empirical regret reduction through experiments.

1. Introduction

A modern paradigm for developing intelligent agents involves pretraining on large quantities of existing data followed by learning from real-time interactions. For instance, to produce a chatbot, one can pretrain a large language model on text gathered from the internet and subsequently improve behavior through learning from interactions with humans (Ziegler et al., 2019; Ouyang et al., 2022). With such an approach, the preexisting text is treated as offline demonstration data that conditions a reinforcement learning agent before it engages in online learning.

It is natural to expect the offline demonstration data to improve performance of the online learning agent. However, the degree of improvement must depend on the quality of the demonstration data. If the data is produced by a *competent* expert, it ought to improve the agent's performance more so than if not. We study using the demonstration data to enhance the performance of an online learning agent in terms of regret minimization, formulating a notion of competence and an approach to learning from the demonstration data in a manner that accounts for this.

As a prototypical model for online learning, we consider multi-armed bandits that offer a simple context for understanding the role of offline data. We focus on Thompson sampling (TS) (Thompson, 1933) which is a popular online learning algorithm, owing to its effectiveness across a wide range of environments and its scalability through the use of approximation methods such as epistemic neural networks (Osband et al., 2021). Moreover, TS offers a coherent way to use the demonstration data by simply following Bayes' rule. In our setting, pretraining amounts to conditioning the distribution used by TS as it initiates online learning. Note that a goal of this paper is to yield insights on how to leverage the demonstration data to improve online learning, and what the potential gains may be. Thus, we use the prototypical setting of a multi-armed bandit and a simple model for the data generation that is mathematically convenient and widely-used in machine learning.

Contributions   Our contribution is three-fold:

(i) We study how the quality of offline demonstration data can improve online learning and propose an *informed* TS algorithm that naturally makes use of the offline data through an *informative* prior. We show that the algorithm's online learning performance improves significantly with the quality of demonstrations as measured by a notion of the expert's competence level that we introduce.

(ii) We establish a *prior-dependent* Bayesian regret bound that offers insight into how pretraining reduces regret and how this reduction depends on the expert's competence level. Previous works (Russo & Van Roy, 2016; Zhang, 2022; Hao & Lattimore, 2022) mostly focus on *prior-free* Bayesian regret bounds and thus cannot characterize the quality of offline data. Our technique extends the information-theoretic

regret decomposition to characterize how an informative prior can reduce both the information ratio and the entropy of the distribution of the optimal arm.

(iii) We propose a practical algorithm that approximates *informed* TS via Bayesian bootstrapping. Through experiments, we show that a *partially-informed* TS algorithm that uses the offline demonstration data naively (i.e., assuming the expert to be naive in that it takes actions uniformly randomly) can only reduce the cumulative regret marginally. However, the *informed* TS algorithm informed by the offline demonstration data and the competence level of the expert, can achieve substantial regret reduction.

Related Work. There is a rich body of literature on learning algorithms for bandits (see Russo et al. (2018); Lattimore & Szepesvári (2020) for a detailed review). Almost all of this literature assumes that the learning agent starts from scratch but this may lead to a long initial learning stage. In fact, offline data is available for many applications, such as training a large language model (Ouyang et al., 2022).

There are some previous attempts that leverage offline data to warm-start an online learning algorithm. For instance, Shivaswamy & Joachims (2012) analyzed a warm start UCB algorithm for the K-armed bandit and Zhang et al. (2019) investigated warm-starting contextual bandits by combining offline supervised feedback that is generated by an uniformly random policy. Banerjee et al. (2022) proposed a meta-algorithm that uses historical data as needed to improve the computation and storage. However, none of these algorithms take the quality of offline data into consideration and thus show little regret reduction.

In the context of reinforcement learning (RL), there are several recent works (Rashidinejad et al., 2021; Xie et al., 2021; Song et al., 2022; Wagenmaker & Pacchiano, 2022) that bridge offline and online RL. However, all of them focus on policy optimization rather than regret minimization. And they require different versions of concentrability coefficient conditions that are hard to be satisfied in practice. The importance of the competence level of the expert was first highlighted by Beliaev et al. (2022) for imitation learning and is modeled through an ϵ-greedy policy. However, their goal is very different from ours since there is no online interaction there.

Offline data can also be viewed as a special form of side information and some other forms of side information are studied for online learning. Degenne et al. (2018) assumed the side information as observations of other arms while Cutkosky et al. (2022) considered some hints about the optimal action for linear bandits.

2. Problem Setting

To offer a coherent formulation – that is, one that conforms with standard axioms of statistical decision theory – we model all unknown quantities as random variables defined with respect to a common probability space $(\Omega, \mathbb{F}, \mathbb{P})$. For a set $\mathcal{S}$, we denote $|\mathcal{S}|$ as its cardinality. For positive integer N, let $[N] := \{1, 2, \ldots, N\}$. The $K \times K$ identity matrix is $\boldsymbol{I}_K$.

We consider a stochastic K-armed linear bandit with action set $\mathcal{A} = \{a_1, a_2, \ldots, a_K\} \subseteq \mathbb{R}^d$. The environment is identified by a random vector $\theta \in \mathbb{R}^d$ with prior distribution $\nu_0(\cdot) = \mathbb{P}(\theta \in \cdot)$. The agent begins with an offline demonstration dataset $\mathcal{D}_0 = \{(\bar{A}_n, \bar{R}_n)\}_{n=1}^N$ consisting of action-reward pairs. Then, at each time $t \in [T]$, the agent chooses an action $A_t \in \mathcal{A}$ and receives a reward

$$R_t = \langle A_t, \theta \rangle + \eta_t \,,$$

where $\langle \cdot, \cdot \rangle$ is the vector inner product and $(\eta_t)_{t=1}^T$ is a sequence of independent standard Gaussian random variables. The experience thus far is recorded in the *online* history $\mathcal{H}_t = \{(A_\tau, R_\tau)\}_{\tau=1}^t$ with $\mathcal{D}_t = (\mathcal{D}_0, \mathcal{H}_{t-1})$ denoting the entire dataset at time t. We write $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot | \mathcal{D}_t)$ as the posterior measure where $\mathbb{P}$ is the probability measure over θ and the history and $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot | \mathcal{D}_t)$.

A (learning) policy $\pi = (\pi_t)_{t \in \mathbb{N}}$ is a sequence of deterministic functions where $\pi_t(\cdot | \mathcal{D}_t)$ specifies a probability distribution over $\mathcal{A}$ conditioned on the dataset $\mathcal{D}_t$. A stationary policy is an element of the probability simplex that does not depend on history. Let $A^* = \operatorname{argmax}_{a \in \mathcal{A}} \theta^\top a$ and we define the Bayesian regret of a policy π as

$$\mathfrak{BR}_T(\pi) := \mathbb{E}\left[\sum_{t=1}^T \langle A^*, \theta \rangle - \sum_{t=1}^T R_t\right],$$

where the expectation is over the environment θ, the interaction sequence induced by the policy and environment and the offline demonstration data $\mathcal{D}_0$.

2.1. Competence

Each action $\bar{A}_n$ in the offline demonstration dataset is generated by an expert and the expert's expertise level is characterized by a notion of *competence*.

In particular, the expert's competence is parameterized by values $\lambda \geq 0$ and $\beta \geq 0$, which represent *knowledgeability* and *deliberateness*. The expert's knowledge takes the form of a vector ϑ, which is distributed as $\mathcal{N}(\theta, \boldsymbol{I}_d / \lambda^2)$ conditioned on θ, and actions are selected according to an expert policy,

$$\phi_{\beta, \lambda}(\bar{A}_n = a | \vartheta) = \frac{\exp\left(\beta a^\top \vartheta\right)}{\sum_{b \in \mathcal{A}} \exp\left(\beta b^\top \vartheta\right)} \,. \qquad (2.1)$$

If λ or β is finite, the expert policy above takes suboptimal actions. One source of suboptimality stems from the agent's knowledgeability λ, which induces error in the agent's knowledge of θ. The other is due to deliberateness β, which drives the agent to choose actions that fail to optimize $\vartheta^\top a$. Together, this knowledgeability and deliberateness determine what we refer to as the agent's *competence*.

The form of the expert policy is simple, mathematically convenient and widely used in reinforcement learning. The particular form we use is also expressive as we now note.

Remark 2.1. *For multi-armed bandits where the actions are the basis vectors, $\phi_{\beta,\lambda}(\bar{A}_n = \cdot|\vartheta)$ is a random variable supported on the whole probability simplex. In other words, if one views β and ϑ as the parameters that parameterize the policy, the softmax policies with structure as in Eq. (2.1) is enough to realize any stationary policy.*

Remark 2.2. *[Beliaev et al. (2022)](#) consider an ϵ-greedy expert policy (simplified from a MDP model): for $\beta \in [0,1]$*

$$\pi_\beta(\bar{A}_n = a|\theta) = \beta\mathbb{I}\left(a = \underset{a \in [K]}{\arg\max}\, a^\top\theta\right) + (1-\beta)/K \,,$$

where $\mathbb{I}(\cdot)$ is the indicator function. However, they assume that the expert has perfect knowledge of the environment, e.g., the knowledgeability parameter $\lambda = \infty$, which is not realistic in practice.

In general we expect that the insights developed for the specific model in Equation (2.1) will generalise to alternative models with similar qualitative properties.

3. Informed Thompson Sampling

We introduce the *informed* Thompson Sampling (iTS) algorithm that uses the offline demonstration data in a coherent way. The details of the iTS algorithm are:

1. It constructs an *informative* prior by the use of offline dataset $\mathcal{D}_0$ and Bayes' rule, and which satisfies

$$\mathbb{P}(\theta \in \cdot|\mathcal{D}_0) \propto \mathbb{P}(\mathcal{D}_0|\theta \in \cdot)\nu_0(\cdot)$$
$$\propto \nu_0(\cdot) \prod_{n=1}^{N} \underbrace{\mathbb{P}(\bar{A}_n|\theta \in \cdot)}_{\text{action likelihood}}\underbrace{\mathbb{P}(\bar{R}_n|\bar{A}_n, \theta \in \cdot)}_{\text{reward likelihood}}, \quad (3.1)$$

where $\nu_0(\cdot)$ is the initial prior for θ.

