Information Directed Sampling for Sparse Linear Bandits

Botao Hao
Deepmind
haobotao000@gmail.com

Tor Lattimore
Deepmind
lattimore@google.com

Wei Deng
Department of Mathematics
Purdue University
weideng056@gmail.com

Abstract

Stochastic sparse linear bandits offer a practical model for high-dimensional online decision-making problems and have a rich information-regret structure. In this work we explore the use of information-directed sampling (IDS), which naturally balances the information-regret trade-off. We develop a class of information-theoretic Bayesian regret bounds that nearly match existing lower bounds on a variety of problem instances, demonstrating the adaptivity of IDS. To efficiently implement sparse IDS, we propose an empirical Bayesian approach for sparse posterior sampling using a spike-and-slab Gaussian-Laplace prior. Numerical results demonstrate significant regret reductions by sparse IDS relative to several baselines.

1 Introduction

Standard linear bandits associate each action with a feature vector and assume the mean reward is the inner product between the feature vector and an unknown parameter vector [Auer, 2002, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Chu et al., 2011, Abbasi-Yadkori et al., 2011]. Sparse linear bandits generalize linear bandits by assuming the unknown parameter vector is sparse [Abbasi-Yadkori et al., 2012, Carpentier and Munos, 2012, Hao et al., 2020b] and is of great practical significance for modeling high-dimensional online decision-making problems [Bastani and Bayati, 2020].

Lattimore and Szepesvári [2020, §24.3] established a $\Omega(\sqrt{sdn})$ regret lower bound for the *data-rich regime*, where n is the horizon, d is the feature dimension, s is the sparsity and data-rich regime refers to the horizon $n \geq d^{\alpha}$ for some $\alpha > 0$. This means polynomial dependence on d is generally not avoidable without additional assumptions. However, this bound hides much of the rich structure of sparse linear bandits by a crude maximisation over all environments.

When the action set admits a well-conditioned exploration distribution, Hao et al. [2020b] discovered the information-regret trade-off phenomenon by establishing an $\Theta(\mathrm{poly}(s)n^{2/3})$ minimax rate for the *data-poor regime*. An interpretation for this optimal rate is that the agent needs to acquire enough information for fast sparse learning by pulling informative actions that even have high regret. Explore-then-commit algorithm can achieve this rate for the data-poor regime but is sub-optimal

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

in the data-rich regime. Therefore, our goal is to develop an efficient algorithm that can adapt to different information-regret structures for sparse linear bandits.

Contributions Our contribution is three-fold:

- We prove that optimism-based algorithms fail to optimally address the information-regret trade-off in sparse linear bandits, which results in a sub-optimal regret bound.
- We provide the first analysis using information theory for sparse linear bandits and derive a class of nearly optimal Bayesian regret bounds for IDS that can adapt to information-regret structures.
- To approximate the information ratio, we develop an empirical Bayesian approach for sparse posterior sampling using spike-and-slab Gaussian-Laplace prior. Through several experiments, we justify the great empirical performance of sparse IDS with an efficient implementation.

2 Preliminary

We first introduce the basic setup of stochastic sparse linear bandits. The agent receives a compact action set $\mathcal{A} \subseteq \mathbb{R}^d$ in the beginning where $|\mathcal{A}| = K$. At each round t, the agent chooses an action $A_t \in \mathcal{A}$ and receives a reward $Y_t = \langle A_t, \theta^* \rangle + \eta_t$, where $(\eta_t)_{t=1}^n$ is a sequence of independent standard Gaussian random variables and $\theta^* \in \mathbb{R}^d$ is the true parameter unknown to the agent. We make the mild boundedness assumption that for all $a \in \mathcal{A}$, $\|a\|_\infty \leq 1$. 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 Bayesian setting where θ^* is a random variable taking values in Θ and denote ρ as the prior distribution. The optimal action is $x^* = \operatorname{argmax}_{a \in \mathcal{A}} \mathbb{E}[\langle a, \theta^* \rangle | \theta^*]$. The agent chooses A_t based on the history $\mathcal{F}_t = (A_1, Y_1, \ldots, A_{t-1}, Y_{t-1})$. Let $\mathcal{D}(\mathcal{A})$ be the space of probability measures over $\mathcal{A}$. A policy $\pi = (\pi_t)_{t \in \mathbb{N}}$ is a sequence of deterministic functions where $\pi_t(\mathcal{F}_t)$ specifies a probability distribution over $\mathcal{A}$. The information-theoretic Bayesian regret of a policy π [Russo and Van Roy, 2014] is defined as

$$\mathfrak{BR}(n; \pi) = \mathbb{E}\left[\sum_{t=1}^n \langle x^*, \theta^* \rangle - \sum_{t=1}^n Y_t \right] ,$$

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

Notation Denote I_d as the $d \times d$ identity matrix. Let $[n] = \{1, 2, \ldots, n\}$. For a vector x and positive semidefinite matrix A, we let $\|x\|_A = \sqrt{x^\top A x}$ be the weighted ℓ_2-norm and $\sigma_{\min}(A)$ be the minimum eigenvalue of A. The relation $x \gtrsim y$ means that x is greater or equal to y up to some universal constant and $\widetilde{O}(\cdot)$ hides modest logarithmic factors and universal constant. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. 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: $\mathbb{P}_{X|Y}(\cdot) = \mathbb{P}(X \in \cdot | Y)$. The mutual information between X and Y is $I(X;Y) = \mathbb{E}[D_{\mathrm{KL}}(\mathbb{P}_{X|Y} || \mathbb{P}_X)]$ 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 $I_t(X;Y) = \mathbb{E}_t[D_{\mathrm{KL}}(\mathbb{P}_{t,X|Y} || \mathbb{P}_{t,X})]$.

3 Related work

Sparse linear bandits 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

Table 1: Comparisons with existing results. APS11, LCS15, HLW20 refer to Abbasi-Yadkori et al. [2012], Lattimore et al. [2015], Hao et al. [2020b] accordingly. Exploratory action set is defined in Definition 5.2 and K is the number of actions. The last lower bound is developed in Hao et al. [2020b].

	Action set	Algorithm	Type	Rate
APS11	arbitrary	online-to-confidence	freq	$\widetilde{O}(\sqrt{sdn})$
LCS15	hypercube	elimination	freq	$\widetilde{O}(s\sqrt{n})$
HLW20	exploratory	explore-then-commit	freq	$\widetilde{O}(s^{2/3}n^{2/3})$
This paper	arbitrary	sparse IDS	Bayesian	$\widetilde{O}(\min(\sqrt{sdn}, \sqrt{dn\log(K)}))$
This paper	arbitrary	sparse TS	Bayesian	$\widetilde{O}(\min(\sqrt{sdn}, \sqrt{dn\log(K)}))$
This paper	exploratory	sparse IDS	Bayesian	$\widetilde{O}(\min(sn^{2/3}, \sqrt{sdn})$
Lower bound	arbitrary	NA	minimax	$\Omega(\sqrt{sdn})$
Lower bound	exploratory	NA	minimax	$\Omega(\min(s^{1/3}n^{2/3}, \sqrt{dn}))$

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. [2020b] introduced the notion of an exploratory action set and proved a $\Theta(\text{poly}(s)n^{2/3})$ minimax rate for the data-poor regime using an explore-then-commit algorithm. Hao et al. [2021] extended this concept to a MDP setting. Carpentier and Munos [2012] considered a special case where the action set is the unit sphere and the noise is vector-valued so that the noise becomes smaller as the dimension grows. We summarize the comparison of existing results with our work in Table 3.

Sparse linear contextual bandits It recently became popular to study the contextual setting, where the action set changes from round to round. These results can not be reduced to our setting since they rely on either careful assumptions on the context distribution [Bastani and Bayati, 2020, Wang et al., 2018, Kim and Paik, 2019, Wang et al., 2020, Ren and Zhou, 2020, Oh et al., 2020] such that classical high-dimensional statistics can be used, or have polynomial dependency on the number of actions [Agarwal et al., 2014, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2020].

Information-directed sampling Russo and Van Roy [2018] introduced IDS and derived Bayesian regret bounds for multi-armed bandits, linear bandits and combinatorial bandits. Liu et al. [2018] studied IDS for bandits with graph-feedback. Kirschner and Krause [2018], Kirschner et al. [2020a] investigated the use of IDS for bandits with heteroscedastic noise and partial monitoring. Kirschner et al. [2020b] proved the asymptotic optimality of frequentist IDS for linear bandits.

