Sparse Optimistic Information Directed Sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many high-dimensional online decision-making problems can be modeled as stochastic sparse linear bandits. Most existing algorithms are designed to achieve optimal worst-case regret in either the data-rich regime, where polynomial dependence on the ambient dimension is unavoidable, or the data-poor regime, where dimension-independence is possible at the cost of worse dependence on the number of rounds. In contrast, the Bayesian approach of Information Directed Sampling (IDS) achieves the best of both worlds: a Bayesian regret bound that has the optimal rate in both regimes simultaneously. In this work, we explore the use of Sparse Optimistic Information Directed Sampling (SOIDS) to achieve the best of both worlds in the worst-case setting, without Bayesian assumptions. Through a novel analysis that enables the use of a time-dependent learning rate, we show that SOIDS can optimally balance information and regret. Our results extend the theoretical guarantees of IDS, providing the first algorithm that simultaneously achieves optimal worst-case regret in both the data-rich and data-poor regimes. We empirically demonstrate the good performance of SOIDS.

1 Introduction

In stochastic linear bandits, one assumes that the mean reward associated with each action is linear in an unknown d-dimensional parameter vector [Abe and Long, 1999, Auer, 2002, Dani et al., 2008, Abbasi-Yadkori et al., 2011]. Under standard conditions, it is known that the minimax regret in this setting is of the order $\mathcal{O}(d\sqrt{T})$ [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010]. Numerous follow-up works have investigated the possibility of reduced regret under various structural assumptions on the unknown parameter vector, the noise, or the shape of the decision set [Valko et al., 2014, Chu et al., 2011, Kirschner and Krause, 2018], [Lattimore and Szepesvári, 2020, Chapter 22]. One such assumption is that the unknown parameter vector is *sparse*, which means that it has only $s \ll d$ non-zero components. This setting is called *sparse linear bandits* and s is referred to as the *sparsity level*. In this setting, previous work has established the existence of algorithms with regret scaling as $\mathcal{O}(\sqrt{sdT})$ [Abbasi-Yadkori et al., 2012]. This result is complemented by a lower bound, which says that this rate cannot be improved as long as $T \geq d^{\alpha}$ for some $\alpha > 0$ [Lattimore and Szepesvári, 2020]. We refer to this scenario as the *data-rich regime*. Since this bound scales polynomially with the dimension d, many researchers have considered this to be a negative result, interpreting it as a sign that sparsity cannot be effectively exploited in linear bandit problems. This interpretation has been challenged by a more recent observation that, when the action set admits an *exploratory distribution*, simple "explore-then-commit" algorithms enjoy regret bounds of order $\mathcal{O}((sT)^{\frac{2}{3}})$ [Hao et al., 2020, Jang et al., 2022]. These bounds scale only logarithmically with the dimension, and constitute a major improvement over the previously mentioned rate in the *data-poor regime*, where $T \ll \left(\frac{d}{s}\right)^3$. Most known algorithms are specialized to either the data-poor or data-rich regime, and perform poorly in the other one. A notable exception is the *sparse Information Directed Sampling* algorithm introduced in Hao et al. [2021], which performs almost optimally in both regimes. However, Hao et al. [2021] only provide *Bayesian* performance guarantees for sparse

Submitted to the 18th European Workshop on Reinforcement Learning (EWRL 2025). Do not distribute.

IDS. These results hold on average, assuming that the problem instance is drawn at random from a known prior distribution.

In this work, we lift this assumption and develop an algorithm that can adapt to both regimes in a "frequentist" sense: we assume that the true parameter is fixed and unknown to the learner, and provide guarantees that hold for any given instance. The algorithm is an adaptation of the recently proposed Optimistic Information Directed Sampling (OIDS) algorithm of Neu, Papini, and Schwartz [2024], which itself is an adaptation of the classic Bayesian IDS algorithm originally proposed by Russo and Van Roy [2017]. Within the Bayesian setting, it has been shown that IDS can exploit various types of problem structure, and adapt to the hardness of the given instance [Hao and Lattimore, 2022, Hao et al., 2022]. These results have been complemented by the recent work of Neu, Papini, and Schwartz [2024], which showed that similar improvements can be achieved without Bayesian assumptions, via a simple adjustment of the standard IDS method. In this paper, we continue this line of work and show that OIDS can achieve a "best-of-both-worlds" guarantee for sparse linear bandits, which has so far remained ellusive outside of the limited Bayesian bandit setting.

Our contribution is as follows:

- We extend the analysis of the optimistic posterior to allow the use of time-dependent learning rates and history-dependent learning rates. This removes the need to know the horizon in advance and allows us to update the learning rate based on data observed by the agent instead of some loose theoretical constant, a necessity for efficient algorithms.
- We demonstrate that the SOIDS algorithm recovers almost optimal rates in both the data-poor and data-rich regimes. This is the first algorithm to do so in a frequentist setting.

2 Preliminaries

Sparse linear bandits. We consider the following decision-making game, in which a learning agent interacts with an environment over a sequence of T rounds. At the start of each round t, the learner selects an action $A_t \in \mathcal{A} \subset \mathbb{R}^d$ according to a randomized policy $\pi_t \in \Delta(\mathcal{A})$. In response, the environment generates a stochastic reward $Y_t = r(A_t) + \epsilon_t$, where $r : \mathcal{A} \to \mathbb{R}$ is a fixed reward function and ϵ_t is zero-mean, conditionally 1-sub-Gaussian noise. We assume that the action set $\mathcal{A}$ is finite, and that the reward function can be written as

$$r(a) = \langle \theta_0, a \rangle \,,$$

where $\theta_0 \in \mathbb{R}^d$ is an unknown parameter vector. We make the mild boundedness assumptions that $\max_{a\in\mathcal{A}} \|a\|_\infty \le 1$ and $\|\theta_0\|_1 \le 1$. We study the special case of this problem in which the parameter vector θ_0 is s-sparse in the sense that at most $s \ll d$ of its components are non-zero. In other words, we assume that θ_0 belongs to the following *sparse parameter space*:

$$\Theta = \left\{ \theta \in \mathbb{R}^d : \textstyle\sum_{j=1}^d \mathbb{I}_{\{\theta_j \neq 0\}} \le s,\ \|\theta\|_1 \le 1 \right\} .$$

We assume that the sparsity level s is known to the agent. The performance of the agent is evaluated in terms of the *regret*, which is defined as

$$R_T = T \max_{a\in\mathcal{A}} \langle \theta_0, a \rangle - \mathbb{E}\left[\sum_{t=1}^T r(A_t, \theta_0) \right] , \tag{1}$$

where the expectation is taken with respect to both the random choices of the agent and the random noise in the observed rewards. We note that the regret is implicitly a function of the true parameter θ_0. Our focus is on proving regret bounds that hold for arbitrary choices of $\theta_0 \in \Theta$.

The data-rich and data-poor regimes. As mentioned in the introduction, it is known there exist algorithms for sparse linear bandits with worst-case regret of the order $\mathcal{O}(\sqrt{sdT})$ [Abbasi-Yadkori et al., 2012]. This regret bound is only meaningful when the dimension d is smaller than the number of rounds T, a situation referred to as the data-rich regime. Under the assumption that there exists an exploratory policy, Hao et al. [2020] showed that there is a simple algorithm that satisfies a problem-dependent regret bound, which can be meaningful in the so-called data-poor regime, where d is much larger than T. Formally, we say that there exists an exploratory policy if the action set $\mathcal{A}$ is such that

$$C_{\min} := \max_{\mu\in\Delta(\mathcal{A})} \sigma_{\min}\left(\int_{\mathcal{A}} aa^T \, d\mu(a) \right) > 0 \,,$$

which is equivalent to the condition that $\mathcal{A}$ spans $\mathbb{R}^d$. The exploratory policy, is the distribution on $\mathcal{A}$ that achieves the maximum (which is guaranteed to exist when $\mathcal{A}$ is finite). The Explore the Sparsity Then Commit (ESTC) algorithm was shown to satisfy a regret bound of the order $\mathcal{O}(s^{2/3}T^{2/3}C_{\min}^{-2/3})$ [Hao et al., 2020]. The transition between the $T^{2/3}$ rate in the data-poor regime and the $\sqrt{T}$ rate in the data-rich regime also appears in an existing lower bound of the order $\Omega(\min(s^{1/3}T^{2/3}C_{\min}^{-1/3}), \sqrt{dT})$ [Hao et al., 2020].

The best of both worlds for sparse linear bandits. Recently, Hao et al. [2021] showed that the sparse Information Directed Sampling (IDS) algorithm achieves a type of "best-of-both-worlds" guarantee. Under the sparse optimal action condition (Definition 1), IDS satisfies a regret bound of the order $\mathcal{O}(\min(\sqrt{dT\Delta}, (sT)^{2/3}\Delta^{1/3}C_{\min}^{-1/3}))$, where $\Delta \propto \min(\log(|\mathcal{A}|), s\log(dT/s))$. This is simultaneously optimal in both the data-rich and data-poor regimes. However, this result is limited to the Bayesian setting. This is because IDS uses the Bayesian posterior to quantify uncertainty, which is only meaningful if θ_0 really is a random draw from the prior.

The sparse optimal action condition. Part of our analysis requires that a certain technical condition is satisfied. This condition comes from prior work [Hao et al., 2021], and is used to bound the regret in the data-poor regime (cf. Lemma 7).

Definition 1. For a given prior Q_1^+, an action set $\mathcal{A}$ has sparse optimal actions if with probability 1 over the random draw of θ from Q_1^+, there exists $a' \in \arg\max_{a\in\mathcal{A}} r(a,\theta)$ such that $\|a'\|_0 \le s$.

We use a prior that only assigns positive probability to s-sparse vectors, which means the sparse optimal action property is satisfied whenever the action set is an ℓ_p-ball. Note that the hard instances in both the $\sqrt{sdT}$ lower bound in Theorem 24.3 of Lattimore and Szepesvári [2020] and the $s^{2/3}T^{2/3}$ lower bound in Theorem 5 of Jang et al. [2022] satisfy the sparse optimal action property[1]. Therefore, imposing this additional condition does not trivialize the problem.

Notation. We conclude this section by introducing some additional notation that will be used in the subsequent sections. For any candidate parameter vector (or model) $\theta \in \mathbb{R}^d$, we let $r(a,\theta) = \langle \theta, a\rangle$ denote the corresponding linear reward function. In addition, we define $a^*(\theta) = \arg\max_{a\in\mathcal{A}} r(a,\theta)$ (with ties broken arbitrarily) and $r^*(\theta) = r(a^*(\theta),\theta)$ to be the optimal action and maximum reward for the model θ. The gap of an action a for a model θ is $\Delta(a,\theta) = r^*(\theta) - r(a,\theta)$. Similarly, the gap for a policy $\pi \in \Delta(\mathcal{A})$ and a model distribution $Q \in \Delta(\Theta)$ is $\Delta(\pi, Q) = \int_{\mathcal{A}\times\Theta} \Delta(a,\theta)\,\mathrm{d}\pi \otimes Q(a,\theta)$, and we let $\Delta_t = \Delta(\pi_t, \theta_0)$ denote the gap of the policy played by the agent in round t under the true model θ_0. Using this notation, the regret can be written as $R_T = \mathbb{E}[\sum_{t=1}^T \Delta_t]$. We define the unnormalized Gaussian likelihood function $p(y|\theta,a) = \exp(-\frac{(y-\langle\theta,a\rangle)^2}{2})$. Finally, we let $\mathcal{F}_t = \sigma(A_1, Y_1, \ldots, A_t, Y_t)$ denote the σ-algebra generated by the interaction between the agent and the environment up to the end of round t.

3 Sparse Optimistic Information Directed Sampling

We develop an extension of the Optimistic Information Directed Sampling (OIDS) algorithm proposed by Neu, Papini, and Schwartz [2024]. The main difference between OIDS and IDS is that the Bayesian posterior is replaced by an appropriately adjusted *optimistic posterior*. For an arbitrary prior $Q_1^+ \in \Delta(\Theta)$, the optimistic posterior is defined by the following update rule:

$$\frac{dQ_{t+1}^+}{dQ_1^+}(\theta) \propto \prod_{s=1}^{t} (p(Y_s \mid \theta, A_s))^\eta \cdot \exp\Big(\lambda_t \sum_{s=1}^{t} \Delta(A_s,\theta)\Big). \tag{2}$$

Here, η is a positive constant that should be thought of as "large", and $(\lambda_t)_t$ is a decreasing sequence of positive real numbers that decays to 0, and should be thought of as "small". We allow λ_t to be computed by the algorithm at the end of the round t. In other words, any $\mathcal{F}_t$-measurable λ_t is admissible. Note that when $\eta = 1$ and $\lambda_t = 0$, the optimistic posterior coincides with the Bayesian posterior. While this construction is closely related to the optimistic posterior update described in Zhang [2022] and Neu, Papini, and Schwartz [2024], there are a few important differences. First,

[1] The optimal actions in the hard instance used to prove Theorem 5 in Jang et al. [2022] are $2s$-sparse, which still allows us to prove the same bound on the surrogate 3-information ratio, up to constant factors.

the $\Delta(A_s, \theta)$ term appearing in the adjustment serves as an alternative to their proposal of using $r^*(\theta)$ for the same purpose. Intuitively this serves to "overestimate" the true gaps with the optimistic posterior, driving exploration towards parameters that promise rewards much higher than whatever would have been accrued by the agent. In contrast, the adjustment of Zhang [2022] drives exploration towards parameters θ with high optimal reward regardless of how well the agent would have performed under the same θ—meaning that it unduly assigns mass to uninteresting parameter choices, where any policy is guaranteed to work well anyway. Intuition aside, this adjustment greatly simplifies our analysis of the optimistic posterior as compared to the analysis of Zhang [2022] and Neu, Papini, and Schwartz [2024]. An important additional novelty is that our update features a time-dependent exploration parameter λ_t, which is crucial for the adaptive regret bounds that we seek in this work. To describe the OIDS algorithm, we must first define the *surrogate information gain* and the *surrogate regret*. For any round t and any policy $\pi \in \Delta(\mathcal{A})$, the surrogate information gain is defined as

$$\overline{\mathrm{IG}}_t(\pi) = \frac{1}{2} \sum_{a \in \mathcal{A}} \pi(a) \int_\Theta \left(\langle \theta - \bar{\theta}(Q_t^+), a \rangle \right)^2 \, \mathrm{d}Q_t^+(\theta) \,,$$

where for any $Q \in \Delta(\Theta)$, $\bar{\theta}(Q) = \mathbb{E}_{\theta \sim Q}[\theta]$ is the mean parameter under distribution Q. The surrogate regret is defined as

$$\widehat{\Delta}_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a) \int_\Theta \Delta(a, \theta) \, \mathrm{d}Q_t^+(\theta) \,.$$

For any policy π and any $\gamma \geq 2$, we define the *surrogate generalized information ratio* as

$$\overline{\mathrm{IR}}_t^{(\gamma)}(\pi) = \frac{(\widehat{\Delta}_t(\pi))^\gamma}{\overline{\mathrm{IG}}_t(\pi)} = 2 \cdot \frac{\left(\sum_{a \in \mathcal{A}} \pi(a) \int_\Theta \langle \theta, a^*(\theta) - a \rangle \, dQ_t^+(\theta) \right)^\gamma}{\sum_{a \in \mathcal{A}} \pi(a) \int_\Theta (\langle \theta - \bar{\theta}(Q_t^+), a \rangle)^2 \, dQ_t^+(\theta)} . \tag{3}$$

We can at last define our algorithm: Sparse Optimistic Information Directed Sampling (SOIDS). In each round t, the policy played by SOIDS is defined to be the distribution on $\mathcal{A}$ that minimizes the 2-information ratio:

$$\pi_t^{(\mathbf{SOIDS})} = \operatorname*{arg\,min}_{\pi \in \Delta(\mathcal{A})} \overline{\mathrm{IR}}_t^{(2)}(\pi) \,. \tag{4}$$

The choice of $\gamma = 2$ is motivated by the remarkable fact that the minimizer of the 2-information ratio is an approximate minimizer of surrogate generalized information ratio for all $\gamma \geq 2$.

Lemma 1. *For all $\gamma \geq 2$,*

$$\overline{IR}_t^{(\gamma)}(\pi_t^{(\mathbf{SOIDS})}) \leq 2^{\gamma - 2} \min_{\pi \in \Delta(\mathcal{A})} \overline{IR}_t^{(\gamma)}(\pi) \,.$$

This fact was discovered for the Bayesian IDS policy by Lattimore and György [2021] and continues to hold within here. We provide a proof in Appendix F.2 for completeness. Finally, we remark that the "sparse" part of the name SOIDS refers to the choice of the prior Q_1^+. We use the subset selection prior from Section 3 of Alquier and Lounici [2011], which is described in Appendix B.2.

4 Main results

In this section, we state our main results. First, we relate the true regret of any policy sequence to the surrogate regret of the same policy sequence. Then, we use the fact that the surrogate regret is controlled by both the 2 and 3-information ratio. This, combined with Lemma 1, allows us to show that with properly tuned parameters, SOIDS has optimal worst-case regret in both the data-poor and data-rich regimes. Finally, we show that SOIDS can be tuned in a data-dependent manner, such that its regret bound scales with the cumulative observed information ratio instead of the time horizon.

4.1 General bound for the Optimistic Posterior

We start with a generic worst-case regret bound relating the true regret of any algorithm to its surrogate regret. Since the surrogate regret is defined with respect to the optimistic posterior, which is known to the learner, it can be easily controlled with standard Bayesian techniques. This result is an extension of the bounds stated in Neu et al. [2024], Zhang [2022]. To our knowledge it is the first result of its kind which is compatible with time-dependent or data-dependent learning rates. The stated result is specialized to the setting of sparse linear bandits, but the techniques used to deal with time-dependent and data-dependent learning rates are applicable beyond this setting.

Theorem 1. *Assume that the optimistic posterior is computed with $\eta = \frac{1}{4}$ and a sequence of decreasing learning rates λ_t satisfying $\forall t \geq 1, \lambda_t \leq \frac{1}{2}$. Set $\lambda_0 = \frac{1}{2}$. If the learning rates do not depend on the history, then the regret of any sequence of policies π_t satisfies*

$$R_T \leq \mathbb{E}\left[\frac{5 + 2s\log\frac{edT}{s}}{\lambda_{T-1}} - \sum_{t=1}^{T}\frac{3}{32}\cdot\frac{\overline{IG}_t(\pi_t)}{\lambda_{t-1}} + 2\sum_{t=1}^{T}\widehat{\Delta}_t(\pi_t)\right]. \tag{5}$$

Otherwise, if the learning rates depend on the history, let $C_{1,T}$ be a deterministic upper bound on $\frac{1}{\lambda_t} - \frac{1}{\lambda_{t-1}}$ valid for all $t \leq T$, and $C_{2,T}$ be a deterministic upper bound on $\frac{1}{\lambda_{T-1}}$. The regret of any sequence of policies π_t satisfies

$$R_T \leq \mathbb{E}\left[\frac{2 + s\log\frac{4e^3d^2T^3C_{1,T}^2C_{2,T}}{s^2}}{\lambda_{T-1}} - \sum_{t=1}^{T}\frac{3}{32}\cdot\frac{\overline{IG}_t(\pi_t)}{\lambda_{t-1}} + 2\sum_{t=1}^{T}\widehat{\Delta}_t(\pi_t)\right] + 2. \tag{6}$$

4.2 Best of both worlds guarantees for Sparse Optimistic Information Directed Sampling

Next, we show that the SOIDS algorithm with properly tuned parameters attains optimal regret rate in both the data-rich and data-poor regimes.

Theorem 2. *Assume that our problem satisfies the spare optimal action condition described in definition 1. Let $\lambda_t^{(2)} = \sqrt{\frac{3C_{t+1}}{128d(t+1)}}$ and $\lambda_t^{(3)} = \frac{1}{4\cdot 6^{\frac{1}{3}}}\left(\frac{C_{t+1}\sqrt{C_{\min}}}{(t+1)\sqrt{s}}\right)^{\frac{2}{3}}$, with $C_t = 5 + 2s\log\frac{edt}{s}$. Now, set $\lambda_t = \min(\frac{1}{2}, \max(\lambda_t^{(2)}, \lambda_t^{(3)}))$, then the regret of SOIDS run with parameter λ_t is upper bounded by*

$$R_T \leq \min\left(27\sqrt{\left(5 + 2s\log\frac{edT}{s}\right)dT}, 30\left(5 + 2s\log\frac{edT}{s}\right)^{\frac{1}{3}}\left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}}\right) + \mathcal{O}(\sqrt{s}\log\frac{d}{\sqrt{s}}) \tag{7}$$

$$= \min\left(\mathcal{O}\left(\sqrt{sdT\log\frac{edT}{s}}\right), \mathcal{O}\left((sT)^{\frac{2}{3}}\left(\log\frac{edT}{s}\right)^{\frac{1}{3}}\right)\right),$$

where $\mathcal{O}(\sqrt{s}\log\frac{d}{\sqrt{s}})$ represents an absolute constant independent of T.