2. The algorithm then uses the *informative* prior to start learning and taking actions in the usual way: At time t, obtain a sample $\widetilde{\theta}_t$ from the current posterior distribution $\mathbb{P}(\theta \in \cdot|\mathcal{D}_t)$ and choose an arm

$$A_t = \arg\max_{a \in \mathcal{A}} a^\top \widetilde{\theta}_t \,.$$

Obtain reward R_t and update the posterior distribution $\mathbb{P}(\theta \in \cdot|\mathcal{D}_t)$.

Due to the action likelihood term, drawing samples from the exact posterior distribution $\mathbb{P}(\theta \in \cdot|\mathcal{D}_t)$ is hard even though we have a conjugate initial prior for θ. In Section 5, we propose an approximate-TS algorithm based on Bayesian bootstrapping.

Remark 3.1. *It is worth emphasizing here that the actions in the offline dataset also carry information about the environment through the action likelihood term $\mathbb{P}(\bar{A}_n|\theta \in \cdot)$ which can incorporate any information about the generative model of the policy used to generate the offline dataset, and thus greatly improve the informativeness of the prior.*

4. An Information-Theoretic Analysis

We now present an information-theoretic regret analysis of the *informed TS* algorithm. In particular, we demonstrate the role of the competence level though a *prior-dependent* Bayesian regret bound.

In the literature, there are two ways to prove Bayesian regret bounds. The first is to introduce confidence sets such that the Bayesian regret bounds of TS match the best possible worst-case regret bounds of the UCB algorithm ([Russo & Van Roy, 2014](#); [Zhang, 2022](#)). However, it is unclear how to use prior information or offline datasets to construct confidence sets in a principled way. There are some exceptions that attempt to show the effect of the prior distribution ([Bubeck & Liu, 2013](#); [Kveton et al., 2021](#); [Simchowitz et al., 2021](#)) but all rely on conjugate priors that do not hold for our setting.

The second is to decompose the Bayesian regret into an *information ratio* term and an *entropy* term and bound them using tools from information theory ([Russo & Van Roy, 2016](#)). Next we show that existing analysis is not sufficient to fully characterize the prior effect.

4.1. Why Existing Analysis Is Not Sufficient

We reproduce the key step of existing information-theoretic analysis ([Russo & Van Roy, 2016](#)). Define the notion of information ratio:

$$\Gamma_t = \frac{(\mathbb{E}_t\left[\langle A^*, \theta\rangle - R_t\right])^2}{\mathbb{I}_t(A^*; (A_t, R_t))} \,,$$

where $\mathbb{I}_t$ is the conditional mutual information[1]. [Russo & Van Roy (2016)](#) bounded the Bayesian regret of the TS algorithm as follows:

$$\mathfrak{BR}_T(\pi^{\text{TS}}) \le \sqrt{\Gamma^* \mathcal{H}(A^*)T} \,,$$

where $\Gamma_t \le \Gamma^*$ almost surely for any $t \in [T]$ and $\mathcal{H}(\cdot)$ is the Shannon entropy. On the one hand, the effect of the prior distribution can be partially characterized through its

---

[1] $\mathbb{I}_t(X;Y) = \mathbb{E}_t[D_{\text{KL}}(\mathbb{P}_{t,X|Y}||\mathbb{P}_{t,X})]$

entropy. But this can only lead to logarithmic improvement since $\mathcal{H}(A^*)$ is always bounded by $\log(K)$.

On the other hand, the upper bound on the information ratio in Russo & Van Roy (2016) is prior-independent. To see this, using the probability matching property of the TS algorithm and Corollary 1 in Russo & Van Roy (2016),

$$
\begin{aligned}
(\mathbb{E}_t[\langle A^*, \theta \rangle - R_t])^2 &= \left(\sum_{a \in \mathcal{A}} \mathbb{P}_t(A^* = a) \Delta_t(a) \right)^2 \\
&\leq 2|\mathcal{A}| \sum_{a \in \mathcal{A}} \mathbb{P}_t(A^* = a)^2 \Delta_t^2(a) \\
&\leq |\mathcal{A}| \mathbb{I}_t(A^*; (A_t, R_t)),
\end{aligned}
\tag{4.1}
$$

where the first inequality follows from the Cauchy–Schwarz inequality and $\Delta_t(a) = \mathbb{E}_t[\langle a, \theta \rangle | A^* = a] - \mathbb{E}_t[\langle a, \theta \rangle]$. However, the use of Cauchy–Schwarz inequality over the whole action set $\mathcal{A}$ is agnostic to the distribution of A^* and thus illuminates the effect of a prior. As far as we know, all the existing upper bound analysis of information ratio (Tossou et al., 2017; Lattimore & Szepesvári, 2019; Hao et al., 2021; Hao & Lattimore, 2022) are prior-independent.

4.2. A Novel Regret Decomposition

We now introduce a novel information-theoretic regret decomposition such that the Bayesian regret bound can reflect the effect of the prior distribution. Our proof template is general and can be used even when a more general form of the expert policy (than in Eq. (2.1)) is used. Further, the proof technique can also be extended for other algorithms beyond TS, such as information-directed sampling (Russo & Van Roy, 2014).

Definition 4.1. Consider a random set $\mathcal{U} \subseteq \mathcal{A}$ which is measurable with respect to the offline dataset $\mathcal{D}_0$. For any $0 \leq \varepsilon \leq 1$, we call $\mathcal{U}$ as $(1 - \varepsilon)$-*informative* if it contains the optimal action A^* with probability at least $1 - \varepsilon$:

$$
\mathbb{P}(A^* \in \mathcal{U}) \geq 1 - \varepsilon. \tag{4.2}
$$

It is easy to see the full action set $\mathcal{A}$ is $(1 - \varepsilon)$-*informative* for any ε. The goal is to find a $\mathcal{U}$ whose expected cardinality is much smaller than $|\mathcal{A}|$.

Given such a $\mathcal{U}$, we can decompose the Bayesian regret based on whether A^* belongs to $\mathcal{U}$ or not:

$$
\begin{aligned}
\mathfrak{BR}_T(\pi^{\text{TS}}) = & \mathbb{E} \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a) \Delta_t(a) \right] \\
& + \mathbb{E} \left[\sum_{t=1}^{T} \sum_{a \notin \mathcal{U}} \mathbb{P}_t(A^* = a) \Delta_t(a) \right].
\end{aligned}
\tag{4.3}
$$

Both terms can be bounded in terms of the expected cardinality of $\mathcal{U}$ and ε as we show in the following lemmas.

Lemma 4.2. *Let $\mathcal{U}$ be an $(1 - \varepsilon)$-informative set defined in Eq. (4.2). Then, the following holds*

$$
\begin{aligned}
& \mathbb{E} \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a) \Delta_t(a) \right] \\
& \leq \sqrt{T \mathbb{E}[|\mathcal{U}|] (\log(\mathbb{E}[|\mathcal{U}|]) + \varepsilon \log(K/\varepsilon))}.
\end{aligned}
\tag{4.4}
$$

The proof is available in Appendix C and the key step is to use Cauchy–Schwarz inequality on a potentially much smaller set $\mathcal{U}$ rather than $\mathcal{A}$. Eq. (4.4) sheds light on how the regret upper bound depends on $\mathbb{E}[|\mathcal{U}|]$. In particular, $\mathbb{E}[|\mathcal{U}|]$ is the upper bound for the *information ratio* while $\log(\mathbb{E}[|\mathcal{U}|]) + \varepsilon \log(K/\varepsilon)$ is the upper bound for the *entropy*. When $|\mathcal{U}| \ll |\mathcal{A}|$, the upper bound on the information ratio is much smaller than the bound in Eq. (4.1).

Lemma 4.3. *Let $\mathcal{U}$ be an $(1 - \varepsilon)$-informative set defined in Eq. (4.2) and suppose the expected reward range $\mathbb{E}[\max a^\top \theta - \min a^\top \theta]$ is bounded by C_1. Then, the following holds*

$$
\mathbb{E} \left[\sum_{t=1}^{T} \sum_{a \notin \mathcal{U}} \mathbb{P}_t(A^* = a) \Delta_t(a) \right] \leq C_1 T \varepsilon.
$$

The proof is available in Appendix B. This is an additive term that captures the regret when the informative set fails to contain the optimal action. This term is always negligible since in most cases ε decays exponentially fast.

Combining Lemmas 4.2 and 4.3 together, we have the following theorem.

Theorem 4.4. *For any $(1 - \varepsilon)$-informative set $\mathcal{U}$, the Bayesian regret of TS algorithm can be upper bounded as*

$$
\begin{aligned}
& \mathfrak{BR}_T(\pi^{\text{TS}}) \\
& \leq \sqrt{T \mathbb{E}[|\mathcal{U}|] (\log(\mathbb{E}[|\mathcal{U}|]) + \varepsilon \log(K/\varepsilon))} + C_1 T \varepsilon.
\end{aligned}
$$

If such a $\mathcal{U}$ is given to the algorithm (not limited to TS) as a prior knowledge, we can easily achieve the bound in Theorem 4.4, e.g., running standard UCB on $\mathcal{U}$ directly. In contrast, TS does not need to know $\mathcal{U}$ and can adapt to different $\mathcal{U}$ automatically. Thus, for TS, the introduction of $\mathcal{U}$ is for *analysis only* rather than used by the algorithm.

Remark 4.5. *Of course, the Bayesian regret bound of TS cannot exceed $O(\sqrt{T|\mathcal{A}| \log(|\mathcal{A}|)})$ by using the standard prior-free analysis (Russo & Van Roy, 2016).*

Next, we use this result to bound the regret for the iTS algorithm for Gaussian bandits by finding such an informative set.

4.3. Prior-Dependent Regret Bound For Informed-TS

Consider a K-armed bandit and assume the prior distribution $\nu_0(\cdot) = \mathcal{N}(0, \boldsymbol{I}_K)$. We define $\mathcal{U}_A$ (that is the set $\mathcal{U}$ we

choose in Section 4.2) as a *set* that contains non-duplicated actions appearing in $\{\bar{A}_1, \ldots, \bar{A}_N\}$ at least once and $\mathcal{U}_A$ has at most K different actions.