Information-theoretic analysis Russo and Van Roy [2014] introduced an information-theoretic analysis of Thompson sampling (TS) and Bubeck and Sellke [2020] strengthened the result with a first-order Bayesian regret analysis. Dong and Roy [2018], Dong et al. [2019] extended the analysis to infinite-many actions and logistic bandits. Lattimore and Szepesvári [2019] explored the use of information-theoretic analysis for partial monitoring. Lu and Van Roy [2019] generalized the analysis to reinforcement learning.

Bayesian sparse linear regression In the Bayesian framework, spike-and-slab methods are commonly used as probabilistic tools for sparse linear regression but most of prior works focus on variable selection and parameter estimation rather than uncertainty quantification [Mitchell and Beauchamp, 1988, George and McCulloch, 1993, Ročková and George, 2018]. Bai et al. [2020] provided a comprehensive overview.

4 Does the optimism optimally balance information and regret?

We demonstrate the necessity of balancing the trade-off between information and regret through a simple sparse linear bandit instance. We give an example where the optimal regret is only possible

by playing actions that are known to be sub-optimal. This phenomenon has disturbing implications for policies based on the principle of optimism, which is that they can never be minimax optimal in certain regime.

Illustrative example Consider a problem instance where $\mathcal{A} = \mathcal{I} \cup \mathcal{U}$ is the union of an informative action set $\mathcal{I}$ and an uninformative action set $\mathcal{U}$:

- $\mathcal{I}$ is a *subset* of the hypercube that has three properties. First, $|\mathcal{I}| = O(s \log(ed/s))$. Second, the last coordinate of actions in $\mathcal{I}$ is always -1. Third, the empirical covariance of the uniform distribution over $\mathcal{I}$ has a restricted minimum eigenvalue (Definition A.1) at least 1/4. We prove such $\mathcal{I}$ does exist in Appendix A.1 through a probabilistic argument.
- $\mathcal{U} = \{x \in \mathbb{R}^d | x_j \in \{-1, 0, 1\} \text{ for } j \in [d-1], \|x\|_1 = s-1, x_d = 0\}$.

The true parameter $\theta^* = (\varepsilon, \ldots, \varepsilon, 0, \ldots, 0, -1)$, where $\varepsilon > 0$ is a small constant.

Information-regret structure Sampling an action uniformly at random from $\mathcal{I}$ ensures the covariance matrix is well-conditioned so that sparse learning such as Lasso [Tibshirani, 1996] can be used for learning θ^* faster than ordinary least squares. This means pulling actions from $\mathcal{I}$ provides more information to infer θ^* than from $\mathcal{U}$. On the other hand, actions from $\mathcal{I}$ lead to high regret due to the last coordinate -1. As a consequence, Hao et al. [2020b] has proven that the minimax regret for this problem is $\Theta(\text{poly}(s) n^{2/3})$ when the horizon is smaller than the ambient dimension.

Sub-optimality of optimism-based algorithms We argue the optimism principle does not take this subtle trade-off into consideration and yields sub-optimal regret in the data-poor regime. In general, optimism-based algorithms choose $A_t = \text{argmax}_{a \in \mathcal{A}} \max_{\widetilde{\theta} \in \mathcal{C}_t} \langle a, \widetilde{\theta} \rangle$, where $\mathcal{C}_t \subseteq \mathbb{R}^d$ is a confidence set that contains the true θ^* with high probability. We assume that there exists a constant $c > 0$ such that $\mathcal{C}_t \subseteq \{\theta : (\widehat{\theta}_t - \theta)^\top V_t (\widehat{\theta}_t - \theta) \leq c\sqrt{s \log(n)}\}$, where $V_t = \sum_{s=1}^t A_s A_s^\top$ and $\widehat{\theta}_t$ is some sparsity-aware estimator. Such confidence set can be constructed through an online-to-confidence set conversion approach [Abbasi-Yadkori et al., 2012]. Define $\mathfrak{R}_\theta(n; \pi) = \mathbb{E}[\sum_{t=1}^n \langle x^*, \theta \rangle - \sum_{t=1}^n Y_t]$ for a fixed θ.

Claim 4.1. Let π^{opt} be such an optimism-based algorithm. There exists a sparse linear bandit instance characterized by θ such that for the data-poor regime, we have

$$\mathfrak{R}_\theta(n; \pi^{\text{opt}}) \gtrsim n/(\log(n) s \log(ed/s)) \,.$$

The proof is deferred to Appendix A.1. The reason is that optimism-based algorithms do not choose actions for which they have collected enough statistics to prove these actions are suboptimal, but in the sparse linear setting it can be worth playing these actions when they are informative about other actions for which the statistics are not yet so clear. This phenomenon has been observed before in linear and structured bandits [Lattimore and Szepesvari, 2017, Combes et al., 2017, Hao et al., 2020a].

5 Information-directed sampling

As shown by Hao et al. [2020b], although the explore-then-commit algorithm can achieve the minimax optimal regret in the data-poor regime, it suffers sub-optimal regret in the data-rich regime. This motivates us consider IDS.

5.1 Design principle of IDS

Unlike the optimism principle, IDS explicitly balances the amount of information it gains about the optimal action and expected single-round regret through minimizing a notion of information ratio. More formally, when playing action a, the *information gain* $I_t(x^*; Y_{t,a})$ is the mutual information

between the optimal action and the reward the agent receives for taking action a, and the expected single-round regret is $\Delta_t(a) := \mathbb{E}_t[\langle x^*, \theta^* \rangle - \langle a, \theta^* \rangle]$. The information ratio is defined as

$$\Psi_{t,\lambda}(\pi) = \frac{(\Delta_t^\top \pi)^\lambda}{I_t^\top \pi}, \tag{5.1}$$

where we write $\Delta_t \in \mathbb{R}^{|\mathcal{A}|}$ and $I_t \in \mathbb{R}^{|\mathcal{A}|}$ as corresponding vectors. Then IDS takes the action according to $\pi_t = \operatorname{argmin}_{\pi \in \mathcal{D}(\mathcal{A})} \Psi_{t,2}(\pi)$.

Remark 5.1. The information ratio defined in Eq. (5.1) is a little more general than what Russo and Van Roy [2018] introduced which specified $\lambda = 2$. As observed by Lattimore and György [2020], the right value of λ depends on the dependence of the regret on the horizon.

5.2 Information-theoretic Bayesian regret bound

In this section, we derive a class of Bayesian regret upper bound for sparse IDS. We first define a notion of exploratory action set.

Definition 5.2 (Exploratory action set)**.** Let $C_{\min}(\mathcal{A}) = \max_{\mu \in \mathcal{D}(\mathcal{A})} \sigma_{\min}(\mathbb{E}_{A \sim \mu}[AA^\top])$. For an action set $\mathcal{A}$, if $C_{\min}(\mathcal{A}) \geq 1$[1], we say $\mathcal{A}$ is exploratory.

We say that $\mathcal{A}$ has sparse optimal actions if the optimal action is s-sparse almost surely with respect to the prior. One can verify the action set of the hard instance developed in Hao et al. [2020b][2] is exploratory and has sparse optimal actions since sampling uniformly from the corner of informative action set shows that $C_{\min}(\mathcal{A}) \geq 1$ and the optimal actions always come from uninformative action set, which is sparse.

Theorem 5.3 (Regret bound for sparse IDS)**.** Suppose $\pi^{\text{IDS}} = (\pi_t)_{t \in \mathbb{N}}$ where $\pi_t = \operatorname{argmin}_\pi \Psi_{t,2}(\pi)$. Let $\Delta = \min(\log(K), 2s \log(Cdn^{1/2}/s))$ for some absolute constant $C > 0$. For an arbitrary action set, the following regret bound holds

$$\mathfrak{BR}(n; \pi^{\text{IDS}}) \leq \sqrt{\frac{1}{2} nd\Delta}.$$

When $\mathcal{A}$ is exploratory and has sparse optimal actions, the following regret bound holds

$$\mathfrak{BR}(n; \pi^{\text{IDS}}) \leq \min\left\{ \sqrt{\frac{1}{2} nd\Delta}, \frac{s^{\frac{2}{3}} n^{\frac{2}{3}} \Delta^{\frac{1}{3}}}{(2C_{\min}(\mathcal{A}))^{\frac{1}{3}}} \right\}.$$

This theorem shows the great adaptivity of IDS for sparse linear bandits in the sense that a single policy adapts to different information-regret structures. We summarize the regret bounds in a variety of of different regimes in Table 5.2.

Remark 5.4. The explore-then-commit algorithm proposed by Hao et al. [2020b] for sparse linear bandits has $O(\text{poly}(s) n^{2/3})$ regret bound when the action set is exploratory and it is known that this $O(n^{2/3})$ rate is not improvable. Thus, it is sub-optimal for data-rich regime comparing with $\Theta(\sqrt{sdn})$ minimax rate. In contrast, IDS is nearly optimal in both regimes.

