Proceedings of Machine Learning Research vol 134:1–42, 2021

Asymptotically Optimal Information-Directed Sampling

Johannes Kirschner JKIRSCHNER@INF.ETHZ.CH
ETH Zurich, Department of Computer Science *

Tor Lattimore LATTIMORE@GOOGLE.COM
DeepMind

Claire Vernade VERNADE@GOOGLE.COM
DeepMind

Csaba Szepesvári SZEPESVA@UALBERTA.CA
University of Alberta & Deepmind

Abstract

We introduce a simple and efficient algorithm for stochastic linear bandits with finitely many actions that is asymptotically optimal and (nearly) worst-case optimal in finite time. The approach is based on the frequentist information-directed sampling (IDS) framework, with a surrogate for the information gain that is informed by the optimization problem that defines the asymptotic lower bound. Our analysis sheds light on how IDS balances the trade-off between regret and information and uncovers a surprising connection between the recently proposed primal-dual methods and the IDS algorithm. We demonstrate empirically that IDS is competitive with UCB in finite-time, and can be significantly better in the asymptotic regime.

1. Introduction

The stochastic linear bandit problem is an iterative game between a learner and an environment played over n rounds. In each round t, the learner chooses an action (or arm) x_t from a finite set of actions $\mathcal{X} \subset \mathbb{R}^d$ and observes a noisy reward $y_t = \langle x_t, \theta^* \rangle + \epsilon_t$ where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector and ϵ_t is zero-mean noise. The learner's goal is to maximize the expected cumulative reward or, equivalently, to minimize the expected regret, which is defined by

$$R_n(\pi, \theta^*) = \mathbb{E}\left[\max_{x \in \mathcal{X}} \sum_{t=1}^{n} \langle x - x_t, \theta^* \rangle\right] , \tag{1}$$

where π is the policy mapping sequences of action/reward pairs to distributions over actions in $\mathcal{X}$ and the expectation is over the randomness in the policy and the rewards. Unlike in the multi-armed bandit setting, the linear structure allows the learner to estimate the reward of an action without directly observing it. In particular, the learner might play an action that it knows to be suboptimal in order to most efficiently identify the optimal action.

The *worst-case regret* $R_n(\pi) = \sup_{\theta \in \mathcal{M}} R_n(\pi, \theta)$ measures the performance of a policy on an adversarially chosen parameter θ in a class of models $\mathcal{M}$. On the other hand, for a fixed instance θ^*, an algorithm can perform much better than the worst-case regret $R_n(\pi)$ suggests, and achieving the optimal instance-dependent regret $R_n(\pi, \theta^*)$ is therefore of significant interest. On a large horizon, the optimal instance-dependent regret, or *asymptotic regret*, is characterized by a convex program, that optimizes the allocated proportion of plays to each action to minimize the regret, subject to the constraint that the policy gathers enough information to infer the best action (Graves and Lai, 1997).

*Research conducted during an internship at DeepMind.