We observe that our algorithm enjoys both the $\widetilde{\mathcal{O}}(\sqrt{sdT})$ and the $\widetilde{\mathcal{O}}((sT)^{\frac{2}{3}})$ regret rates. Unlike the Bayesian regret bound for the sparse IDS algorithm of Hao et al. [2021], our regret bound holds in a "worst-case" sense for any value of $\theta_0 \in \Theta$. To our knowledge, this makes our method the first algorithm to achieve optimal worst-case regret in both the data-poor and data-rich regimes

4.3 Instance dependent guarantees

The bounds presented in the previous sections are minimax in nature, meaning they hold uniformly over all problem instances. We present a bound in which the scaling with respect to the horizon T is replaced with the cumulative surrogate-information ratio, which could be much smaller than T in "easier" instances, leading to better guarantees.

Theorem 3. *Assume that our problem satisfies the sparse optimal action condition described in Definition 1 and that $s \leq \frac{d}{2}$. Let $\lambda_t^{(2)} = \sqrt{\frac{s}{2d + \sum_{s=1}^{t}\overline{IR}_s^{(2)}(\pi_s)}}$ and $\lambda_t^{(3)} = \left(\frac{s}{\frac{3\sqrt{6}s}{\sqrt{C_{\min}}} + \sum_{s=1}^{t}\sqrt{\overline{IR}_s^{(3)}(\pi_s)}}\right)^{\frac{2}{3}}$. Then the regret of SOIDS run with parameter $\lambda_t = \max(\lambda_t^{(3)}, \lambda_t^{(2)})$ satisfies the following regret bound*

$$R_T \leq \left(\frac{2}{s} + \frac{80}{3} + 5\log\frac{edT}{s}\right)\min\left(\sqrt{s\left(2d + \sum_{t=1}^{T-1}\overline{IR}_t^{(2)}(\pi_t)\right)}, s^{\frac{1}{3}}\left(\frac{3\sqrt{6}s}{\sqrt{C_{\min}}} + \sum_{t=1}^{T}\sqrt{\overline{IR}_t^{(3)}(\pi_t)}\right)^{\frac{2}{3}}\right) \tag{8}$$

$$= \mathcal{O}\left(\log\frac{edT}{s}\min\left(\sqrt{s\left(2d + \sum_{t=1}^{T-1}\overline{IR}_t^{(2)}(\pi_t)\right)}, s^{\frac{1}{3}}\left(\frac{3\sqrt{6}s}{\sqrt{C_{\min}}} + \sum_{t=1}^{T}\sqrt{\overline{IR}_t^{(3)}(\pi_t)}\right)^{\frac{2}{3}}\right)\right).$$

This type of result is only possible because our novel analysis of the optimistic posterior (cf. Theorem 1) can handle history-dependent learning rates. A full proof is provided in Appendix D. This result shows that (with appropritate choices of the learning rates) SOIDS is fully adaptive to which of the two regimes is best. Because our analysis requires decreasing learning rates, we are forced to leave the $\log(T)$ terms out of the learning rates, and our logarithmic term has a worse power than in the bound of Theorem 2. An interesting open question is whether it is possible to improve the dependency on this logarithmic term while still using data-dependent learning rates.

5 Analysis

We now provide an outline of the proofs of the main results.

5.1 Proof of Theorem 1

A key observation is that the optimistic posterior can be interpreted as a learner playing an auxiliary online learning game over distributions $\Delta(\Theta)$. The loss of that game is a weighted sum of negative log-likelihood and estimation error losses. We define

$$L_t^{(1)}(\theta) = \sum_{s=1}^{t} \log\left(\frac{1}{p(Y_s|\theta, A_s)}\right) = \sum_{s=1}^{t} \frac{1}{2}\left(\langle\theta, A_s\rangle - Y_s\right)^2$$

to be the *cumulative negative log-likelihood loss* of θ and

$$L_t^{(2)}(\theta) = \sum_{s=1}^{t} -\Delta(A_s, \theta)$$

to be the *cumulative estimation error loss* of θ. In addition, we define the regularizer $\Phi : \Delta(\Theta) \to \mathbb{R}$ by the mapping $P \mapsto \mathcal{D}_{\text{KL}}\left(P \middle\| Q_1^+\right)$, which is the KL-divergence with respect to the prior Q_1^+. With those notations, the optimistic posterior corresponds to an instance of the Follow the Regularized Leader (FTRL) algorithm introduced by Hazan and Kale [2010] and Abernethy et al. [2008]. FTRL is a standard method in online convex optimization that balances cumulative loss minimization with a regularization term to enforce stability and guarantee controlled regret. The update can be reframed as

$$Q_{t+1}^+ = \underset{P\in\Delta(\Theta)}{\arg\min}\langle P, \eta L_t^{(1)} + \lambda_t L_t^{(2)}\rangle + \Phi(P).$$

This formulation enables the application of tools from convex analysis and online learning, such as Fenchel duality, to derive regret bounds for this auxiliary online learning game and to understand the interplay between the two losses under the learning rates η and λ_t. We now focus on the case in which the learning rates λ_t don't depend on the history and relegate the analysis of history-dependent learning rates to Appendix C. The following lemma provides a bound on the average regret when the model θ_0 is drawn from an arbitrary comparator distribution P.

Lemma 2. *Let $P \in \Delta(\Theta)$ be any comparator, then the following bound holds*

$$\sum_{t=1}^{T} \Delta(P, A_t) \le \frac{\mathcal{D}_{KL}\left(P \middle\| Q_1^+\right)}{\lambda_T} + \frac{\Phi^*(\eta(L_T^{(1)}(\theta_T) - L_T^{(1)}(\cdot)) - \lambda_T L_T^{(2)}(\cdot))}{\lambda_T} + \frac{\eta}{\lambda_T}(P\cdot L_T^{(1)} - L_T^{(1)}(\theta_T)).$$

Here $\theta_t = \arg\min_{\theta\in\Theta} L_t^{(1)}(\theta)$ denotes the maximum likelihood estimator at time t, and $\Phi^*(L) = \log\int_\Theta \exp(L(\theta))dQ_1^+(\theta)$ is the Fenchel dual of the regularizer Φ. A complete proof of this result is provided in appendix B.1.1. We aim to chose a comparator P and the prior Q_1^+ such that P is concentrated around θ_0 and the KL divergence $\mathcal{D}_{\text{KL}}\left(P \middle\| Q_1^+\right)$ is controlled. If the parameter space were finite, the natural choice would be to take P as a Dirac on θ_0 and Q_1^+ as a uniform distribution on the whole parameter space; more care is necessary here. Choosing Q_1^+ as a subset-selection prior and P as a uniform distribution on a sparse neighborhood of θ_0 satisfies both requirements.

Lemma 3. *The subset-selection prior $Q_1^+ \in \Delta(\Theta)$ verifies that for any $\epsilon > 0$ and $\theta \in \Theta$, there is a comparator $P(\theta) \in \Delta(\Theta)$ satisfying both*

$$\forall\theta' \in \text{supp}(P(\theta)), \|\theta - \theta'\|_1 \le \epsilon \quad \textit{and} \quad \mathcal{D}_{KL}\left(P(\theta) \middle\| Q_1^+\right) \le s\log\frac{2ed}{\epsilon s}.$$

The proof of this lemma, as well as the exact choice of the prior Q_1^+ and the comparator $P(\theta_0)$, are provided in Appendix B.2. In Appendix **??** (cf. Lemma 21), we establish that both $L_T^{(2)}(\cdot)$ and $\mathbb{E}\left[L_T^{(1)}(\cdot)\right]$ are $2T$-Lipschitz with respect to the ℓ_1-norm. Hence,

$$\mathbb{E}\left[\frac{|P \cdot L_T^{(1)} - L_T^{(1)}(\theta_0)|}{\lambda_T}\right] \leq \frac{2T\epsilon}{\lambda_T}, \quad \text{and} \quad \sum_{t=1}^{T} |\Delta(\theta_0, A_t) - \Delta(P, A_t)| \leq 2T\epsilon.$$

Combining these with Lemma 2, we obtain the following bound on the cumulative regret:

$$\begin{aligned} R_T \leq & \mathbb{E}\left[\frac{s \log \frac{2ed}{\epsilon s} + 2T(\lambda_T + \eta)\epsilon}{\lambda_T} + \frac{\Phi^*(-\eta(L_T^{(1)}(\cdot) - L_T^{(1)}(\theta_T)) - \lambda_T L_T^{(2)}(\cdot))}{\lambda_T}\right] \\ & + \mathbb{E}\left[\frac{\eta}{\lambda_T}(L_T^{(1)}(\theta_0) - L_T^{(1)}(\theta_T))\right]. \end{aligned}$$

The first term balances model complexity and approximation via ϵ. In the usual FTRL analysis, $\lambda \to \frac{\phi^*(\lambda L)}{\lambda}$ is non decreasing for any $L \in \mathbb{R}^\Theta$, and the term involving Φ^* can be telescoped. Things are more complex here because only some part of the loss is weighted by the time varying learning rate λ_T. Through a careful analysis involving the maximum likelihood estimator, we can decompose the Φ^* term into a telescoping sum and a remainder term.

Lemma 4.

$$\begin{aligned} & \frac{\Phi^*(\eta(L_T^{(1)}(\theta_T) - L_T^{(1)}(\cdot)) - \lambda_T L_T^{(2)}(\cdot))}{\lambda_T} \\ & \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\Phi^*(\eta(L_t^{(1)}(\theta_0) L_t^{(1)}(\cdot)) - \lambda_{t-1} L_t^{(2)}(\cdot))}{\lambda_{t-1}} - \frac{\Phi^*(\eta(L_{t-1}^{(1)}(\theta_0) - L_{t-1}^{(1)}(\cdot)) - \lambda_{t-1} L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}}\right] \end{aligned} \tag{9}$$

$$+ \frac{\eta(6 + s \log \frac{edT}{s})}{\lambda_T}. \tag{10}$$

A detailed proof of this result is provided in Appendix B.1.4. Finally, the remaining sum can be handled by looking at the explicit formula for Φ^*. The terms related to the likelihood and the gap estimates can be separated using Hölder's inequality, as is done in Zhang [2022] and Neu, Papini, and Schwartz [2024]. More explicitly, by now choosing $\eta = \frac{1}{4}$, we obtain the following lemma.

Lemma 5.

$$\begin{aligned} & \mathbb{E}\left[\sum_{t=1}^{T} \frac{\Phi^*(\eta(L_t^{(1)}(\theta_0) - L_t^{(1)}(\cdot)) - \lambda_{t-1} L_t^{(2)}(\cdot))}{\lambda_{t-1}} - \frac{\Phi^*(\eta(L_{t-1}^{(1)}(\theta_0) - L_{t-1}^{(1)}(\cdot)) - \lambda_{t-1} L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}}\right] \\ & \quad \leq \mathbb{E}\left[-\sum_{t=1}^{T} \frac{3\overline{IG}_t(\pi_t)}{32\lambda_{t-1}} + 2\sum_{t=1}^{T} \widehat{\Delta}(\pi_t)\right]. \end{aligned} \tag{11}$$

A full proof of this result is provided in Appendix B.1.4. Combining Lemmas 2, 3, 4, 5 and setting $\epsilon = \frac{2}{T}$, we obtain the desired regret bound stated in Theorem 1.

5.2 Proof of Theorem 2

We show how Theorem 1 can be combined with bounds on the surrogate regret to control the true regret. The first important fact is that the surrogate regret of any policy can always be controlled in terms of the 2 or the 3-surrogate information ratio of that policy.

Lemma 6. *Let $\lambda > 0$, then we have that for any policy $\pi \in \Delta(\mathcal{A})$*

$$\widehat{\Delta}_t(\pi) \leq \frac{\overline{IG}_t(\pi)}{\lambda} + \min\left(\frac{1}{4}\lambda \overline{IR}_t^{(2)}(\pi), c_3^* \sqrt{\lambda \overline{IR}_t^{(3)}(\pi)}\right),$$

where $c_3^* < 2$ is an absolute constant defined in Lemma 27.

This is a consequence of a simple generalization of the AM-GM inequality and is proved in Appendix F.1. Combining the previous lemma with $\lambda = \frac{64}{3}\lambda_{t-1}$ and Theorem 1, we can further upper bound the regret of a sequence of policies $(\pi_t)_t$ as

$$
\begin{aligned}
R_T &\le \mathbb{E}\left[\frac{5 + 2s\log\frac{edT}{s}}{\lambda_{T-1}} - \sum_{t=1}^{T}\frac{3\overline{\mathrm{IG}}_t(\pi_t)}{32\lambda_{t-1}} + 2\sum_{t=1}^{T}\widehat{\Delta}_t(\pi_t)\right] \\
&\le \mathbb{E}\left[\frac{C_T}{\lambda_{T-1}} + \sum_{t=1}^{T}\min\left(\frac{32}{3}\lambda_{t-1}\overline{\mathrm{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}\right)\right], \qquad (12)
\end{aligned}
$$

where $C_T = 5 + 2s\log\frac{edT}{s}$. Usually, bounds on the 2-information ratio can be converted to $\mathcal{O}(\sqrt{T})$ bounds and bounds on the 3-information ratio can be converted to $\mathcal{O}(T^{\frac{2}{3}})$ bounds. Hence we will use the 2-information ratio to control the regret in the data-rich regime and the 3-information ratio to control the regret in the data-poor regime. Due to Lemma 1, the SOIDS policy minimizes the 2-information ratio and approximately minimizes the 3-information ratio. As a result, if there exists a "forerunner" algorithm with bounded 2-information ratio or 3-information ratio, SOIDS inherits these bounds automatically. In particular, we can use a different forerunner for each regime and SOIDS will match the regret guarantees of the best forerunner in each regime.

This forerunner-based technique is widely used to analyze IDS based algorithms and has been applied to a variety of Bayesian settings [Russo and Van Roy, 2017, Hao et al., 2021, Hao and Lattimore, 2022] and some frequentist settings [Kirschner and Krause, 2018, Kirschner et al., 2020, 2021]. An advantage of the OIDS framework is that since the surrogate quantities are defined with respect to the optimistic posterior, the analysis of the surrogate information ratio is virtually identical to the corresponding analysis of the information ratio in the Bayesian setting.

The forerunner we consider for the 2-information ratio is the *Feel-Good Thompson Sampling* (FGTS) algorithm of Zhang [2022]. For the 3-information ratio, we consider a mixture of the FGTS policy and an exploratory policy. The following lemma provides bounds on the surrogate information ratios of the SOIDS algorithm.

Lemma 7. *The* 2- *and* 3-*surrogate-information ratio of the SOIDS algorithm satisfy for any* $t \ge 0$

$$\overline{IR}_t^{(2)}(\pi_t^{(\mathbf{SOIDS})}) \le \overline{IR}_t^{(2)}(\pi_t^{(\mathbf{FGTS})}) \le 2d \qquad (13)$$

and

$$\overline{IR}_t^{(3)}(\pi_t^{(\mathbf{SOIDS})}) \le 2\overline{IR}_t^{(3)}(\pi_t^{(\mathbf{mix})}) \le \frac{54s}{C_{\min}}. \qquad (14)$$

The explicit definition of both forerunner algorithms, as well as the proof of this lemma, are deferred to Appendix F.3. Finally, it remains to pick the learning rate λ_t. The following lemma describes the appropriate learning rate for the data-poor and the data-rich regimes separately.

Lemma 8. *The choice of learning rate* $\lambda_t^{(2)} = \sqrt{\frac{3C_{t+1}}{128d(t+1)}}$ *guarantees*

$$\frac{C_T}{\lambda_{T-1}^{(2)}} + \frac{32}{3}\sum_{t=1}^{T}\lambda_{t-1}^{(2)}\overline{IR}_t^{(2)}(\pi_t) \le 16\sqrt{\frac{2}{3}C_T dT}.$$

The choice of learning rate $\lambda_t^{(3)} = \frac{1}{4\cdot 6^{\frac{1}{3}}}\left(\frac{C_{t+1}\sqrt{C_{\min}}}{(t+1)\sqrt{s}}\right)^{\frac{2}{3}}$ *guarantees*

$$\frac{C_T}{\lambda_{T-1}^{(3)}} + \frac{16}{3}c_3^*\sum_{t=1}^{T}\sqrt{3\lambda_{t-1}^{(3)}\overline{IR}_t^{(3)}(\pi_t)} \le 12\cdot 6^{\frac{1}{3}}\left(\frac{s\cdot C_T}{C_{\min}}\right)^{\frac{1}{3}}T^{\frac{2}{3}}.$$

The proof is deferred to Appendix G.2. It remains to analyze what happens when the learning rate $\lambda_t = \min(\frac{1}{2}, \max(\lambda_t^{(2)}, \lambda_t^{(3)}))$ is chosen. We defer this to Appendix G.4.

6 Experiments

We aim to verify that, in both the data-rich and data-poor regimes simultaneously, the regret of SOIDS is comparable with the regret of existing algorithms that achieve near optimal worst-case

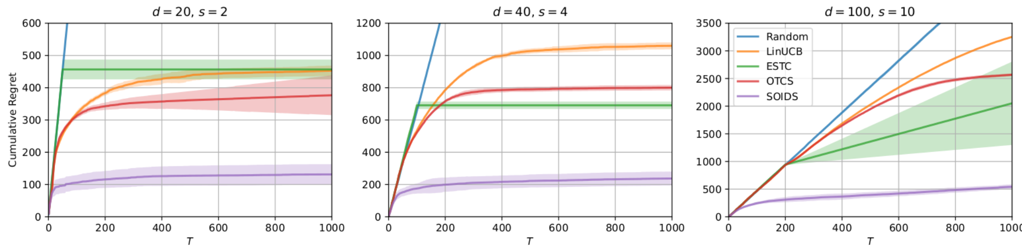

Figure 1: Cumulative regret for $d = 20$ (left) 40 (middle) and 100 (right). We plot the mean $\pm$ standard deviation over 10 repetitions.

regret in either the data-rich or the data-poor regime. Our baseline for the data-rich regime is the online-to-confidence-set (OTCS) method proposed by Abbasi-Yadkori et al. [2012], which has worst case regret of the order $\sqrt{sdT}$. For a tougher comparison, we run this method with the confidence sets from Theorem 4.7 of Clerico et al. [2025], which have much smaller constant factors than those used by Abbasi-Yadkori et al. [2012]. Our baseline for the data-poor regime is the Explore the Sparsity Then Commit (ESTC) algorithm proposed by Hao et al. [2020], which has worst-case regret of the order $(sT)^{2/3}$. For reference, we also compare with LinUCB Abbasi-Yadkori et al. [2011], which does not adapt to sparsity.

It is generally difficult to run the SOIDS algorithm exactly because the surrogate information ratio contains expectations w.r.t. the optimistic posterior. In our implementation of SOIDS, we use the empirical Bayesian sparse sampling procedure of Hao et al. [2021] to draw approximate samples from the optimistic posterior, and then approximate the surrogate information ratio via sample averages. We provide further details regarding the implementations of each method in Appendix J.

For each $d \in \{20, 40, 100\}$, θ_0 is the s-sparse vector in $\mathbb{R}^d$, with $s = d/10$, in which first s components are $10/s$ and the remaining components are zero. The action set consists of 200 random draws from the uniform distribution on $[-1, 1]^d$. The noise variance is 1 and we run each method 10 times. In Figure 1, we report the cumulative regret over $T = 1000$ steps. As d is varied from 20 to 100, we appear to transition from the data-rich regime to the data-poor regime: for $d = 20$, the OTCS method is the best performing baseline, whereas for $d = 100$, ETCS is the best performing baseline. As our theoretical results would suggest, SOIDS performs well in both regimes.

7 Conclusion

There remain several interesting questions that our work leaves open for future research, such as the possibility of improving the logarithmic terms in the data-dependent best-of-both-worlds guarantees (as mentioned earlier in Section 4). We highlight another question below.

In our experiments, we have made use of an approximate implementation of OIDS adapted from Hao et al. [2021]. The initial success we have seen in our experiments suggests that this approximation might be viable in more challenging settings, and worthy of an attempt at a solid theoretical analysis. More broadly, the results indicate a potential advantage of IDS-style methods over DEC-inspired methods [Foster et al., 2022b, Kirschner et al., 2023]. Indeed, we are not aware of any general methods for approximating the optimization problems that the E2D algorithm of Foster et al. [2022b] requires to solve, in contrast to our results that indicate that IDS-inspired algorithms may very well be amenable to practical implementation. Whether the concrete approximation we used in our experiments is the best possible one or not remains to be seen.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. volume 22 of *JMLR Proceedings*, pages 1–9, 2012. URL http://proceedings.mlr.press/v22/abbasi-yadkori12.html.