As a direct application of our analysis, we also include a novel Bayesian regret bound for sparse TS.

Corollary 5.5 (Regret bound for sparse TS)**.** For an arbitrary action set, the following regret bound holds for some absolute constant $C > 0$

$$\mathfrak{BR}(n; \pi^{\text{TS}}) \leq \sqrt{\frac{1}{2} nd \min(\log(K), 2s \log(Cdn^{1/2}/s))}.$$

Proof of Theorem 5.3. We prove our main result in three steps. All the proofs of technical lemmas are deferred to the appendix.

[1] In the definition 1 is simply a convenient constant. More precisely, $C_{\min}(\mathcal{A})$ is a problem-dependent parameter of our regret bound.

[2] The hard instance is almost the same as the one in illustrative example except the informative action set is a full hypercube.

Table 2: Summary of regret bounds of IDS for different regimes. Data-rich regime refers to $n \gtrsim d^3\Delta/s^4$ and large K refers to $K \gtrsim d\exp(s)$.

	Arbitrary action set	Exploratory (data-rich)	Exploratory (data-poor)
Large K	$\widetilde{O}(\sqrt{nds})$	$\widetilde{O}(\sqrt{nds})$	$\widetilde{O}(sn^{2/3})$
Small K	$\widetilde{O}(\sqrt{nd\log(K)})$	$\widetilde{O}(\sqrt{nd\log(K)})$	$\widetilde{O}(s^{2/3}n^{2/3}\log^{1/3}(K))$

Step 1: Generic Bayesian regret upper bound We define $\Psi_{*,\lambda} \in \mathbb{R}$ as the *worse-case information ratio* such that for each $t \in [n]$, $\Psi_{t,\lambda}(\pi_t) \leq \Psi_{*,\lambda}$ almost surely.

Lemma 5.6. Suppose $\pi^{\text{IDS}} = (\pi_t)_{t\in\mathbb{N}}$ where $\pi_t = \operatorname{argmin}_\pi \Psi_{t,2}(\pi)$. Then the following regret bound holds

$$\mathfrak{BR}(n;\pi^{\text{IDS}}) \leq \inf_{\lambda\geq 2} 2^{1-2/\lambda}\Psi_{*,\lambda}^{1/\lambda} I(x^*;\mathcal{F}_{n+1})^{1/\lambda} n^{1-1/\lambda},$$

where $\mathcal{F}_{n+1}$ refers to the history.

This lemma demonstrates the adaptivity of a single IDS for different information ratios. The choice of λ essentially trades off the information-ratio and the horizon.

Step 2: Bounding the worse-case information ratio We bound the worse-case information ratio for different λ. It shows that for certain action sets, the worse-case information ratio with $\lambda = 3$ could be much smaller than the one with $\lambda = 2$.

Lemma 5.7. For an arbitrary action set, we have $\Psi_{*,2} \leq d/2$. For an exploratory action set that has sparse optimal actions, we have $\Psi_{*,3} \leq s^2/(4C_{\min}(\mathcal{A}))$.

The bound of $\Psi_{*,2}$ essentially follows Russo and Van Roy [2014, Proposition 5] that bounds the information ratio of IDS by TS. And $\Psi_{*,3}$ is bounded by the information ratio of a mixture policy $\pi_t^{\text{mix}} = (1-\gamma)\pi_t^{\text{TS}} + \gamma\mu$ where μ is an exploratory policy such that $\sigma_{\min}(\mathbb{E}_{A\sim\mu}[AA^\top])$ is a constant and the mixture rate $\gamma \geq 0$ is optimized to minimize the bound.

Step 3: Bounding the mutual information The mutual information $I(x^*;\mathcal{F}_{n+1})$ quantifies the cumulative information gain about the optimal action. Russo and Van Roy [2014, 2018] naively bound this term by entropy $H(x^*)$, which can be arbitrarily large or even infinite for some priors. Instead, we bound this term by the mutual information between the true parameter and the history through data-processing lemma.

Lemma 5.8. $I(x^*;\mathcal{F}_{n+1}) \leq I(\theta^*;\mathcal{F}_{n+1}) \leq \min\{\log(K), 2s\log(Cdn^{1/2}/s)\}$.

Our proof is based on the metric entropy of the parameter space and square root KL-divergence that is commonly used in information-theoretic lower bound analysis [Yang and Barron, 1999]. As a by-product of our analysis, by setting $s = d$, our analysis recovers the $\widetilde{O}(d\sqrt{n})$ Bayesian regret bound for TS under linear bandits with infinitely many actions without using rate-distortion theory [Dong and Roy, 2018]. Combining Lemmas 5.6-5.8 yields our conclusion. □

6 Computational methods

In this section, we provide an efficient implementation of sparse IDS. The main challenge is to generate posterior samples in a computationally efficient manner to approximate the information ratio. Due to the lack of conjugate prior, we propose an empirical Bayesian approach for sparse sampling with spike-and-slab priors.

6.1 An empirical Bayesian approach for sparse sampling

In the Bayesian framework, the golden standard for modeling θ^* is to place spike-and-slab priors [Mitchell and Beauchamp, 1988]. With a hierarchical structure over the parameter and model space,

vanilla spike-and-slab priors usually have the following form

$$\rho(\theta|\boldsymbol{\gamma},\sigma^2)=\prod_{j=1}^{d}\left[\gamma_j\psi_1(\theta_j,\sigma)+(1-\gamma_j)\psi_0(\theta_j,\sigma)\right],\rho(\boldsymbol{\gamma}|\beta)=\prod_{j=1}^{d}\beta^{\gamma_j}(1-\beta)^{1-\gamma_j}\,, \tag{6.1}$$

where $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_d)^\top$ is an intermediate binary vector that indexs the 2^d possible models and $\beta\in[0,1]$ denotes a priori fraction of relevant variables among all the parameters. In particular, $\psi_1(\theta,\sigma)$ serves as a slab distribution to models relevant variables and $\psi_0(\theta,\sigma)$ is a point mass at zero that serves as a spike distribution to model irrelevant variables.

Remark 6.1. It is typical to assume β follows a Beta prior as $\beta\sim\text{Beta}(a,b)$ and variance σ^2 follows an inverse gamma prior $\rho(\sigma^2)=\text{IG}(\nu/2,\nu\lambda/2)$ with $\nu=1$ and $\lambda=1$ (Rořková and George [2014]). For simplicity, we do not impose those additional layers of priors.

Prior specification. Although the prior in Eq. (6.1) is theoretically sound, exploring the full posterior in high-dimensions over the entire model space using point-mass spike-and-slab priors can be computationally prohibitive. Therefore, the spike distribution is usually relaxed as a small-scale Gaussian distribution [Rořková and George, 2014] or Laplace distribution [Ročková and George, 2018]. Thus, we consider a spike-and-slab Gaussian-Laplace prior that inherits the property of the Lasso [Tibshirani, 1996] for variable selection while the Gaussian component avoids the potential over-shrinkage effect. Each component of the prior is specified as

$$\psi_0(\theta,\sigma)=\frac{1}{2\sigma\lambda_0}\exp\left(-\frac{|\theta|}{\sigma\lambda_0}\right),\psi_1(\theta,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2\lambda_1}}\exp\left(-\frac{\theta^2}{2\sigma^2\lambda_1}\right)\,,$$

where $\lambda_0>0$ denotes a scaling parameter that encourages the shrinkage of irrelevant parameters and λ_1 is often set to a large value for a standard regularization [Rořková and George, 2014].