The optimal worst-case regret rate (up to logarithmic factors) is achieved by various algorithms, including adaptations of the upper confidence bound (UCB) algorithm (Auer, 2003; Dani et al., 2008; Abbasi-Yadkori et al., 2011) and the information-directed sampling (IDS) approach (Russo and Van Roy, 2014; Kirschner and Krause, 2018). A conservative version of Thompson sampling is suboptimal by a factor of $\sqrt{d}$ and logarithmic factors (Agrawal and Goyal, 2013). On the other hand, achieving optimal asymptotic regret has proven to be more challenging. Lattimore and Szepesvári (2017) showed that algorithms based on optimism or Thompson sampling are not asymptotically optimal in the linear setting. They propose an approach based on the explore-then-commit framework that computes an estimate of the optimal allocation and updates the allocation to match the predicted target. Combes et al. (2017) follow a similar plan for the structured bandit setting, which includes the linear setting as a special case. This idea was subsequently extended to the contextual setting by Hao et al. (2019). Unfortunately these algorithms are not at all practical and do not enjoy reasonable minimax regret. More recently, Jun and Zhang (2020) refined this technique in the structured setting with a finite model class to avoid forced exploration and the knowledge of the horizon. Similarly, Van Parys and Golrezaei (2020) use a dual formulation of the lower bound to devise an algorithm that achieves the optimal asymptotic regret up to a constant, and avoids re-solving for the predicted optimal allocation at every round. Degenne et al. (2020) take a different approach and translate the Lagrangian of the lower bound into a fictitious two-player game, where the saddle point corresponds to the asymptotic regret. Using tools from online convex optimization (Hazan et al., 2016; Orabona, 2019), this leads to a family of asymptotically optimal algorithms, which incrementally update the allocation in each round based on primal-dual updates on the Lagrangian of the lower bound. Another primal-dual method is by Tirinzoni et al. (2020), which unlike previous methods is both worst-case and asymptotically optimal and also applies to the contextual case. We explain how IDS relates to primal-dual methods in Section 2.3. Finally, Wagenmaker et al. (2020) combine optimal experimental design with a phased elimination-style algorithm to derive finite-time guarantees that scale with the Gaussian width of the action set.

Contributions Our main contribution is new *conceptual insights* into information-directed sampling (IDS). We show that with an appropriate choice of the information gain, IDS performs *primal-dual updates* on the Lagrangian of the lower bound. The proposed version of IDS for the linear bandit setting is (nearly) *worst-case optimal* in finite time, satisfies an explicit *gap-dependent logarithmic regret* bound and is *asymptotically optimal*. All regret bounds are on *frequentist expected regret* and our analysis is relatively simple, avoiding all but one high-probability bound. The asymptotic analysis uncovers a connection between IDS and recently proposed primal-dual methods (Degenne et al., 2020; Tirinzoni et al., 2020). Moreover, our choice of information gain function approximates the variance based information gain proposed by (Russo and Van Roy, 2014) in the Bayesian setting.

Notation The real numbers are $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$ denotes the positive orthant. The standard Euclidean norm is $\|\cdot\|$ and the Euclidean inner product is $\langle \cdot, \cdot \rangle$. The Euclidean basis in $\mathbb{R}^m$ is $e_1, \dots, e_m$. The identity matrix in $\mathbb{R}^{d\times d}$ is $\mathbf{1}_d$. The diameter of a set $\mathcal{X} \subset \mathbb{R}^d$ is $\mathrm{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} \|x - y\|$. For a positive (semi-)definite, symmetric matrix $A \in \mathbb{R}^{d\times d}$ and a vector $v \in \mathbb{R}^d$, the associated matrix (semi-)norm is $\|v\|_A^2 = \langle v, Av\rangle$. $\mathscr{P}(\mathcal{X})$ is the set of probability measures on a finite set $\mathcal{X}$. Where convenient, we use vector notation, including inner products to denote evaluation of functions $F \in \mathbb{R}^{\mathcal{X}}$, for example $F(x) = \langle e_x, F\rangle$. Functions $F \in \mathbb{R}^{\mathcal{X}}$ are extended linearly to distributions $\mu \in \mathscr{P}(\mathcal{X})$ to denote the expectation $F(\mu) = \langle \mu, F\rangle = \sum_{x\in\mathcal{X}} f(x)\mu(x)$. In this context, we also use e_x for the Dirac measure on $x \in \mathcal{X}$. The reader may refer to Appendix A for a summary of notation.