Let us first denote $\alpha_1 = K \min\{\log(T\beta)/\beta, 1\}$ and $\alpha_2 = \exp(\beta\sqrt{2\log(TK)}/\lambda)$. We further denote

$$f_1 = \frac{3}{T} + \left(1 - \frac{1}{\alpha_2(1 + \alpha_1 + \frac{\log(T)}{\log(1/\alpha_1)} + K/(T\beta))}\right)^N,$$

$$f_2 = \min\left\{\alpha_1 + 1 + \alpha_2\frac{KN}{T\beta} + \frac{1}{T}, K\right\}.$$

Lemma 4.6. *If $K \geq \log_2(T)$, then the set $\mathcal{U}_A$ is $(1 - f_1)$-informative and $\mathbb{E}[|\mathcal{U}_A|] \leq f_2$.*

The proof can be found in Appendix C. Here, α_1 is the price for deliberateness and goes to 0 as the deliberateness level β increases. α_2 is the price the agent pays for the imperfect knowledge of the true environment θ and goes to 1 as the knowledgeability λ tends to infinity. f_1 is the probability that $\mathcal{U}_A$ fails to capture the optimal action and decays exponentially fast as the data size N increases. Note that $K \geq \log_2(T)$ is a technical condition that we hope to relax in the future.

Since $\theta \sim \mathcal{N}(0, \boldsymbol{I}_K)$, we have

$$\mathbb{E}\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right] \leq 2\sqrt{2\log(K)}, \quad (4.5)$$

where the proof of this claim is available in Appendix E. Combining Theorem 4.4, Lemma 4.6 and Eq. (4.5) together, we obtain the final regret bound for the informed TS algorithm, the main theoretical result of this paper.

Theorem 4.7. *The Bayesian regret of the iTS algorithm is bounded as*

$$\mathfrak{BR}_T(\pi^{i\text{-TS}}) \leq \underbrace{2\sqrt{2\log(K)}Tf_1 + 4\sqrt{2\log(K)}}_{\text{remainder term}} + \underbrace{\sqrt{Tf_2\left(\log(f_2) + f_1\log\left(K/f_1\right)\right)}}_{\text{main term}}. \quad (4.6)$$

For the main term, as the deliberateness β increases, the information ratio part (f_2) first drops polynomially until 1 and then the entropy part $(\log(f_2) + f_1\log(K/f_1))$ drops further until $\log(1) = 0$. This implies we must sharpen both the information ratio term and entropy term. Overall, as the competence parameters β, λ of an expert go to infinity, the main term of the Bayesian regret bound goes to 0. Thus, our regret bound Eq. (4.6) can fully characterizes the role of competence.

5. Bayesian Bootstrapping

The posterior update in Eq. (3.1) is computationally challenging due to the loss of conjugacy in the $\mathbb{P}(\bar{A}_n|\theta)$ term

while using the Bayes' rule, which has a sum of exponentials term in the denominator. Thus, we adapt the existing approximate-TS approach based on *Bayesian bootstrapping* (Osband et al., 2019; Lu & Van Roy, 2017). The key idea is to perturb the loss function for the maximum a posterior (MAP) estimate and use the point estimate as a surrogate for the exact posterior sample.

5.1. The Loss Function and Perturbation

We now introduce a loss function whose optimization we show yields the MAP estimate of the model parameters. Suppose $\nu_0(\cdot) = \mathcal{N}(0, \Sigma_0)$. With a bit of notational ambiguity, we view θ and ϑ as parameters rather than random variables in this subsection. We first derive the loss function for the MAP estimate in Lemma 5.1.

Lemma 5.1. *At time t, the MAP estimate for θ, ϑ is equivalent to solving the following optimization problem:*

$$\underset{\theta,\vartheta}{\arg\min}\, \mathcal{L}_1(\theta, \vartheta) + \mathcal{L}_2(\theta, \vartheta) + \mathcal{L}_3(\theta, \vartheta),$$

where $\mathcal{L}_1(\theta, \vartheta) :=$

$$-2\sum_{n=1}^{N}\left(\beta\vartheta^T\bar{A}_n - \log\left(\sum_{b\in\mathcal{A}}\exp\left(\beta\vartheta^T b\right)\right)\right),$$

is the negative log-likelihood contributed by the offline actions and $\mathcal{L}_2(\theta, \vartheta) :=$

$$\sum_{n=1}^{N}\left(\bar{R}_n - \theta^T\bar{A}_n\right)^2 + \sum_{\tau=1}^{t}\left(R_\tau - \theta^T A_\tau\right)^2,$$

is the negative log-likelihood contributed by the rewards and $\mathcal{L}_3(\theta, \vartheta) :=$

$$\lambda^2\|\vartheta - \theta\|_2^2 + \theta^\top\Sigma_0^{-1}\theta,$$

is the log-prior function.

The proof is available in Appendix D. Note that $\mathcal{L}_1(\theta, \vartheta)$ serves a role similar to the imitation learning loss (Ross et al., 2011) since it regularizes the online agent to follow the expert's action and the amount of regularization is guided by the competence level. While there are multiple heuristic choices for the imitation learning loss in literature, ours are derived in a principled way following Bayes' rule and combined with online learning.

Remark 5.2. *When λ is small, the competence level of the expert is low such that the demonstration data is of low-quality. In this case, the online learning agent should not be following the offline action. Fortunately, informed TS can understand this automatically through Bayes' rule. In particular, small λ imposes little regularization on $\|\vartheta - \theta\|_2$ such that estimating θ is independent of the action likelihood.*

Now, we will perturb the loss function. As is standard in the literature (Lu & Van Roy, 2017; Osband et al., 2018; Dwaracherla et al., 2022; Qin et al., 2022), the Gaussian rewards are perturbed by additive Gaussian noise while the offline actions are perturbed by multiplicative random weights. Moreover, the log-prior terms are perturbed by random samples from the prior distribution. Therefore, we denote a set of perturbations as follows and resample all the perturbations at each time:

- *Action perturbation.* Let $w_n \overset{\text{i.i.d}}{\sim} \exp(1)$ be a sequence of Bayesian bootstrap weights. The corresponding perturbed loss function is $\widetilde{\mathcal{L}}_1(\theta, \vartheta) :=$

$$-2 \sum_{n=1}^{N} w_n \left(\beta \vartheta^T \bar{A}_n - \log \left(\sum_{b \in \mathcal{A}} \exp \left(\beta \vartheta^T b \right) \right) \right).$$

- *Reward perturbation.* Let $\xi_n^1, \xi_\tau^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$. The corresponding perturbed loss function is $\widetilde{\mathcal{L}}_2(\theta, \vartheta) :=$

$$\sum_{n=1}^{N} \left(\bar{R}_n + \xi_n^1 - \theta^T \bar{A}_n \right)^2 + \sum_{\tau=1}^{t} \left(R_\tau + \xi_\tau^2 - \theta^T A_\tau \right)^2.$$

- *Prior function perturbation.* Let $\widetilde{\theta}_0 \sim \mathcal{N}(0, \Sigma_0)$ and $\widetilde{\vartheta} \sim \mathcal{N}(0, \boldsymbol{I}_d / \lambda)$. The corresponding perturbed loss function is $\widetilde{\mathcal{L}}_3(\theta, \vartheta) :=$

$$\lambda^2 \| \vartheta - \theta - \widetilde{\vartheta} \|_2^2 + (\theta - \widetilde{\theta}_0)^\top \Sigma_0^{-1} (\theta - \widetilde{\theta}_0).$$

At each time t, let $(\widehat{\theta}, \widehat{\vartheta})$ be the solution of

$$\min_{\theta, \vartheta} \widetilde{\mathcal{L}}_1(\theta, \vartheta) + \widetilde{\mathcal{L}}_2(\theta, \vartheta) + \widetilde{\mathcal{L}}_3(\theta, \vartheta), \qquad (5.1)$$

and use $\widehat{\theta}$ as a surrogate for posterior sampling in the standard TS algorithm. Since the perturbed loss function in Eq. (5.1) is convex, we can solve it by use of standard convex optimization solvers such as CVXPY (Diamond & Boyd, 2016).

Remark 5.3. *1. We would like to mention perturbation-based methods such as Bayesian bootstrapping lead to one kind of approximate-TS algorithms. There are other choices such as MCMC, Laplace approximation or variational inference for sampling from an approximate posterior. For a detailed empirical comparison, we refer the reader to Osband et al. (2022).*

2. While the algorithm as presented above assumes the offline dataset comes from a single expert, it can easily be extended where it comes from multiple experts with different competence parameters $(\beta_j, \lambda_j), j = 1, \cdots, J$, with corresponding (ϑ_j) parameters, and dataset sizes N_j. Namely, there will be J similar terms in the loss function $\widetilde{\mathcal{L}}_1$, one for each expert. Similarly, the first term in the loss function $\widetilde{\mathcal{L}}_3$ will be replaced by J identical terms, one for each expert.

5.2. Estimating Competence Level

Bayesian bootstrapping requires an input for the competence level. In practice, this is often not available and can be estimated only from the offline data. We provide two methods to estimate the deliberateness parameter, β:

1) The first method is based on maximum likelihood estimation (MLE). Similar idea has been proposed to estimate the expertise level in imitation learning (Beliaev et al., 2022). Specifically, we optimize β over the following negative log-likelihood of the offline data:

$$-\sum_{n=1}^{N} \left(\beta \bar{A}_n^\top \widehat{\vartheta}^{\text{LS}} - \log \left(\sum_{b \in \mathcal{A}} \exp \left(\beta b^\top \widehat{\vartheta}^{\text{LS}} \right) \right) \right),$$

where $\widehat{\vartheta}^{\text{LS}}$ is the regularized least square estimate using $\mathcal{D}_0$.

2) The second method is to simply look at the entropy of the empirical distribution of the action in the offline dataset. Suppose the empirical distribution of $\{\bar{A}_n\}_{n=1}^{N}$ is μ_A. Then we use $c_0 / \mathcal{H}(\mu_A)$ as an estimation for β, where $c_0 > 0$ is a hyperparameter. The intuition is that for smaller β, the offline actions tend to be more uniform and thus the entropy will be large. This is an unsupervised approach and agnostic to specific offline data generation process.

