Contextual Information-Directed Sampling

Botao Hao [1] **Tor Lattimore** [1] **Chao Qin** [2]

Abstract

Information-directed sampling (IDS) has recently demonstrated its potential as a data-efficient reinforcement learning algorithm (Lu et al., 2021). However, it is still unclear what is the right form of information ratio to optimize when contextual information is available. We investigate the IDS design through two contextual bandit problems: contextual bandits with graph feedback and sparse linear contextual bandits. We provably demonstrate the advantage of *contextual IDS* over *conditional IDS* and emphasize the importance of considering the context distribution. The main message is that an intelligent agent should invest more on the actions that are beneficial for the future unseen contexts while the conditional IDS can be myopic. We further propose a computationally-efficient version of contextual IDS based on Actor-Critic and evaluate it empirically on a neural network contextual bandit.

1. Introduction

Information-directed sampling (IDS) (Russo & Van Roy, 2018) is a promising approach to balance the exploration and exploitation tradeoff (Lu et al., 2021). Its theoretical properties have been systematically studied in a range of problems (Kirschner & Krause, 2018; Liu et al., 2018a; Kirschner et al., 2020a;b; Hao et al., 2021a). However, the aforementioned analysis[1] is limited to the fixed action set.

In this work, we study the IDS design for contextual bandits (Langford & Zhang, 2007), which can be viewed as a simplified case of full reinforcement learning. The context at each round is generated independently, rather than determined by the actions. Contextual bandits are widely used in real-world applications, such as recommender systems (Li et al., 2010).

A natural generalization of IDS to the contextual case is *conditional IDS*. Given the current context, conditional IDS acts as if it were facing a fixed action set. However, this design ignores the context distribution and can be myopic. We mainly want to address the question of what is the right form of information ratio in the contextual setting and highlight the necessity of utilizing the context distribution.

Contributions Our contribution is four-fold:

- We introduce a version of contextual IDS that takes the context distribution into consideration for contextual bandits with graph feedback and sparse linear contextual bandits.
- For contextual bandits with graph feedback, we prove that conditional IDS suffers $\Omega(\sqrt{\beta(\mathcal{G})n})$ Bayesian regret lower bound for a particular prior and graph while contextual IDS can achieve $\widetilde{O}(\min\{\sqrt{\beta(\mathcal{G})n}, \delta(\mathcal{G})^{1/3}n^{2/3}\})$ Bayesian regret upper bound for any prior. Here, n is the time horizon, $\mathcal{G}$ is a directed feedback graph over the set of actions, $\beta(\mathcal{G})$ is the independence number and $\delta(\mathcal{G})$ is the weak domination number of the graph. In the regime where $\beta(\mathcal{G}) \gtrsim (\lambda(\mathcal{G})^2 n)^{1/3}$, contextual IDS achieves better regret bound than conditional IDS.
- For sparse linear contextual bandits, we prove that conditional IDS suffers $\Omega(\sqrt{nds})$ Bayesian regret lower bound for a particular sparse prior while contextual IDS can achieve $\widetilde{O}(\min\{\sqrt{nds}, sn^{2/3}\})$ Bayesian regret upper bound for any sparse prior. Here, d is the feature dimension and s is the sparsity. In the data-poor regime where $d \gtrsim sn^{1/3}$, contextual IDS achieves better regret bound than conditional IDS.

[*]Equal contribution [1]Deepmind [2]Columbia University. Correspondence to: Botao Hao <haobotao000@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]One exception is Kirschner et al. (2020a) who has extended IDS to contextual partial monitoring.

- We further propose a computationally-efficient algorithm to approximate contextual IDS based on Actor-Critic (Konda & Tsitsiklis, 2000) and evaluate it empirically on a neural network contextual bandit.

2. Related works

Russo & Van Roy (2018) introduced IDS and derived Bayesian regret bounds for multi-armed bandits, linear bandits and combinatorial bandits. Kirschner & Krause (2018); Kirschner et al. (2020a) investigated the use of frequentist IDS for bandits with heteroscedastic noise and partial monitoring. Kirschner et al. (2020b) proved the asymptotic optimality of frequentist IDS for linear bandits. Hao et al. (2021a) developed a class of information-theoretic Bayesian regret bounds for sparse linear bandits that nearly match existing lower bounds on a variety of problem instances. Recently, Lu et al. (2021) proposed a general reinforcement learning framework and used IDS as a key component to build a data-efficient agent. Arumugam & Van Roy (2021) investigated different versions of learning targets when designing IDS. However, there are no concrete regret bounds provided in those works.

Graph feedback Mannor & Shamir (2011); Alon et al. (2015) gave a full characterization of online learning with graph feedback. Tossou et al. (2017) provided the first information-theoretic analysis of Thompson sampling for bandits with graph feedback and Liu et al. (2018a) derived a Bayesian regret bound of IDS. Both bounds depend on the clique cover number which could be much larger than the independence number. Liu et al. (2018b) further improved the analysis with the feedback graph unknown and derived the Bayesian regret bound of a mixture policy in terms of the independence number. In contrast, our work is the first one to consider contextual bandits with graph feedback and derived nearly optimal regret bound in terms of both independence number and domination number for a single algorithm. Very recently, Dann et al. (2020) considered the reinforcement learning with graph feedback. Their $O(\sqrt{n})$-upper bound depends on mas-number rather than independence number. We briefly summarize the comparison in Table 2.

Sparse linear bandits For sparse linear bandits with fixed action set, Abbasi-Yadkori et al. (2012) proposed an inefficient online-to-confidence-set conversion approach that achieves an $\widetilde{O}(\sqrt{sdn})$ upper bound for an arbitrary action set. Lattimore et al. (2015) developed a selective explore-then-commit algorithm that only works when the action set is exactly the binary hypercube and derived an optimal $O(s\sqrt{n})$ upper bound. Hao et al. (2020) introduced the notion of an exploratory action set and proved a $\Theta(\mathrm{poly}(s)n^{2/3})$ minimax rate for the data-poor regime using an explore-then-commit algorithm. Note that the minimax lower bound automatically holds for contextual setting when the context distribution is a Dirac on the hard problem instance. Hao et al. (2021b) extended this concept to a MDP setting.

It recently became popular to study the contextual setting. These results can not be reduced to our setting since they rely on either careful assumptions on the context distribution to achieve $\widetilde{O}(\mathrm{poly}(s)\sqrt{n})$ or $\widetilde{O}(\mathrm{poly}(s)\log(n))$ regret bounds (Bastani & Bayati, 2020; Wang et al., 2018; Kim & Paik, 2019; Wang et al., 2020; Ren & Zhou, 2020; Oh et al., 2021) such that classical high-dimensional statistics can be used. However, minimax lower bounds is missing to understand if those assumptions are fundamental. Another line of works have polynomial dependency on the number of actions (Agarwal et al., 2014; Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2020). We briefly summarize the comparison in Table 2.

3. Problem Setting

We study the stochastic contextual bandit problem with a horizon of n rounds and a finite set of possible contexts $\mathcal{S}$. For each context $m \in \mathcal{S}$, there is an action set $\mathcal{A}^m$. The interaction protocol is as follows. First the environment samples a sequence of independent contexts $(s_t)_{t=1}^n$ from a distribution ξ over $\mathcal{S}$. At the start of round t, the context s_t is revealed to the agent, who may use their observations and possibly an external source of randomness to choose an action $A_t \in \mathcal{A}_t = \mathcal{A}^{s_t}$. Then the agent receives an observation O_{t,A_t} including an immediate reward Y_{t,A_t} as well as some side information. Here

$$Y_{t,a} = f(s_t, a, \theta^*) + \eta_{t,a}\,, \text{ if } A_t = a\,,$$

where f is the reward function, θ^* is the unknown parameter and $\eta_{t,a}$ is 1-sub-Gaussian noise. Sometimes we write $Y_t = Y_{t,A_t}$ for short.

We consider the Bayesian setting in the sense that θ^* is sampled from some prior distribution. Let $\mathcal{A} = \cup_{m\in\mathcal{S}}\mathcal{A}^m$ and the history $\mathcal{F}_t = \{(s_1, A_1, O_1), \ldots, (s_{t-1}, A_{t-1}, O_{t-1})\}$. A policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ is a sequence of (suitably measurable) deterministic mappings from the history $\mathcal{F}_t$ and context space $\mathcal{S}$ to a probability distribution $\mathcal{P}(\mathcal{A})$ with the

Table 1. Comparisons with existing results on regret upper bounds and lower bounds for bandits with graphical feedback. Here, $\beta(\mathcal{G})$ is the independence number, $\delta(\mathcal{G})$ is the weak domination number and $c(\mathcal{G})$ is the clique cover number of the graph. Note that $\beta(\mathcal{G}) \leq c(\mathcal{G})$.

Upper Bound	Regret	Contextual?	Algorithm
Alon et al. (2015)	$\widetilde{O}(\sqrt{\beta(\mathcal{G})n})$	no	Exp3.G for strongly observable graph
Alon et al. (2015)	$\widetilde{O}(\delta(\mathcal{G})^{1/3}n^{2/3})$	no	Exp3.G for weakly observable graph
Tossou et al. (2017)	$\widetilde{O}(\sqrt{c(\mathcal{G})n})$	no	Thompson sampling
Liu et al. (2018a)	$\widetilde{O}(\sqrt{c(\mathcal{G})n})$	no	fixed action-set IDS
This paper	$\widetilde{O}(\min(\sqrt{\beta(\mathcal{G})n}, \delta(\mathcal{G})^{1/3}n^{2/3}))$	yes	contextual IDS
Minimax Lower Bound			
Alon et al. (2015)	$\Omega(\sqrt{\beta(\mathcal{G})n})$	no	strongly observable graph
Alon et al. (2015)	$\Omega(\delta(\mathcal{G})^{1/3}n^{2/3})$	no	weakly observable graph

Table 2. Comparisons with existing results on regret upper bounds and lower bounds for sparse linear bandits. Here, s is the sparsity, d is the feature dimension, n is the number of rounds, K is the number of arms, $C_{\min}$ is the minimum eigenvalue of the data matrix and τ is a problem-dependent parameter that may have a complicated form.