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. pages 263–274, 2008. URL http://colt2008.cs.helsinki.fi/papers/127-Abernethy.pdf.

Pierre Alquier and Karim Lounici. Pac-bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

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

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Sébastien Bubeck and Mark Sellke. First-Order Bayesian Regret Analysis of Thompson Sampling, 2022. URL http://arxiv.org/abs/1902.00681.

Sunrit Chakraborty, Saptarshi Roy, and Ambuj Tewari. Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR, 2023.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. volume 15 of *JMLR Proceedings*, pages 208–214, 2011. URL http://proceedings.mlr.press/v15/chu11a/chu11a.pdf.

Eugenio Clerico, Hamish Flynn, Wojciech Kotowski, and Gergely Neu. Confidence sequences for generalized linear models via regret analysis, 2025. URL https://arxiv.org/abs/2504.16555.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, volume 2, page 3, 2008.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles, 2020. URL http://arxiv.org/abs/2002.04926.

Dylan J. Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. A Note on Model-Free Reinforcement Learning with the Decision-Estimation Coefficient, 2022a. URL http://arxiv.org/abs/2211.14250.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The Statistical Complexity of Interactive Decision Making, 2022b. URL http://arxiv.org/abs/2112.13487.

Dylan J. Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the Complexity of Adversarial Decision Making, 2022c. URL http://arxiv.org/abs/2206.13063.

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *The Journal of Machine Learning Research*, 14(1):729–769, 2013.

Botao Hao and Tor Lattimore. Regret bounds for information-directed reinforcement learning. 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b733cdd80ed2ae7e3156d8c33108c5d5-Abstract-Conference.html.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. 2020. URL https://proceedings.neurips.cc/paper/2020/hash/7a006957be65e608e863301eb98e1808-Abstract.html.

Botao Hao, Tor Lattimore, and Wei Deng. Information directed sampling for sparse linear bandits. pages 16738–16750, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8ba6c657b03fc7c8dd4dff8e45defcd2-Abstract.html.

Botao Hao, Tor Lattimore, and Chao Qin. Contextual Information-Directed Sampling, 2022. URL http://arxiv.org/abs/2205.10895.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine Learning*, 80(2-3):165–188, 2010. doi: 10.1007/S10994-010-5175-X. URL https://doi.org/10.1007/s10994-010-5175-x.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits. *Advances in Neural Information Processing Systems*, 35:2102–2114, 2022.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

Johannes Kirschner and Andreas Krause. Information Directed Sampling and Bandits with Heteroscedastic Noise, 2018. URL http://arxiv.org/abs/1801.09667.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. volume 125 of *Proceedings of Machine Learning Research*, pages 2328–2369, 2020. URL http://proceedings.mlr.press/v125/kirschner20a.html.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. volume 134 of *Proceedings of Machine Learning Research*, pages 2777–2821, 2021. URL http://proceedings.mlr.press/v134/kirschner21a.html.

Johannes Kirschner, Seyed Alireza Bakhtiari, Kushagra Chandak, Volodymyr Tkachuk, and Csaba Szepesvári. Regret minimization via saddle point optimization. 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6eaf8c729af4fbeb18006dc2e6a41d9b-Abstract-Conference.html.

Tor Lattimore and András György. Mirror Descent and the Information Ratio. volume 134 of *Proceedings of Machine Learning Research*, pages 2965–2992, 2021. URL http://proceedings.mlr.press/v134/lattimore21b.html.

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

Gergely Neu, Julia Olkhovskaya, Matteo Papini, and Ludovic Schwartz. Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits, 2022. URL http://arxiv.org/abs/2205.13924.

Gergely Neu, Matteo Papini, and Ludovic Schwartz. Optimistic information directed sampling. volume 247 of *Proceedings of Machine Learning Research*, pages 3970–4006, 2024. URL https://proceedings.mlr.press/v247/neu24a.html.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

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

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17:68:1–68:30, 2016. URL https://jmlr.org/papers/v17/14-087.html.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling, 2017. URL http://arxiv.org/abs/1403.5556.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. volume 32 of *JMLR Workshop and Conference Proceedings*, pages 46–54, 2014. URL http://proceedings.mlr.press/v32/valko14.html.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. 2019. ISBN 978-1-108-49802-9. URL https://books.google.es/books?id=8C8nuQEACAAJ.

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. PMLR, 2018.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022. doi: 10.1137/21M140924X. URL https://doi.org/10.1137/21m140924x.

A Related work

The first algorithms and regret bounds for sparse linear bandits were designed for the data-rich regime. Abbasi-Yadkori et al. [2012] developed an online-to-confidence-set conversion for linear models, which converts any algorithm for online linear regression into a linear bandit algorithm whose regret depends on the regret of the online regression algorithm. When the SeqSEW algorithm [Gerchinovitz, 2013] is used in this conversion, the result is a sparse linear bandit algorithm with a regret bound of the order $\mathcal{O}(\sqrt{sdT})$ (ignoring logarithmic factors). Lattimore and Szepesvári [2020] established a matching lower bound for the data-rich regime, showing that this rate cannot be improved.

More recently, several works have studied the data-poor regime, in which the dimension d is much larger than the number of rounds T. Hao et al. [2020] showed that an explore-then-commit algorithm satisfies a regret bound of the order $O((sT)^{2/3}C_{\min}^{-2/3})$, and established a lower bound of order $\Omega(\min(s^{1/3}T^{2/3}C_{\min}^{-1/3}, \sqrt{dT})$. Subsequently, Jang et al. [2022] proposed the PopArt estimator for sparse linear regression, and showed that an explore-then-commit algorithm that uses this estimator achieves a regret bound of the order $\mathcal{O}(s^{2/3}T^{2/3}H_\star^{2/3})$, where $H_\star$ is another problem-dependent quantity that satisfies $H_\star^2 \leq C_{\min}^{-1}$. In addition, Jang et al. [2022] established a lower bound of order $\Omega(s^{2/3}T^{2/3}C_{\min}^{-1/3})$, showing that the optimal rate for the data-poor regime is $s^{2/3}T^{2/3}$. Hao et al. [2021] showed that sparse IDS has a Bayesian best of both worlds/regimes regret bound.

A number of works have considered the setting of sparse contextual linear bandits, in which the action set $\mathcal{A}$ changes in each round t. In the case where the actions sets are chosen by an adaptive adversary, the upper and lower bounds of the order $\sqrt{sdT}$ by Abbasi-Yadkori et al. [2012] and Lattimore and Szepesvári [2020] respectively still hold. Under the assumption that the action sets are generated randomly, and such that either a uniform or greedy policy is (with high probability) exploratory, several methods have been shown to achieve nearly dimension-free regret bounds Bastani and Bayati [2020], Wang et al. [2018], Kim and Paik [2019], Oh et al. [2021], Chakraborty et al. [2023].

The concept of balancing instantaneous regret and information gain through the information ratio was first introduced by Russo and Roy [2016] in the context of analyzing Thompson Sampling. Building upon this, the Information-Directed Sampling (IDS) algorithm was proposed by Russo and Van Roy [2017] to directly minimize the information ratio, thereby optimizing the trade-off between regret and information gain. These foundational ideas have since been extended and applied to a variety of settings including bandits [Bubeck and Sellke, 2022], contextual bandits [Neu et al., 2022, Hao et al., 2022], reinforcement learning [Hao and Lattimore, 2022], and sparse linear bandits [Hao et al., 2021]. However, these works are primarily situated in the Bayesian framework and focus on Bayesian regret bounds that hold only in expectation with respect to the prior distribution.

A key challenge in extending these methods to the frequentist setting lies in estimating the instantaneous regret and define a meaningful notion of information gain. Both of those things are naturally possible in Bayesian analysis but difficult when the true model is unknown. Moreover, Bayesian posteriors may inadequately represent model uncertainty from a frequentist perspective. We highlight three strands of research that have attempted to address this challenge:

Confidence-set based information ratio approaches: Works such as Kirschner and Krause [2018], Kirschner et al. [2020], and Kirschner et al. [2021] extend the notion of the information ratio to frequentist settings by constructing high-probability confidence sets for the instantaneous regret and information gain. These results are mostly limited to setting with some linear structure.

Distributionally robust and worst-case information-regret trade-offs: The Decision-to-Estimation-Coeffiecient(DEC) line of work of [Foster et al., 2022b, Foster and Rakhlin, 2020, Foster et al., 2022c,a, Kirschner et al., 2023] explores the frequentist setting by analyzing worst-case trade-offs between regret and information gain. One limitation is that the DEC is an inherently worst-case measure of comlexity. Moreover, algorithms based on the DEC usually require solving complex min-max optimization problems at each time step, making their practical implementation challenging and unclear.

Optimistic posterior approaches for frequentist guarantees: The approach most closely related to our work modifies the Bayesian posterior to provide frequentist guarantees. Introduced by Zhang [2022], the optimistic posterior is a modification of the Bayesian posterior which enables frequentist regret bounds for a variant of Thompson Sampling. Subsequently, Neu et al. [2024] studied the optimistic posterior framework in greater depth, defining a frequentist analog of the information ratio to extend IDS to frequentist settings. A notable limitation of these works is their restriction to constant learning rates in the optimistic posterior, which limits adaptivity, an issue that we address in this paper.

B Analysis of the Optimistic posterior

This section provides further details about the prior underlying the optimistic posterior and guarantees on the posterior updates.

B.1 Follow the regularized leader analysis

The main step in our analysis of the optimistic posterior is to leverage the follow the regularized leader formulation of our optimistic posterior update

$$Q_{t+1}^{+} = \operatorname*{arg\,min}_{P\in\Delta(\Theta)} \langle P, \eta L_t^{(1)} + \lambda_t L_t^{(2)} \rangle + \Phi(P).$$

B.1.1 Proof of lemma 2

As is usual in the analysis of the follow the regularized leader algorithm, we introduce the Fenchel conjugate of the regularization function $\Phi = \mathcal{D}_{\mathrm{KL}}\left(\cdot \middle\| Q_1^{+}\right)$ as the function $\Phi^* : \mathbb{R}^\Theta \to \mathbb{R}$ taking values $\Phi^*(L) = \sup_{P\in\Delta(\Theta)} \{\langle P, L\rangle - \phi(P)\}$. The Fenchel–Young inequality guarantees that for any $P \in \Delta(\Theta), L \in \mathbb{R}^\Theta$, we have

$$\langle P, L\rangle \leq \Phi(P) + \Phi^*(L)$$

We now introduce the maximum likelihood estimator $\theta_t = \operatorname*{arg\,min}_{\theta\in\Theta} L_t^{(1)}(\theta)$ and let $L = -\eta(L_T^{(1)}(\cdot) - L_T^{(1)}(\theta_T)) - \lambda_T L_T^{(2)}(\cdot)$. Since λ_T is never used by the algorithm, we can further assume that $\lambda_T = \lambda_{T-1}$. The role of the maximum likelihood estimator is to make sure that the term $L_t^{(1)}(\theta) - L_t^{(1)}(\theta_t)$ is always non-negative. Applying Fenchel–Young to L gives us the following bound:

$$\eta\left(L_T^{(1)}(\theta_T) - \langle P, L_T^{(1)}\rangle\right) - \lambda_T \langle P, L_T^{(2)}\rangle \leq \Phi(P) + \Phi^*\left(-\eta(L_T^{(1)}(\cdot) - L_T^{(1)}(\theta_T)) - \lambda_T L^{(2)}(\cdot)\right)$$

Noticing that $\langle P, L_T^{(1)}\rangle = -\sum_{t=1}^{T} \Delta(P, A_t)$ and rearranging the terms concludes the proof.

B.1.2 Proof of Lemma 4

We start by rewriting the potential function in the form of the following telescopic sum:

$$\begin{aligned}&\frac{\Phi^*(-\eta(L_T^{(1)}(\cdot) - L_T^{(1)}(\theta_T)) - \lambda_T L_T^{(2)}(\cdot))}{\lambda_T}\\ &= \sum_{t=1}^{T} \frac{\Phi^*(-\eta(L_t^{(1)}(\cdot) - L_t^{(1)}(\theta_t)) - \lambda_t L_t^{(2)}(\cdot))}{\lambda_t} - \frac{\Phi^*(-\eta(L_{t-1}^{(1)}(\cdot) - L_{t-1}^{(1)}(\theta_{t-1})) - \lambda_{t-1} L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}}.\end{aligned}$$

In the usual follow-the-regularized-leader analysis, we use the fact that $\lambda \to \frac{\phi^*(\lambda L)}{\lambda}$ is non-decreasing for any $L \in \mathbb{R}^\Theta$. Here however, only some of the linear loss is scaled by λ_t and the usual FTRL analysis fails. Crucially, because we introduced the maximum likelihood estimator θ_t, we have that $L_t^{(1)}(\cdot) - L_t^{(1)}(\theta_t) \geq 0$ and we can instead use the following lemma that guarantees that a scaled and shifted dual is monotonous.

Lemma 9. *Let $\Phi \geq 0, \Phi^*$ be a convex function and its dual as defined previously, $L_1, L_2 \in \mathbb{R}^\Theta$ with $L_1 \geq 0$, then $\lambda \in \mathbb{R}^{+*} \to \frac{\Phi^*(-L_1+\lambda L_2)}{\lambda}$ is a non-decreasing function.*

Proof. By definition, we have

$$\begin{aligned}\frac{\Phi^*(-L_1 + \lambda L_2)}{\lambda} &= \frac{\sup_{P\in\Delta(\Theta)}\langle P, -L_1 + \lambda L_2\rangle - \Phi(P)}{\lambda}\\ &= \sup_{P\in\Delta(\Theta)} \langle P, L_2\rangle - \frac{\langle P, L_1\rangle + \Phi(P)}{\lambda}.\end{aligned}$$

For any $P \in \Delta(\Theta)$, we have that $\Phi(P) + \langle P, L_1 \rangle \geq 0$ and the term inside the supremum is non-decreasing with respect to lambda. Since the supremum of non-decreasing functions is also non-decreasing, this concludes the proof. □

Applying the previous lemma, we upper bound the previous sum by replacing each λ_t factor by λ_{t-1} (using the convention $\lambda_0 = 1/2$), and then we replace the maximum likelihood estimator θ_t by θ_0 inside Φ^* to obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\Phi^*(-\eta(L_t^{(1)}(\cdot) - L_t^{(1)}(\theta_t)) - \lambda_t L_t^{(2)}(\cdot))}{\lambda_t} - \frac{\Phi^*(-\eta(L_{t-1}^{(1)}(\cdot) - L_{t-1}^{(1)}(\theta_{t-1})) - \lambda_{t-1} L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}} \\
&\leq \sum_{t=1}^{T} \frac{\Phi^*(-\eta(L_t^{(1)}(\cdot) - L_t^{(1)}(\theta_t)) - \lambda_t L_t^{(2)}(\cdot))}{\lambda_{t-1}} - \frac{\Phi^*(-\eta(L_{t-1}^{(1)}(\cdot) - L_{t-1}^{(1)}(\theta_{t-1})) - \lambda_t L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}} \\
&= \sum_{t=1}^{T} \frac{\Phi^*(-\eta(L_t^{(1)}(\cdot) - L_t^{(1)}(\theta_0)) - \lambda_t L_t^{(2)}(\cdot))}{\lambda_{t-1}} - \frac{\Phi^*(-\eta(L_{t-1}^{(1)}(\cdot) - L_{t-1}^{(1)}(\theta_0)) - \lambda_t L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}} \\
&+ \frac{\eta}{\lambda_{t-1}} (L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0) + L_{t-1}^{(1)}(\theta_0) - L_{t-1}^{(1)}(\theta_{t-1})).
\end{aligned}$$

It remains to bound the difference of the negative log likelihood of the true parameter and the maximum likelihood estimator. This is done via the following result (whose proof we relegate to appendix E.1.1).

Lemma 10. *For any $t \geq 1$, we have*

$$0 \leq \mathbb{E}\left[L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t)\right] \leq \inf_{\rho} \left\{2\rho t + s \log \frac{ed(1 + 2/\rho)}{s}\right\} \leq 6 + s \log \frac{edt}{s} \tag{15}$$

Using this lemma, we can further bound the previously considered expression as the following telescopic sum:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \frac{\eta}{\lambda_{t-1}} (L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0) + L_{t-1}^{(1)}(\theta_0) - L_{t-1}^{(1)}(\theta_{t-1})) + \frac{\eta}{\lambda_T}(L_T^{(1)}(\theta_0) - L_T^{(1)}(\theta_T))\right] \\
&\quad = \mathbb{E}\left[\sum_{t=1}^{T} \frac{\eta}{\lambda_{t-1}} (L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0)) - \sum_{t=1}^{T} \frac{\eta}{\lambda_t}(L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0))\right] \\
&\quad \leq \eta \cdot \sum_{t=1}^{T} \mathbb{E}\left[(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t))\right]\left(\frac{1}{\lambda_t} - \frac{1}{\lambda_{t-1}}\right) \\
&\quad \leq \frac{\eta(6 + s \log \frac{edT}{s})}{\lambda_T}.
\end{aligned}$$

Here, the first inequality comes from the non-negativity of $L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t)$ by definition of θ_t and the second one is from Lemma 10 just above and a telescoping argument. Finally we obtain the claim of Lemma 4.

B.1.3 Controlling the losses separately

The focus of this section is to understand how to control $\Phi^*(-L)$ where L is either the negative-likelihood loss or the estimation-error loss. We start by analyzing the negative-likelihood loss. As was done in Neu, Papini, and Schwartz [2024], we will relate the negative-likelihood loss to the surrogate information gain.

For this analysis, we define the *true information gain* as

$$\mathrm{IG}_t(\pi) = \frac{1}{2} \sum_{a \in \mathcal{A}} \pi(a) \int_{\Theta} (\langle \theta - \theta_0, a \rangle)^2 \, \mathrm{d}Q_t^+(\theta), \tag{16}$$

and note that, by linearity reward function, the surrogate information gain is always smaller than the true information gain. This is stated formally below.

Proposition 1. *For any policy $\pi \in \Delta(\mathcal{A})$ and any $t \geq 1$ we have that*

$$\overline{IG}_t(\pi) \leq IG_t(\pi) \tag{17}$$

The proof is provided in Appendix I.1. This result can then be used to relate the surrogate and the true information gain to the negative-likelihood loss. This result and its proof are identical to the proof of Lemma 17 in Neu, Papini, and Schwartz [2024].

Lemma 11. *Assume that the noise ϵ_t is conditionnally 1-sub-Gaussian, then for any $t \geq 1, \eta, \alpha \geq 0$ such that $\gamma = \frac{\eta\alpha}{2}(1 - \eta\alpha) > 0$, the following inequality holds*

$$\begin{aligned} \mathbb{E}\left[\log \int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^{\eta\alpha} dQ_t^+(\theta)\right] &\leq -2\gamma(1-2\gamma)\mathbb{E}\left[IG_t(\pi_t)\right] & (18)\\ &\leq -2\gamma(1-2\gamma)\mathbb{E}\left[\overline{IG}_t(\pi_t)\right]. & (19) \end{aligned}$$

In particular, the constant $2\gamma(1-2\gamma)$ can be maximized to the value $\frac{3}{16}$ by the choice $\eta\alpha = \frac{1}{2}$.