The knowledgeability λ is not quite 'estimable' because for a single environment, even though we know the true environment θ and the expert's knowledge ϑ, we only have one pair of observations. Thus, the variance of the estimation for λ could be infinite. However, exact estimation of λ is not often necessary and we show that our algorithm is robust to misspecified λ through experiments in Section 6.2.

We summarize the full (Bayesian bootstrapped) approximate iTS algorithm in Algorithm 1.

Algorithm 1 Approximate iTS

1: **Input:** time horizon T, action set $\mathcal{A}$, parameter λ_0, offline demonstration data $\mathcal{D}_0$;
2: Obtain $\widehat{\beta}$ through either MLE or entropy method described in Section 5.2.
3: **for** $t = 1, \ldots, T$ **do**
4:     Sample a set of perturbations $\{w_n, \xi_n^1, \xi_\tau^2, \widetilde{\theta}_0, \widetilde{\vartheta}\}$ according to Section 5.1.
5:     Solve Eq. (5.1) with competence level $(\widehat{\beta}, \lambda_0)$ and denote the solution as $(\widehat{\theta}_t, \widehat{\vartheta}_t)$.
6:     Take action $A_t = \text{argmax}_{a \in \mathcal{A}} a^\top \widehat{\theta}_t$ and receive R_t.
7: **end for**

---

6. Empirical Results

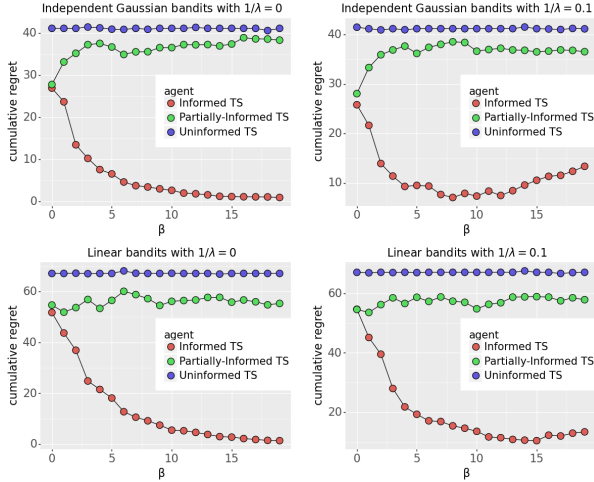We empirically investigate the role of offline demonstration data in terms of regret reduction. We compare the (approxi-

Figure 1. The role of deliberateness β with different knowledge-ability λ.

mate) *informed* TS algorithm with two baseline algorithms:

○ *Uninformed* TS: An algorithm that uses the standard linear Gaussian TS (Russo et al., 2018) and does not use the offline demonstration data $\mathcal{D}_0$;

○ *Partially-informed* TS: An algorithm that uses the offline demonstration data $\mathcal{D}_0$ to update the initial prior but assumes the expert is *naive*, i.e., $\beta = 0$ and hence the action likelihood term $\mathbb{P}(\bar{A}_n | \theta \in \cdot)$ is a constant, e.g.,

$$\mathbb{P}(\theta \in \cdot | \mathcal{D}_0) \propto \nu_0(\cdot) \prod_{n=1}^{N} \mathbb{P}(\bar{R}_n | \bar{A}_n, \theta \in \cdot).$$

Note that this algorithm can use conjugacy in updating the posterior distribution, and hence is computationally simpler.

The environment draws a model θ from the prior distribution $\mathcal{N}(0, \boldsymbol{I}_d)$ and provides a noisy $\vartheta \sim \mathcal{N}(\theta, \boldsymbol{I}_d/\lambda^2)$ to the expert. The expert then generates demonstration data following Eq. (2.1) to obtain a dataset of size N. Each algorithm is run for a horizon $T = 1000$ and we compute the average cumulative regret over 100 independent runs for each algorithm.

6.1. Role of Competence

We first demonstrate the role of the expert's competence level (deliberatness β and knowledability λ). Consider a Gaussian bandit with $K = 5$ independent arms and a linear Gaussian bandit ($K = 20, d = 5$) whose actions are sampled from a d-dimensional unit sphere. The offline demonstration datasize is fixed at $N = 10$. Consider two scenarios: (i) Fix $1/\lambda$ and vary β; and (ii) Fix β and vary $1/\lambda$. The results are shown in Figures 1 and 2.

In Figure 1, for both values of $1/\lambda = 0$ and 0.1, partially-

informed TS only has a marginal regret reduction and its performance is nearly independent of the quality of the offline data. In contrast, informed TS enjoys a significant regret reduction that varies as β, the deliberateness level of the expert increases. In Figure 2, we see that as λ, the knowledgeability of the expert decreases, the amount of regret reduction of informed TS decreases as well. Those empirical results support our main argument that the amount of regret reduction achieved by Algorithm 1 by use of offline data depends on the *quality* of demonstrations.
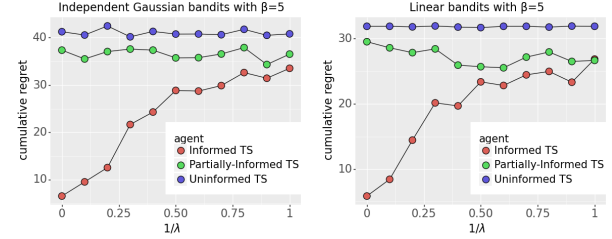

Figure 2. The role of knowledgeability λ with deliberateness $\beta = 5$.

6.2. Robustness to Model Misspecifications

Although the loss function in Lemma 5.1 is derived from a softmax expert policy (Eq. (2.1)), we would like to test if the proposed algorithm is robust to the following types of model misspecifications:

1) *Expert policy used to genetrate offline data.* While the algorithm assumes Eq. (2.1) as its generative model, we generate the dataset by a different ϵ-greedy policy introduced in Remark 2.2: for $\beta \in [0, 1]$,

$$\phi_{\beta,\lambda}(\bar{A}_n = a | \vartheta) = \beta \mathbb{I}\left(a = \operatorname*{argmax}_{a \in [K]} a^\top \vartheta\right) + (1-\beta)/K,$$

where the expert's knowledge still takes the form of a vector $\vartheta \sim \mathcal{N}(\theta, \boldsymbol{I}_d/\lambda^2)$ conditioned on θ. From Figure 3, we can see that although the parametric form of the expert policy is misspecified, the informed TS algorithm still significantly outperforms the two baseline algorithms.

2) *Misspecified competence level.* First, we generate the offline data with the true knowledgeability parameter $\lambda = 0.1$ but the algorithm uses a misspecified λ ranging from 0 to 1. Second, we generate the offline data with the true deliberateness parameter $\beta = 10$ but the algorithm uses a misspecified β ranging from 1 to 20. Figure 4 shows that although the performance of informed TS decreases as the degree of misspecification increases, our algorithm still significantly outperforms the two baseline algorithms.

Figure 3. Misspecified expert policy for linear bandits.
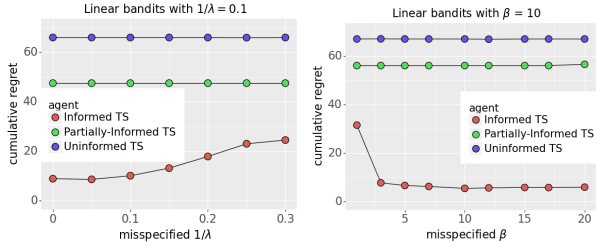
Figure 4. Misspecified λ and β.

6.3. Unknown Competence Level

We evaluate the empirical performance of the competence level estimation methods introduced in Section 5.2. We still consider a linear Gaussian bandit with $1/\lambda = 0$. We compare entropy-based method and MLE based method for informed TS with two baselines: one is to plug-in true β for informed TS; another one is the uninformed TS. As shown in Figure 5, although the performance degrades when we estimate β, our methods are still significantly outperforming the uninformed TS baseline. MLE does not perform well for a large β since it will suffer from unbalanced offline data when computing the regularised least square estimate.
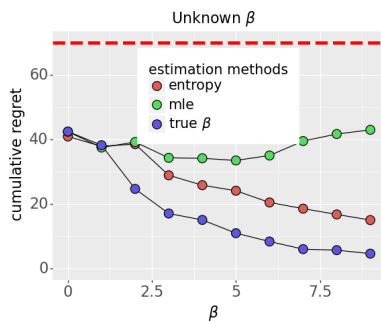
Figure 5. Comparing different methods for estimating β. True β means informed TS with known β. The red horizontal dashed line represents uninformed TS baseline.

6.4. Evaluating the Approximation Quality

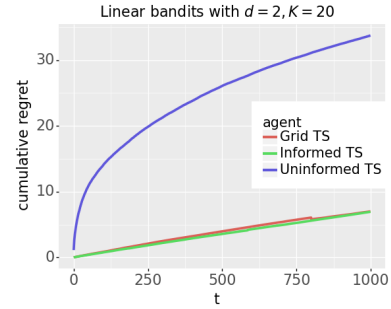Bayesian boostrapping is introduced because it can be computationally challenging to obtain samples from the exact

Figure 6. Evaluating the approximation quality using grid TS.

posterior distribution. That immediately raises the question about how good such an approximation is. To evaluate this, we consider the $d = 2$ case where we can compute a nearly exact posterior distribution by a brute-force method that relies on discretization of the parameter space. We call the algorithm as *Grid-TS*. The results are shown in Figure 6. We see that the regret of iTS is very close to that of Grid-TS (which works with a near exact posterior distribution). Thus the approximation quality of the Bayesian bootstrapping method is reasonably good.

7. Conclusions and Future Work

In this paper, we have investigated how offline demonstrations can be used to improve online learning performance. It is natural to expect that use of offline dataset will result in better online learning performance. *The question is how, and by how much?* Our experimental work shows that if the offline dataset is used in a naive manner, it results in only a marginal regret reduction. However, when the dataset is used by the learning agent in a more "informed manner", i.e., it accounts for varying levels of an expert's competence, a notion we introduce, the reduction in cumulative regret can be substantial: the higher the quality of the demonstration, the more the reduction.