An empirical Bayesian approach. Suppose $\mathcal{L}(\mathcal{F}_{n+1}|\theta,\sigma^2)$ is the likelihood function where Z_n is the historical data. According to the Bayes rule, the full posterior follows

$$p(\theta,\boldsymbol{\gamma}|\mathcal{F}_{n+1},\sigma^2,\beta)\propto\mathcal{L}(\mathcal{F}_{n+1}|\theta,\sigma^2)\rho(\theta|\boldsymbol{\gamma},\sigma^2)\rho(\boldsymbol{\gamma}|\beta)\,. \tag{6.2}$$

To speed up the sampling, we only sample θ and optimize $\boldsymbol{\gamma}$ with respect to the posterior instead. To tackle this issue of the binary vector $\boldsymbol{\gamma}$, we consider a continuous relaxation by introducing a latent vector $\nu\in[0,1]^d$ as the probability of the variable being included in the model. We focus on the simulations of the conditional expectation of the complete posterior:

$$\mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left[\log p(\theta,\boldsymbol{\gamma}|\mathcal{F}_{n+1},\sigma^2,\beta)\right]=\log\mathcal{L}(\mathcal{F}_{n+1}|\theta,\sigma^2)+\mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left[\log\rho(\boldsymbol{\gamma}|\beta)+\log\rho(\theta|\boldsymbol{\gamma},\sigma^2)\right]+C_1\,,$$

where C_1 is the normalizing constant and $\mathbb{E}_{\boldsymbol{\gamma}|\cdot}[\cdot]$ denotes the the conditional expectation with respect to γ given the current parameter θ. Then we compute

$$\mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left[\log\rho(\boldsymbol{\gamma}|\beta)\right]=\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left[\gamma_i\log(\beta)+(1-\gamma_i)\log(1-\beta)\right]=\sum_{i=1}^{d}\log\left(\frac{\beta}{1-\beta}\right)\nu_i+C_2\,,$$

where $C_2=d\log(1-\beta)$ and $\nu_i=\mathbb{E}_{\boldsymbol{\gamma}|\cdot}[\gamma_i]$ denotes a conditional probability. By the Bayes rule, we have

$$\nu_i=\mathbb{P}(\gamma_i=1|\theta_i,\beta)=\frac{\rho(\theta_i|\gamma_i=1)\mathbb{P}(\gamma_i=1|\beta)}{\rho(\theta_i|\gamma_i=1)\mathbb{P}(\gamma_i=1|\beta)+\rho(\theta_i|\gamma_i=0)\mathbb{P}(\gamma_i=0|\beta)}\,. \tag{6.3}$$

For the mixture prior, we optimize the variational lower bound

$$\mathbb{E}_{\boldsymbol{\gamma}|\cdot}[\log\rho(\theta|\boldsymbol{\gamma},\sigma^2)]\geq\sum_{j=1}^{d}-\frac{1-\nu_j}{\lambda_0}\frac{|\theta_j|}{\sigma}-\frac{\nu_j}{\lambda_1}\frac{\theta_j^2}{2\sigma^2}+C_3\,,$$

where the inequality follows by Jensen's inequality and C_3 denotes a trivial constant.

We summarize the full sampling procedure in Algorithm 1. Given a current estimate of $(\theta^{(k)},\nu^{(k)})$ at step k and $\mathcal{F}_{n+1}$, we adapt an empirical Bayesian method [Deng et al., 2019] by iteratively sampling θ based on the negative log-posterior with adaptive priors

$$Q(\theta|\theta^{(k)},\nu^{(k)},\mathcal{F}_{n+1})=-\log\mathcal{L}(\mathcal{F}_{n+1}|\theta,\sigma^2)+\sum_{i=1}^{d}\left(\frac{1-\rho_i}{\lambda_0}\frac{|\theta_i|}{\sigma}+\frac{\rho_i}{\lambda_1}\frac{\theta_i^2}{2\sigma^2}\right)\,,$$

and optimizing the conditional probability ν through stochastic approximation algorithms (Robbins and Monro [1951]) until the equilibrium is achieved.

Remark 6.2. Algorithm 1 will need to output M posterior samples. Suppose we run a Markov Chain for 100 steps ($k = 1, 2, \ldots, 100$) and want 10 posterior samples. Setting the thinning factor $T = 10$ means every 10 steps, we pick a θ^k as the required posterior sample, e.g., $\theta^{10}, \theta^{20}, \ldots, \theta^{100}$.

Algorithm 1 Empirical Bayesian sparse sampling

1: **Input:** dataset $\mathcal{F}_{n+1}$, learning rate (η_k), step size (ω_k), a priori knowledge of σ^2 and β, thinning factor T, number of posterior samples M, regularization parameters λ_0, λ_1.
2: **Initialize:** $\theta^{(0)} \sim N(0, 0.1 I_d)$ and $\nu^{(0)} = (0.5, \ldots, 0.5)^\top$.
3: **for** $k \geq 1$ **do**
4: Sampling step: $\theta^{(k+1)} = \theta^{(k)} - \eta_k \frac{\partial}{\partial \theta} Q(\theta|\theta^{(k)}, \nu^{(k)}, \mathcal{F}_{n+1}) + \sqrt{2\eta_k}\xi^{(k)}$, where $\xi^{(k)}$ is a standard Gaussian random vector.
5: Stochastic approximation step: $\nu^{(k+1)} = (1 - \omega_k)\nu^{(k)} + \omega_k \nu$, where ν is derived from Eq. (6.3).
6: **end for**
7: **Output:** M posterior samples $\theta^{(T)}, \theta^{(2T)}, \ldots, \theta^{(MT)}$.

6.2 Optimize the information ratio

For sparse linear bandits, it is expensive to estimate and optimize the original information ratio that involves the calculation of KL-divergence. Following Section 6.3 in Russo and Van Roy [2018], we optimize a variance-based information ratio instead: $\pi_t = \operatorname{argmin}_\pi (\Delta_t^\top \pi)^2 / (2 v_t^\top \pi)$, where we define $v_t(a) = \mathbb{E}_t[a^\top \mathbb{E}_t[\theta^*|x^*] - a^\top \mathbb{E}_t[\theta^*]]^2$ for each $a \in \mathcal{A}$. With sufficient number of posterior samples of θ^* produced by Algorithm 1, we can accurately estimate $\mathbb{E}_t[\theta^*|x^*], \mathbb{E}_t[\theta^*]$ and the information ratio. The detailed procedure is deferred to Appendix B. For the optimization step, we simply choose the action who can has the minimum per-action variance-based information ratio to accelerate the computation. The overall algorithm can be found in Algorithm 2.

Algorithm 2 Sparse IDS

1: **Input:** time horizon n, action set $\mathcal{A}$, number of posterior samples per round M.
2: **for** $t = 1, \cdots, n$ **do**
3: Obtain M posterior samples $\theta^1, \ldots, \theta^M$ by Algorithm 1.
4: Calculate $\widehat{\Delta}_t \in \mathbb{R}^{|\mathcal{A}|}$ and $\widehat{v}_t \in \mathbb{R}^{|\mathcal{A}|}$ using Algorithm 3 in Appendix B.
5: Take the action $A_t = \operatorname{argmin}_{a \in \mathcal{A}} \widehat{\Delta}_t^2(a) / \widehat{v}_t(a)$ and receive a reward: $Y_t = \langle A_t, \theta^* \rangle + \eta_t$.
6: **end for**

Remark 6.3. While sparse IDS is computationally efficient, we only have samples from approximate posterior distribution rather than the exact one. This will incur a non-negligible approximation error that may depend linearly on the time horizon. This is observed previously by analyzing approximate Thompson sampling [Lu and Van Roy, 2017] or more recently for IDS [Lu et al., 2021].

7 Experiments

First, we evaluate the performance of the empirical Bayesian sparse sampling procedure for generating posterior samples through an offline sparse linear regression. We set $d = 10, s = 3, n = 100$ and the actions are drawn i.i.d from a multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$. The true parameter is $\theta^* = (3, 2, 0, 0, \ldots, 0) \in \mathbb{R}^{10}$. We plot the empirical posterior distributions as well as their posterior mean for the first three covariates in Figure 1. It shows that the posterior distribution concentrates well around the true value and the algorithm identifies the sparse pattern quickly.

Second, we evaluate sparse IDS with several other competitors. In particular, we compare with LinUCB [Abbasi-Yadkori et al., 2011], LinTS with Gaussian prior [Agrawal and Goyal, 2013], IDS

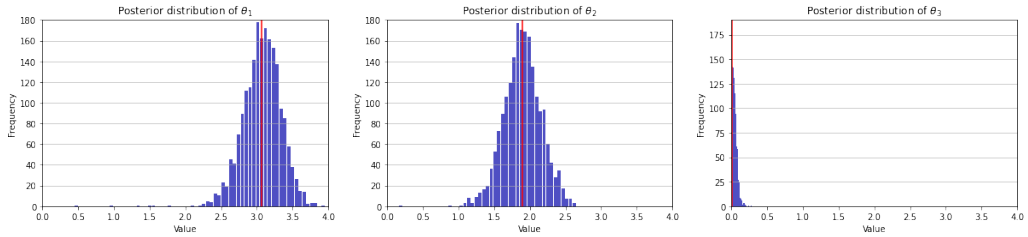

Figure 1: Posterior distributions for the first three coordinates. The red lines are posterior means.