Proof. By the tower rule of expectation and Jensen's inequality applied to the logarithm, we have

$$\begin{aligned} \mathbb{E}\left[-\log \int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^{\eta\alpha}\right] &= \mathbb{E}\left[\mathbb{E}\left[-\log \int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^{\eta\alpha}\middle|\mathcal{F}_t, A_t\right]\right] \\ &\leq \mathbb{E}\left[-\log \mathbb{E}\left[\int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^{\eta\alpha}\middle|\mathcal{F}_t, A_t\right]\right] \\ &= \mathbb{E}\left[-\log \int_\Theta \mathbb{E}\left[\exp\left(-\eta\alpha\left(\frac{(Y_t - \langle\theta, A_t\rangle)^2}{2} - \frac{(Y_t - \langle\theta_0, A_t\rangle)^2}{2}\right)\right)\middle|\mathcal{F}_t, A_t\right]\right]. \end{aligned}$$

Now, we fix some $\theta \in \Theta$ and to simplify the notation, we let $r_0 = \langle\theta_0, A_t\rangle$ and $r = \langle\theta, A_t\rangle$. Using some elementary manipulations and the conditional sub-gaussianity of ϵ_t and $Y_t = r_0 + \epsilon_t$ which implies that for any $(\mathcal{F}_t, A_t)$-measurable ζ_t, $\mathbb{E}\left[\exp\left(Y_t\zeta_t\right)|\mathcal{F}_t, A_t\right] = \exp(r_0\zeta_t)\mathbb{E}\left[\exp\left(\epsilon_t\zeta_t\right)|\mathcal{F}_t, A_t\right] \leq \exp(r_0\zeta_t)\exp\left(\frac{\zeta_t^2}{2}\right)$, we have

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\eta\alpha\left(\frac{(Y_t - r)^2}{2} - \frac{(Y_t - r_0)^2}{2}\right)\right)\middle|\mathcal{F}_t, A_t\right] \\ &\quad= \mathbb{E}\left[\exp\left(-\frac{\eta\alpha}{2}(2Y_t - r - r_0)(r_0 - r)\right)\middle|\mathcal{F}_t, A_t\right] \\ &\quad= \exp\left(\eta\alpha\frac{r_0^2 - r^2}{2}\right)\mathbb{E}\left[\exp\left(\eta\alpha Y_t(r - r_0)\right)|\mathcal{F}_t, A_t\right] \\ &\quad\leq \exp\left(\eta\alpha\frac{r_0^2 - r^2}{2}\right)\cdot\exp\left(\eta\alpha r_0(r - r_0)\right)\exp\left(\frac{\eta^2\alpha^2}{2}(r - r_0)^2\right) \\ &\quad= \exp\left(-(r - r_0)^2\cdot\frac{\eta\alpha}{2}(1 - \eta\alpha)\right). \end{aligned}$$

Further, defining $\gamma = \frac{\eta\alpha}{2}(1 - \eta\alpha)$, we have

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\eta\alpha\left(\frac{(Y_t - r)^2}{2} - \frac{(Y_t - r_0)^2}{2}\right)\right)\middle|\mathcal{F}_t, A_t\right] \\ &\leq \exp(-\gamma(r - r_0)^2) \\ &\leq 1 - \gamma(r - r_0)^2 + \frac{\gamma^2}{2}(r - r_0)^4 \\ &\leq 1 - \gamma(r - r_0)^2 + 2\gamma^2(r - r_0)^2 \\ &\leq 1 - \gamma(1 - 2\gamma)(r - r_0)^2. \end{aligned}$$

Here, we used the elementary inequality $\exp(x) \leq 1 + x + \frac{x^2}{2}$ for $x \leq 0$ and then used $|r - r_0| \leq 2$. Finally, using that $\log x \leq x - 1$ for any $x > 0$, and taking the integral over Θ, we get that

$$\begin{aligned} \mathbb{E}\left[-\log \int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^{\eta\alpha}\right] &\leq -\gamma(1-2\gamma)\mathbb{E}\left[\sum_{a\in\mathcal{A}} \pi_t(A)\int_\Theta (\langle\theta - \theta_0, a\rangle)^2\right] dQ_t^+(\theta) \\ &= -2\gamma(1-2\gamma)\mathbb{E}\left[\mathrm{IG}_t(\pi_t)\right]. \end{aligned}$$

Rearranging and combining the result with Proposition 1 yields the claim of the lemma. □

We now turn our focus to the estimation error loss and relate it to the surrogate regret through the following lemma, whose proof is a straightforward application of Lemma 23.

Lemma 12. *For any $t \geq 1, \beta > 1$, if $\beta\lambda_{t-1} \leq 1$, we have*

$$\mathbb{E}\left[\frac{1}{\beta\lambda_{t-1}} \log \int_\Theta \exp(\beta\lambda_{t-1}\Delta(a_t, \theta))\, dQ_t^+(\theta)\right] \leq \mathbb{E}\left[2\widehat{\Delta}_t(\pi_t)\right]. \tag{20}$$

B.1.4 Separation of the two losses: proof of Lemma 5

We now make use of the fact that the Fenchel dual of Φ can be explicitly written as $\Phi^*(L) = \log \int_\Theta \exp(L(\theta))\, dQ_1(\theta)$. As a result, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \frac{\Phi^*(-\eta(L_t^{(1)}(\cdot) - L_t^{(1)}(\theta_0)) - \lambda_{t-1}L_t^{(2)}(\cdot))}{\lambda_{t-1}} - \frac{\Phi^*(-\eta(L_{t-1}^{(1)}(\cdot) - L_{t-1}^{(1)}(\theta_0)) - \lambda_{t-1}L_{t-1}^{(2)}(\cdot))}{\lambda_{t-1}}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \frac{1}{\lambda_{t-1}} \log \frac{\int_\Theta \left(\frac{p(Y_t|\theta,a_t)}{p(Y_t|\theta_0,A_t)}\right)^\eta \exp\left(\lambda_{t-1}\Delta(A_t,\theta)\right) \exp\left(-\eta L_{t-1}^{(1)}(\theta) - \lambda_{t-1}L_{t-1}^{(2)}(\theta)\right) dQ_1(\theta)}{\int_\Theta \exp\left(-\eta L_{t-1}^{(1)}(\theta) - \lambda_{t-1}L_{t-1}^{(2)}(\theta)\right) dQ_1(\theta)}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \frac{1}{\lambda_{t-1}} \log \int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^\eta \exp\left(\lambda_{t-1}\Delta(A_t,\theta)\right)\, dQ_t^+(\theta)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \frac{1}{\alpha\lambda_{t-1}} \log \int_\Theta \left(\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\right)^{\eta\alpha} + \frac{1}{\beta\lambda_{t-1}} \log \int_\Theta \exp\left(\beta\lambda_{t-1}\Delta(A_t,\theta)\right)\, dQ_t^+(\theta)\right],
\end{aligned}$$

where the last equality is by definition of the optimistic posterior and the last inequality follows from using Hölder's inequality with the two real numbers $\alpha, \beta > 1$ that satisfy $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Combining Lemma 11 and Lemma 12 with the choice $\alpha = \beta = 2$, the fact that $\eta = \frac{1}{4}$ and the last inequality yields the claim of the Lemma. ☐

B.2 Choice of the prior and comparator distribution: proof of Lemma 3

In order to construct the prior Q_1 and the comparator P for the regret analysis, we need to take into account two criteria: that $\mathcal{D}_{\text{KL}}(P\|Q_1)$ be controlled and that $|\langle P, L\rangle - L(\theta_0)|$ be small. Note that the comparator should be a function of the unknown parameter θ_0, and thus we denote it by $P(\theta_0)$. As for the prior, it should take into account the sparsity level of the unknown θ_0, but should have no access to its support.

For the prior, we first design a distribution Π over the set of all subsets of $[d] = \{1, \ldots, d\}$, which have cardinality at most s. We choose the distribution such that: a) the probability assigned to each subset depends only on its cardinality; b) the probability assigned to the set of all subsets of size k is proportional to 2^{-k}, where $1 \leq k \leq s$. In other words, we prefer smaller subsets and have no preference over which indices in $[d]$ are included. The distribution that satisfies these requirements is

$$\Pi(S) = \frac{2^{-|S|}}{\binom{d}{|S|}\sum_{k=1}^s 2^{-k}}. \tag{21}$$

For $S = \emptyset$, we set $\Pi(S) = 0$. Doing so only complicates matters if the support of θ_0 is empty (i.e., $\theta_0 = 0$). However, in this case, the reward function is 0 everywhere, which means any algorithm would have 0 regret. We therefore continue under the assumption that $\theta_0 \neq 0$. The most important property of this distribution, which we will use later, is that for any subset S of cardinality s, $\log(1/\Pi(S)) \leq s\log(2ed/s)$. For each subset S, we define Q_S to be the uniform distribution on Θ_S. The prior is defined to be

$$Q_1 = \sum_{S\subset[d]:|S|\leq s} \Pi(S)Q_S\,.$$

As for the comparator distribution $P(\theta_0)$, we would ideally like to take a Dirac measure on θ_0, but this would make the KL divergence appearing in the bound blow up. Thus, we pick a comparator P which dilutes its mass around θ_0. For any $\bar{\theta} \in \Theta$, with support $\bar{S}$, and any $\epsilon \in (0,1)$, we define the set $(1-\epsilon)\bar{\theta} + \epsilon\Theta_{\bar{S}} = \{(1-\epsilon)\bar{\theta} + \epsilon\theta' : \theta' \in \Theta_{\bar{S}}\} \subset \Theta_{\bar{S}}$. We will choose P to be the uniform distribution on $(1-\epsilon)\theta_0 + \epsilon\Theta_{S_0}$. We now bound $\Phi(P) = \mathcal{D}_{\text{KL}}(P\|Q_1)$ for this choice of P in the following lemma, from which the claim of Lemma 3 then directly follows.

Lemma 13. *For any $\bar\theta \in \Theta$, let $\bar S$ denote its support, and let $|\bar S| = s$. If, for $\epsilon \in (0,1)$, $P = \mathcal{U}((1-\epsilon)\bar\theta + \epsilon\Theta_{\bar S})$ and $Q_1 = \sum_{S\subset[d]:|S|=s} \Pi(S)Q_S$, then $\mathcal{D}_{KL}(P\|Q_1) \le s\log\frac{2ed}{\epsilon s}$.*

Proof. We notice that $(1-\epsilon)\bar\theta + \epsilon\Theta_{\bar S}$ is an s-dimensional L1 ball of radius ϵ, which is contained in $\Theta_{\bar S}$. Therefore, on the support of P, $\frac{\mathrm{d}P}{\mathrm{d}Q_{\bar S}}$ is equal to the ratio of the volumes of a unit L1 ball and an L1 ball of radius ϵ, which is $(1/\epsilon)^s$. Thus,

$$\mathcal{D}_{\mathrm{KL}}(P\|Q_1) = \int \log\frac{\mathrm{d}P}{\sum_S \Pi(S)\mathrm{d}Q_S}\mathrm{d}P \le \int \log\frac{\mathrm{d}P}{\Pi(\bar S)\mathrm{d}Q_{\bar S}}\mathrm{d}P \le s\log\frac{1}{\epsilon} + \log\frac{1}{\Pi(\bar S)}\,.$$

Using the definition of Π and the bound $\binom{d}{s} \le (\frac{ed}{s})^s$ on the binomial coefficient, we have

$$\log\frac{1}{\Pi(\bar S)} = \log\binom{d}{s} + s\log(2) + \log\sum_{k=1}^{s} 2^{-k} \le s\log\frac{2ed}{s}\,.$$

Combining everything, we obtain

$$\mathcal{D}_{\mathrm{KL}}(P\|Q_1) \le s\log\frac{1}{\epsilon} + s\log\frac{2ed}{s} = s\log\frac{2ed}{\epsilon s}\,, \tag{22}$$

as advertised. □

C Proof of the history-dependent part of Theorem 1

We now focus on the case in which λ_t is allowed to depend on the history. Following the original analysis, we arrive again at equation 2

$$\Delta(P,a_t) \le \frac{\mathcal{D}_{\mathrm{KL}}(P\|Q_1)}{\lambda_T} + \frac{\Phi^*(-\eta L_T^{(1)}(\cdot) + \eta L_T^{(1)}(\theta_T) + \lambda_T L_T^{(2)}(\cdot))}{\lambda_T} + \frac{\eta}{\lambda_T}(P\cdot L_T^{(1)} - L_T^{(1)}(\theta_T)),$$

where $P \in \Delta(\Theta)$ can be any comparator distribution. Lemma 3 is still valid and we can chose the same prior as before. We can still choose a comparator distribution supported on an ϵ-ball around θ_0. However, because λ_t depends on the history, we can no longer upper bound $\mathbb{E}\left[\frac{|P\cdot L_T^{(1)} - L_T^{(1)}(\theta_0)|}{\lambda_{T-1}}\right]$ by $\mathbb{E}\left[\frac{2T\epsilon}{\lambda_T}\right]$. Using Lemma 21, we still have that $L_T^{(2)}(\cdot)$ is $2T$-Lipschitz and $\mathbb{E}\left[L_T^{(1)}(\cdot)\right]$ is $2T$-Lipschitz. Hence,

$$\mathbb{E}\left[\frac{|P\cdot L_T^{(1)} - L_T^{(1)}(\theta_0)|}{\lambda_{T-1}}\right] \le 2T\epsilon C_{2,T}, \quad \text{and} \quad \sum_{t=1}^{T}|\Delta(\theta_0,a_t) - \Delta(P,a_t)| \le 2T\epsilon,$$

where we used $C_{2,T}$, a deterministic upper bound on $\frac{1}{\lambda_{T-1}}$. Exactly the same telescoping of Φ^* can be done, however because the learning rate is history-dependent, the difference between the negative log likelihood of θ_0 and θ_t must be treated with more care. We have the following lemma

Lemma 14. *Let $C_{1,T}$ be a deterministic upper bound on $\left(\frac{1}{\lambda_{t+1}} - \frac{1}{\lambda_t}\right)$ that holds for all $t < T$, then*

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta}{\lambda_{t-1}}(L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0) + L_{t-1}^{(1)}(\theta_0) - L_{t-1}^{(1)}(\theta_{t-1})) + \frac{\eta}{\lambda_T}(L_T^{(1)}(\theta_0) - L_T^{(1)}(\theta_T))\right]\\ &\le \mathbb{E}\left[\frac{\eta(12 + 3s\log\frac{2e^2dT^2C_{1,T}^2}{s})}{2\lambda_{T-1}}\right].\end{aligned} \tag{23}$$

A complete proof of that result can be found in appendix E.2.1.

Finally, as was the case in the history independent version the telescoping sum can be handled by looking at the explicit formula for Φ^* and Lemma 5 still holds. Applying Lemma 5 and setting $\epsilon = \frac{1}{TC_{2,T}}$ yields the claim of the theorem.

D Proof of Theorem 3

We turn our attention to data-dependent bounds (that will scale with the cumulative information ratio rather than the time horizon). Combining the second part of Theorem 1 with Lemma 6 and the choice $\lambda = \frac{64}{3}\lambda_{t-1}$, we have that for any non-increasing sequence of learning rates λ_t satisfying $\lambda_0 \leq \frac{1}{2}$, the following holds

$$R_T \leq \mathbb{E}\left[\frac{C_T}{\lambda_{T-1}} + \min\left(\sum_{t=1}^{T} \frac{32}{3}\lambda_{t-1}\overline{\mathrm{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}\right)\right], \tag{24}$$

where $C_T = 2 + s\log\frac{4e^3d^2T^3C_{1,T}^2C_{2,T}}{s^2}$ and $C_{1,T}$, respectively $C_{2,T}$ are deterministic upper bounds on $\frac{1}{\lambda_t} - \frac{1}{\lambda_{t-1}}$, respectively $\frac{1}{\lambda_{T-1}}$.

We let $\lambda_t^{(2)} = \sqrt{\frac{s}{2d+\sum_{s=1}^{t}\overline{\mathrm{IR}}_s^{(2)}(\pi_s)}}$ and $\lambda_t^{(3)} = \left(\frac{s}{\frac{3\sqrt{6}s}{\sqrt{C_{\min}}}+\sum_{s=1}^{t}\sqrt{\overline{\mathrm{IR}}_s^{(3)}(\pi_s)}}\right)^{\frac{2}{3}}$, and verify that $\lambda_t = \max(\lambda_t^{(2)}, \lambda_t^{(3)})$ is decreasing and always smaller than $\frac{1}{2}$. We also verify that $C_{1,T} = C_{2,T} = \sqrt{\frac{dT}{s}}$ are valid upper bounds. As a result, we have the following upper bound

$$C_T = 2 + s\log\frac{4e^3d^2T^3C_{1,T}^2C_{2,T}}{s^2} \leq 2 + s\log 4e^3T^{4.5}\left(\frac{d}{s}\right)^{3.5} \leq 2 + 5s\log(\frac{edT}{s}). \tag{25}$$

We know focus on bounding the sum containing the information ratios. Applying Lemma 7, we obtain that for all $t \geq 1$, $\overline{\mathrm{IR}}_t^{(2)}(\pi_t) \leq 2d$ and for any $T \geq 1$

$$\begin{aligned}
\sum_{t=1}^{T}\lambda_{t-1}^{(2)}\overline{\mathrm{IR}}_t^{(2)}(\pi) &= \sqrt{s}\sum_{t=1}^{T}\frac{\overline{\mathrm{IR}}_t^{(2)}(\pi_t)}{\sqrt{2d+\sum_{s=1}^{t-1}}} \\
&\leq \sqrt{s}\sum_{t=1}^{T}\frac{\overline{\mathrm{IR}}_t^{(2)}(\pi_t)}{\sqrt{\sum_{s=1}^{t}\overline{\mathrm{IR}}_s^{(2)}(\pi_s)}} \\
&\leq 2\sqrt{s\sum_{t=1}^{T}\overline{\mathrm{IR}}_t^{(2)}(\pi_t)} \\
&\leq 2\sqrt{s\left(2d+\sum_{t=1}^{T-1}\overline{\mathrm{IR}}_t^{(2)}(\pi_t)\right)},
\end{aligned}$$

where we applied Lemma 19 with the function $f(x) = \frac{1}{\sqrt{x}}$ and $a_i = \overline{\mathrm{IR}}_i^{(2)}(\pi_i)$ to get the second inequality. This can be seen as a generalization of the usual $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ inequality. We now define $R_T^{(2)} = \sqrt{s\left(2d+\sum_{t=1}^{T-1}\overline{\mathrm{IR}}_t^{(2)}(\pi_t)\right)}$, the constant-free regret rate associated to the 2-surrogate-information ratio.

We now turn our attention to the 3-information ratio. Applying Lemma 7 we obtain that for all $t \geq 1$, $\overline{\text{IR}}_t^{(3)}(\pi_t) \leq 54\frac{s}{C_{\min}} \leq 54\frac{s^2}{C_{\min}}$ and for any $T \geq 1$

$$\begin{aligned}\sum_{t=1}^{T}\sqrt{\lambda_{t-1}^{(3)}\overline{\text{IR}}_t^{(3)}(\pi_t)} &= s^{\frac{1}{3}}\sum_{t=1}^{T}\frac{\sqrt{\overline{\text{IR}}_t^{(3)}(\pi_t)}}{\left(\frac{3\sqrt{6}s}{\sqrt{C_{min}}}+\sum_{s=1}^{t-1}\sqrt{\overline{\text{IR}}_s^{(3)}(\pi_s)}\right)^{\frac{1}{3}}}\\ &\leq s^{\frac{1}{3}}\sum_{t=1}^{T}\frac{\sqrt{\overline{\text{IR}}_t^{(3)}(\pi_t)}}{\left(\sum_{s=1}^{t}\sqrt{\overline{\text{IR}}_s^{(3)}(\pi_s)}\right)^{\frac{1}{3}}}\\ &\leq \frac{3}{2}s^{\frac{1}{3}}\left(\sum_{t=1}^{T}\sqrt{\overline{\text{IR}}_t^{(3)}(\pi_t)}\right)^{\frac{2}{3}}\\ &\leq \frac{3}{2}s^{\frac{1}{3}}\left(\frac{3\sqrt{6}s}{\sqrt{C_{min}}}+\sum_{t=1}^{T-1}\sqrt{\overline{\text{IR}}_t^{(3)}(\pi_t)}\right),\end{aligned}$$

where we applied Lemma 19 with the function $f(x) = \frac{1}{x^{\frac{1}{3}}}$ and $a_i = \sqrt{\overline{\text{IR}}_i^{(3)}(\pi_i)}$ to get the second inequality. This can be seen as a generalization of the usual $\sum_{t=1}^{T}\frac{1}{t^{\frac{1}{3}}} \leq \frac{3}{2}T^{\frac{2}{3}}$. We now define $R_T^{(3)} = s^{\frac{1}{3}}\left(\frac{3\sqrt{6}s}{\sqrt{C_{min}}}+\sum_{t=1}^{T-1}\sqrt{\overline{\text{IR}}_t^{(3)}(\pi_t)}\right)^{\frac{2}{3}}$, the constant-free regret rate associated to the 3-surrogate-information ratio. We now consider the last time that the learning rates $\lambda_t^{(3)}$ and $\lambda_t^{(2)}$ have been used. More specifically, we denote $T_2 = \max\{t \leq T, \lambda_{t-1}^{(2)} \geq \lambda_{t-1}^{(3)}\}$, and $T_3 = \max\{t \leq T, \lambda_{t-1}^{(3)} \geq \lambda_{t-1}^{(2)}\}$. Coming back to the bound of Equation 24 and using the definition $\lambda_t = \max(\lambda_t^{(2)}, \lambda_t^{(3)})$), the following bound holds

$$\begin{aligned}&R_T\\ &\leq \mathbb{E}\left[\frac{C_T}{\lambda_{T-1}}+\sum_{t=1}^{T}\min\left(\frac{32}{3}\lambda_{t-1}\overline{\text{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}\overline{\text{IR}}_t^{(3)}(\pi_t)}\right)\right]\\ &\leq \mathbb{E}\left[C_T\min\left(\frac{1}{\lambda_{T-1}^{(2)}},\frac{1}{\lambda_{T-1}^{(3)}}\right)+\sum_{t=1}^{T}\min\left(\frac{32}{3}\max(\lambda_{t-1}^{(2)},\lambda_{t-1}^{(3)})\overline{\text{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\max(\lambda_{t-1}^{(2)},\lambda_{t-1}^{(3)})\overline{\text{IR}}_t^{(3)}(\pi_t)}\right)\right].\end{aligned}$$

We can now separate the sum obtained at the last line based on which learning rate was used at time t.

$$\begin{aligned}&\sum_{t=1}^{T}\min\left(\frac{32}{3}\max(\lambda_{t-1}^{(2)},\lambda_{t-1}^{(3)})\overline{\text{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\max(\lambda_{t-1}^{(2)},\lambda_{t-1}^{(3)})\overline{\text{IR}}_t^{(3)}(\pi_t)}\right)\\ &\leq \sum_{\lambda_{t-1}^{(2)}\geq\lambda_{t-1}^{(3)}}\frac{32}{3}\lambda_{t-1}^{(2)}\overline{\text{IR}}_t^{(2)}(\pi_t)+\sum_{\lambda_{t-1}^{(3)}\geq\lambda_{t-1}^{(2)}}\frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}^{(3)}\overline{\text{IR}}_t^{(3)}(\pi_t)}\\ &\leq \sum_{t=1}^{T_2}\frac{32}{3}\lambda_{t-1}^{(2)}\overline{\text{IR}}_t^{(2)}(\pi_t)+\sum_{t=1}^{T_3}\frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}^{(3)}\overline{\text{IR}}_t^{(3)}(\pi_t)}.\end{aligned}$$

We further bound $\sum_{t=1}^{T_2}\frac{32}{3}\lambda_{t-1}^{(2)}\overline{\text{IR}}_t^{(2)}(\pi_t) \leq \frac{64}{3}R_{T_2}^{(2)}$ and $\sum_{t=1}^{T_3}\frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}^{(3)}\overline{\text{IR}}_t^{(3)}(\pi_t)} \leq \frac{16}{3}R_{T_3}^{(3)}$ (Using the explicit value $c_3^* = \frac{2}{3^{\frac{3}{2}}}$).