The goal of this paper is to yield insights on how to use the offline data to improve online learning. To that end, we have used a prototypical model of online learning - finite multi-armed bandits, and assumed a simple and mathematically convenient generative model for the expert policy. However, we note that the potential to generalize to more sophisticated settings and models exists. This includes linear bandits, MDPs, and more general generative models of expert policies. We will explore such directions in future work, and hope that it will inspire other researchers to explore further development of this very interesting and new research direction.

References

Banerjee, S., Sinclair, S. R., Tambe, M., Xu, L., and Yu, C. L. Artificial replay: A meta-algorithm for harnessing historical data in bandits. *arXiv*, pp. 2210.00025, 2022.

Beliaev, M., Shih, A., Ermon, S., Sadigh, D., and Pedarsani, R. Imitation learning by estimating expertise of demonstrators. *Proceedings of the 39th International Conference on Machine Learning*, 162:1732–1748, 2022.

Bubeck, S. and Liu, C.-Y. Prior-free and prior-dependent regret bounds for thompson sampling. *Advances in neural information processing systems*, 26, 2013.

Cutkosky, A., Dann, C., Das, A., and Zhang, Q. Leveraging initial hints for free in stochastic linear bandits. In *International Conference on Algorithmic Learning Theory*, pp. 282–318. PMLR, 2022.

Degenne, R., Garcelon, E., and Perchet, V. Bandits with side observations: Bounded vs. logarithmic regret. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dwaracherla, V., Wen, Z., Osband, I., Lu, X., Asghari, S. M., and Van Roy, B. Ensembles for uncertainty estimation: Benefits of prior functions and bootstrapping. *arXiv preprint arXiv:2206.03633*, 2022.

Hao, B. and Lattimore, T. Regret bounds for information-directed reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=1pHC-yZfaTK.

Hao, B., Lattimore, T., and Deng, W. Information directed sampling for sparse linear bandits. *Advances in Neural Information Processing Systems*, 34:16738–16750, 2021.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-w., Mladenov, M., Boutilier, C., and Szepesvari, C. Meta-thompson sampling. In *International Conference on Machine Learning*, pp. 5884–5893. PMLR, 2021.

Lattimore, T. and Szepesvári, C. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pp. 2111–2139. PMLR, 2019.

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

Lu, X. and Van Roy, B. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Osband, I., Aslanides, J., and Cassirer, A. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Osband, I., Van Roy, B., Russo, D. J., Wen, Z., et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

Osband, I., Wen, Z., Asghari, M., Ibrahimi, M., Lu, X., and Van Roy, B. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2021.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Lu, X., Ibrahimi, M., Lawson, D., Hao, B., O'Donoghue, B., and Roy, B. V. The neural testbed: Evaluating joint predictions. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=JyTT03dqCFD.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Qin, C., Wen, Z., Lu, X., and Van Roy, B. An analysis of ensemble sampling. *arXiv preprint arXiv:2203.01303*, 2022.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

Russo, D. and Van Roy, B. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

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

Shivaswamy, P. and Joachims, T. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pp. 1046–1054. PMLR, 2012.

Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D. J., Lykouris, T., Dudik, M., and Schapire, R. E. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34:26382–26394, 2021.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Tossou, A. C., Dimitrakakis, C., and Dubhashi, D. Thompson sampling for stochastic bandits with graph feedback. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wagenmaker, A. and Pacchiano, A. Leveraging offline data in online reinforcement learning. *arXiv preprint arXiv:2211.04974*, 2022.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Zhang, C., Agarwal, A., Daumé III, H., Langford, J., and Negahban, S. N. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301*, 2019.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Proof of Lemma 4.2

We first define an information ratio with respect to a set $\mathcal{U}$:

$$\Gamma_t(\mathcal{U}) = \frac{\left(\sum_{a \in \mathcal{U}} \mathbb{P}_t\left(A^* = a\right) \Delta_t(a)\right)^2}{\mathbb{I}_t(A^*; (A_t, R_t))} \, .$$

Applying Cauchy–Schwarz inequality, we have

$$\mathbb{E}\left[\sum_{t=1}^T \sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a)\Delta_t(a)\right] = \mathbb{E}\left[\sum_{t=1}^T \frac{\sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a)\Delta_t(a)}{\sqrt{\mathbb{I}_t(A^*; (A_t, R_t))}} \sqrt{\mathbb{I}_t(A^*; (A_t, R_t))}\right]$$

$$\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^T \frac{\left(\sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a)\Delta_t(a)\right)^2}{\mathbb{I}_t(A^*; (A_t, R_t))}\right] \sum_{t=1}^T \mathbb{E}\left[\mathbb{I}_t(A^*; (A_t, R_t))\right]} \, .$$

Using the chain rule of mutual information,

$$\sum_{t=1}^T \mathbb{E}\left[\mathbb{I}_t(A^*; (A_t, R_t))\right] = \mathbb{I}(A^*; \mathcal{D}_T) \leq \mathcal{H}(A^*) \, .$$

Then we have

$$\mathbb{E}\left[\sum_{t=1}^T \sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a)\Delta_t(a)\right] \leq \sqrt{\sum_{t=1}^T \mathbb{E}[\Gamma_t(\mathcal{U})]\mathcal{H}(A^*)} \, . \tag{A.1}$$

For the information ratio $\mathbb{E}[\Gamma_t(\mathcal{U})]$, applying Cauchy–Schwarz inequality over the set $\mathcal{U}$ and following the step in Eq. (4.1),

$$\left(\sum_{a \in \mathcal{U}} \mathbb{P}_t(A^* = a)\Delta_t(a)\right)^2 \leq |\mathcal{U}| \frac{\mathbb{I}_t(A^*; (A_t, R_t))}{2} \, ,$$

which implies $\mathbb{E}[\Gamma_t(\mathcal{U})] \leq \mathbb{E}[|\mathcal{U}|]$ for any $t \in [T]$. For the entropy part $\mathcal{H}(A^*)$, by the definition of Shannon entropy,

$$\mathcal{H}(A^*) = \mathbb{E}\left[\mathcal{H}(\mathbb{P}(A^* \in \cdot | \mathcal{U}))\right]$$
$$= \mathbb{E}\left[-\sum_{a \in \mathcal{U}} \mathbb{P}(A^* = a) \log(\mathbb{P}(A^* = a)) - \sum_{a \notin \mathcal{U}} \mathbb{P}(A^* = a) \log(\mathbb{P}(A^* = a))\right] \, . \tag{A.2}$$

- The first term can be bounded by the uniform bound on the entropy of a probability distribution:

$$\mathbb{E}\left[-\sum_{a \in \mathcal{U}} \mathbb{P}(A^* = a) \log(\mathbb{P}(A^* = a))\right] \leq \mathbb{E}\left[\log(|\mathcal{U}|)\right] \leq \log\left(\mathbb{E}[|\mathcal{U}|]\right) \, . \tag{A.3}$$

where the second inequality uses Jenson's inequality.

- For the second term, we use the following fact and the definition of $\mathcal{U}$ in Eq. (4.2):

$$\mathbb{E}\left[-\sum_{a \notin \mathcal{U}} \mathbb{P}(A^* = a) \log(\mathbb{P}(A^* = a))\right]$$
$$= \mathbb{E}\left[\varepsilon \sum_{a \notin \mathcal{U}} \frac{\mathbb{P}(A^* = a)}{\varepsilon} \log\left(\frac{\varepsilon}{\mathbb{P}(A^* = a)}\right)\right] - \mathbb{E}\left[\sum_{a \notin \mathcal{U}} \mathbb{P}(A^* = a) \log(\varepsilon)\right]$$
$$\leq \mathbb{E}\left[\varepsilon \log(K - |\mathcal{U}|)\right] + \mathbb{E}\left[\sum_{a \notin \mathcal{U}} \mathbb{P}(A^* = a)\right] \log(1/\varepsilon) \tag{A.4}$$
$$\leq \varepsilon \log(K) + \varepsilon \log(1/\varepsilon) = \varepsilon \log(K/\varepsilon) \, .$$

11

Putting Eqs. (A.2)-(A.4) together,

$$\mathcal{H}(A^*) \leq \log\left(\mathbb{E}[|\mathcal{U}|]\right) + \varepsilon \log\left(K/\varepsilon\right) .$$

Plugging the above bounds into Eq. (A.1),

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{U}}\mathbb{P}_t(A^* = a)\Delta_t(a)\right] \leq \sqrt{T\mathbb{E}[|\mathcal{U}|]\left(\log\left(\mathbb{E}[|\mathcal{U}|]\right) + \varepsilon \log\left(K/\varepsilon\right)\right)}.$$

We now reach the conclusion.

B. Proof of Lemma 4.3

By the definition of $\Delta_t(a)$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\notin\mathcal{U}}\mathbb{P}_t\left(A^* = a\right)\Delta_t(a)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\notin\mathcal{U}}\mathbb{P}_t\left(A^* = a\right)\left(\mathbb{E}_t[\langle a, \theta\rangle | A^* = a] - \mathbb{E}_t[\langle a, \theta\rangle]\right)\right] .$$

First, we note that

$$\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\mathbb{E}_t\left[\langle a, \theta\rangle | A^* = a\right] \leq \sum_{a\in\mathcal{A}}\mathbb{P}_t(A^* = a)\mathbb{E}_t\left[\langle a, \theta\rangle | A^* = a\right]$$

$$= \mathbb{E}_t\left[\langle A^*, \theta\rangle\right] = \mathbb{E}_t\left[\max_a\langle a, \theta\rangle\right] .$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\notin\mathcal{U}}\mathbb{P}_t\left(A^* = a\right)\Delta_t(a)\right] \leq \sum_{t=1}^{T}\mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\mathbb{E}_t\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]\right] .$$

By Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\mathbb{E}_t\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]\right] \leq \sqrt{\mathbb{E}\left[\left(\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\right)^2\right]\mathbb{E}\left[\left(\mathbb{E}_t\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]\right)^2\right]}$$

$$\leq \sqrt{\left(\mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\right]\right)^2\left(\mathbb{E}\left[\mathbb{E}_t\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]\right]\right)^2}$$

$$= \mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\right]\mathbb{E}\left[\mathbb{E}_t\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]\right]$$

where we use Jenson's inequality for the second inequality. According to the tower property of conditional expectation, we have

$$\mathbb{E}\left[\mathbb{E}_t\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]\right] = \mathbb{E}\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right] ,$$

and

$$\mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{P}_t(A^* = a)\right] = \mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{E}\left[\mathbb{I}\left(A^* = a\right)|\mathcal{D}_t\right]\right]$$

$$= \mathbb{E}\left[\sum_{a\notin\mathcal{U}}\mathbb{I}\left(A^* = a\right)\right] = \mathbb{P}(A^* \notin \mathcal{U}).$$

When $\mathbb{E}\left[\max_a\langle a, \theta\rangle - \min_a\langle a, \theta\rangle\right]$ is bounded by C_1, we reach the conclusion.