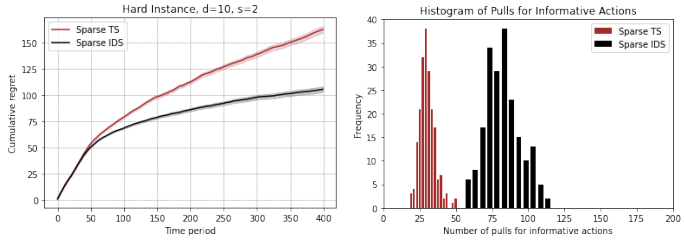

Figure 2: The left panel is the cumulative regret and the right panel is the histogram of number of pulls for informative actions. It's clear that sparse TS does not value information enough as it should.

for linear bandits (Algorithm 6 in Russo and Van Roy [2018]) and ESTC [Hao et al., 2020b]. Note that the first three algorithms are not sparsity-aware. Our sparse sampling procedure naturally induces a sparse TS algorithm so we include it into comparison.

Setting All the true parameters are randomly generated from a multivariate normal distribution, truncated to be sparse and normalized to have square norm 1. The noise variance is fixed to be 2 and we replicate the experiments over 200 trials. We plot the empirical cumulative Bayesian regret. Each Bayesian algorithm will take 10000 posterior samples. We use the TS without blow-up factor for the variance and tune the length of confidence interval of LinUCB over a candidate set.

Hard sparse linear bandits instance Consider the hard problem instance introduced in Section 4 that includes informative and uninformative action sets and set $d = 10, s = 2$. For each trial, we record the number of pulls of sparse TS and sparse IDS for informative actions. We draw the histogram of number of pulls during 200 trial in Figure 2. It is clear that IDS tends to invest more on the informative action sets but suffer less regret than TS if there exists an information-regret trade-off phenomenon.

Multivariate Gaussian action set We consider a more general case where each action is generated from multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$. The number of actions K is fixed to be 200 and the level of sparsity s/d is fixed to be 0.1. We report the results in Figure 3 for $d = 20, 40, 100$. It is obvious that sparse IDS consistently outperforms other algorithms and the improvement increases as the feature dimension increases.

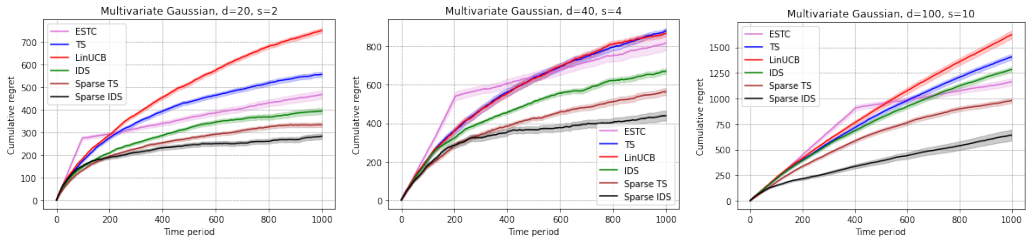

Figure 3: Cumulative regret for $d = 20, 40, 100$.

8 Conclusion

In this work, we investigate the theoretic and practical applicability of information-directed sampling for sparse linear bandits. An interesting future direction is to extend similar ideas to sparse linear contextual bandits.

Acknowledgement

We thank Yoan Russac, Dorian Baudry and Alexandre Filiot for making their code of information-directed sampling public.

References

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

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9, 2012.

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

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Ray Bai, Veronika Rockova, and Edward I George. Spike-and-slab meets lasso: A review of the spike-and-slab lasso. *arXiv preprint arXiv:2010.06451*, 2020.

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

Sébastien Bubeck and Mark Sellke. First-order bayesian regret analysis of thompson sampling. In *Algorithmic Learning Theory*, pages 196–233. PMLR, 2020.

Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198, 2012.

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

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2017.

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

Wei Deng, Xiao Zhang, Faming Liang, and Guang Lin. An adaptive empirical bayesian method for sparse deep learning. *Advances in neural information processing systems*, 2019:5563, 2019.

Shi Dong and Benjamin Van Roy. An information-theoretic analysis for thompson sampling with many actions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4161–4169, 2018.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160. PMLR, 2019.

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

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. *AISTATS*, 2020a.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020b.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pages 5869–5879, 2019.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR, 2020a.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. *arXiv preprint arXiv:2011.05944*, 2020b.

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

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017.

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

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

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 964–972, 2015.

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

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3260–3268, 2017.

Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. *arXiv preprint arXiv:1911.09724*, 2019.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*, 2021.

T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. *arXiv preprint arXiv:2007.08477*, 2020.

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22: 400–407, 1951.

Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

Veronika Rořková and Edward I. George. EMVS: The EM Approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

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

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

Roman Vershynin. On the role of sparsity in compressed sensing and random matrix theory. In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE, 2009.

Xue Wang, Mingcheng Wei, and Tao Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208, 2018.

Yining Wang, Yi Chen, Ethan X Fang, Zhaoran Wang, and Runze Li. Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *arXiv preprint arXiv:2009.02003*, 2020.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

A.1 Proof of Claim 4.1

We first define the notion of restricted minimum eigenvalue.

Definition A.1 (Restricted minimum eigenvalue)**.** Given a symmetric matrix $H \in \mathbb{R}^{d\times d}$ and integer $s \geq 1$, and $L > 0$, the restricted minimum eigenvalue of H is defined as

$$\phi^2(H, s, L) := \min_{\mathcal{S}\subset[d],|\mathcal{S}|\leq s} \min_{\theta\in\mathbb{R}^d} \left\{ \frac{\langle \theta, H\theta\rangle}{\|\theta_{\mathcal{S}}\|_2^2} : \theta \in \mathbb{R}^d, \|\theta_{\mathcal{S}^c}\|_1 \leq L\|\theta_{\mathcal{S}}\|_1 \right\}.$$

Suppose $\{x^{(t)}\}_{t=1}^k \subseteq \mathbb{R}^d$ are k independent random vectors who first $d-1$ coordinates are drawn uniformly from $\{-1, 1\}$ and the last coordinate is 1. Denote $\widehat{\Sigma} = \sum_{t=1}^k x^{(t)}x^{(t)\top}$. It is easy to see $\mathbb{E}[\widehat{\Sigma}] = I_d$ and $\sigma_{\min}(\mathbb{E}[\widehat{\Sigma}]) = 1$. From the definition of restricted minimum eigenvalue, we have for any $L > 0$,

$$\phi^2(\mathbb{E}[\widehat{\Sigma}], s, L) \geq \sigma_{\min}^2(\mathbb{E}[\widehat{\Sigma}]) = 1.$$

According to Theorem 10 in Javanmard and Montanari [2014] (essentially from Theorem 6 in Rudelson and Zhou [2013]), if the population covariance matrix satisfies the restricted eigenvalue condition, the empirical covariance matrix satisfies it as well with high probability. Specifically, when $k = C_1 s\log(ed/s)$ for some large constant $C_1 > 0$, the following holds:

$$\mathbb{P}\left(\phi^2(\widehat{\Sigma}, s, 3) \geq \frac{1}{4}\right) \geq 1 - 2\exp(-k/C_1) \geq 0.5.$$

According to probabilistic argument, there exists a set of fixed actions $\{x^{(1)}, \ldots, x^{(k)}\}$ with $k = C_1 s\log(ed/s)$ such that if we pull uniformly at random from them, the restricted minimum eigenvalue of the resulting covariance matrix is at least 1/4.

Next we compute how many rounds at most the optimism-based algorithm will choose from informative action set $\mathcal{I}$. Let $N_{t-1}(a)$ as the number of pulls for action a until round t. Since we have $\theta^* \in \mathcal{C}_t$ with high probability, then

$$\max_{a\in\mathcal{U}} \max_{\widetilde{\theta}\in\mathcal{C}_t} \langle a, \widetilde{\theta}\rangle \geq \max_{a\in\mathcal{U}} \langle a, \theta^*\rangle \geq s\varepsilon.$$

On the other hand for any action $a \in \mathcal{I}$,

$$\begin{aligned}
\max_{\widetilde{\theta}\in\mathcal{C}_t} \langle a, \widetilde{\theta}\rangle &= \max_{\widetilde{\theta}\in\mathcal{C}_t} \langle a, \widetilde{\theta} - \theta^*\rangle + \langle a, \theta^*\rangle \leq \max_{\widetilde{\theta}\in\mathcal{C}_t} \langle a, \widetilde{\theta} - \theta^*\rangle + \max_{a\in\mathcal{I}} \langle a, \theta^*\rangle \\
&= \max_{\widetilde{\theta}\in\mathcal{C}_t} \langle a, \widetilde{\theta} - \theta^*\rangle + s\varepsilon - 1 \leq 2c\sqrt{\|a\|_{V_t^{-1}} s\log(n)} + s\varepsilon - 1 \\
&\leq 2c\sqrt{\frac{s\log(n)}{N_{t-1}(a)}} + s\varepsilon - 1.
\end{aligned}$$