The crucial observation is that which of $\lambda_T^{(3)}$ or $\lambda_T^{(2)}$ is bigger will determine whether $R_T^{(2)}$ or $R_T^{(3)}$ is the term of leading order (up to some constants). More specifically, Let T be such that

$\lambda^{(2)}_{T-1} \geq \lambda^{(3)}_{T-1}$ which means that $\sqrt{\frac{s}{2d+\sum_{t=1}^{T-1}\overline{\mathrm{IR}}^{(2)}_t(\pi_t)}} \geq \left(\frac{s}{\frac{3\sqrt{6}s}{\sqrt{C_{\min}}}+\sum_{t=1}^{T-1}\sqrt{\overline{\mathrm{IR}}^{(3)}_t(\pi_t)}}\right)^{\frac{2}{3}}$. Rearranging, this implies that $\sqrt{s}\left(2d+\sum_{s=1}^{T-1}\overline{\mathrm{IR}}^{(2)}_t(\pi_t)\right) \leq s^{\frac{2}{3}}\left(\frac{3\sqrt{6}s}{\sqrt{C_{\min}}}+\sum_{t=1}^{T-1}\sqrt{\overline{\mathrm{IR}}^{(3)}_t(\pi_t)}\right)^{\frac{2}{3}}$, which means that $R^{(2)}_T \leq R^{(3)}_T$. Following the exact same steps, we also have that $\lambda^{(3)}_{T-1} \geq \lambda^{(2)}_{T-1}$ implies that $R^{(3)}_T \leq R^{(2)}_T$. We apply this to the time T_2 in which $\lambda^{(2)}_{T_2-1} \geq \lambda^{(3)}_{T_2-1}$ by definition. we have that $R^{(2)}_{T_2} \leq R^{(3)}_{T_2}$ and putting this together with the previous bound, we have

$$\begin{aligned}
R_T &\leq \mathbb{E}\left[\frac{C_T}{\lambda^{(3)}_{T-1}} + \frac{64}{3}R^{(2)}_{T_2} + \frac{16}{3}R^{(3)}_{T_3}\right] \\
&\leq \mathbb{E}\left[\frac{C_T}{s}R^{(3)}_T + \frac{64}{3}R^{(2)}_{T_2} + \frac{16}{3}R^{(3)}_{T_3}\right] \\
&\leq \mathbb{E}\left[\frac{C_T}{s}R^{(3)}_T + \frac{64}{3}R^{(3)}_{T_2} + \frac{16}{3}R^{(3)}_{T_3}\right] \\
&\leq \mathbb{E}\left[\frac{C_T}{s}R^{(3)}_T + \frac{64}{3}R^{(3)}_{T} + \frac{16}{3}R^{(3)}_{T}\right] \\
&\leq \mathbb{E}\left[\left(\frac{C_T}{s} + \frac{80}{3}\right)R^{(3)}_T\right],
\end{aligned}$$

where we use the fact that $T \to R^{(2)}_T$ and $T \to R^{(3)}_T$ are non-decreasing and $T_2 \leq T, T_3 \leq T$ Similarly by definition of T_3, we have that $\lambda^{(3)}_{T_3-1} \geq \lambda^{(2)}_{T_3-1}$ and we can conclude that $R^{(3)}_{T_3} \leq R^{(2)}_{T_3}$. Putting this together, with the previous bound, we have

$$\begin{aligned}
R_T &\leq \mathbb{E}\left[\frac{C_T}{\lambda^{(3)}_{T-1}} + \frac{64}{3}R^{(2)}_{T_2} + \frac{16}{3}R^{(3)}_{T_3}\right] \\
&\leq \mathbb{E}\left[\frac{C_T}{s}R^{(2)}_T + \frac{64}{3}R^{(2)}_{T_2} + \frac{16}{3}R^{(3)}_{T_3}\right] \\
&\leq \mathbb{E}\left[\frac{C_T}{s}R^{(2)}_T + \frac{64}{3}R^{(2)}_{T_2} + \frac{16}{3}R^{(2)}_{T_3}\right] \\
&\leq \mathbb{E}\left[\frac{C_T}{s}R^{(2)}_T + \frac{64}{3}R^{(2)}_{T} + \frac{16}{3}R^{(2)}_{T}\right] \\
&\leq \mathbb{E}\left[(\frac{C_T}{s} + \frac{80}{3})R^{(2)}_T\right],
\end{aligned}$$

where we use the fact that $T \to R^{(2)}_T$ and $T \to R^{(3)}_T$ are non-decreasing and $T_2 \leq T, T_3 \leq T$. Putting both of those bounds together with Equation 25 yields the claim of the Theorem.

E Maximum likelihood estimation

The focus of this section is to bound the difference between the log-likelihoods associated with the true parameter and the maximum likelihood estimator (MLE). We start by establishing an upper bound that holds in expectation which suffices to handle history-independent learning rates. Then, we move on to high-probability bounds that will allow us to deal with data-dependent learning rates.

E.1 Bound in expectation

We start with the case in which the maximum likelihood estimator is computed on a finite subset of the parameter space Θ.

Lemma 15. *Let $t \geq 1$, and Θ' be a finite subset of Θ, we define the MLE over Θ' as*

$$\theta_{MLE,t}(\Theta') = \underset{\theta\in\Theta'}{arg\,min}\, L^{(1)}_t(\theta).$$

Then,

$$\mathbb{E}\left[L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_{MLE,t}(\Theta'))\right] \leq \log|\Theta'| \tag{26}$$

Proof. By the concavity of the logarithm and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left[L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_{\text{MLE},t}(\Theta'))\right] &\leq \log \mathbb{E}\left[\prod_{s=1}^{t} \frac{p(Y_s|\theta_{\text{MLE},t}(\Theta'), A_s)}{p(Y_s|\theta_0, A_s)}\right] \\ &= \log \mathbb{E}\left[\max_{\theta\in\Theta'} \prod_{s=1}^{t} \frac{p(Y_s|\theta, A_s)}{p(Y_s|\theta_0, A_s)}\right] \leq \log \mathbb{E}\left[\sum_{\theta\in\Theta'} \prod_{s=1}^{t} \frac{p(Y_s|\theta, A_s)}{p(Y_s|\theta_0, A_s)}\right] \\ &= \log \sum_{\theta\in\Theta'} \mathbb{E}\left[\prod_{s=1}^{t} \frac{p(Y_s|\theta, A_s)}{p(Y_s|\theta_0, A_s)}\right]\end{aligned}$$

By Lemma 25, we have that $\exp\left(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta)\right) = \prod_{s=1}^{t} \frac{p(Y_s|\theta,A_s)}{p(Y_s|\theta_0,A_s)}$ is a non-negative supermartingale with respect to the filtration $\mathcal{F}_t' = \sigma(\mathcal{F}_{t-1}, A_t)$. That implies that each term in the sum is upper bounded by 1. Hence,

$$\mathbb{E}\left[L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_{\text{MLE},t}(\Theta'))\right] \leq \log \sum_{\theta\in\Theta'} 1 = \log|\Theta'|,$$

which proves the claim. □

To extend the previous bound to the full parameter space, we use a covering argument. A subset $\Theta' \subset \Theta$ is said to be a valid ρ-covering of Θ with respect to the ℓ_1 norm if for every $\theta \in \Theta$, there exists a $\theta' \in \Theta'$ such that $\|\theta - \theta'\|_1 \leq \rho$. We denote by $\mathcal{N}(\Theta, \|\cdot\|_1, \rho)$ the smallest possible cardinality of a valid ρ covering. We have the following bound on this quantity.

Lemma 16. *For every $\rho > 0$,*

$$\log \mathcal{N}(\Theta, \|\cdot\|_1, \rho) \leq \log \binom{d}{s}(1 + \tfrac{2}{\rho})^s \leq s \log \frac{ed(1 + 2/\rho)}{s}.$$

Proof. For each subset $S \subset [d]$ of cardinality $|S| = s$, there is a surjective isometric embedding from $(\Theta_S, \|\cdot\|_1)$ to $(\mathbb{B}_1^s(1), \|\cdot\|_1)$. In particular, to embed $\theta \in \Theta_S$ into $\mathbb{B}_1^s(1)$, one can simply remove all the components of θ corresponding to indices not in S. Therefore, for every $\rho > 0$, $\mathcal{N}(\Theta_S, \|\cdot\|_1, \rho) \leq \mathcal{N}(\mathbb{B}_1^s(1), \|\cdot\|_1, \rho)$. Moreover, via a standard argument, we have $\mathcal{N}(\mathbb{B}_1^s(1), \|\cdot\|_1, \rho) \leq (1 + \frac{2}{\rho})^s$ (see, e.g., Lemma 5.7 in Wainwright, 2019). Now, let $\Theta_{S,\rho}$ denote any minimal ρ-covering of Θ_S and notice that for an arbitrary $\theta \in \Theta$ with support S, there exists a subset $\tilde{S}$ such that $S \subseteq \tilde{S}$ and $|\tilde{S}| = s$. Therefore, there exists $\tilde{\theta} \in \Theta_{\tilde{S},\rho}$ such that $\|\theta - \tilde{\theta}\|_1 \leq \rho$. Hence, $\cup_{S\subset[d]:|S|=s}\Theta_{S,\rho}$ forms a valid ρ-covering of Θ and its cardinality is bounded by

$$\mathcal{N}(\Theta, \|\cdot\|_1, \rho) \leq \left|\cup_{S\subset[d]:|S|=s}\Theta_{S,\rho}\right| \leq \sum_{S\subset[d]:|S|=s} \left(1 + \tfrac{2}{\rho}\right)^s = \binom{d}{s}\left(1 + \tfrac{2}{\rho}\right)^s.$$

and we conclude by the elementary inequality $\binom{d}{s} \leq \left(\frac{de}{s}\right)^s$. □

E.1.1 Proof of Lemma 10

We bound the difference between the log-likelihood of the true parameter and that of the maximum likelihood estimator on the full parameter space. To this end, let $\rho > 0$ and Θ' be a minimal valid ρ-cover of Θ as is defined in Lemma 16, and $\theta' \in \Theta'$ be such that $\|\theta' - \theta_t\| \leq \rho$, which exists by definition of a ρ-covering. Then,

$$\begin{aligned}\mathbb{E}\left[L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t)\right] =& \mathbb{E}\left[L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_{\text{MLE},t}(\Theta'))\right] \\ &+ \mathbb{E}\left[L_t^{(1)}(\theta_{\text{MLE},t}(\Theta')) - L^{(1)}(\theta')\right] \\ &+ \mathbb{E}\left[L_t^{(1)}(\theta') - L^{(1)}(\theta_t)\right] \\ \le& \log(\mathcal{N}(\Theta, \|\cdot\|_1, \rho) + 0 + 2\rho t,\end{aligned}$$

where the first term is bounded by Lemma 26, the second term is non-positive by definition of the maximum likelihood estimator because $\theta' \in \Theta'$ and the third term is bounded because the mapping $\theta \mapsto \mathbb{E}\left[L_t^{(1)}(\theta)\right]$ is $2t$-Lipschitz with respect to the 1-norm by Lemma 21. Finally applying Lemma 16 and setting $\rho = \frac{2}{t}$ yields the desired bound. □

E.2 High-probability bounds

We begin with the case where the maximum likelihood estimator is computed over a finite subset of the parameter space Θ and provide a corresponding high-probability bound.

Lemma 17. *Let Θ' be a finite subset of Θ, we define $\theta_{MLE,t}(\Theta') = \arg\min_{\theta \in \Theta'} L_t^{(1)}(\theta)$. Then*

$$\mathbb{P}\left[\exists t \ge 1, L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_{MLE,t}(\Theta')) \ge \log \frac{|\Theta'|}{\delta}\right] \le \delta. \tag{27}$$

Proof. Fix $\theta \in \Theta'$. By Lemma 25, we have that $\exp\left(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta)\right) = \prod_{s=1}^t \frac{p(Y_s|\theta, A_s)}{p(Y_s|\theta_0, A_s)}$ is a non-negative supermartingale with respect to the filtration $\mathcal{F}'_t = \sigma(\mathcal{F}_{t-1}, A_t)$, allowing us to invoke Ville's inequality to get the following guarantee:

$$\mathbb{P}\left[\exists t \ge 1, \exp(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta)) \ge \frac{1}{\delta}\right] \le \delta.$$

Taking the logarithm and a union bound on Θ' yields the desired result. □

We now provide a bound on the expected product of a bounded random variable with the differenece in log-likelihood between the true parameter and the maximum likelihood estimator.

Lemma 18. *Let $B \in \mathbb{R}$ and X be a random variable satisfying $0 \le X \le B$ almost surely. Then for any $t \ge 1$,*

$$\begin{aligned}\mathbb{E}\left[X(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t))\right] &\le \inf_{\delta,\rho>0}\left\{\mathbb{E}\left[Xs\log\frac{ed(1+\frac{2}{\rho})}{s\delta^{\frac{1}{s}}}\right] + 2B\rho t + B\delta s \log\frac{e^{1+\frac{1}{s}}d(1+\frac{2}{\rho})}{s\delta^{\frac{1}{s}}}\right\} \\ &\le 4 + s\log\frac{2e^2dT^2B^2}{s}\mathbb{E}\left[X + \frac{1}{T}\right].\end{aligned} \tag{28}$$

Proof. Let $\delta, \rho > 0$ and Θ' be a minimal valid ρ-cover of Θ as defined in Lemma 16, $N = |\Theta'|$, let $\theta' = \theta_{\text{MLE},t}(\Theta')$ and let $\bar{\theta} \in \Theta'$ be such that $\|\bar{\theta} - \theta_t\| \le \rho$, which exists by definition of a valid ρ-cover. We have the following decomposition:

$$\begin{aligned}\mathbb{E}\left[X(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t))\right] \le& \mathbb{E}\left[X(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta'))\mathbf{1}_{\{L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta') \le \log\frac{N}{\delta}\}}\right] \\ &+ B\mathbb{E}\left[(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta'))\mathbf{1}_{\{L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta') > \log\frac{N}{\delta}\}}\right] \\ &+ B\mathbb{E}\left[(L_t^{(1)}(\bar{\theta}) - L_t^{(1)}(\theta_t))\right] + B\mathbb{E}\left[(L_t^{(1)}(\theta') - L_t^{(1)}(\bar{\theta}))\right].\end{aligned}$$

The first term is upper bounded by $\mathbb{E}\left[X \log\frac{N}{\delta}\right]$, the third term is upper bounded by $2B\rho t$ because $\mathbb{E}\left[L_t^{(1)}(\cdot)\right]$ is $2t$-Lipschitz by Lemma 21. The fourth term is non-positive because θ' minimizes the negative log likelihood on Θ'. Finally, we turn our attention to the second term. To simplify the computations, we let $Y = L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta')$, and compute $\mathbb{E}\left[Y\mathbf{1}_{\{Y > \log\frac{N}{\delta}\}}\right]$. Conditionting on wheter ϵ is larger or smaller than $\log\frac{N}{\delta}$ yields the following identity

$$\mathbb{P}\left[Y\mathbf{1}_{\{Y\geq\log\frac{N}{\delta}\}} \geq \epsilon\right] = \begin{cases}\mathbb{P}\left[Y \geq \epsilon\right] & \text{if } \epsilon \geq \log\frac{N}{\delta}, \\ \mathbb{P}\left[Y \geq \log\frac{N}{\delta}\right] & \text{otherwise.}\end{cases}$$

We can now upper bound the expectation as follows

$$\begin{aligned}\mathbb{E}\left[Y\mathbf{1}_{\{Y\geq\log\frac{N}{\delta}\}}\right] &= \int_0^\infty \mathbb{P}\left[Y\mathbf{1}_{\{Y\geq\log\frac{N}{\delta}\}} \geq \epsilon\right] d\epsilon \\ &= \log\frac{N}{\delta}\mathbb{P}\left[Y \geq \log\frac{N}{\delta}\right] + \int_{\log\frac{N}{\delta}}^\infty \mathbb{P}\left[Y \geq \epsilon\right] d\epsilon \\ &= \log\frac{N}{\delta}\mathbb{P}\left[Y \geq \log\frac{N}{\delta}\right] + \int_0^\delta \frac{1}{\delta'}\mathbb{P}\left[Y \geq \log\frac{N}{\delta'}\right] d\delta' \\ &\leq \delta\log\frac{N}{\delta} + \delta,\end{aligned}$$

where we used the change of variable $\epsilon = \log\frac{N}{\delta'}$ and used $\mathbb{P}\left[Y \geq \log\frac{N}{\delta}\right] \leq \delta$ by Lemma 17. Finally, putting everything together and using $N \leq \mathcal{N}(\Theta, \|\cdot\|_1, \rho) \leq \left(\frac{ed(1+\frac{2}{\rho})}{s}\right)^s$, by Lemma 16, we get

$$\mathbb{E}\left[X(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t))\right] \leq \mathbb{E}\left[Xs\log\frac{ed(1+\frac{2}{\rho})}{s\delta^{\frac{1}{s}}}\right] + 2B\rho t + B\delta s\log\frac{e^{1+\frac{1}{s}}d(1+\frac{2}{\rho})}{s\delta^{\frac{1}{s}}}.$$

To balance the trade-off between the approximation error and the covering complexity, we choose $\rho = \frac{2}{BT}$, and $\delta = \frac{1}{BT}$ which yields the desired form of the logarithmic factors. Subsituting these into the bound completes the proof. □

E.2.1 Proof of Lemma 14

As was noted in the analysis, since λ_T is not used by the algorithm, we can replace λ_T by λ_{T-1} in our computations. We have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T \frac{\eta}{\lambda_{t-1}}(L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0) + L_{t-1}^{(1)}(\theta_0) - L_{t-1}^{(1)}(\theta_{t-1})) + \frac{\eta}{\lambda_T}(L_T^{(1)}(\theta_0) - L_T^{(1)}(\theta_T))\right] \\ &\quad= \mathbb{E}\left[\sum_{t=1}^T \frac{\eta}{\lambda_{t-1}}(L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0)) - \sum_{t=1}^T \frac{\eta}{\lambda_t}(L_t^{(1)}(\theta_t) - L_t^{(1)}(\theta_0))\right] \\ &\quad= \eta\cdot\sum_{t=1}^T \mathbb{E}\left[(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t))\left(\frac{1}{\lambda_t} - \frac{1}{\lambda_{t-1}}\right)\right].\end{aligned}$$