1.1. Setting

Let $\mathcal{X} \subset \mathbb{R}^d$ be a finite set of k actions. We assume that $\mathcal{X}$ spans $\mathbb{R}^d$ and $\operatorname{diam}(\mathcal{X}) \le 1$. Denote by $\theta^* \in \mathcal{M}$ an unknown parameter vector, where $\mathcal{M} \subset \mathbb{R}^d$ is a known convex polytope with $\operatorname{diam}(\mathcal{M}) \le 1$. In each round $t = 1, \dots, n$, the learner chooses a distribution μ_t over $\mathcal{X}$. Then x_t is sampled from μ_t and the learner observes the reward $y_t = \langle x_t, \theta \rangle + \epsilon_t$ where ϵ_t is sampled independently from a Gaussian with zero mean and unit variance. All our upper bounds hold without modification for conditionally 1-subgaussian noise. The objective is to minimize the expected cumulative regret $R_n = R_n(\pi, \theta^*)$ defined in Eq. (1), where $\pi = (\mu_t)_{t=1}^n$ is the policy chosen by the learner. The dependency of the regret on θ^* and π is mostly omitted when there is no ambiguity. The expectation conditioned on previous observations is $\mathbb{E}_s[\,\cdot\,] = \mathbb{E}[\,\cdot\,|(x_l, y_l)_{l=1}^{s-1}]$. In line with all previous work focusing on the asymptotic setting, we assume that the optimal action $x^* = x^*(\theta^*) = \arg\max_{x \in \mathcal{X}} \langle x, \theta^* \rangle$ is unique. Eliminating this assumption is left as a delicate and possibly non-trivial challenge for the future. The sub-optimality gap of an action $x \in \mathcal{X}$ is $\Delta(x) = \langle x^* - x, \theta^* \rangle$ and $\Delta_{\min} = \min_{x \neq x^*} \Delta(x)$ denotes the smallest non-zero gap. For actions $x, z \in \mathcal{X}$, we denote by $\mathcal{H}_x^z = \{\nu \in \mathcal{M} : \langle x - z, \nu \rangle \ge 0\}$ the (convex) set of parameters where the reward of x is at least the reward of z. The set of *alternative parameters* is $\mathcal{C}^* = \cup_{x \neq x^*} \mathcal{H}_x^{x^*}$.

Asymptotic Lower Bound For an allocation $\alpha \in \mathbb{R}_{\ge 0}^{\mathcal{X}}$ over actions we define the associated covariance matrix $V(\alpha) = \sum_{x \in \mathcal{X}} \alpha(x) x x^\top$. Let c^* be the solution to the following convex program,

$$c^* \triangleq \inf_{\alpha \in \mathbb{R}_{\ge 0}^{\mathcal{X}}} \sum_{x \in \mathcal{X}} \alpha(x) \langle x^* - x, \theta^* \rangle \qquad \text{s.t.} \qquad \min_{\nu \in \mathcal{C}^*} \tfrac{1}{2} \|\nu - \theta^*\|_{V(\alpha)}^2 \ge 1\,. \tag{2}$$

The optimization minimizes the regret over (unbounded) allocations α that collect sufficient statistical evidence to reject all parameters $\nu \in \mathcal{C}^*$ for which an action $x \neq x^*$ is optimal. Note that for a fixed $\nu \in \mathbb{R}^d$, the constraints are linear in the allocation, $\|\nu - \theta\|_{V(\alpha)}^2 = \sum_{x \in \mathcal{X}} \alpha(x) \langle \nu - \theta, x \rangle^2$. The next lemma is a well-known result, which relates the asymptotic regret to the solution of (2). A policy π is called *consistent* if for all $\theta \in \mathcal{M}$ and $p > 0$ it holds that $R_n(\theta, \pi) = o(n^p)$. Assuming consistency is required to rule out policies that are defined to always play a fixed action x^*, which incurs zero regret when x^* is indeed optimal, but linear regret on other instances.