Upper Bound	Regret	Contextual?	Assumption	Algorithm
Abbasi-Yadkori et al. (2012)	$\widetilde{O}(\sqrt{sdn})$	yes	none	UCB
Sivakumar et al. (2020)	$\widetilde{O}(\sqrt{sdn})$	yes	adver. + Gaussian noise	greedy
Bastani & Bayati (2020)	$\widetilde{O}(\tau K s^2(\log(n))^2)$	yes	compatibility condition	Lasso greedy
Lattimore et al. (2015)	$\widetilde{O}(s\sqrt{n})$	no	action set is hypercube	explore-then-commit
Hao et al. (2021a)	$\widetilde{O}(\min(\sqrt{sdn}, sn^{2/3}/C_{\min}))$	no	none	fixed action-set IDS
This paper	$\widetilde{O}(\min(\sqrt{sdn}, sn^{2/3}/C_{\min}))$	yes	none	contextual IDS
Minimax Lower Bound				
Hao et al. (2020)	$\Omega(\min(\sqrt{sdn}, C_{\min}^{-1/3}s^{1/3}n^{2/3}))$	no	N.A	N.A.

constraint that for each context m, $\pi_t(\cdot|m)$ can only put mass over $\mathcal{P}(\mathcal{A}^m)$. Then the *Bayesian regret* of a policy π is defined as

$$\mathfrak{BR}(n;\pi) = \mathbb{E}\left[\sum_{t=1}^{n} \max_{a\in\mathcal{A}_t} f(s_t, a, \theta^*) - \sum_{t=1}^{n} Y_t\right], \quad (3.1)$$

where the expectation is over the interaction sequence induced by the agent and environment, the context distribution and the prior distribution over θ^*.

Notations Given a measure $\mathbb{P}$ and jointly distributed random variables X and Y we let $\mathbb{P}_X$ denote the law of X and we let $\mathbb{P}_{X|Y}$ be the conditional law of X given Y such that: $\mathbb{P}_{X|Y}(\cdot) = \mathbb{P}(X \in \cdot|Y)$. The mutual information between X and Y is $\mathbb{I}(X;Y) = \mathbb{E}[D_{\mathrm{KL}}(\mathbb{P}_{X|Y}||\mathbb{P}_X)]$ where D_{KL} is the relative entropy. We write $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|\mathcal{F}_t)$ as the posterior measure where $\mathbb{P}$ is the probability measure over θ^* and the history and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. Denote $\mathbb{I}_t(X;Y) = \mathbb{E}_t[D_{\mathrm{KL}}(\mathbb{P}_{t,X|Y}||\mathbb{P}_{t,X})]$. We write $\mathbb{1}\{\cdot\}$ as an indicator function. For a positive semidefinite matrix A, we let $\sigma_{\min}(A)$ be the minimum eigenvalue of A. Denote $\mathcal{P}(\mathcal{A})$ be the space of probability measures over a set $\mathcal{A}$ with the Borel σ-algebra.

4. Conditional IDS versus Contextual IDS

We introduce the formal definition of conditional IDS and contextual IDS for general contextual bandits. Suppose that the optimal action associated with context s_t is $a_t^* = \mathrm{argmax}_{a\in\mathcal{A}_t} f(s_t, a, \theta^*)$, which in Bayesian setting is a random variable. We first define the one-step expected regret of taking action $a \in \mathcal{A}_t$ as

$$\Delta_t(a|s_t) := \mathbb{E}_t\left[f(s_t, a_t^*, \theta^*) - f(s_t, a, \theta^*)\Big|s_t\right], \quad (4.1)$$

and write $\Delta_t(s_t) \in \mathbb{R}^{|\mathcal{A}_t|}$ as the corresponding vector.

4.1. Conditional IDS

A natural generalization of the standard IDS design (Russo & Van Roy, 2018) to contextual setting is *conditional IDS*. It has been investigated empirically in Lu et al. (2021) for the full reinforcement learning setting.

Conditional on the current context, we define the information gain of taking action a with respect to the current opti-

mal action a_t^* as $\mathbb{I}_t(a_t^*; O_{t,a}|s_t) :=$

$$\mathbb{E}_t\left[D_{\mathrm{KL}}(\mathbb{P}_t(a_t^* \in \cdot|s_t, O_{t,a})||\mathbb{P}_t(a_t^* \in \cdot|s_t))\Big|s_t\right],$$

and write $\mathbb{I}_t(a_t^*, s_t) \in \mathbb{R}^{|\mathcal{A}_t|}$ such that $[\mathbb{I}_t(a_t^*, s_t)]_a = \mathbb{I}_t(a_t^*; O_{t,a}|s_t)$.

We introduce the (α, λ)-*conditional information ratio* (CIR) as $\Gamma_{t,\alpha}^{\lambda} : \mathcal{P}(\mathcal{A}) \to \mathbb{R}$ such that

$$\Gamma_{t,\alpha}^{\lambda}(\pi(\cdot|s_t)) = \frac{\max\left(0, \Delta_t(s_t)^\top \pi(\cdot|s_t) - \alpha\right)^\lambda}{\mathbb{I}_t(a_t^*, s_t)^\top \pi(\cdot|s_t)}, \quad (4.2)$$

for some parameters $\alpha, \lambda > 0$. Conditional IDS minimizes (α, λ)-CIR to find a probability distribution over the action space:

$$\pi_t(\cdot|s_t) = \operatorname*{argmin}_{\pi(\cdot|s_t) \in \mathcal{P}(\mathcal{A}_t)} \Gamma_{t,\alpha}^{\lambda}(\pi(\cdot|s_t)).$$

Remark 4.1. *In comparison to the standard information ratio by Russo & Van Roy (2018) who specified $\lambda = 2, \alpha = 0$ in the non-contextual setting, we consider a generalized version introduced by Lattimore & György (2020). As observed by Lattimore & György (2020), the right value of λ depends on the dependence of the regret on the horizon. The parameter α is always chosen at order $O(1/\sqrt{n})$ for certain problems such that its regret is negligible. Their role will be more clear in Sections 5&6.*

4.2. Contextual IDS

A possible limitation of conditional IDS is that the CIR does not take the context distribution into consideration. As we show later, this can result in sub-optimality in certain problems. *Contextual IDS*, which was first proposed by Kirschner et al. (2020a) for partial monitoring, is introduced to remedy this shortcoming.

Let $\pi^* : \mathcal{S} \to \mathcal{A}$ be the optimal policy such that for each $m \in \mathcal{S}$, $\pi^*(m) = \operatorname{argmax}_{a \in \mathcal{A}^m} f(m, a, \theta^*)$ with the tie broken arbitrarily. We define the information gain of taking action a with respect to π^* as:

$$\mathbb{I}_t(\pi^*; O_{t,a}) := \mathbb{E}_t\left[D_{\mathrm{KL}}(\mathbb{P}_t(\pi^* \in \cdot|O_{t,a})||\mathbb{P}_t(\pi^* \in \cdot))\right],$$

and write $\mathbb{I}_t(\pi^*) \in \mathbb{R}^{|\mathcal{A}_t|}$ as the corresponding vector. Define a policy class $\Pi = \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}), \mathrm{supp}(\pi(\cdot|m)) \subset \mathcal{A}^m\}$ where $\mathrm{supp}(\cdot)$ is the support of a vector. We introduce the (α, λ)-*marginal information ratio* (MIR) as $\Psi_{t,\alpha}^{\lambda} : \Pi \to \mathbb{R}$ such that

$$\Psi_{t,\alpha}^{\lambda}(\pi) = \frac{\max\left(0, (\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi(\cdot|s_t)] - \alpha)\right)^\lambda}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top \pi(\cdot|s_t)]}, \quad (4.3)$$

where the expectation is taken with respect to the context distribution. Contextual IDS minimizes the (α, λ)-MIR to find a mapping from the context space to the action space:

$$\pi_t = \operatorname*{argmin}_{\pi \in \Pi} \Psi_{t,\alpha}^{\lambda}(\pi).$$

When receiving context s_t at round t, the agent plays actions according to $\pi_t(\cdot|s_t)$.

We highlight the key difference between conditional IDS and contextual IDS as follows:

- Conditional IDS only optimizes a probability distribution for the current context while contextual IDS optimizes a full mapping from the context space to action space.
- Conditional IDS only seeks information about the current optimal action while contextual IDS seeks information for the whole optimal policy.

Remark 4.2. *We can also define the information gain in conditional IDS and contextual IDS with respect to the unknown parameter θ^*. This could bring in certain computational advantage when approximating the information gain. However, θ^* contains much more information to learn than the optimal action or policy such that the agent may suffer more regret.*

4.3. Why conditional IDS could be myopic?

Conditional IDS myopically balances exploration and exploitation without taking the context distribution into consideration. This has the advantage of being simple to implement but sometimes leads both under and over exploration, which the following examples illustrate. In both examples there are two contexts arriving with equal probability.

- **Example 1** [UNDER EXPLORATION] Consider a noiseless case. Context set 1 contains k actions where one is the optimal action and the remaining $k-1$ actions yield regret 1. Context set 2 contains a revealing action with regret 1 and one action with no regret. The revealing action provides an observation of the rewards for all the k actions in context set 1. When context set 2 arrives, conditional IDS will never play the revealing action since it incurs high immediate regret with no useful information for the current context set. However, this ignores the fact that the revealing action could be informative for the unseen context set 1. Conditional IDS *under-explores* and suffers $O(k)$ regret. In contrast, contextual IDS exploits the context

distribution and plays the revealing action in context 2 and only suffers $O(1)$ regret.

- **Example 2** [OVER EXPLORATION] Context set 1 contains a single revealing action (hence no regret). Context set 2 has k actions. The first is a revealing action and has a (known) regret of $\Theta(\sqrt{k}\Delta)$ with $\Delta = \Theta(1/\sqrt{n})$. Of the remaining actions, one is optimal (zero regret) and the others have regret Δ, with the prior such that the identify of the optimal action is unknown. Contextual IDS will avoid the revealing action in context set 2 because it understands that this information can be obtained more cheaply in context set 1. Its regret is $O(\sqrt{n})$. Meanwhile, if the constants are tuned appropriately, then conditional IDS will play the revealing action in context set 2 and suffer regret $\Omega(\sqrt{nk})$.

Remark 4.3. *Similar shortcoming could also hold for the class of UCB and Thompson sampling algorithms since they have no explicit way to encode the information of context distribution.*

4.4. Generic regret bound for contextual IDS

Before we step into specific examples, we first provide a generic bound for contextual IDS. Denote $\mathcal{I}_{\alpha,\lambda}$ as the worst-case MIR such that for any $t \in [n]$, $\Psi_{t,\alpha}^{\lambda}(\pi_t) \leq \mathcal{I}_{\alpha,\lambda}$ almost surely.

Theorem 4.4. *Let $\pi^{MIR} = (\pi_t)_{t\in[n]}$ be the contextual IDS policy such that π_t minimizes $(\alpha, 2)$-MIR at each round. Then the following regret upper bound holds*

$$\mathfrak{BR}(n;\pi^{MIR}) \leq n\alpha + \inf_{\lambda\geq 2} 2^{1-2/\lambda}\mathcal{I}_{\alpha,\lambda}^{1/\lambda} n^{1-1/\lambda}\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right]^{\frac{1}{\lambda}}.$$

The proof is deferred to Appendix A.1. An interesting consequence of Theorem 4.4 is that contextual IDS minimizing $\lambda = 2$ can adapt to $\lambda > 2$. This shows the power of contextual IDS to adapt to different information-regret structures.