Let $C_{1,T}$ be a deterministic upper bound on $\left(\frac{1}{\lambda_{t+1}} - \frac{1}{\lambda_t}\right)$. Applying Lemma 28 to $X = \left(\frac{1}{\lambda_{t+1}} - \frac{1}{\lambda_t}\right)$ and telescoping, we get

$$\begin{aligned}&\eta\cdot\sum_{t=1}^T \mathbb{E}\left[(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta_t))\left(\frac{1}{\lambda_t} - \frac{1}{\lambda_{t-1}}\right)\right] \\ .&\leq \eta\left(4 + s\log\frac{2e^2dt^2C_{1,T}^2}{s}\right)\sum_{t=1}^T \mathbb{E}\left[\left(\frac{1}{\lambda_t} - \frac{1}{\lambda_{t-1}}\right) + \frac{1}{T}\right] \\ &\leq \eta\left(4 + s\log\frac{2e^2dt^2C_{1,T}^2}{s}\right)\mathbb{E}\left[\left(\frac{1}{\lambda_T} + 1\right)\right] \\ &\leq \mathbb{E}\left[\frac{\eta(12 + 3s\log\frac{2e^2dt^2C_{1,T}^2}{s})}{2\lambda_{T-1}}\right],\end{aligned}$$

where in the last step, we used $1 \leq \frac{1}{2\lambda_T}$ which implies $\frac{1}{\lambda_T} + 1 \leq \frac{3}{2\lambda_T}$. This finishes the proof. $\square$

F Bounding the surrogate information ratio

F.1 Proof of Lemma 6

The surrogate regret of a policy is directly related to its 2- and 3-information ratio by definition

$$\widehat{\Delta}_t(\pi) = \sqrt{\overline{\text{IG}}_t(\pi)\overline{\text{IR}}_t^{(2)}(\pi)} = \left(\overline{\text{IG}}_t(\pi)\overline{\text{IR}}_t^{(3)}(\pi)\right)^{\frac{1}{3}}.$$

By the AM-GM inequality, we have that for any $\lambda > 0$, the surrogate regret is controlled as follows

$$\widehat{\Delta}_t(\pi) \leq \frac{\overline{\text{IG}}_t(\pi)}{\lambda} + \frac{\lambda}{4}\overline{\text{IR}}_t^{(2)}(\pi).$$

Similarly, by Lemma 27 which generalizes the AM-GM inequality, we can obtain the following regret bound

$$\widehat{\Delta}_t(\pi) \leq \frac{\overline{\text{IG}}_t(\pi)}{\lambda} + c_3^* \sqrt{\lambda \overline{\text{IR}}_t^{(3)}(\pi)},$$

where $c_3^* < 2$ is an absolute constant defined in Lemma 27. This concludes the proof. $\square$

F.2 Proof of Lemma 1

The proof of Lemma 1 is essentially the same as the proof of Lemma 5.6 in Hao et al. [2021], but we state it here for completeness. Throughout this proof, we use $\langle p, f\rangle = \sum_{a\in\mathcal{A}} p(a)f(a)$ to denote the inner product between a signed measure p on $\mathcal{A}$ and a function $f : \mathcal{A} \to \mathbb{R}$. Using this notation, we can, for example, write the generalized surrogate information ratio as $\overline{\text{IR}}_t^{(\gamma)}(\pi) = \langle \pi, \overline{\text{IR}}_t^{(\gamma)}\rangle$. We define $\pi_t^{(\gamma)} \in \arg\min_{\pi\in\Delta(\mathcal{A})} \overline{\text{IR}}_t^{(\gamma)}(\pi)$ to be any minimizer of the generalized surrogate information ratio with parameter $\gamma \geq 2$. First, we observe that

$$\nabla_\pi \overline{\text{IR}}_t^{(2)}(\pi) = \frac{2\langle \pi, \widehat{\Delta}_t\rangle \widehat{\Delta}_t}{\langle \pi, \overline{\text{IG}}_t\rangle} - \frac{(\langle \pi, \widehat{\Delta}_t\rangle)^2 \overline{\text{IG}}_t}{(\langle \pi, \overline{\text{IG}}_t\rangle)^2}.$$

Therefore, from the first-order optimality condition for convex constrained minimization (and the fact that $\overline{\text{IR}}_t^{(2)}$ is convex on $\Delta(\mathcal{A})$), we have

$$\forall \pi \in \Delta(\mathcal{A}),\ 0 \leq \langle \pi - \pi_t^{(\mathbf{SOIDS})}, \nabla_\pi \overline{\text{IR}}_t^{(2)}(\pi_t^{(\mathbf{SOIDS})})\rangle.$$

In particular,

$$0 \leq \frac{2\langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle \langle \pi_t^{(\gamma)} - \pi^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle}{\langle \pi_t^{(\mathbf{SOIDS})}, \overline{\text{IG}}_t\rangle} - \frac{(\langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle)^2 \langle \pi_t^{(\gamma)} - \pi^{(\mathbf{SOIDS})}, \overline{\text{IG}}_t\rangle}{(\langle \pi_t^{(\mathbf{SOIDS})}, \overline{\text{IG}}_t\rangle)^2}.$$

This inequality is equivalent to

$$2\langle \pi_t^{(\gamma)}, \widehat{\Delta}_t\rangle \geq \langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle \left(1 + \frac{\langle \pi_t^{(\gamma)}, \overline{\text{IG}}_t\rangle}{\langle \pi_t^{(\mathbf{SOIDS})}, \overline{\text{IG}}_t\rangle}\right) \geq \langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle.$$

From this inequality, we obtain

$$\begin{aligned}
\frac{(\langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle)^\gamma}{\langle \pi_t^{(\mathbf{SOIDS})}, \overline{\text{IG}}_t\rangle} &= \frac{(\langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle)^2 (\langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle)^{\gamma-2}}{\langle \pi_t^{(\mathbf{SOIDS})}, \overline{\text{IG}}_t\rangle} \\
&\leq \frac{(\langle \pi_t^{(\gamma)}, \widehat{\Delta}_t\rangle)^2 (\langle \pi_t^{(\mathbf{SOIDS})}, \widehat{\Delta}_t\rangle)^{\gamma-2}}{\langle \pi_t^{(\gamma)}, \overline{\text{IG}}_t\rangle} \\
&\leq 2^{\gamma-2} \frac{(\langle \pi_t^{(\gamma)}, \widehat{\Delta}_t\rangle)^\gamma}{\langle \pi_t^{(\gamma)}, \overline{\text{IG}}_t\rangle} = 2^{\gamma-2} \min_{\pi\in\Delta(\mathcal{A})} \overline{\text{IR}}_t^{(\gamma)}(\pi),
\end{aligned}$$

thus proving the claim. $\square$

F.3 Proof of Lemma 7

This section is focused on bounding the information ratios of the sparse optimistic information directed sampling policy. As is widely done in the information directed sampling literature, we will introduce a "forerunner" algorithm with controlled surrogate information ratio. By Lemma 1, the sOIDS policy will then automatically inherit the bound of the forerunner.

As one of our forerunners, we will make use of the "Feel-Good Thompson Sampling" first introduced by Zhang [2022]. Letting $\widetilde{\theta}_t \sim Q_t^+$, the FGTS policy is defined as

$$\pi_t^{(\mathbf{FGTS})}(a) = \mathbb{P}_t\left[a^*(\widetilde{\theta}_t) = a\right]. \tag{29}$$

Which can be seen as the policy obtained by sampling a parameter $\widetilde{\theta}_t \sim Q_t^+$ and then picking the optimal action under this parameter. Compared to the usual Thompson Sampling policy, this boils down to replacing the Bayesian posterior by the optimistic posterior. Whenever the optimal action for θ is non-unique, we define $a^*(\theta)$ to be any optimal action with minimal 0-norm. If there are multiple optimal actions with minimal 0-norm, ties can be broken arbitrarily.

For the bound on the surrogate 3-information ratio, we assume that the prior Q_1^+ and the action set $\mathcal{A}$ are such that for all θ in the support of the prior, there exists $a' \in \arg\max_{a\in\mathcal{A}} r(a,\theta)$ such that $\|a'\|_0 \le s$. We refer to this as the sparse optimal action property. Since the support of our prior Q_1^+ only contains s-sparse vectors, the sparse optimal action property is satisfied whenever the action set is a a unit ℓ_p ball. Note also that the hard instances in both the $\sqrt{sdT}$ lower bound in Theorem 24.3 of Lattimore and Szepesvári [2020] and the $s^{2/3}T^{2/3}$ lower bound in Theorem 5 of Jang et al. [2022] satisfy the sparse optimal action property[2]. Therefore, even with this addtional assumption, the lower bounds for both the data-rich and data-poor regimes remain meaninful. Whenever the optimal action for θ is non-unique, we define $a^*(\theta)$ to be any optimal action with minimal 0-norm, with ties broken arbitrarily.

F.3.1 Bounding the two information ratio

We will now prove the first part of lemma 7, by showing that the information ratio of the FGTS policy is bounded by the dimension. The proof is exactly the same as in the Bayesian setting as is done in Proposition 5 of Russo and Roy [2016], Lemma 7 of Lemma 7 in Neu et al. [2022] or in Lemma 5.7 of Hao et al. [2021], except the Bayesian posterior is replaced with the optimistic posterior. We provide the proof here for completeness.

Since we defined the surrogate information gain in terms of the model θ, as opposed to the optimal action $a^*(\theta)$, we follow the proof of Lemma 7 in Neu et al. [2022]. For brevity, we let $\alpha_a = \pi_t^{(\mathbf{FGTS})}(a) = \mathbb{P}_t\left[a^*(\widetilde{\theta}_t) = a\right]$. We define the $|\mathcal{A}| \times |\mathcal{A}|$ matrix M by

$$M_{a,a'} = \sqrt{\alpha_a \alpha_{a'}}(\mathbb{E}_t[r(a,\widetilde{\theta}_t)|a^*(\widetilde{\theta}_t) = a'] - r(a,\bar{\theta}(Q_t^+)))\,.$$

Next, we relate the surrogate information gain and the surrogate regret to the Frobenius norm and the trace of M. First, we can lower bound the surrogate information gain of FGTS as

$$\begin{aligned}
\overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{FGTS})}) &= \frac{1}{2}\sum_{a\in\mathcal{A}} \alpha_a \int_\Theta (r(a,\bar{\theta}(Q_t^+)) - r(a,\theta))^2 \mathrm{d}Q_t^+(\theta) \\
&= \frac{1}{2}\sum_{a\in\mathcal{A}} \alpha_a \int_\Theta \sum_{a'\in\mathcal{A}} \mathbf{1}_{\{a^*(\theta)=a'\}} (r(a,\bar{\theta}(Q_t^+)) - r(a,\theta))^2 \mathrm{d}Q_t^+(\theta) \\
&= \frac{1}{2}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} \alpha_a \int_\Theta \mathbf{1}_{\{a^*(\theta)=a'\}} \mathrm{d}Q_t^+(\theta) \mathbb{E}_t[(r(a,\bar{\theta}(Q_t^+)) - r(a,\widetilde{\theta}_t)|a^*(\widetilde{\theta}_t) = a'] \\
&\ge \frac{1}{2}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} \alpha_a\alpha_{a'} \left(r(a,\bar{\theta}(Q_t^+)) - \mathbb{E}_t[r(a,\widetilde{\theta}_t)|a^*(\widetilde{\theta}_t) = a']\right)^2 \\
&= \frac{1}{2}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} M_{a,a'}^2 = \frac{1}{2}\|M\|_F^2\,.
\end{aligned}$$

[2]The optimal actions in the hard instance used to prove Theorem 5 in Jang et al. [2022] are $2s$-sparse, which still allows us to prove the same bound on the surrogate 3-information ratio, up to constant factors.

Next, we can re-write the surrogate regret of FGTS as

$$\begin{aligned}\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) &= \int_\Theta r(a^*(\theta),\theta)\mathrm{d}Q_t^+(\theta) - \sum_{a\in\mathcal{A}}\alpha_a \int_\Theta r(a,\theta)\mathrm{d}Q_t^+ \\ &= \int_\Theta \sum_{a\in\mathcal{A}} \mathbf{1}_{\{a^*(\theta)=a\}} r(a^*(\theta),\theta)\mathrm{d}Q_t^+(\theta) - \sum_{a\in\mathcal{A}}\alpha_a r(a,\bar\theta(Q_t^+)) \\ &= \sum_{a\in\mathcal{A}}\alpha_a \mathbb{E}_t[r(a,\widetilde\theta_t)|a^*(\widetilde\theta_t)=a] - \sum_{a\in\mathcal{A}}\alpha_a r(a,\bar\theta(Q_t^+)) \\ &= \mathrm{tr}(M)\,.\end{aligned} \tag{30}$$

Using Fact 10 from Russo and Roy [2016], we bound $\overline{\mathrm{IR}}_t^{(2)}(\pi_t^{(\mathbf{FGTS})})$ as

$$\overline{\mathrm{IR}}_t^{(2)}(\pi_t^{(\mathbf{FGTS})}) = \frac{(\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}))^2}{\overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{FGTS})})} \le \frac{2(\mathrm{tr}(M))^2}{\|M\|_F^2} \le 2\cdot\mathrm{rank}(M)\,.$$

All the remains is to show that M has rank at most d. Enumerate the actions as $\mathcal{A}=\{a_1,\dots,a_{|\mathcal{A}|}\}$, and let $\mu_i = \mathbb{E}_t[\widetilde\theta_t|a^*(\widetilde\theta_t)=a_i]$. By linearity of expectation (and of the reward function), we can write

$$M_{i,j} = \sqrt{\alpha_i\alpha_j}\langle \mu_i - \bar\theta(Q_t^+), a_j\rangle\,.$$

Therefore, M can be factorised as

$$M = \begin{bmatrix} \sqrt{\alpha_1}(\mu_1 - \bar\theta(Q_t^+))^\top \\ \vdots \\ \sqrt{\alpha_{|\mathcal{A}|}}(\mu_{|\mathcal{A}|} - \bar\theta(Q_t^+))^\top \end{bmatrix} \begin{bmatrix}\sqrt{\alpha_1}a_1 & \cdots & \sqrt{\alpha_{|\mathcal{A}|}}a_{|\mathcal{A}|}\end{bmatrix}\,.$$

Since M is the product of a $K\times d$ matrix and a $d\times K$ matrix, it must have rank at most $\min(K,d)$.

F.3.2 Bounding the three information ratio

To bound the 3 information ratio we follow Hao et al. [2021] and we introduce the exploratory policy

$$\mu = \operatorname*{arg\,max}_{\pi\in\Delta(\mathcal{A})} \sigma_{\min}\left(\sum_{a\in\mathcal{A}}\pi(a)aa^T\right). \tag{31}$$

We define the mixture policy $\pi_t^{(\mathbf{mix})} = (1-\gamma)\pi_t^{(\mathbf{FGTS})} + \gamma\mu$ where $\gamma\ge 0$ will be determined later. First, we lower bound the surrogate information gain of the mixture policy in the same way that we lower bounded the surrogate information gain of the FGTS policy previously. This time, we obtain the lower bound

$$\begin{aligned}\overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{mix})}) &\ge \frac{1}{2}\sum_{a\in\mathcal{A}}\pi_t^{(\mathbf{mix})}(a)\sum_{a'\in\mathcal{A}}\mathbb{P}_t(a^*(\widetilde\theta_t)=a')(r(a,\bar\theta(Q_t^+)) - \mathbb{E}_t[r(a,\widetilde\theta_t)|a^*(\widetilde\theta_t)=a'])^2 \\ &= \frac{1}{2}\sum_{a\in\mathcal{A}}\pi_t^{(\mathbf{mix})}(a)\sum_{a'\in\mathcal{A}}\mathbb{P}_t(a^*(\widetilde\theta_t)=a')\langle\mu_{a'} - \bar\theta(Q_t^+),a\rangle^2\,,\end{aligned}$$

where $\mu_{a'} = \mathbb{E}_t[\widetilde\theta_t|a^*(\widetilde\theta_t)=a']$. From the inequality $\pi_t^{(\mathbf{mix})}(a)\ge\gamma\mu(a)$, and the definition of $C_{\min}$, we have

$$\begin{aligned}\overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{mix})}) &\ge \frac{\gamma}{2}\sum_{a'\in\mathcal{A}}\mathbb{P}_t(a^*(\widetilde\theta_t)=a')\sum_{a\in\mathcal{A}}\mu(a)(\mu_{a'}-\bar\theta(Q_t^+))^\top aa^\top(\mu_{a'}-\bar\theta(Q_t^+)) \\ &\ge \frac{\gamma}{2}\sum_{a'\in\mathcal{A}}\mathbb{P}_t(a^*(\widetilde\theta_t)=a')C_{\min}\|\mu_{a'}-\bar\theta(Q_t^+)\|_2^2\,.\end{aligned}$$

Using the expression for the surrogate regret of FGTS in (30), we obtain

$$\begin{aligned}\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) &= \sum_{a\in\mathcal{A}}\mathbb{P}_t(a^*(\widetilde\theta_t)=a)(\mathbb{E}_t[\langle\widetilde\theta_t),a\rangle|a^*(\widetilde\theta_t)=a] - \langle\bar\theta(Q_t^+),a\rangle) \\ &\le \sqrt{\sum_{a\in\mathcal{A}}\mathbb{P}_t(a^*(\widetilde\theta_t)=a)(\mathbb{E}_t[\langle\widetilde\theta_t,a\rangle|a^*(\widetilde\theta_t)=a] - \langle\bar\theta(Q_t^+),a\rangle)^2}\,,\end{aligned}$$

where in the last the line we used the Cathy-Schwarz inequality. Due to the sparse optimal action property, all actions for which $\mathbb{P}_t(a^*(\widetilde{\theta}_t) = a) > 0$ have at most s non-zero elements. Therefore,

$$\sum_{a\in\mathcal{A}} \mathbb{P}_t(a^*(\widetilde{\theta}_t) = a)(\mathbb{E}_t[\langle\widetilde{\theta}_t, a\rangle | a^*(\widetilde{\theta}_t) = a] - \langle\bar{\theta}(Q_t^+), a\rangle)^2 \leq \sum_{a\in\mathcal{A}} \mathbb{P}_t(a^*(\widetilde{\theta}_t) = a)s\|\mu_a - \bar{\theta}(Q_t^+)\|_2^2 \,.$$

This, combined with the lower bound on $\overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{mix})})$ means that

$$\begin{aligned}
\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) &\leq \sqrt{\sum_{a\in\mathcal{A}} \mathbb{P}_t(a^*(\widetilde{\theta}_t) = a)s\|\mu_a - \bar{\theta}(Q_t^+)\|_2^2} \\
&= \sqrt{\frac{2s}{\gamma C_{\min}} \frac{\gamma}{2} \sum_{a\in\mathcal{A}} \mathbb{P}_t(a^*(\widetilde{\theta}_t) = a) C_{\min}\|\mu_a - \bar{\theta}(Q_t^+)\|_2^2} \\
&\leq \sqrt{\frac{2s}{\gamma C_{\min}} \overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{mix})})} \,.
\end{aligned}$$

Choosing $\gamma = 1$, this tells us that

$$(\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}))^2 \leq \frac{2s}{C_{\min}} \overline{\mathrm{IG}}_t(\mu) \,.$$

We bound the information ratio in three cases. First, suppose that $\widehat{\Delta}_t(\mu) \leq \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})$. In this case,

$$\overline{\mathrm{IR}}_t^{(3)}(\mu) = \frac{\widehat{\Delta}_t(\mu)(\widehat{\Delta}_t(\mu))^2}{\overline{\mathrm{IG}}_t(\mu)} \leq \frac{2(\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}))^2}{\overline{\mathrm{IG}}_t(\mu)} \leq \frac{4s}{C_{\min}} \,.$$

Next, we consider the case where $\widehat{\Delta}_t(\mu) > \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})$. For any $\gamma \in (0, 1]$,

$$\overline{\mathrm{IR}}_t^{(3)}(\pi_t^{(\mathbf{mix})}) = \frac{((1-\gamma)\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) + \gamma\widehat{\Delta}_t(\mu))^3}{(1-\gamma)\overline{\mathrm{IG}}_t(\pi_t^{(\mathbf{FGTS})}) + \gamma\overline{\mathrm{IG}}_t(\mu)} \leq \frac{((1-\gamma)\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) + \gamma\widehat{\Delta}_t(\mu))^3}{\gamma\overline{\mathrm{IG}}_t(\mu)} \,.$$

We define $f(\gamma) = ((1-\gamma)\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) + \gamma\widehat{\Delta}_t(\mu))^3/(\gamma\overline{\mathrm{IG}}_t(\mu))$ to be the RHS of the previous equation. One can verify that the derivative of $f(\gamma)$ is

$$f'(\gamma) = \frac{((1-\gamma)\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) + \gamma\widehat{\Delta}_t(\mu))^2}{\gamma^2\overline{\mathrm{IG}}_t(\mu)} \left[2\gamma(\widehat{\Delta}_t(\mu) - \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})) - \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})\right] \,,$$

and that $f(\gamma)$ is minimised w.r.t. $\gamma > 0$ at $\widehat{\gamma}$, where $\widehat{\gamma}$ is the positive solution of $f'(\widehat{\gamma}) = 0$, which is

$$\widehat{\gamma} = \frac{\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})}{2(\widehat{\Delta}_t(\mu) - \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}))} \,.$$

That $\widehat{\gamma}$ is always positive follows from the fact that $\widehat{\Delta}_t(\mu) > \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})$. If $\widehat{\gamma} \leq 1$, then we can take the forerunner to be the mixture policy with $\gamma = \widehat{\gamma}$. In this case,

$$\begin{aligned}
\overline{\mathrm{IR}}_t^{(3)}(\pi_t^{(\mathbf{mix})}) &= \frac{(\frac{3}{2})^3 2(\widehat{\Delta}_t(\mu) - \widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}))\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})})^2}{\overline{\mathrm{IG}}_t(\mu)} \\
&\leq \frac{(\frac{3}{2})^3 8s}{C_{\min}} = \frac{27s}{C_{\min}} \,.
\end{aligned}$$

Otherwise, if $\widehat{\gamma} > 1$, then

$$\widehat{\Delta}_t(\mu) \leq \frac{3}{2}\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}) \,.$$

In this case, we can take the forerunner to be μ. The surrogate 3-information ratio can then be upper bounded as

$$\overline{\mathrm{IR}}_t^{(3)}(\mu) = \frac{\widehat{\Delta}_t(\mu)(\widehat{\Delta}_t(\mu))^2}{\overline{\mathrm{IG}}_t(\mu)} \leq \frac{2(\frac{3}{2})^2(\widehat{\Delta}_t(\pi_t^{(\mathbf{FGTS})}))^2}{\overline{\mathrm{IG}}_t(\mu)} \leq \frac{(\frac{3}{2})^2 4s}{C_{\min}} = \frac{9s}{C_{\min}} \,.$$

Therefore, one can always find a value of $\gamma \in (0, 1]$ such that

$$\overline{\mathrm{IR}}_t^{(3)}(\pi_t^{(\mathbf{mix})}) \leq \frac{27s}{C_{\min}} \,.$$

G Choosing the learning rates

This section is focused on the choice of the learning rates required to obtain the bound of Theorem 2.