C. Proof of Lemma 4.6

We split the proof of Lemma 4.6 into two parts: prove $\mathcal{U}_A$ is $(1 - f_1)$-informative and prove $\mathbb{E}[|\mathcal{U}_A|] \leq f_2$.

C.1. Prove $\mathcal{U}_A$ is $(1 - f_1)$-informative

Based on the assumption of the offline data generating process, we know that conditioned on ϑ, $\bar{A}_n$ is independent of $\bar{A}_{n'}$ for any $n \neq n'$. This implies

$$\mathbb{P}\left(A^* \notin \mathcal{U}_A\right) = \mathbb{P}\left(\bar{A}_n \neq A^*, \forall n \in [N]\right) = \mathbb{P}\left(\bigcap_{n=1}^{N}\left\{\bar{A}_n \neq A^*\right\}\right)$$

$$= \mathbb{E}\left[\mathbb{P}\left(\bigcap_{n=1}^{N}\left\{\bar{A}_n \neq A^*\right\}\Big|\theta, \vartheta\right)\right] = \mathbb{E}\left[\prod_{n=1}^{N}\mathbb{P}(\bar{A}_n \neq A^*|\theta, \vartheta)\right] \tag{C.1}$$

$$= \mathbb{E}\left[\prod_{n=1}^{N}\left(1 - \mathbb{P}\left(\bar{A}_n = A^*\big|\theta, \vartheta\right)\right)\right],$$

where A^* is a function of θ and thus a random variable as well. According to the definition of the softmax expert policy in Eq. (2.1),

$$\mathbb{P}(\bar{A}_n = A^*|\theta, \vartheta) = \frac{\exp(\beta\vartheta^\top A^*)}{\sum_{b\in\mathcal{A}}\exp(\beta\vartheta^\top b)} = \frac{1}{\sum_{b\in\mathcal{A}}\exp(-\beta\langle A^* - b, \vartheta\rangle)}$$

$$= \left(\sum_{b\in\mathcal{A}}\exp\left(\beta\langle A^* - b, \theta - \vartheta\rangle - \beta\langle A^* - b, \theta\rangle\right)\right)^{-1}$$

$$\geq \left(\sum_{b\in\mathcal{A}}\exp\left(\beta\|A^* - b\|_1\|\vartheta - \theta\|_\infty - \beta\langle A^* - b, \theta\rangle\right)\right)^{-1}.$$

For multi-armed bandits, $\|A^* - b\|_1 \leq 1$ almost surely for any $b \in \mathcal{A}$ so

$$\mathbb{P}(\bar{A}_n = A^*|\theta, \vartheta) \geq \left(\sum_{b\in\mathcal{A}}\exp\left(\beta\|\vartheta - \theta\|_\infty - \beta\langle A^* - b, \theta\rangle\right)\right)^{-1}.$$

Since $\vartheta - \theta \sim N(0, \boldsymbol{I}_K/\lambda^2)$, using standard Hoeffding's bound (Vershynin, 2010) implies

$$\mathbb{P}\left(\|\vartheta - \theta\|_\infty \geq t\right) \leq K\exp\left(-\frac{t^2\lambda^2}{2}\right).$$

Set $t = \sqrt{2\log(TK)}/\lambda$ and define an event $\mathcal{E}_1 := \{\|\vartheta - \theta\|_\infty \leq \sqrt{2\log(TK)}/\lambda\}$ such that $\mathbb{P}(\mathcal{E}_1^c) \leq 1/T$. We decompose Eq. (C.1) according to $\mathcal{E}_1$:

$$\mathbb{P}\left(A^* \notin \mathcal{U}_A\right) \leq \mathbb{E}\left[\prod_{n=1}^{N}\left(1 - \mathbb{P}\left(\bar{A}_n = A^*\big|\theta, \vartheta\right)\right)\mathbb{I}(\mathcal{E}_1)\right] + \mathbb{P}(\mathcal{E}_1^c)$$

$$\leq \mathbb{E}\left[\prod_{n=1}^{N}\left(1 - \left(\exp\left(\frac{\beta\sqrt{2\log(TK)}}{\lambda}\right)\sum_{b\in\mathcal{A}}\exp\left(-\beta\langle A^* - b, \theta\rangle\right)\right)^{-1}\right)\right] + \frac{1}{T}. \tag{C.2}$$

Let us define a set

$$\mathcal{B} = \{a : \langle A^* - a, \theta\rangle \leq \Delta\}, \tag{C.3}$$

where Δ will be chosen later. Then we can further decompose Eq. (C.2) as

$$\mathbb{P}\left(A^* \notin \mathcal{U}_A\right)$$

$$\leq \mathbb{E}\left[\prod_{n=1}^{N}\left(1 - \left(\exp\left(\frac{\beta\sqrt{2\log(TK)}}{\lambda}\right)\left(\sum_{b\in\mathcal{B}}\exp\left(-\beta\langle A^* - b, \theta\rangle\right) + \sum_{b\notin\mathcal{B}}\exp\left(-\beta\langle A^* - b, \theta\rangle\right)\right)\right)^{-1}\right)\right] + \frac{1}{T}.$$

For $b \notin \mathcal{B}$, we have $\exp(-\beta\langle A^* - b, \theta\rangle) \leq \exp(-\beta\Delta)$. For $b \in \mathcal{B}$, we have $\exp(-\beta\langle A^* - b, \theta\rangle) \leq \exp(-\beta 0) = 1$. Putting the above together, we can have

$$\sum_{b \in \mathcal{B}} \exp(-\beta\langle A^* - b, \theta\rangle) + \sum_{b \notin \mathcal{B}} \exp(-\beta\langle A^* - b, \theta\rangle) \leq |\mathcal{B}| + (K - |\mathcal{B}|)\exp(-\beta\Delta) \leq |\mathcal{B}| + K\exp(-\beta\Delta), \quad \text{(C.4)}$$

where $|\mathcal{B}| = \sum_b \mathbb{I}(\langle A^* - b, \theta\rangle \leq \Delta)$ is a random variable. Putting Eqs. (C.2)-(C.4) together implies

$$\mathbb{P}\left(A^* \notin \mathcal{U}_A\right) \leq \mathbb{E}\left[\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{|\mathcal{B}| + K\exp(-\beta\Delta)}\right)^N\right] + \frac{1}{T}$$

$$= \sum_{k=0}^{K-1} \mathbb{P}\left(|\mathcal{B}| - 1 = k\right)\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{k + K\exp(-\beta\Delta)}\right)^N + \frac{1}{T} \quad \text{(C.5)}$$

$$= \sum_{k=0}^{K-1} \sum_{a \in \mathcal{A}} \mathbb{P}\left(|\mathcal{B}| - 1 = k | A^* = a\right)\mathbb{P}\left(A^* = a\right)\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{k + K\exp(-\beta\Delta)}\right)^N + \frac{1}{T},$$

where we use the change-of-variables formula for the push-forward measure for the first equation.

Next we study the distribution of $|\mathcal{B}| - 1$ conditional on A^*. Without loss of generality, we first consider the conditional distribution with conditioning on $A^* = a_1$:

$$\mathbb{P}\left(|\mathcal{B}| - 1 = k | A^* = a_1\right) = \mathbb{P}\left(\sum_{a \in \mathcal{A}} \mathbb{I}(\langle A^* - a, \theta\rangle \leq \Delta) - 1 = k \Big| A^* = a_1\right)$$

$$= \frac{1}{\mathbb{P}(A^* = a_1)}\mathbb{P}\left(\sum_{a \in \mathcal{A}} \mathbb{I}(\langle A^* - a, \theta\rangle \leq \Delta) - 1 = k, A^* = a_1\right)$$

$$= \frac{1}{\mathbb{P}(A^* = a_1)}\mathbb{P}\left(\sum_{a \in \mathcal{A}} \mathbb{I}(\theta_a \geq \theta_1 - \Delta) - 1 = k, \bigcap_{a \in \mathcal{A}}\{\theta_1 \geq \theta_a\}\right) \quad \text{(C.6)}$$

$$= \frac{1}{\mathbb{P}(A^* = a_1)}\int_R \binom{K-1}{k}\left[\int_{\theta_1 - \Delta}^{\theta_1} d\rho(\theta)\right]^k \left[\int_{-\infty}^{\theta_1 - \Delta} d\rho(\theta)\right]^{K-1-k} d\rho(\theta_1),$$

where $\rho(\cdot)$ is the univariate Gaussian distribution and $\theta_a = a^\top\theta$ for any $a \in \mathcal{A}$. By defining

$$F(\theta_1) = \int_{-\infty}^{\theta_1} (2\pi)^{-1/2}\exp(-x^2/2)\,dx\,,$$

we have

$$\int_{\theta_1 - \Delta}^{\theta_1} \frac{d\rho(\theta)}{F(\theta_1)} + \int_{\infty}^{\theta_1 - \Delta} \frac{d\rho(\theta)}{F(\theta_1)} = 1\,.$$

We further define

$$q(\theta_1) = \int_{\theta_1 - \Delta}^{\theta_1} \frac{d\rho(\theta)}{F(\theta_1)}\,. \quad \text{(C.7)}$$