Theorem 1 (Graves and Lai (1997); Combes et al. (2017)) *Any consistent algorithm π for the linear bandit setting with Gaussian noise has regret $R_n(\theta^*, \pi)$ at least*

$$\liminf_{n \to \infty} \frac{R_n(\theta^*, \pi)}{\log(n)} \ge c^*(\theta^*)\,.$$

2. Asymptotically Optimal Information-Directed Sampling

The information-directed sampling (IDS) principle was introduced by Russo and Van Roy (2014) in the Bayesian setting. Our work is based on the frequentist version of this approach, developed by Kirschner and Krause (2018). The central idea is to compute a distribution over the actions that optimizes the following trade-off between a gap estimate $\hat{\Delta}_s(x)$ and an information gain $I_s(x)$, defined at step $s \ge 1$ for each $x \in \mathcal{X}$:

$$\mu_s = \operatorname*{arg\,min}_{\mu \in \mathscr{P}(\mathcal{X})} \left\{ \Psi_s(\mu) \triangleq \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)} \right\}. \tag{3}$$

Algorithm 1: Asymptotically Optimal Information-Directed Sampling

$s \leftarrow 1$
for $t = 1, 2, 3, \dots$ **do**
  $V_s \leftarrow \sum_{i=1}^{s-1} x_i x_i^\top + \mathbf{1}_d$
  $\hat{\theta}_s \leftarrow V_s^{-1} \sum_{i=1}^{s-1} x_i y_i$ `// least squares estimate`
  $\hat{x}_s \leftarrow \arg\max_{x \in \mathcal{X}} \langle x, \hat{\theta}_s \rangle$ `// empirically best action`
  $\beta_{s,1/\delta} \leftarrow (\sqrt{2 \log \delta^{-1} + \log \det(V_s)} + 1)^2$
  $\hat{\Delta}_s(x) \leftarrow \big(\max_{z \in \mathcal{X}} \langle z, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|z\|_{V_s^{-1}}\big) - \langle x, \hat{\theta}_s \rangle$ `// gap estimates`
  $\hat{\nu}_s(z) \leftarrow \arg\min_{\nu \in \mathcal{H}_z^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ `// see Eq. (10)`
  $m_s \leftarrow \min_{z \neq \hat{x}_s} \frac{1}{2} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2$
  $\eta_s \leftarrow \min_{l \leq s} m_l^{-1/2} \log(k)$
  $q_s(z) \leftarrow \exp(-\eta_s \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2)$
  $I_s(x) \leftarrow \frac{1}{2} \sum_{z \neq \hat{x}_s} q_s(z) \big(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}}\big)^2$ `// information gain†`
  if $m_s \geq \frac{1}{2} \beta_{s,t \log(t)}$ **then**
    Choose $\hat{x}_s$ `// exploitation (disregard data)`
  else
    $\mu_s \leftarrow \arg\min_{\mu \in \mathscr{P}(\mathcal{X})} \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)}$ `// IDS distribution`
    Sample $x_s \sim \mu_s$, observe $y_s = \langle x_s, \theta^* \rangle + \epsilon_s$
    $s \leftarrow s + 1$ `// exploration step counter`

† For the analysis, we normalize the q-weights, but this is not necessary to compute the IDS distribution.

Intuitively, this objective requires to sample actions that have either small regret or large information gain. The *information ratio* Ψ_t is a convex function of the distribution (Russo and Van Roy, 2014, Prop. 6) and can be minimized efficiently as we explain below. In *exploration rounds*, indexed by s, IDS samples the action x_s from the IDS distribution μ_s. Otherwise, in *exploitation rounds*, x^* is identified with high probability, and the algorithm plays the action it believes to be optimal, denoted by $\hat{x}_s$ (where s is the index of the last exploration round). The interaction with the environment, described in Algorithm 1, is over rounds $t = 1, \dots, n$ on a *horizon* n, which is unknown a priori. Exploration rounds are counted separately by s, inducing an implicit mapping $s \mapsto t_s \leq t$. The number of exploration rounds up to time t is s_t. We refer to s and t as *local* and *global time* respectively, and to s_n as the *effective horizon*. The convention is that an s-index refers to the local time quantities, whereas a t-index refers to global time quantities. For example, the action chosen in exploration round s at global time t_s is x_s and the observed reward is y_s. Similarly, an action x_s at local time s has a global time correspondence $x_t = x_{t_s}$.