5. Contextual Bandits with Graph Feedback

We consider a Bayesian formulation of contextual k-armed bandits with graph feedback, which generalizes the setup in Alon et al. (2015). Let $\mathcal{G} = (\mathcal{A}, \mathcal{E})$ be a directed feedback graph over the set of actions $\mathcal{A}$ with $|\mathcal{A}| = k$. For each $i \in \mathcal{A}$, we define its in-neighborhood and out-neighborhood as $N^{\text{in}}(i) = \{j \in \mathcal{A} : (j,i) \in \mathcal{E}\}$ and $N^{\text{out}}(i) = \{j \in \mathcal{A} : (i,j) \in \mathcal{E}\}$. The feedback graph $\mathcal{G}$ is fixed and known to the agent.

Let the unknown parameter $\theta^* \in \mathbb{R}^k$. At each round t, the agent receives an observation characterized by $\mathcal{G}$ in the sense that taking action $a \in \mathcal{A}_t$, the observation $O_{t,a}$ contains the rewards $Y_{t,a'} = \theta^*_{a'} + \eta_{t,a'}$ for each action $a' \in N^{\text{out}}(a)$. We review two standard quantities to describe the graph structure used in Alon et al. (2015).

Definition 5.1 (Independence number)**.** An independent set of a graph $\mathcal{G} = (\mathcal{A}, \mathcal{E})$ is a subset $\mathcal{S} \subseteq \mathcal{A}$ such that no two different $i, j \in \mathcal{A}$ are connected by an edge in $\mathcal{E}$. The cardinality of a largest independent set is the *independence number* of $\mathcal{G}$, denoted by $\beta(\mathcal{G})$.

A vertex a is observable if $N^{\text{in}}(a) \neq \emptyset$. A vertex is strongly observable if it has either a self-loop or incoming edges from all other vertices. A vertex is weakly observable if it is observable but not strongly observable. A graph $\mathcal{G}$ is observable if all its vertices are observable and it is strongly observable if all its vertices are strongly observable. A graph is weakly observable if it is observable but not strongly observable.

Definition 5.2 (Weak domination number)**.** In a directed graph $\mathcal{G} = (\mathcal{A}, \mathcal{E})$ with a set of weakly observable vertices $\mathcal{W} \subseteq \mathcal{A}$, a weakly dominating set $\mathcal{D} \subseteq \mathcal{A}$ is a set of vertices that dominates $\mathcal{W}$. That means for any $w \in \mathcal{W}$ there exists $d \in \mathcal{D}$ such that $w \in N^{\text{out}}(d)$. The *weak domination number* of $\mathcal{G}$, denoted by $\delta(\mathcal{G})$, is the size of the smallest weakly dominating set.

5.1. Lower bound for conditional IDS

We first prove that conditional IDS suffers $\Omega(\sqrt{n\beta(\mathcal{G})})$ Bayesian regret lower bound for a particular prior and show later the optimal rate for this prior could be much smaller when $\beta(\mathcal{G})$ is very large (Section 5.2).

Theorem 5.3. *Let π^{CIR} be a conditional IDS policy. There exists a contextual bandit with graph feedback instance such that $\beta(\mathcal{G}) = k$, $\delta(\mathcal{G}) = 1$ and*

$$\mathfrak{BR}(n;\pi^{CIR}) \geq \frac{1}{16}\sqrt{n(\beta(\mathcal{G})-3)}.$$

Proof. Let us construct the hard problem instance that generalizes Example 1 in Section 4.3. Suppose $\{x_1, \ldots, x_k\} \subseteq \mathbb{R}^k$ is the set of basis vectors that corresponds to k arms. The first $k-1$ arms form a standard multi-armed Gaussian bandit with unit variance. When kth arm is pulled, the agent always suffers a regret of 1, but obtains samples for the first $k-1$ arms. In terms of the

language of graph feedback, the first $k-1$ arms only contain self-loop while the out-neighborhood of the last arm contains all the first $k-1$ arms. One can verify $\beta(\mathcal{G}) = k$, $\delta(\mathcal{G}) = 1$.

There are two context sets and the arrival probability of each context is 1/2. Context set 1 consists of $\{x_{k-1}, x_k\}$ while context set 2 consists of $\{x_1, \ldots, x_{k-2}\}$. One can verify $\beta(\mathcal{G}) = k$, $\delta(\mathcal{G}) = 1$.

For each $i \in \{2, \ldots, k-2\}$, let $\theta^{(i)} \in \mathbb{R}^k$ with $\theta^{(i)}_j = \mathbb{1}\{i = j\}\gamma$ for $j \in \{2, \ldots, k-2\}$ and $\theta^{(i)}_1 = 0$, $\theta^{(i)}_{k-1} = 0$, $\theta^{(i)}_k = \gamma - 1$ where $\gamma > 0$ is the gap that will be chosen later. Assume the prior of θ^* is uniformly distributed over $\{\theta^{(2)}, \ldots, \theta^{(k-2)}\}$.

We write $\mathbb{E}_i[\cdot] = \mathbb{E}_{\theta^{(i)}}[\cdot]$ for short and the expectation is taken with respect to the measures on the sequence of outcomes $(A_1, Y_1, \ldots, A_n, Y_n)$ determined by $\theta^{(i)}$. Define the cumulative regret of policy π interacting with bandit $\theta^{(i)}$ as

$$R_{\theta^{(i)}}(n;\pi) = \sum_{t=1}^{n} \mathbb{E}_i\left[\left(\langle x^*_{s_t}, \theta^{(i)}\rangle - Y_t\right)\mathbb{1}(s_t = 1)\right] + \sum_{t=1}^{n} \mathbb{E}_i\left[\left(\langle x^*_{s_t}, \theta^{(i)}\rangle - Y_t\right)\mathbb{1}(s_t = 2)\right],$$

such that $\mathfrak{BR}(n;\pi) = \frac{1}{k-3}\sum_{i=2}^{k-2} R_{\theta^{(i)}}(n;\pi)$.

Step 1. Fix a conditional IDS policy π. From the definition of conditional IDS in Eq. (4.2), when context set 1 is arriving, conditional IDS will always pull x_{k-1} for this prior since x_{k-1} is always the optimal arm for context set 1. This means conditional IDS will suffer no regret for context set 1, which implies for any $i \in \{2, \ldots, k-2\}$,

$$\sum_{t=1}^{n} \mathbb{E}_i\left[\left(\langle x^*_{s_t}, \theta^{(i)}\rangle - Y_t\right)\mathbb{1}(s_t = 1)\right] = 0\,.$$

Step 2. Define $T_i(n)$ as the number of pulls for arm i over n rounds. On the other hand,

$$\mathbb{E}_i\left[\sum_{t=1}^{n}\left(\langle x^*_{s_t}, \theta^{(i)}\rangle - Y_t\right)\mathbb{1}(s_t = 2)\right] = \mathbb{E}_i\left[\sum_{j=1, j\neq i}^{k-2} T_j(n)\right]\gamma = \mathbb{E}_i\left[n_2 - T_i(n)\right]\gamma\,,$$

where the first equation comes from the fact that the context sets 1 and 2 are disjoint and n_2 is the number of times that context 2 arrives over n rounds. Since the context is generated independently with respect to the learning process, we have $\mathbb{E}_i[n_2] = n/2$. By Pinsker's inequality and the divergence decomposition lemma (Lattimore & Szepesvári, 2020, Lemma 15.1), we know that with $\gamma = \frac{1}{2}\sqrt{k/n}$,

$$\sum_{i=2}^{k-2}\mathbb{E}_i[T_i(n)] \le \sum_{i=2}^{k-2}\mathbb{E}_1[T_i(n)] + \frac{1}{4}\sum_{i=2}^{k-2}\sqrt{nk\mathbb{E}_1(T_i(n))} \le n + \frac{1}{4}kn\,.$$

Therefore,

$$\mathfrak{BR}(n;\pi) = \frac{1}{k-3}\sum_{i=2}^{k-2}\mathbb{E}_i\left[n_2 - T_i(n)\right]\gamma = \frac{1}{k-3}\left(\frac{k-3}{2}n - \sum_{i=2}^{k-2}\mathbb{E}_i[T_i(n)]\right)\gamma \ge \frac{1}{16}\sqrt{n(k-3)}\,.$$

This finishes the proof. □

Remark 5.4. *We can strengthen the lower bound such that it holds for any graph by replacing the standard multi-armed bandit instance in context set 2 by the hard instance in Alon et al. (2017, Theorem 5).*

5.2. Upper bound achieved by contextual IDS

In this section, we derive a Bayesian regret upper bound achieved by contextual IDS. According to Theorem 4.4, it suffices to bound the worst-case MIR $\mathcal{I}_{\alpha,\lambda}$ for $\lambda = 2, 3$ as well as the cumulative information gain respectively. Let $R_{\max}$ be an upper bound on the maximum expected reward.

Lemma 5.5 (Squared information ratio)**.** *For any $\epsilon \in [0, 1]$ and a strongly observable graph $\mathcal{G}$, the $(2\epsilon R_{\max}, 2)$-MIR can be bounded by*

$$\mathcal{I}_{2\epsilon R_{\max},2} \le \frac{4R^2_{\max} + 4}{1-\epsilon}\beta(\mathcal{G})\log\left(\frac{4k^2}{\beta(\mathcal{G})\epsilon}\right).$$

The proof is deferred to Appendix B.1 and the proof idea is to bound the worst-case squared information ratio by the squared information ratio of Thompson sampling.

Next we will bound $\mathcal{I}_{\alpha,\lambda}$ for $\lambda = 3$ in terms of an *explorability constant*.

Definition 5.6 (Explorability constant)**.** We define the explorability constant as $\vartheta(\mathcal{G}, \xi) :=$

$$\max_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})}\min_{d\in\mathcal{A}}\mathbb{E}_{s\sim\xi, a\sim\pi(\cdot|s)}\left[\mathbb{1}\left\{d \in N^{\text{out}}(a)\right\}\right].$$

Lemma 5.7 (Cubic information ratio). *For any $\epsilon \in [0,1]$ and a weakly observable graph $\mathcal{G}$, the $(2\epsilon R_{\max}, 3)$-MIR can be bounded by*

$$\mathcal{I}_{2\epsilon R_{\max},3} \leq \frac{R_{\max}^3 + R_{\max}}{\vartheta(\mathcal{G},\xi)}\,.$$

The proof is deferred to Appendix B.2 and the proof idea is to bound the worst-case cubic information ratio by the cubic information ratio of a policy that mixes Thompson sampling and a policy maximizing the explorability constant.

Lemma 5.8. *The cumulative information gain can be bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right] \leq \mathbb{H}(\pi^*) \leq M\log(k)\,,$$

where $\mathbb{H}(\cdot)$ is the entropy and M is the number of available contexts.

The proof is deferred to Appendix B.3. Combining the results in Lemmas 5.5-5.8 and the generic regret bound in Theorem 4.4, we obtain the follow theorem.