If $N_{t-1}(a) > 4c^2 s\log(n)$ for $a \in \mathcal{I}$, then we have $\max_{\widetilde{\theta}\in\mathcal{C}_t} \langle a, \widetilde{\theta}\rangle < s\varepsilon$. Based on the optimism principle, the algorithm will switch to pull uninformative actions. This leads to the fact that optimism-based algorithm will pull at most $|\mathcal{I}|4c^2 s\log(n)$ rounds of information actions. According to the proof of minimax lower bound in Hao et al. [2020b], we have when $\sum_{a\in\mathcal{I}} N_n(a) < 1/(s\varepsilon^2)$, there exists another sparse parameter θ' such that

$$R_\theta(n) + R_{\theta'}(n) \gtrsim ns\varepsilon\exp\left(-\frac{2n\varepsilon^2 s^2}{d}\right). \tag{A.1}$$

By choosing $\varepsilon = \sqrt{1/(s^2\log(n)4c^2|\mathcal{I}|)}$, we have for $d \geq n/(s\log(n)\log(ed/s))$

$$R_\theta(n) + R_{\theta'}(n) \gtrsim \frac{n}{\sqrt{\log(n)|\mathcal{I}|}}. \tag{A.2}$$

Note that $|\mathcal{I}| = O(s\log(d/s))$ as we proved before. Then we can argue there exists a sparse linear bandit instance such that optimism-based algorithm will suffer linear regret for a data-poor regime. This ends the proof.

A.2 Proof of Lemma 5.6

We decompose the Bayesian regret in terms of the instantaneous regret:

$$\begin{aligned}\mathfrak{BR}(n;\pi^{\text{IDS}}) &= \mathbb{E}\left[\sum_{t=1}^{n}\langle x^*,\theta\rangle - \sum_{t=1}^{n} Y_t\right] = \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[\langle x^*,\theta^*\rangle - Y_t\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\mathbb{E}_t\left[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\right]\pi_t(a)\right] = \mathbb{E}\left[\sum_{t=1}^{n}\langle \pi_t,\Delta_t\rangle\right],\end{aligned} \tag{A.3}$$

where the third equation is due to the zero mean of the noise.

We then bound one-step instantaneous regret. From the definition of π_t, we have

$$\pi_t = \operatorname*{argmin}_{\pi\in\mathcal{D}(\mathcal{A})} \frac{\langle \pi,\Delta_t\rangle^2}{\langle \pi, I_t\rangle}. \tag{A.4}$$

In addition, we denote

$$q_{\lambda,t} = \operatorname*{argmin}_{\pi\in\mathcal{D}(\mathcal{A})} \Psi_{t,\lambda}(\pi) = \operatorname*{argmin}_{\pi\in\mathcal{D}(\mathcal{A})} \frac{\langle \pi,\Delta_t\rangle^\lambda}{\langle \pi, I_t\rangle}. \tag{A.5}$$

Note that

$$\nabla_\pi \Psi_{t,2}(\pi) = \frac{2\langle \pi,\Delta_t\rangle\Delta_t}{\langle \pi, I_t\rangle} + \frac{\langle \pi,\Delta_t\rangle^2 I_t}{\langle \pi, I_t\rangle^2}.$$

By the first-order optimality condition in Lemma C.1,

$$0 \le \langle \nabla_\pi \Psi_{t,2}(\pi_t), q_{\lambda,t}-\pi_t\rangle = \frac{2\langle q_{\lambda,t}-\pi_t,\Delta_t\rangle\langle \pi_t,\Delta_t\rangle}{\langle \pi_t, I_t\rangle} - \frac{\langle q_{\lambda,t}-\pi_t, I_t\rangle\langle \pi_t,\Delta_t\rangle^2}{\langle \pi_t, I_t\rangle^2}.$$

This further implies

$$2\langle q_{\lambda,t},\Delta_t\rangle \ge \langle \pi_t,\Delta_t\rangle\Big(1+\frac{\langle q_{\lambda,t}, I_t\rangle}{\langle \pi_t, I_t\rangle}\Big) \ge \langle \pi_t,\Delta_t\rangle.$$

Based on the above equation, we can bound the generalized information ratio as follows:

$$\begin{aligned}\frac{\langle \pi_t,\Delta_t\rangle^\lambda}{\langle \pi_t, I_t\rangle} &= \frac{\langle \pi_t,\Delta_t\rangle^2\langle \pi_t,\Delta_t\rangle^{\lambda-2}}{\langle \pi_t, I_t\rangle} \le \frac{2^{\lambda-2}\langle \pi_t,\Delta_t\rangle^2\langle q_{\lambda,t},\Delta_t\rangle^{\lambda-2}}{\langle \pi_t, I_t\rangle} \\ &\le \frac{2^{\lambda-2}\langle q_{\lambda,t},\Delta_t\rangle^{\lambda-2}\langle q_{\lambda,t},\Delta_t\rangle^2}{\langle q_{\lambda,t}, I_t\rangle} = 2^{\lambda-2}\min_{\pi\in\mathcal{D}(\mathcal{A})}\frac{\langle \pi,\Delta_t\rangle^\lambda}{\langle \pi, I_t\rangle},\end{aligned}$$

where the first inequality is from Eq. (A.4) and the second inequality is from Eq. (A.5). According to the definition of $\Psi_{*,\lambda}$, we have

$$\langle \pi_t,\Delta_t\rangle \le 2^{1-2/\lambda}\langle \pi_t, I_t\rangle^{1/\lambda}\Psi_{*,\lambda}^{1/\lambda}.$$

Next we prove $\langle \pi_t, I_t\rangle = I_t(x^*;(A_t,Y_t))$. By the chain rule of mutual information,

$$\begin{aligned}I_t(x^*;(A_t,Y_t)) &= I_t(x^*;A_t) + \mathbb{E}_t[I_t(x^*;Y_t|A_t)] = \mathbb{E}_t[I_t(x^*;Y_t|A_t)] \\ &= \sum_{a\in\mathcal{A}}\pi_t(a) I_t(x^*;Y_t|A_t=a),\end{aligned}$$

where we use the fact that A_t and x^* are independent. If Z is independent of X and Y, then we have $I(X;Y|Z)=I(X;Y)$. Since A_t is independent of x^* and Y_t conditional on $\mathcal{F}_t$, then

$$\sum_{a\in\mathcal{A}}\pi_t(a) I_t(x^*;Y_t|A_t=a) = \sum_{a\in\mathcal{A}}\pi_t(a) I_t(x^*;Y_{t,a}) = \langle \pi_t, I_t\rangle.$$

This proves the previous claim. Combining with Eq. (A.3),

$$\begin{aligned}\mathfrak{BR}(n;\pi^{\text{IDS}}) &= \mathbb{E}\left[\sum_{t=1}^{n}\langle \pi_t, \Delta_t\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^{n} 2^{1-2/\lambda} I_t(x^*;(A_t,Y_t))^{1/\lambda}\Psi_{*,\lambda}^{1/\lambda}\right] \\ &= 2^{1-2/\lambda}\Psi_{*,\lambda}^{1/\lambda}\mathbb{E}\left[\sum_{t=1}^{n} I_t(x^*;(A_t,Y_t))^{1/\lambda}\right] \\ &\le 2^{1-2/\lambda}\Psi_{*,\lambda}^{1/\lambda} n^{1-1/\lambda}\mathbb{E}\left[\sum_{t=1}^{n} I_t(x^*;(A_t,Y_t))\right]^{1/\lambda},\end{aligned} \tag{A.6}$$

where the last inequality is from Holder's inequality with $p = \lambda/(\lambda-1)$ and $q = \lambda$.

In the end, we bound the cumulative information gain using the chain rule of mutual information,

$$\sum_{t=1}^{n}\mathbb{E}[I_t(x^*;(A_t,Y_t))] = \sum_{t=1}^{n} I(x^*;(A_t,Y_t)|\mathcal{F}_t) = I(x^*;\mathcal{F}_{n+1}).$$

Combining with Eq. (A.6), we have

$$\mathfrak{BR}_n(\pi,\rho) \le 2^{1-2/\lambda}(\Psi_{*,\lambda} I(x^*;\mathcal{F}_{n+1}))^{1/\lambda} n^{1-1/\lambda}.$$

This ends the proof.

A.3 Proof of Lemma 5.8

Denote $Z_1 = (A_1, Y_1), \ldots, Z_n = (A_n, Y_n)$ such that $Z^n = (Z_1, \ldots, Z_n)$. When the number of actions K is small, we could directly bound it by

$$I(x^*;Z^n) = H(x^*) - H(x^*|Z^n) \le H(x^*) \le \log|\mathcal{A}| = \log(K),$$

where for the first inequality we use the non-negativity of Shannon entropy.