Gap Estimates All estimated quantities are defined using data collected in exploration rounds, whereas observation data from exploitation rounds is discarded. To justify this choice intuitively, note that with high probability, in exploration rounds the algorithm samples the optimal action x^*, thereby accumulating exponentially more data points on the optimal actions compared to suboptimal actions. Ignoring data from exploitation rounds leads to a much more balanced data set.

Let $\hat{\theta}_s \triangleq V_s^{-1} \sum_{i=1}^{s-1} x_i y_i$ be the regularized least squares estimator with covariance matrix $V_s \triangleq \sum_{i=1}^{s-1} x_i x_i^\top + \mathbf{1}_d$, computed with data $(x_1, y_1), \ldots, (x_{s-1}, y_{s-1})$. The *empirically best action* is $\hat{x}_s \triangleq \arg\max_{x \in \mathcal{X}} \langle x, \hat{\theta}_s \rangle$. We assume that the learner has a *concentration coefficient* $\beta_{s,1/\delta}$ that satisfies

$$\mathbb{P}[\exists\, s \geq 1 \text{ with } \|\hat{\theta}_s - \theta^*\|_{V_s}^2 \geq \beta_{s,1/\delta}] \leq \delta\,. \tag{4}$$

For concreteness, we use the choice derived by Abbasi-Yadkori et al. (2011), which is

$$\beta_{s,1/\delta}^{1/2} \triangleq \sqrt{2 \log \delta^{-1} + \log \det(V_s)} + 1\,. \tag{5}$$

The reader might worry about the log determinant term, which is known to create an asymptotically suboptimal dependence on the dimension, and can be improved with a different choice of the confidence coefficient (Lattimore and Szepesvári, 2017). Since $\beta_{s,1/\delta} = 2 \log \frac{1}{\delta} + \mathcal{O}(d \log(s))$, we circumvent this shortcoming by limiting the amount of data the algorithm collects to $s_n = \mathcal{O}\big(\mathrm{poly}(\log(n))\big)$, which implies $\beta_{s_n,1/\delta} = 2 \log \frac{1}{\delta} + \mathcal{O}(d \log\log(n))$. We also exploit this property for other steps in the analysis, but it is unclear whether or not it is essential.

For all $z \neq \hat{x}_s$, let $\hat{\nu}_s(z) = \arg\min_{\nu \in \mathcal{H}_z^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ be the closest parameter to $\hat{\theta}_s$ in V_s-norm for which z is better than $\hat{x}_s$. This is a strongly convex objective over the convex set $\mathcal{H}_z^{\hat{x}_s}$, hence $\hat{\nu}_s(z)$ can be computed efficiently. In practice, we can drop the constraints on the parameter set (i.e. set $\mathcal{M} = \mathbb{R}^d$), in which case $\hat{\nu}_s(z)$ can be computed in closed form, see (10) below. Exploitation rounds are defined by the *exploitation condition*,

$$m_s \triangleq \frac{1}{2} \min_{x \neq \hat{x}_s} \|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2 \geq \frac{1}{2} \beta_{s,t\log(t)}\,, \tag{E}$$

which guarantees that with confidence level $\beta_{s,t\log(t)}$ there exists no plausible alternative parameter $\nu \neq \hat{\theta}_s$, such that an action $x \neq \hat{x}_s$ is optimal for ν. At local time s, the *gap estimate* is

$$\hat{\Delta}_s(x) \triangleq \max_{z \in \mathcal{X}} \langle z - x, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|z\|_{V_s^{-1}}\,.$$

Note that we use a different confidence level in the definition of the gap estimate, and in fact the only explicit dependence on the global time t is in the exploitation condition. The gap estimate is an upper bound on the true gap, provided $\hat{\theta}_s$ is well concentrated, i.e. $\|\hat{\theta}_s - \theta^*\|_{V_s}^2 \leq \beta_{s,s^2}$,

$$\Delta(x) \leq \max_{y \in \mathcal{X}} \langle y, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|y\|_{V_s^{-1}} - (\langle x, \hat{\theta}_s \rangle - \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}}) \leq 2\hat{\Delta}_s(x)\,. \tag{6}$$