Theorem 5.9 (Regret bound for graph contextual IDS). *Suppose $\pi^{MIR} = (\pi_t)_{t\in\mathbb{N}}$ where π_t minimizes $(2R_{\max}/\sqrt{n}, 2)$-MIR at each round. If $\mathcal{G}$ is strongly observable, we have*

$$\mathfrak{BR}(n;\pi^{MIR}) \leq CR_{\max}\min\left(\left(\frac{2M\log(k)}{\vartheta(\mathcal{G},\xi)}\right)^{\frac{1}{3}} n^{\frac{2}{3}}, \sqrt{\beta(\mathcal{G})\log\left(\frac{4k^2\sqrt{n}}{\beta(\mathcal{G})}\right)nM\log(k)}\right),$$

where C is an absolute constant.

Ignoring the constant and logarithmic terms, the Bayesian regret upper bound in Theorem 5.9 can be simplified to

$$\widetilde{O}\left(\min\left(\sqrt{\beta(\mathcal{G})Mn}, (M/\vartheta(\mathcal{G},\xi))^{1/3}n^{2/3}\right)\right).$$

When the independence number is large enough such that $\beta(\mathcal{G}) \geq (\vartheta(\mathcal{G},\xi),\xi^2 n/M)^{1/3}$, this bound is dominated by $\widetilde{O}((M/\vartheta(\mathcal{G},\xi))^{1/3}n^{2/3})$ that is independent of the independence number. Together with Theorem 5.3, we can argue that conditional IDS is sub-optimal comparing with contextual IDS in this regime.

Remark 5.10 (Connection between $\vartheta(\mathcal{G},\xi)$ and $\delta(\mathcal{G})$). *Let $\mathcal{D} \subseteq \mathcal{A}$ be the smallest weakly dominating set of the full graph with $|\mathcal{D}| = \delta(\mathcal{G})$. Consider a policy μ such that $\mu(\cdot|s_t)$ is uniform over $\mathcal{D}\cap\mathcal{A}_t$. Then we have*

$$\min_{d\in\mathcal{A}} \mathbb{E}_{s\sim\xi, a\sim\mu(\cdot|s)}\left[\mathbb{1}\left\{d \in N^{out}(a)\right\}\right] \geq \frac{1}{M\delta(\mathcal{G})}\,,$$

which implies $1/\vartheta(\mathcal{G},\xi) \leq M\delta(\mathcal{G})$. When $M = 1$, Alon et al. (2015) demonstrated that the minimax optimal rate for weakly observable graph is $\widetilde{\Theta}(\delta(\mathcal{G})^{1/3}n^{2/3})$ which implies $1/\vartheta(\mathcal{G}) \gtrsim \delta(\mathcal{G})$ up to some logarithm factors.

Remark 5.11 (Open problem). *For the fixed action set setting, Alon et al. (2015) proved that the independence number and weakly domination number are the fundamentally quantity to describe the graph structure. However, our regret upper bound has the number of contexts M appearing. It is interesting to investigate if M can be removed for contextual bandits with graph feedback.*

Note that one can also bound $\mathbb{H}(\pi^*) \leq \mathbb{H}(\theta^*) \leq k$ such that the dependency on M does not appear. But it is unclear the right dependency on M and k for the optimal regret rate in the contextual setting.

6. Sparse Linear Contextual Bandits

We consider a Bayesian formulation of sparse linear contextual bandits. The notion of sparsity can be defined through the parameter space Θ:

$$\Theta = \left\{\theta \in \mathbb{R}^d \middle| \sum_{j=1}^{d} \mathbb{1}\{\theta_j \neq 0\} \leq s, \|\theta\|_2 \leq 1\right\}.$$

We assume s is known and it can be relaxed by putting a prior on it. We consider the case where $\mathcal{A}^m = \mathcal{A}$ for all $m \in \mathcal{S}$. Suppose $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ is a feature map known to the agent. In sparse linear contextual bandits, the reward of taking action a is $f(s_t,a,\theta^*) = \phi(s_t,a)^\top\theta^* + \eta_{t,a}$, where θ^* is a random variable taking values in Θ and denote ρ as the prior distribution.

We first define a quantity to describe the explorability of the feature set.

Definition 6.1 (Explorability constant). We define the explorability constant as $C_{\min}(\phi,\xi) :=$

$$\max_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \sigma_{\min}\left(\mathbb{E}_{s\sim\xi}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}[\phi(s,a)\phi(s,a)^\top]\right]\right).$$

Remark 6.2. *The explorability constant is a problem-dependent constant that has previously been introduced in Hao et al. (2020; 2021a) for a non-contextual sparse linear bandits. For an easy instance when $\{\phi(s_t,a)\}_{a\in\mathcal{A}}$ is*

a full hypercube for every $s_t \in \mathcal{S}$, $C_{\min}(\phi, \xi)$ could be as small as 1 where the policy is to sample uniformly from the corner of the hypercube no matter which context arrives.

6.1. Lower bound for conditional IDS

We prove an algorithm-dependent lower bound for a sparse linear contextual bandits instance.

Theorem 6.3. *Let π^{CIR} be a conditional IDS policy. There exists a sparse linear contextual bandit instance whose explorability constant is 1/2 such that*

$$\mathfrak{BR}(n; \pi^{CIR}) \geq \frac{1}{16}\sqrt{nds}\,.$$

How about the principle of optimism? Hao et al. (2021a, Section 4) has shown that even for non-contextual case, optimism-based algorithms such as UCB or Thompson sampling fail to optimally balance information and regret and result in a sub-optimal regret bound.

6.2. Upper bound achieved by contextual IDS

In this section, we will prove that contextual IDS can achieve a nearly dimension-free regret bound. We say that the feature set has sparse optimal actions if the optimal action is s-sparse for each context almost surely with respect to the prior.

Lemma 6.4 (Bounds for information ratio)**.** *We have $\mathcal{I}_{0,2} \leq d/2$. When the feature set has sparse optimal actions, we have $\mathcal{I}_{0,3} \leq s^2/(4C_{\min}(\phi, \xi))$.*

From the definition of $\pi^*(m) = \operatorname{argmin}_{a \in \mathcal{A}^m} a^\top \theta^*$, we know that π^* is a deterministic function of θ^*. By the data processing lemma, we have $\mathbb{I}(\pi^*; \mathcal{F}_{n+1}) \leq \mathbb{I}(\theta^*; \mathcal{F}_{n+1})$. According to Lemma 5.8 in Hao et al. (2021a), we have the following lemma.

Lemma 6.5. *The cumulative information gain can be bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right] \leq 2s\log(dn^{1/2}/s)\,.$$

Combining the results in Lemmas 6.4-6.5 with the generic regret bound in Theorem 4.4, we obtain the following regret bound.

Theorem 6.6 (Regret bound for sparse contextual IDS)**.** *Suppose $\pi^{MIR} = (\pi_t)_{t \in \mathbb{N}}$ where $\pi_t = \operatorname{argmin}_\pi \Psi^2_{t,0}(\pi)$. When the feature set has sparse optimal actions, the following regret bound holds*

$$\mathfrak{BR}(n; \pi^{MIR}) \lesssim \min\left\{\sqrt{nds\log(dn^{1/2}/s)}, \frac{sn^{\frac{2}{3}}(\log(dn^{1/2}/s))^{\frac{1}{3}}}{(C_{\min}(\phi, \xi))^{\frac{1}{3}}}\right\}.$$

In the data-poor regime where $d \gtrsim n^{1/3}s/C_{\min}^{2/3}$, contextual IDS achieves $\widetilde{O}(sn^{2/3})$ regret bound that is tighter than the lower bound for conditional IDS. This rate matches the minimax lower bound derived in Hao et al. (2020) up to a s factor in the data-poor regime.

7. Practical Algorithm

As shown in Section 4.1, Conditional IDS only needs to optimize over probability distributions. As proved by Russo & Van Roy (2018, Proposition 6), conditional IDS suffices to randomize over two actions at each round and the computational complexity is further improved to $O(|\mathcal{A}|)$ by Kirschner (2021, Lemma 2.7). However, contextual IDS in Section 4.2 needs to optimize over a mapping from the context space to action space at each round which in general is much more computationally demanding.

To obtain a practical and scalable algorithm, we approximate the contextual IDS using Actor-Critic (Konda & Tsitsiklis, 2000). We parametrize the policy class by a neural network and optimize the information ratio by multi-steps stochastic gradient decent.

Consider a parametrized policy class $\Pi_\theta = \{\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$. The policy, which is a conditional probability distribution π_θ, can be parameterized with a neural network. This neural network maps (deterministically) from a context s to a probability distribution over $\mathcal{A}$. We further parametrize the critic which is the reward function by a value network Q_θ.

To avoid additional approximation errors, we assume we can sample from the true context distribution. At each round, the parametrized contextual IDS minimizes the following empirical MIR loss through SGD:

$$\min_{\pi \in \Pi_\theta} \frac{\sum_{i=1}^{w}[\Delta_t(s_t^{(i)})^\top \pi(\cdot|s_t^{(i)})]^2}{\sum_{i=1}^{w}[\mathbb{I}_t(\pi^*)^\top \pi(\cdot|s_t^{(i)})]}\,, \tag{7.1}$$

where $\{s_t^{(1)}, \ldots, s_t^{(w)}\}$ are the independent samples of contexts. We use the Epistemic Neural Networks (ENN) (Osband et al., 2021a) to quantify the posterior uncertainty of the value network. For each given $s_t^{(i)}$, following the procedure in Section 6.3.3 and 6.3.4 of Lu et al. (2021), we

can approximate the one-step expected regret and the information gain efficiently using the samples outputted by ENN.

8. Experiment

We conduct some preliminary experiments to evaluate the parametrized contextual IDS through a neural network contextual bandit.

At each round t, the environment independently generates an observation in the form of d-dimensional contextual vector x_t from some distributions. Let $f_{\theta^*} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{A}|}$ be a neural network link function. When the agent takes action a, she will receive a reward in the form of $Y_{t,a} = [f_{\theta^*}(x_t)]_a + \eta_{t,a}$. This is the not sharing parameter formulation for contextual bandits which means each arm has its own parameter.

We set the generative model f_{θ^*} being a 2-hidden-layer ReLU MLP with 10 hidden neurons. The number of actions is 5. The contextual vector $x_t \in \mathbb{R}^{10}$ is sampled from $N(0, I_{10})$ and the noise is sampled from standard Gaussian distribution.

We compare contextual IDS with conditional IDS and Thompson sampling. For a fair comparison, we use the same ENN architecture to obtain posterior samples for three agents. As reported by Osband et al. (2021b), ensemble sampling with randomized prior function tends to be the best ENN that balances the computation and accuracy so we use 10 ensembles in our experiment. With 200 posterior samples, we use the same way described by Lu et al. (2021) to approximate the one-step regret and information gain for both conditional IDS and contextual IDS. We sample 20 independent contextual vectors at each round. Both the policy network and value network are using 2-hidden-layer ReLU MLP with 20 hidden neurons and optimized by Adam with learning rate 0.001. We report our results in Figure 1 where parametrized contextual IDS achieves reasonably good regret.