G.1 Technical tools

We start by a collection of technical results to help with choosing a time-dependent learning rate.

Lemma 19. *Let $a_i \geq 0$ and $f : [0, \infty) \to [0, \infty)$ be a nonincreasing function. Then*

$$\sum_{t=1}^{T} a_t f\left(\sum_{i=0}^{t} a_i\right) \leq \int_{a_0}^{\sum_{t=0}^{T} a_t} f(x)\, dx. \tag{32}$$

The proof follows from elementary manipulations comparing sums and integrals. The result is taken from Lemma 4.13 of Orabona [2019], where a complete proof is also supplied. The following lemma ensures that the learning rates are non-increasing.

Lemma 20. *Let $C_1 > e, C_2 > 0$ and define $\lambda_t = \frac{\log(C_1 t)}{C_2 t}$, then λ_t is a non-decreasing sequence.*

Proof. Let $t > 0$, we have

$$\frac{\log(C_1(t+1))}{\log(C_1 t)} = \frac{\log\left(C_1 t\left(\frac{t+1}{t}\right)\right)}{\log(C_1 t)} = \frac{\log(C_1 t) + \log\left(\frac{t+1}{t}\right)}{\log(C_1 t)} \leq 1 + \frac{1}{t \log(C_1 t)} \leq 1 + \frac{1}{t},$$

where the first inequality uses $\log(1 + x) \leq x$ for any $x > -1$ and the second inequality uses $\log(C_1 t) \geq \log(C_1) \geq 1$ because we assumed $C_1 \geq e$. Since $\frac{C_2(t+1)}{C_2 t} = 1 + \frac{1}{t}$, we can conclude that the sequence λ_t is non-increasing. □

G.2 Data-rich regime: Proof of Lemma 8

We start by focusing on the data rich regime, and we bound the following part of the regret bound given in Equation (12):

$$\frac{C_T}{\lambda_{T-1}} + \frac{32}{3}\sum_{t=1}^{T} \lambda_{t-1} \overline{\mathrm{IR}}_t^{(2)}(\pi_t).$$

Here, $C_T = 5 + 2s \log \frac{edT}{s}$. To proceed, we let $\lambda_t = \alpha\sqrt{\frac{C_{t+1}}{d(t+1)}}$, where $\alpha > 0$ is a constant that we will optimize later. Because $t \to C_t$ is increasing, we get that $\lambda_{t-1} \leq \alpha\sqrt{\frac{C_T}{dt}}$. By Lemma 7, we know that for all $t \geq 1$, $\overline{\mathrm{IR}}_t^{(2)}(\pi_t) \leq 2d$, hence

$$\begin{aligned}
\frac{C_T}{\lambda_{T-1}} + \frac{32}{3}\sum_{t=1}^{T} \lambda_{t-1} \overline{\mathrm{IR}}_t^{(2)}(\pi_t) &\leq \frac{1}{\alpha}\sqrt{C_T dT} + \frac{64}{3}\alpha\sqrt{C_T}\sum_{t=1}^{T}\frac{d}{\sqrt{dt}} \\
&\leq \frac{1}{\alpha}\sqrt{C_T dT} + \frac{128}{3}\alpha\sqrt{C_T dT} \\
&\leq \left(\frac{1}{\alpha} + \frac{128}{3}\alpha\right)\sqrt{C_T dT} \\
&\leq 16\sqrt{\frac{2}{3} C_T dT},
\end{aligned}$$

where the second line uses the standard inequality $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, and the last line is obtained by optimizing the expression $\left(\frac{1}{\alpha} + \frac{128}{3}\alpha\right)$ with the optimal choice $\alpha = \sqrt{\frac{3}{128}}$ which yields the value $16\sqrt{\frac{2}{3}}$. This concludes the proof of the claim. □

G.3 Data-poor regime: proof of Lemma 8

We now focus on the data-poor regime and specifically on bounding the following part of the bound given in Equation (12):

$$\frac{C_T}{\lambda_{T-1}} + \frac{16}{3} c_3^* \sum_{t=1}^{T} \sqrt{3\lambda_{t-1}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}.$$

Here, $C_T = 5 + 2s \log \frac{edT}{s}$. Now, we let $\lambda_t = \alpha \left(\frac{C_{t+1}\sqrt{C_{\min}}}{(t+1)\sqrt{s}}\right)^{\frac{2}{3}}$, where $\alpha > 0$ is a constant that we will optimize later. Because $t \to C_t$ is increasing, we get that $\lambda_{t-1} \le \alpha \left(\frac{C_T\sqrt{C_{\min}}}{ts}\right)^{\frac{2}{3}}$. By Lemma 7, the 3-surrogate-information ratio is bounded for all $t \ge 1$ as $\overline{\mathrm{IR}}_t^{(3)}(\pi_t) \le \frac{54s}{C_{\min}}$. Hence, the following holds:

$$\begin{aligned}
\frac{C_T}{\lambda_{T-1}} + \frac{16}{3} c_3^* \sum_{t=1}^{T} \sqrt{3\lambda_{t-1}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)} &\le \frac{1}{\alpha}(C_T)^{\frac{1}{3}} \left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}} + 48c_3^*\sqrt{2\alpha}(C_T)^{\frac{1}{3}} \left(\frac{\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}} \sum_{t=1}^{T} \frac{1}{t^{\frac{1}{3}}} \\
&\le \frac{1}{\alpha}(C_T)^{\frac{1}{3}} \left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}} + 72c_3^*\sqrt{2\alpha}(C_T)^{\frac{1}{3}} \left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}} \\
&\le \left(\frac{1}{\alpha} + 72c_3^*\sqrt{2\alpha}\right)(C_T)^{\frac{1}{3}} \left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}} \\
&\le 12 \cdot 6^{\frac{1}{3}}(C_T)^{\frac{1}{3}} \left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}}.
\end{aligned}$$

Here, we have applied Lemma 19 with the function $f(x) = x^{\frac{1}{3}}$and $a_i = 1$ to bound $\sum_{t=1}^{T} t^{-1/3} \le \frac{3}{2}T^{\frac{2}{3}}$ in the second line, the last line comes from the choice $\alpha = \frac{1}{4 \cdot 6^{\frac{1}{3}}}$ which optimizes the constant $\left(\frac{1}{\alpha} + 144c_3^*\sqrt{2\alpha}\right)$ (as per Lemma 27). This proves the statement. □

G.4 Joint learning rates, end of the proof of Theorem 2

In the section below, we present the technical derivation related to choosing the choice of learning rate $\lambda_t = \min(\frac{1}{2}, \max(\lambda_t^{(2)}, \lambda_t^{(3)}))$, where $\lambda_t^{(2)} = \sqrt{\frac{3C_{t+1}}{128d(t+1)}}$ and $\lambda_t^{(3)} = \frac{1}{4 \cdot 6^{\frac{1}{3}}} \left(\frac{C_{t+1}\sqrt{C_{\min}}}{(t+1)\sqrt{s}}\right)^{\frac{2}{3}}$, with $C_t = 5 + 2s \log \frac{edt}{s}$. This choice interpolates between the data-rich and data-poor regimes. As a first step, we start by confirming via Lemma 20 that both $\lambda_t^{(2)}$ and $\lambda_t^{(3)}$ are non-increasing and the bound of Theorem 1 holds with our choice of λ_t.

First, note that our choice of learning rates ensures that $\lambda_t \le \frac{1}{2}$ holds as long as T is larger than an absolute constant, and thus we focus on this case here (and relegate the complete details of establishing this absolute constant to Appendix G.5). To proceed, we define the (constant-free) regret rates $R_t^{(2)} = \sqrt{C_t dt}$ and $R_t^{(3)} = \left(t\sqrt{s\frac{C_t}{C_{\min}}}\right)^{\frac{2}{3}}$ and note that they correspond to the regret bounds obtained when using the respective learning rates $\lambda_t^{(2)}$ and $\lambda_t^{(3)}$, as per Lemma 8.

We now consider the last time that the learning rates $\lambda_t^{(3)}$ and $\lambda_t^{(2)}$ have been used. More specifically, we denote $T_2 = \max\{t \le T, \lambda_{t-1}^{(2)} \ge \lambda_{t-1}^{(3)}\}$, and $T_3 = \max\{t \le T, \lambda_{t-1}^{(3)} \ge \lambda_{t-1}^{(2)}\}$. Combining the bound of Equation 12 and using the definition $\lambda_t = \min(\frac{1}{2}, \max(\lambda_t^{(2)}, \lambda_t^{(3)}))$, the following bound holds

$$
\begin{aligned}
&R_T \\
&\le \mathbb{E}\left[\frac{C_T}{\lambda_{T-1}} + \sum_{t=1}^{T} \min\left(\frac{32}{3}\lambda_{t-1}\overline{\mathrm{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}\right)\right] \\
&= \mathbb{E}\left[\frac{C_T}{\min(\frac{1}{2}, \max(\lambda_{T-1}^{(2)}, \lambda_{T-1}^{(3)}))}\right. \\
&\left. + \sum_{t=1}^{T} \min\left(\frac{32}{3}\min(\frac{1}{2}, \max(\lambda_{t-1}^{(2)}, \lambda_{t-1}^{(3)}))\overline{\mathrm{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\min(\frac{1}{2}, \max(\lambda_{t-1}^{(2)}, \lambda_{t-1}^{(3)}))\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}\right)\right] \\
&\le \mathbb{E}\left[C_T \min\left(\frac{1}{\lambda_{T-1}^{(2)}}, \frac{1}{\lambda_{T-1}^{(3)}}\right) + \sum_{t=1}^{T} \min\left(\frac{32}{3}\max(\lambda_{t-1}^{(2)}, \lambda_{t-1}^{(3)})\overline{\mathrm{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\max(\lambda_{t-1}^{(2)}, \lambda_{t-1}^{(3)})\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}\right)\right].
\end{aligned}
$$

We can now separate the sum obtained at the last line based on which learning rate was used at time t.

$$
\begin{aligned}
&\sum_{t=1}^{T} \min\left(\frac{32}{3}\max(\lambda_{t-1}^{(2)}, \lambda_{t-1}^{(3)})\overline{\mathrm{IR}}_t^{(2)}(\pi_t), \frac{16}{3}c_3^*\sqrt{3\max(\lambda_{t-1}^{(2)}, \lambda_{t-1}^{(3)})\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}\right) \\
&\le \sum_{\lambda_t^{(2)} \ge \lambda_t^{(3)}} \frac{32}{3}\lambda_{t-1}^{(2)}\overline{\mathrm{IR}}_t^{(2)}(\pi_t) + \sum_{\lambda_t^{(3)} \ge \lambda_t^{(2)}} \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}^{(3)}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)} \\
&\le \sum_{t=1}^{T_2} \frac{32}{3}\lambda_{t-1}^{(2)}\overline{\mathrm{IR}}_t^{(2)}(\pi_t) + \sum_{t=1}^{T_3} \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}^{(3)}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)}.
\end{aligned}
$$

Following exactly the same step as in the proof of Lemma 8, we further bound $\sum_{t=1}^{T_2} \frac{32}{3}\lambda_{t-1}^{(2)}\overline{\mathrm{IR}}_t^{(2)}(\pi_t) \le 8\sqrt{\frac{2}{3}}R_{T_2}^{(2)}$ and $\sum_{t=1}^{T_3} \frac{16}{3}c_3^*\sqrt{3\lambda_{t-1}^{(3)}\overline{\mathrm{IR}}_t^{(3)}(\pi_t)} \le 8 \cdot 6^{\frac{1}{3}} R_{T_3}^{(3)}$.

The crucial observation is that which of $\lambda_T^{(3)}$ or $\lambda_T^{(2)}$ is bigger will determine whether $R_T^{(2)}$ or $R_T^{(3)}$ is the term of leading order (up to some constants). More specifically, Let T be such that $\lambda_{T-1}^{(2)} \ge \lambda_{T-1}^{(3)}$ which means that $\sqrt{\frac{3C_T}{128dT}} \ge \frac{1}{4 \cdot 6^{\frac{1}{3}}}\left(\frac{C_T\sqrt{C_{\min}}}{T\sqrt{s}}\right)^{\frac{2}{3}}$. Rearraging, this implies that $\sqrt{C_T dT} \le \frac{6^{\frac{5}{6}}}{4}\left(T\sqrt{s\frac{C_T}{C_{\min}}}\right)^{\frac{2}{3}}$, which means that $R_T^{(2)} \le \frac{6^{\frac{5}{6}}}{4}R_T^{(3)}$. Following the exact same steps, we also have that $\lambda_{T-1}^{(3)} \ge \lambda_{T-1}^{(2)}$ implies that $R_T^{(3)} \le \frac{4}{6^{\frac{5}{6}}}R_T^{(2)}$. We apply this to the time T_2 in which $\lambda_{T_2-1}^{(2)} \ge \lambda_{T_2-1}^{(3)}$ by definition. we have that $R_{T_2}^{(2)} \le \frac{6^{\frac{5}{6}}}{4}R_{T_2}^{(3)}$ and putting this together with the previous bound, we have

$$
\begin{aligned}
R_T &\le \frac{C_T}{\lambda_{T-1}^{(3)}} + 8\sqrt{\frac{2}{3}}R_{T_2}^{(2)} + 8 \cdot 6^{\frac{1}{3}} R_{T_3}^{(3)} \\
&\le 4 \cdot 6^{\frac{1}{3}} R_T^{(3)} + 8\sqrt{\frac{2}{3}} \cdot \frac{6^{\frac{5}{6}}}{4} R_{T_2}^{(2)} + 8 \cdot 6^{\frac{1}{3}} R_{T_3}^{(3)} \\
&\le 4 \cdot 6^{\frac{1}{3}} R_T^{(3)} + 4 \cdot 6^{\frac{1}{3}} R_{T_2}^{(3)} + 8 \cdot 6^{\frac{1}{3}} R_{T_3}^{(3)} \\
&\le 4 \cdot 6^{\frac{1}{3}} R_T^{(3)} + 4 \cdot 6^{\frac{1}{3}} R_T^{(3)} + 8 \cdot 6^{\frac{1}{3}} R_T^{(3)} \\
&\le 16 \cdot 6^{\frac{1}{3}} R_T^{(3)},
\end{aligned}
$$

where we use the fact that $T \to R_T^{(3)}$ is increasing and $T_2 \le T, T_3 \le T$.

Using the same argument as before, we have that $\lambda_{T_3-1}^{(3)} \ge \lambda_{T_3-1}^{(2)}$, and we can conclude that $R_{T_3}^{(3)} \le \frac{4}{6^{\frac{5}{6}}}R_{T_3}^{(2)}$.

Putting this together, with the previous bound, we have

$$\begin{aligned}
R_T &\leq \frac{C_T}{\lambda^{(2)}_{T-1}} + 8\sqrt{\frac{2}{3}}R^{(2)}_{T_2} + 8\cdot 6^{\frac{1}{3}}R^{(3)}_{T_3} \\
&\leq 8\sqrt{\frac{2}{3}}R^{(2)}_{T} + 8\sqrt{\frac{2}{3}}R^{(2)}_{T_2} + 8\cdot 6^{\frac{1}{3}}\cdot\frac{4}{6^{\frac{5}{6}}}R^{(3)}_{T_3} \\
&\leq 8\sqrt{\frac{2}{3}}R^{(2)}_{T} + 8\sqrt{\frac{2}{3}}R^{(2)}_{T_2} + 16\sqrt{\frac{2}{3}}R^{(2)}_{T_3} \\
&\leq 8\sqrt{\frac{2}{3}}R^{(2)}_{T} + 8\sqrt{\frac{2}{3}}R^{(2)}_{T} + 16\sqrt{\frac{2}{3}}R^{(2)}_{T} \\
&\leq 32\sqrt{\frac{2}{3}}R^{(2)}_{T},
\end{aligned}$$

where we use the fact that $T \to R^{(3)}_T$ is increasing and $T_2 \leq T, T_3 \leq T$. Evaluating the constants numerically yields $16 \cdot 6^{\frac{1}{3}} \approx 29.07 \leq 30$ and $32\sqrt{\frac{2}{3}} \approx 26.13 \leq 27$.

G.5 Upper bound on the learning rates

We now consider the case where the learning rates exceed $\frac{1}{2}$, and show that this only holds for small values of T. First, we have that $\lambda^{(2)}_{T-1} \leq \frac{1}{2}$ if

$$\sqrt{\frac{3C_T}{128dT}} \leq \frac{1}{2}.$$

Rearranging the inequality and recalling $C_T = 5 + 2s\log\frac{edT}{s}$, this is equivalent to

$$T \geq \frac{15}{32d} + \frac{3s}{16d}\log\frac{edT}{s}.$$

Using the loose inequality $\log\frac{edT}{s} \leq \frac{dT}{s}$, we get that this condition is satisfied for any $T \geq 1$.