For a fixed θ_1, let $X_{\theta_1} \sim \text{Binomial}(K - 1, q(\theta_1))$. Together with Eq. (C.6),

$$\mathbb{P}\left(|\mathcal{B}| - 1 = k | A^* = a_1\right) = \int_R \mathbb{P}(X_{\theta_1} = k)\frac{F(\theta_1)^{K-1}}{\mathbb{P}(A^* = a_1)}\,d\rho(\theta_1) = \int_R \mathbb{P}(X_{\theta_1} = k)\,d\mu(\theta_1), \quad \text{(C.8)}$$

where we denote

$$d\mu(\theta_1) = \frac{F(\theta_1)^{K-1}}{\mathbb{P}(A^* = a_1)}\,d\rho(\theta_1)\,.$$

Thus, $|\mathcal{B}| - 1$ follows a mixture of binomial distribution. Plugging Eq. (C.8) into Eq. (C.5),

$$\mathbb{P}\left(A^* \notin \mathcal{U}_A\right) \leq \sum_{k=0}^{K-1} \sum_{a \in \mathcal{A}} \int_R \mathbb{P}(X_{\theta_a} = k) \, \mathrm{d}\mu(\theta_a) \mathbb{P}\left(A^* = a\right) \left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{k + K\exp(-\beta\Delta)}\right)^N + \frac{1}{T}$$

$$= \sum_{a \in \mathcal{A}} \int_R \sum_{k=0}^{K-1} \mathbb{P}(X_{\theta_a} = k) \left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{k + K\exp(-\beta\Delta)}\right)^N \mathrm{d}\mu(\theta_a) \mathbb{P}\left(A^* = a\right) + \frac{1}{T}$$

$$= \sum_{a \in \mathcal{A}} \int_R \mathbb{E}\left[\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + X_{\theta_a} + K\exp(-\beta\Delta)}\right)^N \bigg| \theta_a\right] \mathrm{d}\mu(\theta_a) \mathbb{P}\left(A^* = a\right) + \frac{1}{T}$$

$$= \int_R \mathbb{E}\left[\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + X_{\theta_1} + K\exp(-\beta\Delta)}\right)^N \bigg| \theta_1\right] \mathrm{d}\mu(\theta_1) + \frac{1}{T},$$

where the last equation is due to $\mathbb{P}(A^* = a) = 1/K$ for any $a \in \mathcal{A}$.

For fixed θ_1, using the Stirling's approximation for binomial tail bound,

$$\mathbb{P}\left(X_{\theta_1} \geq (K-1)q(\theta_1) + \frac{\log(T)}{\log(1/((K-1)q(\theta_1)))} \bigg| \theta_1\right) \leq \frac{1}{T}.$$

Define an event

$$\mathcal{E}_2 = \left\{X_{\theta_1} \leq (K-1)q(\theta_1) + \frac{\log(T)}{\log(1/((K-1)q(\theta_1)))}\right\},$$

such that $\mathbb{P}(\mathcal{E}_2^c|\theta_1) \leq 1/T$. We decompose the above based on $\mathcal{E}_2$:

$$\mathbb{P}\left(A^* \notin \mathcal{U}_A\right) \leq \int_R \mathbb{E}\left[\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + X_{\theta_1} + K\exp(-\beta\Delta)}\right)^N \mathbb{I}(\mathcal{E}_1) \bigg| \theta_1\right] \mathrm{d}\mu(\theta_1) + \frac{2}{T}$$

$$\leq \int_R \mathbb{E}\left[\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + (K-1)q(\theta_1) + \frac{\log(T)}{\log(1/((K-1)q(\theta_1)))} + K\exp(-\beta\Delta)}\right)^N \mathbb{I}(\mathcal{E}_1) \bigg| \theta_1\right] \mathrm{d}\mu(\theta_1) + \frac{2}{T},$$

$$\tag{C.9}$$

where the expectation is with respect to θ_1 under measure $\mu(\cdot)$. It remains to derive an upper bound $q(\theta_1)$ which is defined in Eq. (C.7).

- First, we have

$$\mathbb{P}\left(\theta_1 \leq x | A^* = a_1\right) = \frac{\mathbb{P}\left(\theta_1 \leq x, A^* = a_1\right)}{\mathbb{P}\left(A^* = a_1\right)}$$

$$= \frac{\mathbb{P}\left(\theta_1 \leq x, \bigcap_{a \neq a_1}\{\theta_1 \geq \theta_a\}\right)}{\mathbb{P}\left(A^* = a_1\right)}$$

$$= \frac{\int_{-\infty}^x \prod_{a \neq a_1} \int_{-\infty}^{\theta_1} \mathrm{d}\rho(\theta_a) \, \mathrm{d}\rho(\theta_1)}{\mathbb{P}\left(A^* = a_1\right)}$$

$$= \frac{\int_{-\infty}^x F(\theta_1)^{K-1} \, \mathrm{d}\rho(\theta_1)}{\mathbb{P}\left(A^* = a_1\right)} = \int_{-\infty}^x \mathrm{d}\mu(\theta_1).$$

Define another event $\mathcal{E}_3 = \{F(\theta_1) \geq 1/2\}$. Using the assumption $K \geq \log_2(T)$,

$$\mathbb{P}\left(\theta_1 \leq 0 | A^* = a_1\right) = \mathbb{P}\left(\max_a \theta_a \leq 0\right) = 2^{-K} \leq \frac{1}{T}.$$

Thus, we have $\mathbb{P}(\mathcal{E}_3^c) \leq 1/T$ under the measure μ.

- Second, we have

$$\int_{\theta_1-\Delta}^{\theta_1} \mathrm{d}\rho(\theta) \leq \Delta \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\theta_1-\Delta)^2}{2}\right) \leq \frac{\Delta}{\sqrt{2\pi}}.$$

Note that from the definition of $q(\theta_1)$ in Eq. (C.7), $q(\theta_1)$ cannot exceed 1. Therefore, under event $\mathcal{E}_3$, $q(\theta_1)$ is upper bounded by $\min(2\Delta/\sqrt{2\pi}, 1) \leq \min(\Delta, 1)$. We decompose Eq. (C.9) based on $\mathcal{E}_3$,

$$\mathbb{P}(A^* \notin \mathcal{U}_A) \leq \int_R \mathbb{E}\left[\left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + K\min(\Delta, 1) + \frac{\log(T)}{\log(1/((K-1)\min(\Delta,1)))} + K\exp(-\beta\Delta)}\right)^N \mathbb{I}(\mathcal{E}_3)\Big|\theta_1\right] \mathrm{d}\mu(\theta_1) + \frac{3}{T}$$

$$\leq \left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + K\min(\Delta, 1) + \frac{\log(T)}{\log(1/((K-1)\min(\Delta,1)))} + K\exp(-\beta\Delta)}\right)^N + \frac{3}{T}.$$

(C.10)

With the choice of $\Delta = \log(T\beta)/\beta$, we have

$$\mathbb{P}(A^* \notin \mathcal{U}_A) \leq \left(1 - \frac{\exp\left(-\beta\sqrt{2\log(TK)}/\lambda\right)}{1 + K\min(\log(T\beta)/\beta, 1) + \frac{\log(T)}{\log(1/((K-1)\min(\log(T\beta)/\beta,1)))} + K/(T\beta)}\right)^N + \frac{3}{T}.$$

This ends the proof.

C.2. Prove $\mathbb{E}[|\mathcal{U}_A|] \leq f_2$

We first observe that

$$\mathbb{E}[|\mathcal{U}_A|] = \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{I}(a \in \mathcal{U}_A)\right] = \mathbb{E}\left[\sum_{a\in\mathcal{B}} \mathbb{I}(a \in \mathcal{U}_A) + \sum_{a\notin\mathcal{B}} \mathbb{I}(a \in \mathcal{U}_A)\right]$$

$$\leq \mathbb{E}[|\mathcal{B}|] + \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{I}(a \in \mathcal{U}_A, a \notin \mathcal{B})\right],$$

(C.11)

where $\mathcal{B}$ is defined in Eq. (C.3). For the second term in Eq. (C.11),

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{I}(a \in \mathcal{U}_A, a \notin \mathcal{B})\right] = \sum_{a\in\mathcal{A}} \mathbb{P}(a \in \mathcal{U}_A, a \notin \mathcal{B})$$

$$= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{P}\left(a \in \mathcal{U}_A, a \notin \mathcal{B}\Big|\theta, \vartheta\right)\right]$$

$$\leq \mathbb{E}\left[\sum_{n=1}^N \sum_{a\in\mathcal{A}} \mathbb{P}\left(\bar{A}_n = a, \langle A^* - a, \theta\rangle \geq \Delta\Big|\theta, \vartheta\right)\right]$$

$$= \mathbb{E}\left[\sum_{n=1}^N \sum_{a\in\mathcal{A}} \mathbb{P}\left(\bar{A}_n = a, \langle A^* - a, \theta - \vartheta\rangle + \langle A^* - a, \vartheta\rangle \geq \Delta\Big|\theta, \vartheta\right)\right]$$

$$\leq \mathbb{E}\left[\sum_{n=1}^N \sum_{a\in\mathcal{A}} \mathbb{P}\left(\bar{A}_n = a, \|A^* - a\|_1\|\theta - \vartheta\|_\infty + \langle A^* - a, \vartheta\rangle \geq \Delta\Big|\theta, \vartheta\right)\right].$$