9. Discussion and Future Work

In this work, we investigate the right form of information ratio for contextual bandits and emphasize the importance of utilizing context distribution information through contextual bandits with graph feedback and sparse linear contextual bandits. For linear contextual bandits with moderately small number of actions, one future work is to see if contextual IDS can achieve $O(\sqrt{dn\log(k)})$ regret bound.

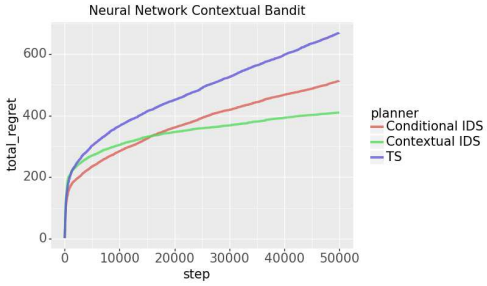

Figure 1. Cumulative regret for a neural network contextual bandit.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pp. 1–9, 2012.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pp. 23–35. PMLR, 2015.

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Arumugam, D. and Van Roy, B. The value of information when deciding what to learn. *Advances in Neural Information Processing Systems*, 34, 2021.

Bastani, H. and Bayati, M. Online decision making with high-dimensional covariates. *Operations Research*, 68 (1):276–294, 2020.

Dann, C., Mansour, Y., Mohri, M., Sekhari, A., and Sridharan, K. Reinforcement learning with feedback graphs. *Advances in Neural Information Processing Systems*, 33, 2020.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Hao, B., Lattimore, T., and Wang, M. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

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

Hao, B., Lattimore, T., Szepesvári, C., and Wang, M. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 316–324. PMLR, 2021b.

Kim, G.-S. and Paik, M. C. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pp. 5869–5879, 2019.

Kirschner, J. *Information-Directed Sampling-Frequentist Analysis and Applications*. PhD thesis, ETH Zurich, 2021.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

Kirschner, J., Lattimore, T., and Krause, A. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pp. 2328–2369. PMLR, 2020a.

Kirschner, J., Lattimore, T., Vernade, C., and Szepesvári, C. Asymptotically optimal information-directed sampling. *arXiv preprint arXiv:2011.05944*, 2020b.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014, 2000.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lattimore, T. and György, A. Mirror descent and the information ratio. *arXiv preprint arXiv:2009.12228*, 2020.

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

Lattimore, T., Crammer, K., and Szepesvári, C. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pp. 964–972, 2015.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Liu, F., Buccapatnam, S., and Shroff, N. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a.

Liu, F., Zheng, Z., and Shroff, N. Analysis of thompson sampling for graphical bandits without the graphs. *arXiv preprint arXiv:1805.08930*, 2018b.

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*, 2021.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24:684–692, 2011.

Oh, M.-h., Iyengar, G., and Zeevi, A. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pp. 8271–8280. PMLR, 2021.

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

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Hao, B., Ibrahimi, M., Lawson, D., Lu, X., O'Donoghue, B., and Van Roy, B. Evaluating predictive distributions: Does bayesian deep learning work? *arXiv preprint arXiv:2110.04629*, 2021b.

Ren, Z. and Zhou, Z. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*, 2020.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

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

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN*, 2020.

Sivakumar, V., Wu, Z. S., and Banerjee, A. Structured linear contextual bandits: A sharp and geometric smoothed analysis. *International Conference on Machine Learning*, 2020.

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.

Wang, X., Wei, M., and Yao, T. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pp. 5200–5208, 2018.

Wang, Y., Chen, Y., Fang, E. X., Wang, Z., and Li, R. Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *arXiv preprint arXiv:2009.02003*, 2020.

A. General Results

A.1. Proof of Theorem 4.4

Proof. From the definitions in Eqs. (3.1) and (4.1), we could decompose the Bayesian regret bound of contextual IDS as

$$\mathfrak{BR}(n;\pi^{\text{MIR}}) = n\alpha + \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_{s_t}\left[(\Delta_t(s_t)-\alpha)^\top \pi_t(\cdot|s_t)\right]\right]. \tag{A.1}$$

Recall that at each round contextual IDS follows

$$\pi_t = \operatorname*{argmin}_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \Psi^2_{t,\alpha}(\pi) = \operatorname*{argmin}_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \frac{\max\left(0,(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi(\cdot|s_t)]-\alpha)\right)^2}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi(\cdot|s_t)]}.$$

Moreover, we define

$$q_{t,\lambda} = \operatorname*{argmin}_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \Psi^\lambda_{t,\alpha}(\pi) = \operatorname*{argmin}_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \frac{\max\left(0,(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi(\cdot|s_t)]-\alpha)\right)^\lambda}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi(\cdot|s_t)]}.$$

Suppose for a moment that $\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha>0$. Let M be the number of contexts. Then the derivative can be written as

$$\nabla_{\pi(\cdot|m)}\Psi^2_{t,\alpha}(\pi_t) = \frac{2p_m\Delta_t(m)(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi(\cdot|s_t)]-\alpha)}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi(\cdot|s_t)]} - \frac{p_m\mathbb{I}_t(m)\left(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi(\cdot|s_t)]-\alpha\right)^2}{\left(\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi(\cdot|s_t)]\right)^2},$$

where p_m is the arrival probability for context m. Using the first-order optimality condition for each $m\in[M]$,

$$\mathbb{E}_{s_t}\left[\langle\nabla_{\pi(\cdot|s_t)}\Psi_{t,2}(\pi_t), q_{t,\lambda}(\cdot|s_t)-\pi_t(\cdot|s_t)\rangle\right] = \sum_{m=1}^{M} p_m\langle\nabla_{\pi(\cdot|m)}\Psi^2_{t,\alpha}(\pi_t), q_{t,\lambda}(\cdot|m)-\pi_t(\cdot|m)\rangle \geq 0.$$

This implies

$$\begin{aligned}
&\frac{2\mathbb{E}_{s_t}\left[\Delta_t(s_t)^\top(q_{t,\lambda}(\cdot|s_t)-\pi_t(\cdot|s_t))\right](\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha)}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi_t(\cdot|s_t)]}\\
&\geq \frac{\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top(q_{t,\lambda}(\cdot|s_t)-\pi_t(\cdot|s_t))\right]\left(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha\right)^2}{\left(\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi_t(\cdot|s_t)]\right)^2}.
\end{aligned}$$

Since we assume $\mathbb{E}_{s_t}[\Delta_t(s_t)^\top(\pi_t(\cdot|s_t)]-\alpha>0$ and the information gain is always non-negative, we have

$$2\mathbb{E}_{s_t}\left[\Delta_t(s_t)^\top(q_{t,\lambda}(\cdot|s_t)-\pi_t(\cdot|s_t))\right] \geq \frac{\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top(q_{t,\lambda}(\cdot|s_t)-\pi_t(\cdot|s_t))\right]\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha)}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi_t(\cdot|s_t)]}.$$

which implies

$$\begin{aligned}
2\left(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top q_{t,\lambda}(\cdot|s_t)]-\alpha\right) &\geq \left(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha\right)\left(1+\frac{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top q_{t,\lambda}(\cdot|s_t)]}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi_t(\cdot|s_t)]}\right)\\
&\geq \mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha.
\end{aligned}$$

Putting the above results together,

$$\begin{aligned}
\frac{(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha)^\lambda}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi(\cdot|s_t)]} &= \frac{(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha)^2(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi(\cdot|s_t)]-\alpha)^{\lambda-2}}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi(\cdot|s_t)]}\\
&\leq \frac{2^{\lambda-2}(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top\pi_t(\cdot|s_t)]-\alpha)^2(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top q_{t,\lambda}(\cdot|s_t)]-\alpha)^{\lambda-2}}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top\pi_t(\cdot|s_t)]}\\
&\leq \frac{2^{\lambda-2}(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top q_{t,\lambda}(\cdot|s_t)]-\alpha)^2(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top q_{t,\lambda}(\cdot|s_t)]-\alpha)^{\lambda-2}}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top q_{t,\lambda}(\cdot|s_t)]}\\
&= \frac{2^{\lambda-2}(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top q_{t,\lambda}(\cdot|s_t)]-\alpha)^\lambda}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top q_{t,\lambda}(\cdot|s_t)]}.
\end{aligned}$$

Recall that $\mathcal{I}_{\alpha,\lambda}$ is the worst-case information ratio such that for any $t \in [n]$, $\Psi_{t,\alpha}^{\lambda}(\pi_t) \leq \mathcal{I}_{\alpha,\lambda}$ almost surely. Therefore,

$$\mathbb{E}\left[\Delta_t(s_t)^\top \pi_t(\cdot|s_t) - \alpha\right] \leq 2^{1-2/\lambda}\left(\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right)^{1/\lambda} \mathcal{I}_{\alpha,\lambda}^{1/\lambda}\,, \tag{A.2}$$

which is obvious when $\mathbb{E}[\Delta_t(s_t)^\top(\pi_t(\cdot|s_t)] - \alpha \leq 0$. Combining Eqs. (A.1) and (A.2) together and using Holder's inequality, we obtain

$$\mathfrak{BR}(n;\pi^{\text{IDS}}) \leq n\alpha + 2^{1-2/\lambda}\mathcal{I}_{\alpha,\lambda}^{1/\lambda} n^{1-1/\lambda}\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right]^{1/\lambda}.$$

This ends the proof. □

B. Contextual Bandits with Graph Feedback

B.1. Proof of Lemma 5.5

Proof. We bound the squared information ratio in terms of independence number $\beta(\mathcal{G})$. Let $\pi_t^{\text{TS}}(\cdot|s_t) = \mathbb{P}_t(a_t^* = \cdot)$ and consider a mixture policy $\pi_t^{\text{mix}} = (1-\epsilon)\pi_t^{\text{TS}} + \epsilon/k$ with some mixing parameter $\epsilon \in [0,1]$ that will be determined later.