When the number of actions is large or infinite, we will bound it through the following information-theoretic argument. Recall that $x^* = \operatorname{argmin}_{a\in\mathcal{A}} x^\top\theta^*$ so x^* can be viewed as a deterministic function θ^*. By the data processing lemma (Lemma C.2), we have $I(x^*;Z^n) \le I(\theta^*;Z^n)$. In other words, we bound the information gain regarding the optimal action by the information gain regarding the true parameter.

Recall that we assume the prior distribution of θ^* is $\rho(\theta^*)$ that takes the value in Θ. From Vershynin [2009] , we know Θ enjoys an ε-net $\mathcal{N}_\varepsilon$ under ℓ_2-norm and its cardinality at most $(Cd/s\varepsilon)^s$ where C is a constant. Hence, its metric entropy satisfies

$$\log|\mathcal{N}_\varepsilon| \le s\log(Cd/s\varepsilon). \tag{A.7}$$

Suppose the Bayes mixture density $p_\rho(z^n) = \int_{\theta\in\Theta} p(z^n|\theta)d\rho(\theta)$. According to the definition of mutual information,

$$\begin{aligned}I(\theta^*;Z^n) &= \mathbb{E}_{\theta^*}\left[D_{\text{KL}}(\mathbb{P}_{Z^n|\theta^*}||\mathbb{P}_{Z^n})\right] \\ &= \int_{\theta^*\in\Theta}\int p(z^n|\theta^*)\log\left(\frac{p(z^n|\theta^*)}{p_w(z^n)}\right)\mu(dz^n)d\rho(\theta^*) \\ &\le \int_{\theta\in\Theta}\int p(z^n|\theta^*)\log\left(\frac{p(z^n|\theta^*)}{q(z^n)}\right)\mu(dz^n)d\rho(\theta^*) \\ &= \int_{\theta\in\Theta} D_{\text{KL}}(\mathbb{P}_{Z^n|\theta^*}||\mathbb{Q}_{Z^n})d\rho(\theta^*).\end{aligned} \tag{A.8}$$

where the inequality is due to the fact that Bayes mixture density $p_\rho(z^n)$ minimizes the average KL divergences over any choice of densities $q(z^n)$. Then we choose ρ_1 as an uniform distribution over $\mathcal{N}_\varepsilon$ such that $q(z^n) = p_{\rho_1}(z^n) = \int_{\theta\in\Theta} p(z^n|\theta)d\rho_1(\theta)$ and we denote $\mathbb{Q}_{Z^n}$ as the corresponding

probability measure. Since $\mathcal{N}_\varepsilon$ is an ε-net over Θ under ℓ_2-norm, for each $\theta \in \Theta$, there exists $\widetilde{\theta} \in \Theta$ such that $\|\theta - \widetilde{\theta}\|_2 \le \varepsilon$.

To bound the KL-divergence term, we follow

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_{Z^n|\theta}||\mathbb{Q}_{Z^n}) &= \mathbb{E}\left[\log \frac{p(z^n|\theta^*)}{(1/|\mathcal{N}_\varepsilon|)\sum_{\widetilde{\theta}\in\mathcal{N}_\varepsilon} p(z^n|\widetilde{\theta})}\right] \\ &\le \mathbb{E}\left[\log \frac{p(z^n|\theta^*)}{(1/|\mathcal{N}_\varepsilon|)p(z^n|\widetilde{\theta})}\right] \\ &\le \log|\mathcal{N}_\varepsilon| + D_{\mathrm{KL}}(\mathbb{P}_{Z^n|\theta}||\mathbb{P}_{Z^n|\widetilde{\theta}}). \end{aligned} \tag{A.9}$$

By the chain rule of KL-divergence,

$$D_{\mathrm{KL}}(P_{Z^n|\theta}||\mathbb{P}_{Z^n|\widetilde{\theta}}) \le \mathbb{E}\left[\sum_{t=1}^n D_{\mathrm{KL}}(\mathbb{P}_{Y_t|A_t,Z^{t-1},\theta^*}||\mathbb{P}_{Y_t|A_t,Z^{t-1},\widetilde{\theta}})\right],$$

where we define $Z^0 = \emptyset$. Under linear model and bandit θ, we know $Y_t \sim N(A_t^\top\theta, 1)$. A straightforward computation leads to

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_{Y_t|A_t,Z^{t-1},\theta^*}||\mathbb{P}_{Y_t|A_t,Z^{t-1},\widetilde{\theta}}) &= \frac{1}{2\sigma^2}\|A_t^\top\theta^* - A_t^\top\widetilde{\theta}\|_2^2 \\ &\le \frac{1}{2\sigma^2}\|A_t\|_\infty^2\|\theta^* - \widetilde{\theta}\|_1^2 \\ &\le \frac{1}{2\sigma^2}s\|\theta^* - \widetilde{\theta}\|_2^2 \\ &\le \frac{s}{2\sigma^2}\varepsilon^2, \end{aligned} \tag{A.10}$$

where the first inequality we use the fact that $\|a\|_\infty \le 1$ and the parameters are sparse. Here actually we only require $\|a\|_\infty$ for $a \in \mathcal{A}$ being bounded by a constant since evetually it will only appears inside the logarithm term. Putting Eqs. (A.7)-(A.10) together, we have

$$I(\theta^*; Z^n) \le \int_{\theta^*\in\Theta}\Big(s\log(Cd/s\varepsilon) + \frac{ns}{2\sigma^2}\varepsilon^2\Big)d\theta^* = s\log(Cd/s\varepsilon) + \frac{ns}{2\sigma^2}\varepsilon^2.$$

With the choice of $\varepsilon = 1/\sqrt{n}$, we finally have

$$I(\theta^*; Z^n) \le 2s\log(Cdn^{1/2}/s).$$

This ends the proof.

A.4 Proof of Lemma 5.7

For any particular policy $\widetilde{\pi}$, if one can derive an worse-case bound of $\Psi_{t,\lambda}(\widetilde{\pi})$, we get an upper bound for $\Psi_{*,\lambda}$ automatically. The remaining step is to choose proper policy $\widetilde{\pi}$.

First, we bound the information ratio with $\lambda = 2$ that essentially follows Proposition 5 in Russo and Van Roy [2014] and Lemma 3 in Russo and Van Roy [2014] for a Gaussian noise. By the definition of mutual information, for any $a \in \mathcal{A}$, we have

$$\begin{aligned} I_t(x^*; Y_{t,a}) &= D_{\mathrm{KL}}\left(\mathbb{P}_t((x^*, Y_{t,a}))||\mathbb{P}_t(x^* \in \cdot)\mathbb{P}_t\big(Y_{t,a} \in \big)\right) \\ &= \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(x^* = a^*) D_{\mathrm{KL}}\left(\mathbb{P}_t(Y_{t,a} = \cdot|x^* = a^*)||\mathbb{P}_t(Y_{t,a} = \cdot)\right). \end{aligned} \tag{A.11}$$

Define $R_{\max}$ as 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 and Van Roy [2014], we have

$$I_t(x^*; Y_{t,a}) \ge \frac{2}{R_{\max}^2 + 1}\sum_{a^*\in\mathcal{A}} \mathbb{P}_t(x^* = a^*)\Big(\mathbb{E}_t[Y_{t,a}|x^* = a^*] - \mathbb{E}_t[Y_{t,a}]\Big)^2. \tag{A.12}$$

We bound the information ratio of IDS by the information ratio of TS:

$$\Psi_{*,2} \leq \max_{t\in[n]} \frac{\langle \pi_t^{\text{TS}}, \Delta_t\rangle^2}{\langle \pi_t^{\text{TS}}, I_t\rangle}.$$

Using the matrix trace rank trick described in Proposition 5 in Russo and Van Roy [2014], we have $\Psi_{*,2} \leq (R_{\max}^2+1)d/2$ in the end.