Similarly, we have that $\lambda^{(3)}_{T-1} \leq \frac{1}{2}$ if

$$\frac{1}{4\cdot 6^{\frac{1}{3}}}\left(\frac{C_T\sqrt{C_{\min}}}{T\sqrt{s}}\right)^{\frac{2}{3}} \leq \frac{1}{2}.$$

We note that

$$C_{\min} = \max_{\mu\in\Delta(A)}\sigma_{\min}(\mathbb{E}_{A\sim\mu}\left[AA^T\right]) \leq \max_{\mu\in\Delta(A)}\frac{\text{Tr}(\mathbb{E}_{A\sim\mu}\left[AA^T\right])}{d} \leq 1,$$

where the first inequality uses that the trace of a matrix is always bigger than d-times its smallest eigenvalue and the second inequality uses the fact that for any matrix A, we have $\text{Tr}(AA^T) = \sum_{i=1}^d a_i^2 \leq d\max_i |a_i| \leq d$ because we assumed that all the actions are bounded in infinity norm. Hence the previous inequality will be satisfied if

$$\frac{1}{4\cdot 6^{\frac{1}{3}}}\left(\frac{C_T}{T\sqrt{s}}\right)^{\frac{2}{3}} \leq \frac{1}{2}.$$

Rearranging the inequality, this is equivalent to

$$T \geq 4\sqrt{\frac{3}{s}}C_t = 8\sqrt{3s}\log(eT) + \sqrt{3s}\left(\frac{20}{s} + 8\log\frac{d}{s}\right).$$

Applying Lemma 24 with $a = 8\sqrt{3s}$ and $b = \sqrt{3s}\left(\frac{20}{s} + 8\log(\frac{d}{s})\right)$, we find that the previous inequality is satisfied for all

$$T \geq 2a\log ea + 2b = 40\sqrt{\frac{3}{s}} + 16\sqrt{3s}\log\frac{8e\sqrt{3}d}{\sqrt{s}}.$$

Thus, letting $T_{\min} = 40\sqrt{\frac{3}{s}} + 16\sqrt{3s}\log\frac{8e\sqrt{3}d}{\sqrt{s}}$ be the constant given above, both learning rates stay upper bounded by $\frac{1}{2}$ for all $T \geq T_{\min}$ and the upper bound on the regret given the previous subsection holds. Otherwise, we upper bound the instantaneous regret by 2 and this leads to an additional $2T_{\min} = \mathcal{O}(\sqrt{s}\log\frac{d}{\sqrt{s}})$ in the regret. Putting this together with the bound proved in the previous section, we thus have that the following regret bound is valid for any $T \geq 1$:

$$R_T \leq \min\left(27\sqrt{\left(5 + 2s\log\frac{edT}{s}\right)dT}, 30\left(5 + 2s\log\frac{edT}{s}\right)^{\frac{1}{3}}\left(\frac{T\sqrt{s}}{\sqrt{C_{\min}}}\right)^{\frac{2}{3}}\right) + \mathcal{O}\left(\sqrt{s}\log\frac{d}{\sqrt{s}}\right).$$

This concludes the proof of Theorem 2. □

I Technical Results

In this section, we state and prove the remaining technical results.

Lemma 21. *Let $\pi \in \Delta(\mathcal{A})$, the function $\theta \to \Delta(\pi, \theta)$ is 2-Lipschitz with respect to the 1 norm. Let $t \geq 1$, the function $\theta \to \mathbb{E}\left[\log\left(\frac{1}{p_t(Y_t|\theta, A_t)}\right)\right]$ is 2-Lipschitz with respect to the 1 norm.*

Proof. Let $\theta, \theta' \in \Theta$, we have

$$\begin{aligned}
|r(\pi,\theta) - r(\pi,\theta')| &= \left|\sum_{a\in\mathcal{A}} \pi(a)\langle\theta - \theta', a\rangle\right| \\
&\leq \sum_{a\in\mathcal{A}} \pi(a)|\langle\theta - \theta', a\rangle| \\
&\leq \sum_{a\in\mathcal{A}} \pi(a)\left\|\theta - \theta'\right\|_1 \|a\|_\infty \\
&\leq \left\|\theta - \theta'\right\|_1 .
\end{aligned}$$

Similarly,

$$|r^*(\theta) - r^*(\theta')| = |\max_{a\in\mathcal{A}} r(a,\theta) - \max_{a\in\mathcal{A}} r(a,\theta')| \leq \max_{a\in\mathcal{A}} |r(\theta,a) - r(a,\theta')| \leq \left\|\theta - \theta'\right\|_1 .$$

Finally

$$|\Delta(\pi,\theta) - \Delta(\pi,\theta')| = |r^*(\theta) - r^*(\theta') + r(\pi,\theta') - r(\pi,\theta)| \leq 2\left\|\theta - \theta'\right\|_1 .$$

For the negative log-likelihood, for simplicity, we let $r = \langle\theta, A_t\rangle$, $r' = \langle\theta', A_t\rangle$ and $r_0 = \langle\theta_0, A_t\rangle$,

$$\begin{aligned}
\mathbb{E}\left[\log\left(\frac{1}{p(Y_t|\theta, A_t)}\right) - \log\left(\frac{1}{p(Y_t|\theta', A_t)}\right)\right] &= \frac{1}{2}\mathbb{E}\left[(\langle\theta, A_t\rangle - Y_t)^2 - (\langle\theta', A_t\rangle - Y_t)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[(r - Y_t)^2 - (r' - Y_t)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[(r - r')(r + r' - 2Y_t)\right] \\
&= \frac{1}{2}\mathbb{E}\left[(r - r')(r + r' - 2r_0)\right] \\
&\leq 2\left\|\theta - \theta'\right\|_1 .
\end{aligned}$$

□

Lemma 22. *(Hoeffding's Lemma) Let X be a bounded real random variable such that $X \in [a, b]$ almost surely. Let $\eta \neq 0$, then we have*

$$\frac{1}{\eta}\log\mathbb{E}\left[\exp\left(\eta X\right)\right] \leq \mathbb{E}\left[X\right] + \frac{\eta(b-a)^2}{8}. \tag{33}$$

Proof. See for instance Chapter 2 in Boucheron et al. [2013]. □

We now provide a data dependent version of Hoeffding's lemma that is used in the analysis of the gaps in the optimistic posterior.

Lemma 23. *(A data dependent version of Hoeffding's Lemma) Let X be a real random variable and $\eta \neq 0$ be such that $\eta X \leq 1$ almost surely, then we have*

$$\frac{1}{\eta}\log \mathbb{E}\left[\exp\left(\eta X\right)\right] \leq \mathbb{E}\left[X\right] + \eta \mathbb{E}\left[X^2\right] \leq 2\mathbb{E}\left[X\right]. \tag{34}$$

Proof. Using the elementary inequalities $\log(x) \leq x - 1$ for $x > 0$ and $e^x \leq 1 + x + x^2$ for $x \leq 1$, we get that

$$\begin{aligned}\frac{1}{\eta}\log \mathbb{E}\left[\exp\left(\eta X\right)\right] &\leq \frac{1}{\eta}\mathbb{E}\left[\exp(\eta X) - 1\right] \\ &\leq \frac{1}{\eta}\mathbb{E}\left[\eta X + \eta^2 X^2\right] \\ &\leq \mathbb{E}\left[X\right] + \eta\mathbb{E}\left[X^2\right].\end{aligned}$$

□

The following lemmas help us to analyze when the learning rates are smaller or bigger than $\frac{1}{2}$.

Lemma 24. *Let $a \geq 1, b \geq 0$, then, the equation $t \geq a \log et + b$ is verified for any $t \geq 2a\log ea + 2b$* .

Proof. We let $f(t) = t - a\log et - b$, we have that $f'(t) \geq 0$ on $[a, +\infty)$ and $f(a) \leq 0$. Hence $f(t) = 0$ has a unique solution α on $[a, \infty)$ such that $f(t) \geq 0$ if $t \geq \alpha$. We now focus on upper bounding α. The equation $f(\alpha) = 0$ is equivalent to

$$\log \alpha = \frac{\alpha - b}{a} - 1.$$

Now taking the exponential and reordering this is also equivalent to

$$\frac{-\alpha}{a}\exp\left(\frac{-\alpha}{a}\right) = \frac{\exp\left(-\frac{a+b}{a}\right)}{a}.$$

Let

$$\begin{aligned} g : (-\infty, -1] &\longrightarrow [-\frac{1}{e}, 0) \\ x &\longmapsto xe^x. \end{aligned}$$

The previous equation can be rewritten $g\left(\frac{-\alpha}{a}\right) = -\frac{\exp\left(-\frac{a+b}{a}\right)}{a}$.

We define $W_{-1} : [-\frac{1}{e}, 0) \longrightarrow (-\infty, 1]$ as the(functional) inverse of g. g is the -1 branch of the Lambert W function.

We have that for any $x \leq -1$, $W_{-1}(xe^x) = x$ and that for any $y \geq e$, $-W_{-1}(-\frac{1}{y}) \leq 2\log(y)$. Since g is decreasing on its domain, W_{-1} is well-defined and decreasing. Moreover, for any $x \leq -1$, $W_{-1}(g(x)) = x$. In particular, we have that $\alpha = aW_{-1}\left(-\frac{\exp\left(-\frac{a+b}{a}\right)}{a}\right)$. We will use that formulation to find an upper bound on α.

We fix some $y \geq e$. We have $-2\log(y) \leq -1$ hence $W_{-1}\left(-2\log(y)e^{(-2\log(y))}\right) = -2\log(y)$, which means that $2\log(y) = -W_{-1}(-\frac{1}{y^*})$ where $y^* = \frac{e^{(2\log(y))}}{2\log(y)} = \frac{y^2}{2\log(y)}$.

Because of the elementary inequality $2\log(x) \leq x$ for $x > 0$, we conclude that $y \leq y^*$. Since $y \longrightarrow -W_{-1}(-\frac{1}{y})$ is an increasing function we finally have that for any $y \geq e$

$$W_{-1}\left(-\frac{1}{y}\right) \leq W_{-1}\left(-\frac{1}{y^*}\right) = 2\log(y).$$

Applying this to $y = a \exp\left(\frac{a+b}{a}\right) \geq e$, we get

$$\alpha = W_{-1}\left(\frac{-1}{y}\right) \leq 2\log(y) = 2a \log ea + 2b.$$

Since any $t \geq \alpha$ will satisfy $f(t) \geq 0$, this concludes our proof.

□

Lemma 25. *Let $\theta \in \Theta$, then $M_t = \exp(L_t^{(1)}(\theta_0) - L_t^{(1)}(\theta)) = \prod_{s=1}^t \frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}$ is a supermartingale with respect to the filration $\mathcal{F}_t$.*

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\middle|\mathcal{F}_{t-1}, A_t\right] &= \mathbb{E}\left[\exp\left(\frac{(\langle\theta_0, A_t\rangle - Y_t)^2 - (\langle\theta, A_t\rangle - Y_t^2)}{2}\right)\middle|\mathcal{F}_{t-1}, A_t\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\epsilon_t^2 - (\langle\theta - \theta_0, A_t\rangle - \epsilon_t)^2}{2}\right)\middle|\mathcal{F}_{t-1}, A_t\right] \\
&= \exp\left(-\frac{(\langle\theta - \theta_0, A_t\rangle)^2}{2}\right)\mathbb{E}\left[\exp\left(\epsilon_t\langle\theta - \theta_0, A_t\rangle\right)|\mathcal{F}_{t-1}, A_t\right] \\
&\leq \exp\left(-\frac{(\langle\theta - \theta_0, A_t\rangle)^2}{2}\right)\cdot\exp\left(\frac{(\langle\theta - \theta_0, A_t\rangle)^2}{2}\right) \\
&= 1,
\end{aligned}$$

where the inequality comes from the conditional subgaussianity of ϵ_t. Finally, by the tower rule of conditional expectations

$$\mathbb{E}\left[\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\middle|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{p(Y_t|\theta, A_t)}{p(Y_t|\theta_0, A_t)}\middle|\mathcal{F}_{t-1}, A_t\right]\middle|\mathcal{F}_{t-1}\right] \leq 1.$$

□

I.1 Proof of Proposition 1

This is coming from the fact that the mean is the constant minimizing the mean squared error. We remind the reader of the definition of the surrogate information gain and the true information gain for a policy $\pi \in \Delta(\mathcal{A})$

$$\overline{\mathrm{IG}}_t(\pi) = \frac{1}{2}\sum_{a\in\mathcal{A}}\pi(a)\int_\Theta(\langle\theta - \bar{\theta}(Q_t^+), a\rangle)^2\, dQ(\theta), \tag{35}$$

where $\bar{\theta}(Q_t^+) = \mathbb{E}_{\theta\sim Q_t^+}[\theta]$ is the mean parameter under the optimistic posterior Q_t^+.

$$\mathrm{IG}_t(\pi) = \frac{1}{2}\sum_{a\in\mathcal{A}}\pi(a)\int_\Theta(\langle\theta, a\rangle - \langle\theta_0, a\rangle)^2\, dQ_t^+(\theta), \tag{36}$$

Let's fix $a \in \mathcal{A}$, we have that

$$\begin{aligned}
(\langle\theta - \theta_0, a\rangle)^2 &= (\langle\theta - \bar{\theta}(Q_t^+) + \bar{\theta}(Q_t^+) - \theta_0, a\rangle)^2 \\
&= (\langle\theta - \bar{\theta}(Q_t^+), a\rangle)^2 + 2\langle\theta - \bar{\theta}(Q_t^+), a\rangle\langle\bar{\theta}(Q_t^+) - \theta_0, a\rangle + (\langle\bar{\theta}(Q_t^+) - \theta_0, a\rangle)^2 \\
&\geq (\langle\theta - \bar{\theta}(Q_t^+), a\rangle)^2 + 2\langle\theta - \bar{\theta}(Q_t^+), a\rangle\langle\bar{\theta}(Q_t^+) - \theta_0, a\rangle
\end{aligned}$$

Now using that $\bar{\theta}(Q_t^+) = \int_\Theta \theta\, dQ_t^+(\theta)$ and integrating, we get

$$\int_\Theta(\langle\theta - \theta_0, a\rangle)^2\, dQ_t^+(\theta) \geq \int_\Theta(\langle\theta - \bar{\theta}(Q_t^+), a\rangle)^2\, dQ_t^+(\theta).$$

Multiplying by $\pi(a)$ and summing over actions, we get the claim of the lemma.

I.2 Generalization of the AM-GM inequality

Dealing with the generalized information ratio requires bounding the cubic root of products. While one could use Hölder's inequality to deal directly with products, we find it more flexible to use a variational form of this inequality. In all that follows, we let $p > 1$ be a real number and q be such that $\frac{1}{p} + \frac{1}{q} = 1$. It is not hard to check that $q = \frac{p}{p-1}$. We start by stating a direct consequence of the Fenchel-Young Inequality which can be seen as an extension of the AM-GM inequality.

Lemma 26. *Let $x, y \geq 0$, then*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}. \tag{37}$$

With equality if and only if $px^{p-1} = y$

Proof. One can check that the Fenchel dual of the function

$$\begin{aligned} f : &\mathbb{R}^+ \longrightarrow \mathbb{R} \\ x &\longmapsto \frac{x^p}{p} \end{aligned}$$

is exactly $f^*(y) = \frac{1}{q}|y|^q sgn(y)$. Then the Lemma is a direct consequence of the Fenchel Young inequality and of its equality case. $\square$

Refining a bit this Lemma, we get the following variational form of the previous inequality :

Lemma 27. *Let $x, y \geq 0, \lambda > 0$, then*

$$\sqrt[p]{xy} \leq \frac{x}{\lambda} + c_p^*(\lambda y)^{\frac{1}{p-1}} \tag{38}$$

where $c_p^* = (p-1)\frac{1}{p}^{\frac{p}{p-1}}$ with equality if and only if $x = y = 0$ or $\lambda = p\frac{x^{\frac{p-1}{p}}}{y^{\frac{1}{p}}}$.

Proof. We apply the previous lemma to $\sqrt[p]{\frac{px}{\lambda}}$ and $\sqrt[p]{\frac{\lambda y}{p}}$. $\square$

In order to go from the variational form to the product form, we may use the following result.

Lemma 28. *Let $\alpha, \beta > 0$, then*

$$\inf_{\lambda > 0} \frac{\alpha}{\lambda} + \beta\lambda^{\frac{1}{p-1}} = c_p \alpha^{\frac{1}{p}} \beta^{\frac{p-1}{p}}, \tag{39}$$

where $c_p = p\frac{1}{p-1}^{\frac{p-1}{p}}$ satisfies $c_p \cdot c_p^{*\frac{p-1}{p}} = 1$, and the minimum is reached at $\lambda^* = (p-1)^{\frac{p-1}{p}} \frac{\alpha^{\frac{p-1}{p}}}{\beta^{\frac{p-1}{p}}}$.

Proof. Applying the previous Lemma to $x = \alpha$ and $y = c_p^{\frac{p}{p-1}} \beta^{p-1}$ yields the result. $\square$

Remark An alternative is to pick λ to make both terms equals resulting in the same result but with 2 as a leading constant. Now

$$\begin{aligned} c_p &= p^{\frac{1}{p}} \frac{p}{p-1}^{\frac{p-1}{p}} \\ &= \exp\left(\frac{1}{p}\log p + \frac{p-1}{p}\log\frac{p}{p-1}\right) \\ &\leq \frac{1}{p} \cdot p + \frac{p-1}{p} \cdot \frac{p}{p-1} \\ &= 2. \end{aligned}$$

With equality if and only if $p = 2$. So, the choice of c_p always yields a better leading constant. However, $c_3 \simeq 1.88$ so one could argue that the gain is small. Since we will usually use Lemma 27, c_p^* will naturally appear and c_p will cancel it, ultimately making the leading constant as simple as possible.

J Experimental details

Here, we describe our implementation of the SOIDS algorithm in more detail, as well as the hyperparameters of all the methods used in our experiments. To run the SOIDS algorithm, one must minimise $\overline{\mathrm{IR}}_t^{(2)}(\pi)$ w.r.t. π in each round t. This is not straightforward, because $\overline{\mathrm{IR}}_t^{(2)}(\pi)$ contains expectations w.r.t. the optimistic posterior Q_t^+. When we use the Spike-and-Slab prior in Appendix B.2, we are not aware of any efficient method that can be used to maximise $\overline{\mathrm{IR}}_t^{(2)}(\pi)$. Instead, we draw (approximate) samples $\theta^{(1)}, \dots, \theta^{(M)}$ from Q_t^+ to produce the estimates $\widetilde{\Delta}_t(\pi)$ and $\widetilde{\mathrm{IG}}_t(\pi)$ for the surrogate regret and the surrogate information respectively, where

$$\widetilde{\Delta}_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a) \frac{1}{M} \sum_{i=1}^{M} \Delta(a, \theta^{(i)}), \qquad \widetilde{\mathrm{IG}}_t(\pi) = \frac{1}{2} \sum_{a \in \mathcal{A}} \pi(a) \frac{1}{M} \sum_{i=1}^{M} \left(\langle \theta^{(i)} - \bar{\theta}_M, a \rangle\right)^2 .$$

Here, $\bar{\theta}_M$ is the sample mean $\frac{1}{M}\sum_{i=1}^{M} \theta^{(i)}$. We then maximimse the approximate surrogate information ratio $\widetilde{\mathrm{IR}}_t^{(2)}(\pi)$, where

$$\widetilde{\mathrm{IR}}_t^{(2)}(\pi) = \frac{(\widetilde{\Delta}_t(\pi))^2}{\widetilde{\mathrm{IG}}_t(\pi)} .$$

To draw the samples $\theta^{(1)}, \dots, \theta^{(M)}$, we use the empirical Bayesian sparse sampling procedure proposed by Hao et al. [2021], which is designed to draw samples from the Bayesian posterior. To sample from the optimistic posterior, we incorporate the optimistic adjustment into the likelihood. This method replaces the theoretically sound spike-and-slab prior with a relaxation in which the "spikes" are Laplace distributions with small variance, and the "slabs" are Gaussian distributions with large variance. In particular, the density of this prior is

$$q_1(\theta) = \sum_{\gamma \in \{0,1\}^d} p(\gamma) \prod_{j=1}^{d} [\gamma_j \psi_1(\theta_j) + (1 - \gamma_j)\psi_0(\theta_j)] .$$

Here, $\psi_1(\theta)$ is the density function of a univariate Gaussian distribution, with mean 0 and variance ρ_1, and ψ_0 is the density function of a univariate Laplace distribution, with mean 0 and scale parameter ρ_0. $p(\gamma)$ is a product of Bernoulli distributions with mean β. In our experiments, we always use $\rho_1 = 10$, $\rho_0 = 0.1$ and $\beta = 0.1$. Also, we set the learning rates to $\eta = 1/2$ and $\lambda_t = \min(\frac{1}{2}, \frac{1}{10}\max(\sqrt{\frac{s\log(edt/s)}{dt}}, (\frac{\log(edt/s)}{t})^{2/3}))$.

Implementing the OTCS baseline exactly would require us to compute the means of the distributions played by an exponentially weighted average forecaster with a sparsity prior. These distributions are the same as the optimistic posterior, except $\lambda_t = 0$ (i.e. there is no optimistic adjustment). In our implementation of the OTCS baseline, we draw samples using the same empirical Bayesian sparse sampling procedure, and then replace the exact means with the sample means. We use the same choices for the parameters η, ρ_1, ρ_0 and β. We set the radii of the confidence sets to the values given in Theorem 4.7 of Clerico et al. [2025]

For the LinUCB baseline, we set the radii of the confidence sets to the values given in Theorem 2 of Abbasi-Yadkori et al. [2011]. For the ESTC baseline, we set the exploration length T_1 to 50 when $d = 20$, 100 when $d = 40$ and $d = 100$. These values were chosen based on a small amount of trial and error. The theoretically motivated values in Theorem 4.2 of Hao et al. [2020] are much larger than these values. Also for ESTC, we set the LASSO regularisation parameter to $\lambda = 4\sqrt{\log(d)/T_1}$, which is the value given in Theorem 4.2 of Hao et al. [2020].