First, we will derive an upper bound for one-step regret. From the definition of one-step regret,

$$\begin{aligned}\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t) &= \sum_{a\in\mathcal{A}_t}\pi_t^{\text{mix}}(a|s_t)\mathbb{E}_t\left[Y_{t,a_t^*} - Y_{t,a}|s_t\right] \\ &= (1-\epsilon)\sum_{a\in\mathcal{A}_t}\pi_t^{\text{TS}}(a|s_t)\mathbb{E}_t\left[Y_{t,a_t^*} - Y_{t,a}|s_t\right] + \epsilon\sum_{a\in\mathcal{A}_t}\frac{1}{k}\mathbb{E}_t\left[Y_{t,a_t^*} - Y_{t,a}|s_t\right].\end{aligned} \tag{B.1}$$

Recall that $R_{\max}$ is the upper bound of maximum expected reward and let $d_t(a,a') = D_{\text{KL}}(\mathbb{P}_t(Y_{t,a}\in\cdot|a_t^*=a')||\mathbb{P}_t(Y_{t,a}\in\cdot)))$. It is easy to see $Y_{t,a}$ is a $\sqrt{R_{\max}^2+1}$ sub-Gaussian random variable. For the first term of Eq. (B.1), according to Lemma 3 in Russo & Van Roy (2014), we have

$$\sum_{a\in\mathcal{A}_t}\pi_t^{\text{TS}}(a|s_t)\mathbb{E}_t\left[Y_{t,a_t^*} - Y_{t,a}|s_t\right] \leq \sum_{a\in\mathcal{A}_t}\pi_t^{\text{TS}}(a|s_t)\sqrt{\frac{R_{\max}^2+1}{2}d_t(a,a)}\,. \tag{B.2}$$

For the second term of Eq. (B.1), we directly bound it by

$$\epsilon\sum_{a\in\mathcal{A}_t}\frac{1}{k}\mathbb{E}_t\left[Y_{t,a_t^*} - Y_{t,a}|s_t\right] \leq 2\epsilon R_{\max}\,. \tag{B.3}$$

Putting Eqs. (B.1)-(B.3) together, we have

$$\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t) - 2\epsilon R_{\max} \leq \sum_{a\in\mathcal{A}_t}\pi_t^{\text{TS}}(a|s_t)\sqrt{\frac{R_{\max}^2+1}{2}d_t(a,a)}\,.$$

Second, we will derive a lower bound for the information gain. Let $E = (a_{ij})_{1\leq i,j\leq|\mathcal{A}|}$ be the adjacency matrix that represents the graph feedback structure $\mathcal{G}$. Then $a_{ij} = 1$ if there exists an edge $(i,j)\in\mathcal{E}$ and $a_{ij} = 0$ otherwise. Since for any $a_i, a_j \in N^{\text{out}}(a)$, Y_{t,a_i} and Y_{t,a_j} are mutually independent,

$$\mathbb{I}_t\left(\pi^*; (Y_{t,a'})_{a'\in N^{\text{out}}(a)}\right) \geq \sum_{a'\in N^{\text{out}}(a)}\mathbb{I}_t(\pi^*; Y_{t,a'}) = \sum_{a'\in\mathcal{A}_t}E(a,a')\mathbb{I}(\pi^*; Y_{t,a'})\,.$$

Therefore,

$$\begin{aligned}
\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t) &\geq \sum_{a\in\mathcal{A}_t} \mathbb{I}_t\left(\pi^*; (Y_{t,a'})_{a'\in N^{\text{out}}(a)}\right) \pi_t^{\text{mix}}(a|s_t) \\
&\geq \sum_{a\in\mathcal{A}_t}\sum_{a'\in\mathcal{A}_t} E(a,a')\mathbb{I}(\pi^*; Y_{t,a'})\pi_t^{\text{mix}}(a|s_t) \\
&= \sum_{a\in\mathcal{A}_t}\left(\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)} \pi_t^{\text{mix}}(a'|s_t)\right)\mathbb{I}(\pi^*; Y_{t,a}) \\
&= \sum_{a\in\mathcal{A}_t}\left(\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)} \pi_t^{\text{mix}}(a'|s_t)\right)\sum_{a'\in\mathcal{A}_t} \pi_t^{\text{TS}}(a'|s_t) d_t(a,a') \\
&\geq \sum_{a\in\mathcal{A}_t}\left(\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)} \pi_t^{\text{mix}}(a'|s_t)\right)\pi_t^{\text{TS}}(a|s_t) d_t(a,a) \\
&= \sum_{a\in\mathcal{A}_t} \frac{\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)} \pi_t^{\text{mix}}(a'|s_t)}{\pi_t^{\text{TS}}(a|s_t)} \pi_t^{\text{TS}}(a|s_t)^2 d_t(a,a)\,.
\end{aligned}$$

Let's denote

$$U_{a,t} = \frac{\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}_{s_t}(a)} \pi_t^{\text{mix}}(a'|s_t)}{\pi_t^{\text{TS}}(a|s_t)}\,.$$

Putting the above together,

$$\begin{aligned}
\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t) - 2\epsilon R_{\max} &\leq \sqrt{\frac{R_{\max}^2+1}{2}}\sqrt{\sum_{a\in\mathcal{A}_t}\frac{1}{U_{a,t}}}\sqrt{\sum_{a\in\mathcal{A}_t} U_{a,t}\pi_t^{\text{TS}}(a|s_t)^2 d_t(a,a)} \\
&\leq \sqrt{\frac{R_{\max}^2+1}{2}}\sqrt{\sum_{a\in\mathcal{A}_t}\frac{1}{U_{a,t}}}\sqrt{\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)}\,,
\end{aligned}$$

where the first inequality is due to Cauchy–Schwarz inequality. This implies

$$\begin{aligned}
\frac{\left(\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t) - 2\epsilon R_{\max}\right)^2}{\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)} &\leq \frac{R_{\max}^2+1}{2}\sum_{a\in\mathcal{A}_t}\frac{\pi_t^{\text{TS}}(a|s_t)}{\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)}\pi_t^{\text{mix}}(a'|s_t)} \\
&\leq \frac{R_{\max}^2+1}{2(1-\epsilon)}\sum_{a\in\mathcal{A}_t}\frac{\pi_t^{\text{mix}}(a|s_t)}{\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)}\pi_t^{\text{mix}}(a'|s_t)}\,,
\end{aligned}$$

where we use $\pi_t^{\text{mix}} \geq (1-\epsilon)\pi_t^{\text{TS}}$. Using Jenson's inequality,

$$\begin{aligned}
\frac{\left(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t)] - 2\epsilon R_{\max}\right)^2}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)]} &\leq \mathbb{E}_{s_t}\left[\frac{(\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t) - \alpha)^2}{\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)}\right] \\
&\leq \mathbb{E}_{s_t}\left[\frac{R_{\max}^2+1}{2(1-\epsilon)}\sum_{a\in\mathcal{A}_t}\frac{\pi_t^{\text{mix}}(a|s_t)}{\pi_t^{\text{mix}}(a|s_t) + \sum_{a'\in N^{\text{in}}(a)}\pi_t^{\text{mix}}(a'|s_t)}\right]\,.
\end{aligned}$$

Next we restate Lemma 5 from Alon et al. (2015) to bound the right hand side.

Lemma B.1. *Let $\mathcal{G} = (\mathcal{A}, \mathcal{E})$ be a directed graph with $|\mathcal{A}| = k$. Assume a distribution π over $\mathcal{A}$ such that $\pi(i) \geq \eta$ for all $i \in \mathcal{A}$ with some constant $0 < \eta < 0.5$. Then we have*

$$\sum_{i\in\mathcal{A}}\frac{\pi(i)}{\pi(i) + \sum_{j\in N^{in}(i)}\pi(j)} \leq 4\beta(G)\log\left(\frac{4k}{\beta(\mathcal{G})\eta}\right)\,.$$

Using Lemma B.1 with $\eta = \epsilon/k$,

$$\frac{(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t)] - 2\epsilon R_{\max})^2}{\mathbb{E}_{s_t}[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)]} \leq \frac{2R_{\max}^2 + 2}{1-\epsilon} 2\beta(\mathcal{G}) \log\left(\frac{4k^2}{\beta(\mathcal{G})\alpha}\right).$$

According to the definition of contextual IDS, this ends the proof. □

B.2. Proof of Lemma 5.7

Proof. We bound the worst-case marginal information ratio with $\lambda = 3$. Define an exploratory policy $\mu : \mathcal{S} \to \mathcal{A}$ such that

$$\mu = \operatorname*{argmax}_{\pi} \min_{d} \mathbb{E}_{s,a\sim\pi(\cdot|s)} \left[\mathbb{1}\left\{d \in N^{\text{out}}(a)\right\}\right].$$

Consider a mixture policy $\pi_t^{\text{mix}} = (1-\epsilon)\pi_t^{\text{TS}} + \epsilon\mu$ for some mixing parameter $\epsilon \in [0,1]$. Since for any $a_i, a_j \in N^{\text{out}}(a)$, Y_{t,a_i} and Y_{t,a_j} are mutually independent, we have

$$\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right] = \mathbb{E}_{s_t,a\sim\pi_t^{\text{mix}}(\cdot|s_t)}\left[\mathbb{I}_t\left(\pi^*; (Y_{t,a'})_{a'\in N^{\text{out}}(a)}\right)\right] \geq \mathbb{E}_{s_t,a\sim\pi_t^{\text{mix}}(\cdot|s_t)}\left[\sum_{a'\in N^{\text{out}}(a)} \mathbb{I}_t\left(\pi^*; Y_{t,a'}\right)\right].$$

From the definition of the mixture policy, $\pi^{\text{mix}}(a|s_t) \geq \epsilon\mu(a|s_t)$ for any $a \in \mathcal{A}_t$. Then we have

$$\begin{aligned}
\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right] &\geq \epsilon\mathbb{E}_{s_t,a\sim\mu(\cdot|s_t)}\left[\sum_{a'\in N^{\text{out}}(a)} \mathbb{I}_t\left(\pi^*; Y_{t,a'}\right)\right] \\
&\geq \epsilon\mathbb{E}_{s_t,a\sim\mu(\cdot|s_t)}\left[\sum_{a'\in N^{\text{out}}(a)} \mathbb{I}_t\left(\pi^*; Y_{t,a'}\right)\mathbb{1}\left\{a' \in N^{\text{out}}(a)\right\}\right] \\
&\geq \epsilon\mathbb{E}_{s_t,a\sim\mu(\cdot|s_t)}\left[\sum_{a'\in \mathcal{A}_t} \mathbb{I}_t\left(\pi^*; Y_{t,a'}\right)\mathbb{1}\left\{a' \in N^{\text{out}}(a)\right\}\right].
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right] &\geq \epsilon\mathbb{E}_{s_t,a\sim\mu(\cdot|s_t)}\left[\sum_{a'\in \mathcal{A}_t} \mathbb{I}_t(\pi^*; Y_{t,a'})\right] \min_{a}{}' \mathbb{E}_{s_t,a\sim\mu(\cdot|s_t)}\left[\mathbb{1}\left\{a' \in N^{\text{out}}(a)\right\}\right] \\
&\geq \epsilon\vartheta(\mathcal{G},\xi)\mathbb{E}_{s_t}\left[\sum_{a'\in \mathcal{A}_t} \mathbb{I}_t(\pi^*; Y_{t,a'})\right] \\
&\geq \frac{2\epsilon\vartheta(\mathcal{G},\xi)}{R_{\max}^2+1}\mathbb{E}_{s_t}\left[\sum_{\zeta}\mathbb{P}_t(\pi^*=\zeta)\sum_{a\in\mathcal{A}_t}\left(\mathbb{E}_t[Y_{t,a}|\pi^*=\zeta] - \mathbb{E}_t[Y_{t,a}]\right)^2\right],
\end{aligned} \tag{B.4}$$

where the last inequality is from Pinsker's inequality.