Similar to the proof in Appendix C.1, we decompose the above based on $\mathcal{E}_1$:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{a\in\mathcal{A}}\mathbb{I}\left(a\in\mathcal{U}_A,a\notin\mathcal{B}\right)\right] &= \mathbb{E}\left[\sum_{n=1}^{N}\sum_{a\in\mathcal{A}}\mathbb{P}\left(\bar{A}_n=a,\langle A^*-a,\theta-\vartheta\rangle+\langle A^*-a,\vartheta\rangle\geq\Delta\Big|\theta,\vartheta\right)\mathbb{I}(\mathcal{E}_1)\right]+\mathbb{P}(\mathcal{E}_1^c)\\
&\leq \mathbb{E}\left[\sum_{n=1}^{N}\sum_{a\in\mathcal{A}}\mathbb{P}\left(\bar{A}_n=a,\langle A^*-a,\vartheta\rangle\geq\Delta-\sqrt{2\log(TK)}/\lambda\Big|\theta,\vartheta\right)\mathbb{I}(\mathcal{E}_1)\right]+\frac{1}{T}\\
&\leq N\mathbb{E}\left[\sum_{a\in\mathcal{A},\langle A^*-a,\vartheta\rangle\geq\Delta-\sqrt{2\log(TK)}/\lambda}\frac{\exp(\beta\langle a,\vartheta\rangle)}{\sum_{b\in\mathcal{A}}\exp(\beta\langle b,\vartheta\rangle)}\right]+\frac{1}{T}\\
&= N\mathbb{E}\left[\sum_{a\in\mathcal{A},\langle A^*-a,\vartheta\rangle\geq\Delta-\sqrt{2\log(TK)}/\lambda}\frac{1}{\sum_{b\in\mathcal{A}}\exp(\beta\langle b-a,\vartheta\rangle)}\right]+\frac{1}{T}\\
&\leq N\mathbb{E}\left[\sum_{a\in\mathcal{A},\langle A^*-a,\vartheta\rangle\geq\Delta-\sqrt{2\log(TK)}/\lambda}\frac{1}{\exp(\beta\langle A^*-a,\vartheta\rangle)}\right]+\frac{1}{T}\\
&\leq N\mathbb{E}\left[\sum_{a\in\mathcal{A},\langle A^*-a,\vartheta\rangle\geq\Delta-\sqrt{2\log(TK)}/\lambda}\frac{1}{\exp(\beta(\Delta-\sqrt{2\log(TK)}/\lambda))}\right]+\frac{1}{T}\\
&\leq NK\exp\left(-\beta(\Delta-\sqrt{2\log(TK)}/\lambda)\right)+\frac{1}{T}\,.
\end{aligned}
$$

$$\text{(C.12)}$$

Combining Eqs. (C.11)-(C.12) together,

$$
\mathbb{E}[|\mathcal{U}_A|]\leq\mathbb{E}[|\mathcal{B}|]+NK\exp\left(-\beta(\Delta-\sqrt{2\log(TK)}/\lambda)\right)+\frac{1}{T}\,. \tag{C.13}
$$

Now we start to bound $\mathbb{E}[|\mathcal{B}|]$. By the definition,

$$
\begin{aligned}
\mathbb{E}\left[|\mathcal{B}|-1\right] &= \sum_{a\in\mathcal{A}}\mathbb{E}\left[|\mathcal{B}|-1|A^*=a\right]\mathbb{P}(A^*=a)\\
&= \sum_{a\in\mathcal{A}}\sum_{k=0}^{K-1}k\mathbb{P}\left(|\mathcal{B}|-1=k|A^*=a\right)\mathbb{P}(A^*=a)\,.
\end{aligned}
$$

Using Eq. (C.8),

$$
\begin{aligned}
\mathbb{E}\left[|\mathcal{B}|-1\right] &= \sum_{a\in\mathcal{A}}\sum_{k=0}^{K-1}k\int_{R}\mathbb{P}(X_{\theta_a}=k)\,\mathrm{d}\mu(\theta_a)\mathbb{P}(A^*=a)\\
&= \sum_{a\in\mathcal{A}}\int_{R}\sum_{k=0}^{K-1}k\mathbb{P}(X_{\theta_a}=k)\,\mathrm{d}\mu(\theta_a)\mathbb{P}(A^*=a)=\sum_{a\in\mathcal{A}}\int_{R}\mathbb{E}[X_{\theta_a}]\,\mathrm{d}\mu(\theta_a)\mathbb{P}(A^*=a)\\
&= \sum_{a\in\mathcal{A}}\int_{R}(K-1)q(\theta_a)\,\mathrm{d}\mu(\theta_a)\mathbb{P}(A^*=a)\,.
\end{aligned}
$$

Bounding $q(\theta_a)$ in a similar way, it implies

$$
\begin{aligned}
\mathbb{E}\left[|\mathcal{B}|-1\right] &\leq \sum_{a\in\mathcal{A}}\frac{2\Delta(K-1)}{\sqrt{2\pi}}\int_{R}\exp\left(-\frac{(\theta_a-\Delta)^2}{2}\right)\mathrm{d}\mu(\theta_a)\mathbb{P}(A^*=a)\\
&\leq \frac{2}{\sqrt{2\pi}}(K-1)\Delta\leq(K-1)\Delta\,.
\end{aligned}
$$

Note that $|\mathcal{B}| - 1$ is at most $K - 1$ such that

$$\mathbb{E}\left[|\mathcal{B}| - 1\right] \leq (K - 1)\min(\Delta, 1) \leq K\min(\Delta, 1).$$

Together with Eq. (C.13), we can show that

$$\mathbb{E}[|\mathcal{U}_A|] \leq K\min(\Delta, 1) + 1 + NK\exp\left(-\beta(\Delta - \sqrt{2\log(TK)}/\lambda)\right) + \frac{1}{T}. \tag{C.14}$$

With the choice of $\Delta = \log(T\beta)/\beta$, we have

$$\mathbb{E}[|\mathcal{U}_A|] \leq K\min\left(\frac{\log(T\beta)}{\beta}, 1\right) + 1 + \frac{NK}{T\beta}\exp\left(\beta\sqrt{2\log(TK)}/\lambda\right) + \frac{1}{T}.$$

This ends the proof.

D. Proof of Lemma 5.1

At time t, to obtain the MAP estimate, we can solve the following optimization problem:

$$\operatorname*{argmax}_{\theta, \vartheta} \underbrace{\log P(\mathcal{D}_t|\theta, \vartheta)}_{\text{log-likelihood function}} + \underbrace{\log f(\theta, \vartheta)}_{\text{log-prior function}},$$

where the log-prior function has the form

$$\begin{aligned}
\log f(\theta, \vartheta) &= \log f(\vartheta|\theta) + \log f(\theta) \\
&= -\frac{d}{2}\log(2\pi/\lambda^2) - \frac{\lambda^2}{2}\|\vartheta - \theta\|_2^2 - \frac{1}{2}\log(\det(2\pi\Sigma_0)) - \frac{1}{2}\theta^\top\Sigma_0^{-1}\theta,
\end{aligned} \tag{D.1}$$

and the log-likelihood function can be defined in three steps:

- First, we write this as sum of two terms, one involving the offline dataset $\mathcal{D}_0$ and the other involving the online data $\mathcal{H}_t$:

$$\log P(\mathcal{D}_t|\theta, \vartheta) = \log P(\mathcal{D}_0|\theta, \vartheta) + \log P(\mathcal{H}_t|\mathcal{D}_0, \theta, \vartheta).$$

- Second, we decompose the log-likelihood function for the offline dataset $\mathcal{D}_0$ into a sum of action likelihood and reward likelihood functions:

$$\begin{aligned}
\log P(\mathcal{D}_0|\theta, \vartheta) &= \sum_{n=1}^{N}\left(\log P\left(\bar{A}_n|\theta, \vartheta\right) + \log P\left(\bar{R}_n|\bar{A}_n, \theta, \vartheta\right)\right) \\
&= \sum_{n=1}^{N}\left(\beta\vartheta^T\bar{A}_n - \log\left(\sum_{b\in\mathcal{A}}\exp\left(\beta\vartheta^T b\right)\right)\right) - \frac{1}{2}\sum_{n=1}^{N}\left(\bar{R}_n - \theta^T\bar{A}_n\right)^2 - \frac{N}{2}\log(2\pi),
\end{aligned} \tag{D.2}$$

where the second equation follows Eq. (2.1) and Gaussian noise assumption.

- Third, as per the TS algorithm, A_t is independent of θ conditioned on $\mathcal{U}_{t-1}$, which implies

$$\begin{aligned}
\log P(\mathcal{H}_t|\mathcal{D}_0, \theta, \vartheta) &= \sum_{\tau=1}^{t}\log P\left(A_\tau|\mathcal{D}_{\tau-1}\right) + \sum_{\tau=1}^{t}\log P\left(R_\tau|A_\tau, \theta, \vartheta\right) \\
&= -\frac{1}{2}\sum_{\tau=1}^{t}\left(R_\tau - \theta^T A_\tau\right)^2 - \frac{t}{2}\log(2\pi) + \text{const.}
\end{aligned} \tag{D.3}$$

Putting Eqs. (D.1)-(D.3) together, our loss function to obtain a MAP estimate can be simplified to

$$\begin{aligned}
\mathcal{L}(\theta, \vartheta) = &-\sum_{n=1}^{N}\left(\beta\vartheta^T\bar{A}_n - \log\left(\sum_{b\in\mathcal{A}}\exp\left(\beta\vartheta^T b\right)\right)\right) \\
&+ \frac{1}{2}\sum_{n=1}^{N}\left(\bar{R}_n - \theta^T\bar{A}_n\right)^2 + \frac{1}{2}\sum_{\tau=1}^{t}\left(R_\tau - \theta^T A_\tau\right)^2 + \frac{\lambda^2}{2}\|\vartheta - \theta\|_2^2 + \frac{1}{2}\theta^\top\Sigma_0^{-1}\theta.
\end{aligned}$$

This ends the proof.

E. Proof of Claim 4.5

We just need to prove an upper bound for $\mathbb{E}[\max_k X_k]$ where $X_k \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. By the Jenson's inequality,

$$\exp\left(t\mathbb{E}\left[\max_k X_k\right]\right) \leq \mathbb{E}\left[\exp\left(t\max_k X_k\right)\right] = \mathbb{E}\left[\max_k \exp\left(tX_k\right)\right] \leq \sum_{k=1}^{K} \mathbb{E}\left[\exp\left(tX_k\right)\right] = K\exp(t^2/2),$$

where the last equality follows from the definition of the Gaussian moment generating function. This implies

$$\mathbb{E}\left[\max_k X_k\right] \leq \frac{\log(K)}{t} + \frac{t}{2}.$$

Letting $t = \sqrt{2\log(K)}$, we have

$$\mathbb{E}[\max_k X_k] \leq \sqrt{2\log(K)}.$$

This ends the proof.