Second, we bound the information ratio with $\lambda = 3$. Recall that the exploratory policy μ is defined as

$$\max_{\mu\in\mathcal{D}(\mathcal{A})} \sigma_{\min}\Big(\int_{x\in\mathcal{A}} xx^\top d\mu(x)\Big).$$

Consider a mixture policy $\pi_t^{\text{mix}} = (1-\gamma)\pi_t^{\text{TS}} + \gamma\mu$ where the mixture rate $\gamma \geq 0$ will be decided later. Then we will bound the following in two steps.

$$\Psi_{t,3}(\pi_t^{\text{mix}}) = \frac{\langle \pi_t^{\text{mix}}, \Delta_t\rangle^3}{\langle \pi_t^{\text{mix}}, I_t\rangle}.$$

Step 1: Bound the information gain According the lower bound of information gain in Eq. (A.12),

$$\begin{aligned}\langle \pi_t^{\text{mix}}, I_t\rangle &\geq \frac{2}{(R_{\max}^2+1)} \sum_{a\in\mathcal{A}} \pi_t^{\text{mix}}(a) \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(x^*=a^*)\left(\mathbb{E}_t[Y_{t,a}|x^*=a^*] - \mathbb{E}_t[Y_{t,a}]\right)^2 \\ &= \frac{2}{(R_{\max}^2+1)} \sum_{a\in\mathcal{A}} \pi_t^{\text{mix}}(a) \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(x^*=a^*)\left(a^\top\mathbb{E}_t[\theta^*|x^*=a^*] - a^\top\mathbb{E}_t[\theta^*]\right)^2.\end{aligned}$$

By the definition of the mixture policy, we know that $\pi_t(a) \geq \gamma\mu(a)$ for any $a \in \mathcal{A}$. Then we have

$$\begin{aligned}\langle \pi_t^{\text{mix}}, I_t\rangle \geq& \frac{2}{(R_{\max}^2+1)}\gamma \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(x^*=a^*) \\ &\cdot \sum_{a\in\mathcal{A}} \mu(a)(\mathbb{E}_t[\theta^*|x^*=a^*] - \mathbb{E}_t[\theta^*])^\top aa^\top(\mathbb{E}_t[\theta^*|x^*=a^*] - \mathbb{E}_t[\theta^*]).\end{aligned}$$

From the definition of minimum eigenvalue, we have

$$\langle \pi_t^{\text{mix}}, I_t\rangle \geq \frac{2\gamma}{(R_{\max}^2+1)} \sum_{a\in\mathcal{A}} \mathbb{P}_t(x^*=a) C_{\min} \left\|\mathbb{E}_t[\theta^*|x^*=a^*] - \mathbb{E}_t[\theta^*]\right\|_2^2.$$

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

$$\begin{aligned}&\langle \pi_t^{\text{mix}}, \Delta_t\rangle \\ &= \sum_a \mathbb{E}_t\Big[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\Big]\pi_t^{\text{mix}}(a), \\ &= (1-\gamma)\sum_a \pi_t^{\text{TS}}(a)\mathbb{E}_t\Big[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\Big] + \gamma\sum_a \mathbb{E}_t\Big[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\Big]\mu(a) \\ &= (1-\gamma)\sum_a \mathbb{P}_t(x^*=a)\mathbb{E}_t\Big[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\Big] + \gamma\sum_a \mathbb{E}_t\Big[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\Big]\mu(a)\end{aligned} \tag{A.13}$$

Since $R_{\max}$ is the upper bound of maximum expected reward, the second term can be bounded $2R_{\max}\gamma$. Next we bound the first term as follows:

$$\begin{aligned}&\sum_a \mathbb{P}_t(x^*=a)\mathbb{E}_t\Big[\langle x^*,\theta^*\rangle - \langle a,\theta^*\rangle\Big] \\ &= \sum_a \mathbb{P}_t(x^*=a)\Big(\mathbb{E}_t[\langle a,\theta^*\rangle|x^*=a] - \mathbb{E}_t[\langle a,\theta^*\rangle]\Big) \\ &= \sum_a \mathbb{P}_t^{1/2}(x^*=a)\mathbb{P}_t^{1/2}(x^*=a)\Big(\mathbb{E}_t[\langle a,\theta^*\rangle|x^*=a] - \mathbb{E}_t[\langle a,\theta^*\rangle]\Big) \\ &\leq \sqrt{\sum_a \mathbb{P}_t(x^*=a)\Big(\mathbb{E}_t[\langle a,\theta^*\rangle|x^*=a] - \mathbb{E}_t[\langle a,\theta^*\rangle]\Big)^2},\end{aligned}$$

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

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

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

$$\begin{aligned}
&\sum_a \mathbb{P}_t(x^* = a)\mathbb{E}_t\Big[\langle x^*, \theta^*\rangle - \langle a, \theta^*\rangle\Big] \\
&\leq \sqrt{\sum_a \mathbb{P}_t(x^* = a)s^2 \left\|\mathbb{E}_t[\theta^*|x^* = a^*] - \mathbb{E}_t[\theta^*]\right\|_2^2} \\
&= \sqrt{\frac{s^2(R_{\max}^2+1)}{2\gamma C_{\min}}\frac{2\gamma}{(R_{\max}^2+1)}\sum_a \mathbb{P}_t(x^* = a)C_{\min}\left\|\mathbb{E}_t[\theta^*|x^* = a^*] - \mathbb{E}_t[\theta^*]\right\|_2^2} \\
&\leq \sqrt{\frac{s^2(R_{\max}^2+1)}{2\gamma C_{\min}}\langle \pi_t^{\text{mix}}, I_t\rangle}.
\end{aligned} \tag{A.14}$$

Putting Eq. (A.13) and (A.14) together, we have

$$\langle \pi_t^{\text{mix}}, \Delta_t\rangle \leq \sqrt{\frac{s^2(R_{\max}^2+1)}{2\gamma C_{\min}}\langle \pi_t^{\text{mix}}, I_t\rangle} + 2R_{\max}\gamma.$$

By optimizing the mixture rate γ, we have

$$\frac{\langle \pi_t^{\text{mix}}, \Delta_t\rangle^3}{\langle \pi_t^{\text{mix}}, I_t\rangle} \leq \frac{s^2(R_{\max}^2+1)}{8R_{\max}^2 C_{\min}} \leq \frac{s^2}{4C_{\min}}.$$

This ends the proof.

B Detailed algorithms

For each $a \in \mathcal{A}$, we expand $v_t(a)$ as follows:

$$\begin{aligned}
v_t(a) &= \text{Var}_t(\mathbb{E}_t[a^\top\theta|x^*]) = \mathbb{E}_t\Big[a^\top\mathbb{E}_t[\theta|x^*] - \mathbb{E}_t\big[a^\top\mathbb{E}_t[\theta|x^*]\big]\Big]^2 \\
&= \mathbb{E}_t\Big[a^\top\mathbb{E}_t[\theta|x^*] - a^\top\mathbb{E}_t[\theta]\Big]^2 = a^\top\mathbb{E}_t[(\mathbb{E}_t[\theta|x^*] - \mathbb{E}_t[\theta])(\mathbb{E}_t[\theta|x^*] - \mathbb{E}_t[\theta])^\top]a.
\end{aligned}$$

We denote $\mu_t = \mathbb{E}_t[\theta]$ as the posterior mean and $\mu_t^a = \mathbb{E}_t[\theta|x^* = a]$. We let $\Phi \in \mathbb{R}^{|\mathcal{A}|\times d}$ as the feature matrix where each row of Φ represent each action in $\mathcal{A}$. We summarize the procedure of estimating Δ_t, I_t in Algorithm 3.

Algorithm 3 Approximate Δ_t, v_t based on posterior samples

1: **Input:** M posterior samples $\theta^1, \ldots, \theta^M$ from Eq. (6.2), action set $\mathcal{A}$.
2: Calculate $\widehat{\mu}_t = \sum_m \theta^m/M$.
3: **for** $a \in \mathcal{A}$ **do**
4: Find $\widehat{\Theta}_a = \{m \in [M] : (\Phi\theta^m)_a = \max_{a'\in\mathcal{A}}(\Phi\theta^m)_{a'}\}$.
5: Calculate $\widehat{p}_a^* = |\widehat{\Theta}_a|/M$.
6: Calculate $\widehat{\mu}_t^a = \sum_{m\in\widehat{\Theta}_a}\theta^m/|\widehat{\Theta}_a|$.
7: Calculate

$$\widehat{v}_t(a) = a^\top\sum_a \widehat{p}_a^*(\widehat{\mu}_t^a - \widehat{\mu}_t)(\widehat{\mu}_t^a - \widehat{\mu}_t)^\top a, \widehat{\Delta}_t(a) = \sum_{a\in\mathcal{A}}\widehat{p}_a^* a^\top\widehat{\mu}_t^a - a^\top\widehat{\mu}_t.$$

8: **end for**
9: **Output:** $\widehat{v}_t, \widehat{\Delta}_t$.

C Supporting lemmas

Lemma C.1 (First-order optimality condition.)**.** Suppose that f_0 in a convex optimization problem is differentiable. Let $\mathcal{X}$ denote the feasible set. Then x is optimal if and only if $x \in \mathcal{X}$ and $\nabla f_0(x)^\top (y - x) \geq 0, \forall y \in \mathcal{X}$.

Lemma C.2 (Data processing lemma)**.** If $Z = g(Y)$ for a deterministic function g, then $I(X;Y) \geq I(X;Z)$.