On the other hand, by the definition of mixture policy and Jenson's inequality,

$$\begin{aligned}
\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t)] &\leq (1-\epsilon)\mathbb{E}_{s_t,a\sim\pi_t^{\text{TS}}(\cdot|s_t)}\mathbb{E}_t[Y_{t,a_t^*} - Y_{t,a}] + 2\epsilon R_{\max} \\
&\leq \mathbb{E}_{s_t,a\sim\pi_t^{\text{TS}}(\cdot|s_t)}\left[\mathbb{E}_t[Y_{t,a}|a_t^*=a] - \mathbb{E}_t[Y_{t,a}]\right] + 2\epsilon R_{\max} \\
&\leq \sqrt{\mathbb{E}_{s_t}\left[\sum_{a\in\mathcal{A}_t}\mathbb{P}_t(a_t^*=a)\left(\mathbb{E}_t[Y_{t,a}|a_t^*=a] - \mathbb{E}_t[Y_{t,a}]\right)^2\right]} + 2\epsilon R_{\max} \\
&= \sqrt{\mathbb{E}_{s_t}\left[\sum_{\zeta}\mathbb{P}_t(\pi^*=\zeta)\sum_{a\in\mathcal{A}_t}\left(\mathbb{E}_t[Y_{t,a}|\pi^*=\zeta] - \mathbb{E}_t[Y_{t,a}]\right)^2\right]} + 2\epsilon R_{\max}.
\end{aligned}$$

Combining with the lower bound of the information gain in Eq. (B.4),

$$\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t)] \le \sqrt{\frac{R_{\max}^2+1}{2\epsilon\vartheta(\mathcal{G},\xi)}\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right]} + 2\epsilon R_{\max}\,.$$

By choosing

$$\epsilon = \left(\frac{R_{\max}^2+1}{R_{\max}^2}\frac{1}{8\vartheta(\mathcal{G},\xi)}\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right]\right)^{1/3},$$

we reach

$$\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t)] \le 2R_{\max}\left(\frac{R_{\max}^2+1}{R_{\max}^2}\frac{1}{8\vartheta(\mathcal{G},\xi)}\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right]\right)^{1/3}.$$

This implies

$$\frac{\left(\mathbb{E}_{s_t}[\Delta_t(s_t)^\top \pi_t^{\text{mix}}(\cdot|s_t)]\right)^3}{\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right]} \le \frac{R_{\max}^3+R_{\max}}{\vartheta(\mathcal{G},\xi)}\,.$$

This ends the proof. □

B.3. Proof of Lemma 5.8

Proof. According the the definition of cumulative information gain,

$$\mathbb{E}\left[\sum_{t=1}^n \mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right] = \mathbb{E}_{t,s_t}[\mathbb{I}_t(\pi^*; O_t|A_t,s_t)]\,. \tag{B.5}$$

Note that

$$\mathbb{I}_t(\pi^*;(s_t,A_t,O_t)) = \mathbb{E}_{t,s_t}[\mathbb{I}_t(\pi^*;O_t|A_t,s_t)] + \mathbb{E}_{t,s_t}[\mathbb{I}_t(\pi^*;A_t|s_t)] + \mathbb{I}_t(\pi^*;s_t) = \mathbb{E}_{t,s_t}[\mathbb{I}_t(\pi^*;O_t|A_t,s_t)]\,.$$

By the chain rule,

$$\mathbb{I}(\pi^*;\mathcal{F}_{n+1}) = \sum_{t=1}^n \mathbb{E}\left[\mathbb{I}_t(\pi^*;(s_t,A_t,O_t))\right] = \sum_{t=1}^n \mathbb{E}\left[\mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right]\,.$$

On the other hand,

$$\mathbb{I}(\pi^*;\mathcal{F}_{n+1}) \le \mathbb{H}(\pi^*) \le M\log(k)\,,$$

where $\mathbb{H}(\cdot)$ is the entropy. Putting the above together,

$$\mathbb{E}\left[\sum_{t=1}^n \mathbb{E}_{s_t}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right] \le M\log(k).$$

This ends the proof. □

Remark B.2. *We analyze the upper bound $\mathbb{H}(\pi^*)$ under the following examples. Denote the number of contexts by M, which could be infinite.*

1. *The possible choices of π^* is at most k^M, and thus $\mathbb{H}(\pi^*) \le \log(k^M) = M\log(k)$. When $M=\infty$, this bound is vacuous.*
2. *We can divide the context space into fixed and disjoint clusters such that for each parameter in the parameter space, the optimal policy maps all the contexts in a single cluster to a single action. For example, a naive partition is that each cluster contains a single context, and in this case, the number of clusters denoted equals the number of contexts M. Let P be the minimum number of such clusters, so $P \le M$. The possible choice of π^* is k^P, and thus $\mathbb{H}(\pi^*) \le \log(k^P) = P\log(k)$. When $M=\infty$ but $P<\infty$, we obtain a valid upper bound instead of ∞.*

B.4. Proof of Theorem 5.9

Proof. According to Theorem 4.4, we have

$$\mathfrak{BR}(n;\pi^{\text{MIR}}) \le 2R_{\max}\sqrt{n} + \inf_{\lambda\ge 2} 2^{1-2/\lambda}\mathcal{I}^{1/\lambda}_{2R_{\max}/\sqrt{n},\lambda} n^{1-1/\lambda}\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top \pi_t(\cdot|s_t)\right]\right]^{1/\lambda}.$$

Using Lemma 5.5 with $\epsilon = 1/\sqrt{n}$, we have

$$\mathcal{I}_{2R_{\max}/\sqrt{n},2} \le \frac{2R_{\max}^2+2}{1-1/\sqrt{n}} 2\beta(\mathcal{G})\log\left(\frac{4k^2\sqrt{n}}{\beta(\mathcal{G})}\right) \le 16(R_{\max}^2+1)\beta(\mathcal{G})\log\left(\frac{4k^2\sqrt{n}}{\beta(\mathcal{G})}\right).$$

Combining Lemmas 5.7-5.8 together, we have

$$\begin{aligned}
&\mathfrak{BR}(n;\pi^{\text{MIR}}) \le 2R_{\max}\sqrt{n} + \\
&\min\left(\left(16(R_{\max}^2+1)\beta(\mathcal{G})\log\left(\frac{4k^2\sqrt{n}}{\beta(\mathcal{G})}\right)nM\log(k)\right)^{1/2}, \left(\frac{R_{\max}^3+R_{\max}}{\vartheta(\mathcal{G},\xi)}2M\log(k)\right)^{\frac{1}{3}} n^{\frac{2}{3}}\right) \\
&\ge CR_{\max}\min\left(\sqrt{\beta(\mathcal{G})\log\left(\frac{4k^2\sqrt{n}}{\beta(\mathcal{G})}\right)nM\log(k)}, \left(\frac{2M\log(k)}{\vartheta(\mathcal{G},\xi)}\right)^{\frac{1}{3}} n^{\frac{2}{3}}\right),
\end{aligned}$$

for some absolute constant $C > 0$. This ends the proof. □

C. Sparse Linear Contextual Bandits

C.1. Proof of Theorem 6.3

Proof. We assume there are two available contexts. Context set 1 consists of a single action $x_0 = (0,\dots,0)^\top \in \mathbb{R}^{d+1}$ and an informative action set $\mathcal{H}$ as follows:

$$\mathcal{H} = \left\{x \in \mathbb{R}^{d+1} \middle| x_j \in \{-\kappa,\kappa\} \text{ for } j\in[d], x_d = 1\right\}, \tag{C.1}$$

where $0 < \kappa \le 1$ is a constant. Let $d = sp$ for some integer $p \ge 2$. Context set 2 consists of a multi-task bandit action set $\mathcal{A} = \{(\{e_i \in \mathbb{R}^p : i\in[p]\}^s, 0)\} \subset \mathbb{R}^{d+1}$ whose element's last coordinate is always 0. The arriving probability of each context is 1/2. One can verify for this feature set, the explorability constant $C_{\min}(\phi,\xi) = 1/2$ achieved by a policy that uniformly samples from $\mathcal{H}$ when context set 1 arrives.

Fix a conditional IDS policy π^{CIR}. Let $\Delta > 0$ and $\Theta = \{\Delta e_i : i\in[p]\} \subset \mathbb{R}^p$. Given $\theta \in \{(\Theta^s, 1)\} \subset \mathbb{R}^{d+1}$ and $i \in [s]$, let $\theta^{(i)} \in \mathbb{R}^p$ be defined by $\theta^{(i)}_k = \theta_{(i-1)s+k}$, which means that

$$\theta^\top = [\theta^{(1)\top},\dots,\theta^{(s)\top},1].$$

Assume the prior of θ^* is uniformly distributed over $\{(\Theta^s,1)\}$. Define the cumulative regret of policy π interacting with bandit θ as

$$\begin{aligned}
R_\theta(n;\pi) &= \sum_{t=1}^{n}\mathbb{E}_\theta\left[\langle x^*_{s_t},\theta\rangle - Y_t\right] \\
&= \sum_{t=1}^{n}\mathbb{E}_\theta\left[\left(\langle x^*_{s_t},\theta\rangle - Y_t\right)\mathbb{1}(s_t=1)\right] + \sum_{t=1}^{n}\mathbb{E}_\theta\left[\left(\langle x^*_{s_t},\theta\rangle - Y_t\right)\mathbb{1}(s_t=2)\right],
\end{aligned} \tag{C.2}$$

where we write $x^*_{s_t} = \phi(s_t, a^*_t)$ for short. Therefore,

$$\mathfrak{BR}(n;\pi) = \frac{1}{|\Theta|^s}\sum_{\theta\in\{(\Theta^s,1)\}} R_\theta(n;\pi).$$

Note that when context set 1 arrives, the action from $\mathcal{H}$ suffers at least $1-s\Delta$ regret and thus since x_0 is always the optimal action for context set 1. From the definition of conditional IDS in Eq. (4.2), when context set 1 is arriving and $s\Delta < 1$, conditional IDS will always pull x_0 for this prior. That means conditional IDS will suffer no regret for context set 1 and implies for any $\theta \in \{(\Theta^s, 1)\}$,

$$\sum_{t=1}^{n} \mathbb{E}_\theta \left[\left(\langle x_{s_t}^*, \theta \rangle - Y_t \right) \mathbb{I}(s_t = 1) \right] = \sum_{t=1}^{n} \mathbb{E}_\theta \left[\langle x_1^*, \theta \rangle - \langle \pi^{\text{CIR}}(i|1), \theta \rangle | s_t = 1 \right] \mathbb{P}(s_t = 1) = 0 \,.$$

It remains to bound the second term in Eq. (C.2). It essentially follows the proof of Theorem 24.3 in Lattimore & Szepesvári (2020). From the proof of multi-task bandit lower bound, we have

$$\frac{1}{|\Theta|^s} \sum_{\theta \in \{(\Theta^s, 1)\}} \sum_{t=1}^{n} \mathbb{E}_\theta \left[\left(\langle x_{s_t}^*, \theta \rangle - Y_t \right) \mathbb{1}(s_t = 2) \right] \geq \frac{1}{16} \sqrt{dsn} \,.$$

This ends the proof. □

C.2. Proof of Lemma 6.4

Proof. If one can derive a worst-case bound of $\Psi_{t,\alpha}^{\lambda}(\widetilde{\pi})$ for a particular policy $\widetilde{\pi}$,, we can have an upper bound for $\mathcal{I}_{\alpha,\lambda}$ automatically. The remaining step is to choose a proper policy $\widetilde{\pi}$ for $\lambda = 2, 3$ separately.

First, we bound the information ratio with $\lambda = 2$. By the definition of mutual information, for any $a \in \mathcal{A}_t$, we have

$$\mathbb{I}_t(\pi^*; Y_{t,a}) = \sum_{\zeta \in \mathcal{A}^1 \times \cdots \times \mathcal{A}^M} \mathbb{P}_t(\pi^* = \zeta) D_{\text{KL}} \left(\mathbb{P}_t(Y_{t,a} = \cdot | \pi^* = \zeta) || \mathbb{P}_t(Y_{t,a} = \cdot) \right). \tag{C.3}$$

et $(\Omega, \mathcal{F})$ be a measurable space and let $P, Q : \mathcal{F} \to [0, 1]$ be two probability measures. Based on Pinsker's inequality,

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} (P(A) - Q(A)) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P||Q)}.$$

Let $a < b$ and $X : \Omega \to [a, b]$. Exercise 14.4 in (Lattimore & Szepesvári, 2020) indicates that

$$\int_\Omega X(w) dP(w) - \int_\Omega X(w) dQ(w) \leq (b-a)\delta(P, Q) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P||Q)}.$$

Overall,

$$D_{\text{KL}}(P||Q) \geq \frac{2}{(b-a)^2} \left(\int_\Omega X(w) dP(w) - \int_\Omega X(w) dQ(w) \right).$$

Recall that $R_{\max}$ is the upper bound of maximum expected reward. It is easy to see $Y_{t,a}$ is a $\sqrt{R_{\max}^2 + 1}$ sub-Gaussian random variable. According to Lemma 3 in Russo & Van Roy (2014), we have

$$\mathbb{I}_t(\pi^*; Y_{t,a}) \geq \frac{2}{R_{\max}^2 + 1} \sum_{\zeta \in \mathcal{A}^1 \times \cdots \times \mathcal{A}^M} \mathbb{P}_t(\pi^* = \zeta) \Big(\mathbb{E}_t[Y_{t,a} | \pi^* = \zeta] - \mathbb{E}_t[Y_{t,a}] \Big)^2. \tag{C.4}$$

The MIR of contextual IDS can be bounded by the MIR of Thompson sampling:

$$\mathcal{I}_{0,2} \leq \max_{t \in [n]} \frac{\max\left(0, (\mathbb{E}[\Delta_t(s_t)^\top \pi_t^{\text{TS}}(\cdot|s_t)])\right)^2}{\mathbb{E}[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{TS}}(\cdot|s_t)]} \leq \max_{t \in [n]} \mathbb{E}\left[\frac{(\Delta_t(s_t)^\top \pi_t^{\text{TS}}(\cdot|s_t))^2}{\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{TS}}(\cdot|s_t)} \right],$$

where the second inequality is from Jenson's inequality. Using the matrix trace rank trick described in Proposition 5 in Russo & Van Roy (2014), we have $\mathcal{I}_{0,2} \leq (R_{\max}^2 + 1)d/2$ in the end.

Second, we bound the information ratio with $\lambda = 3$. Let's define an exploratory policy μ such that

$$\mu = \operatorname*{argmax}_{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})} \sigma_{\min} \left(\mathbb{E}_{s \sim \xi} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} [\phi(s, a) \phi(s, a)^\top] \right] \right). \tag{C.5}$$

Consider a mixture policy $\pi_t^{\text{mix}} = (1 - \epsilon)\pi_t^{\text{TS}} + \epsilon\mu$ where the mixture rate $\epsilon \geq 0$ will be decided later.

Step 1: Bound the information gain According the lower bound of information gain in Eq. (C.4),

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top \pi_t^{\text{mix}}(\cdot|s_t)\right] \\
\geq &\frac{2}{(R_{\max}^2+1)}\mathbb{E}_{s_t\sim\xi,a\sim\pi_t^{\text{mix}}(\cdot|s_t)}\left[\sum_{\zeta\in\mathcal{A}^1\times\cdots\times\mathcal{A}^M}\mathbb{P}_t(\pi^*=\zeta)\left(\mathbb{E}_t[Y_{t,a}|\pi^*=\zeta]-\mathbb{E}_t[Y_{t,a}]\right)^2\right] \\
= &\frac{2}{(R_{\max}^2+1)}\mathbb{E}_{s_t\sim\xi,a\sim\pi_t^{\text{mix}}(\cdot|s_t)}\left[\sum_{\zeta\in\mathcal{A}^1\times\cdots\times\mathcal{A}^M}\mathbb{P}_t(\pi^*=\zeta)\left(\phi(s_t,a)^\top\mathbb{E}_t[\theta^*|\pi^*=\zeta]-\phi(s_t,a)^\top\mathbb{E}_t[\theta^*]\right)^2\right].
\end{aligned}$$

By the definition of the mixture policy, we know that $\pi_t^{\text{mix}}(a|s_t)\geq\epsilon\mu(a|s_t)$ for any $a\in\mathcal{A}_t$. Then we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right] \geq &\frac{2\epsilon}{R_{\max}^2+1}\sum_{\zeta\in\mathcal{A}^1\times\cdots\times\mathcal{A}^M}\mathbb{P}_t(\pi^*=\zeta) \\
\cdot\,\mathbb{E}_{s_t\sim\xi,a\sim\mu(\cdot|s_t)}&\left[(\mathbb{E}_t[\theta^*|\pi^*=\zeta]-\mathbb{E}_t[\theta^*])^\top\phi(s_t,a)\phi(s_t,a)^\top(\mathbb{E}_t[\theta^*|\pi^*=\zeta]-\mathbb{E}_t[\theta^*])\right].
\end{aligned}$$

From the definition of μ in Eq. (C.5), we have

$$\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right]\geq\frac{2\epsilon}{R_{\max}^2+1}\sum_{\zeta\in\mathcal{A}^1\times\cdots\times\mathcal{A}^M}\mathbb{P}_t(\pi^*=\zeta)C_{\min}(\phi,\xi)\left\|\mathbb{E}_t[\theta^*|\pi^*=\zeta]-\mathbb{E}_t[\theta^*]\right\|_2^2.$$

Step 2: Bound the instant regret We decompose the regret by the contribution from the exploratory policy and the one from TS:

$$\begin{aligned}
\mathbb{E}\left[\Delta_t(s_t)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right] = &(1-\epsilon)\mathbb{E}_{s_t}\left[\sum_a\mathbb{P}_t(a_t^*=a)\Big(\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\big|a_t^*=a\right]-\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\right]\Big)\right] \\
&+\epsilon\mathbb{E}_{s_t}\left[\sum_a\mathbb{E}_t\Big[\phi(s_t,a_t^*)^\top\theta^*-\phi(s_t,a)^\top\theta^*\Big]\mu(a|s_t)\right].
\end{aligned}\tag{C.6}$$

Since $R_{\max}$ is the upper bound of maximum expected reward, the second term can be bounded $2R_{\max}\epsilon$. Using Jenson's inequality, the first term can be bounded by

$$\begin{aligned}
&\mathbb{E}_{s_t}\left[\sum_a\mathbb{P}_t(a_t^*=a)\Big(\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\big|a_t^*=a\right]-\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\right]\Big)\right] \\
\leq &\sqrt{\mathbb{E}_{s_t}\left[\sum_a\mathbb{P}_t(a_t^*=a)\Big(\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\big|a_t^*=a\right]-\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\right]\Big)^2\right]}.
\end{aligned}$$

Since all the optimal actions are sparse, any action a with $\mathbb{P}_t(a_t^*=a)>0$ must be sparse. Then we have

$$\Big(\phi(s_t,a)^\top(\mathbb{E}_t[\theta^*|a_t^*=a]-\mathbb{E}_t[\theta^*])\Big)^2\leq s^2\left\|\mathbb{E}_t[\theta^*|a_t^*=a]-\mathbb{E}_t[\theta^*]\right\|_2^2,$$

for any action a with $\mathbb{P}_t(a_t^* = a) > 0$. This further implies

$$
\begin{aligned}
&\mathbb{E}_{s_t}\left[\sum_a \mathbb{P}_t(a_t^*=a)\Big(\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\big|a_t^*=a\right]-\mathbb{E}_t\left[\phi(s_t,a)^\top\theta^*\right]\Big)\right]\\
&\le\sqrt{\mathbb{E}_{s_t}\left[\sum_a \mathbb{P}_t(a_t^*=a)s^2\Big\|\mathbb{E}_t[\theta^*|a_t^*=a]-\mathbb{E}_t[\theta^*]\Big\|_2^2\right]}\\
&=\sqrt{\mathbb{E}_{s_t}\left[\sum_{\zeta\in\mathcal{A}^1\times\cdots\times\mathcal{A}^M} \mathbb{P}_t(\pi^*=\zeta)s^2\Big\|\mathbb{E}_t[\theta^*|\pi^*=\zeta]-\mathbb{E}_t[\theta^*]\Big\|_2^2\right]} \qquad \text{(C.7)}\\
&=\sqrt{\frac{s^2(R_{\max}^2+1)}{2\epsilon C_{\min}(\phi,\xi)}\frac{2\epsilon C_{\min}(\phi,\xi)}{(R_{\max}^2+1)}\mathbb{E}_{s_t}\left[\sum_{\zeta\in\mathcal{A}^1\times\cdots\times\mathcal{A}^M} \mathbb{P}_t(\pi^*=\zeta)s^2\Big\|\mathbb{E}_t[\theta^*|\pi^*=\zeta]-\mathbb{E}_t[\theta^*]\Big\|_2^2\right]}\\
&\le\sqrt{\frac{s^2(R_{\max}^2+1)}{2\epsilon C_{\min}(\phi,\xi)}\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right]}\,.
\end{aligned}
$$

Putting Eq. (C.6) and (C.7) together, we have

$$
\mathbb{E}\left[\Delta_t(s_t)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right]\le\sqrt{\frac{s^2(R_{\max}^2+1)}{2\epsilon C_{\min}(\phi,\xi)}\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right]}+2R_{\max}\epsilon\,.
$$

By optimizing the mixture rate ϵ, we have

$$
\mathcal{I}_{0,3}\le\frac{\left(\mathbb{E}\left[\Delta_t(s_t)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right]\right)^3}{\mathbb{E}\left[\mathbb{I}_t(\pi^*)^\top\pi_t^{\text{mix}}(\cdot|s_t)\right]}\le\frac{s^2(R_{\max}^2+1)}{8R_{\max}^2C_{\min}}\le\frac{s^2}{4C_{\min}(\phi,\xi)}\,.
$$

This ends the proof. □