The first inequality follows from the definition of the confidence scores, and the second inequality uses $\hat{\Delta}_s(x) \geq \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}}$. The gap estimate of the empirically best action $\hat{x}_s$ is $\delta_s \triangleq \hat{\Delta}_s(\hat{x}_s)$. Importantly, the gap estimate can be written as $\hat{\Delta}_s(x) = \langle \hat{x}_s - x, \hat{\theta}_s \rangle + \delta_s$, and therefore we also refer to δ_s as the *estimation error*. The UCB action is $x_s^{\mathrm{UCB}} \triangleq \arg\max_{x \in \mathcal{X}} \langle x, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}}$.

Information Gain Recall that $\hat{\nu}_s(z) = \arg\min_{\nu \in \mathcal{H}_z^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ is the closest alternative parameter to $\hat{\theta}_s$ in V_s-norm for which $\hat{x}_s$ is not optimal. The *information gain* is set to

$$I_s(x) \triangleq \frac{1}{2} \sum_{z \neq \hat{x}_s} q_s(z) \left(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}} \right)^2\,, \tag{7}$$

where the mixing distribution $q_s \in \mathcal{P}(\mathcal{X})$ is defined so that

$$q_s(z) \propto \begin{cases} 0 & \text{if } z = \hat{x}_s \\ \exp\left(-\frac{\eta_s}{2}\|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2\right) & \text{otherwise}\,. \end{cases} \tag{8}$$

The learning rate is $\eta_s \triangleq \min_{l \le s} m_l^{-1/2} \log(k)$, where $m_s \triangleq \frac{1}{2} \min_{z \ne \hat{x}_s} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2$. The weights q_s can be interpreted as a soft-min approximation of the minimum constraint value where the learning rate controls the lower order term (Lemma 22),

$$m_s \le \frac{1}{2} \sum_{z \ne \hat{x}_s} q_s(z) \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2 \le m_s + \frac{\log(k)}{\eta_s}\,. \tag{9}$$

Computational Complexity There are three kinds of operations in the algorithm. First, using elementary matrix operations, we can update V_s^{-1}, $\det(V_s)$ and $\hat{\theta}_s$ incrementally, and note that the s-index terms only need to be updated after exploration rounds. It can be checked that $\mathcal{O}(kd^2 s_n)$ operations are needed over all n rounds to compute this part. Second, the IDS distribution (3) is defined as a minimizer of the convex objective $\Psi_s(\mu)$ and always admits a solution supported on two actions (Russo and Van Roy, 2014), see Lemma 7. Hence, we can obtain the IDS distribution by computing the optimal trade-off between all $\mathcal{O}(k^2)$ pairs of actions (Lemma 8). A closer inspection of the regret bounds reveals that it always suffices to optimize the trade-off between the greedy action $\hat{x}_s$ and some other (informative) action, which reduces the computational complexity to $\mathcal{O}(k)$. Third, the optimization problem that defines the alternative parameters $\hat{\nu}_s(z)$ is a quadratic program with d variables and linear constraints $\langle \hat{\nu}_s(z), z - \hat{x}_s \rangle \ge 0$ and $\hat{\nu}_s(z) \in \mathcal{M}$. Such optimization problems can be solved efficiently in practice and in $O(ld^3)$ time in the worst case for model sets $\mathcal{M}$ with l constraints. Note, the analysis suggests that we can tolerate an additive numerical error on the information gain of order $\mathcal{O}(s^{-2})$. In practice, we can drop the constraints on $\mathcal{M}$, in which case

$$\hat{\nu}_s(z) = \hat{\theta}_s - \frac{\langle \hat{\theta}_s, \hat{x}_s - z\rangle}{\|\hat{x}_s - z\|_{V_s^{-1}}^2} V_s^{-1}(\hat{x}_s - z)\,. \tag{10}$$

With these improvements, the overall computation complexity is $\mathcal{O}(n + kd^2 s_n)$ over n rounds, where the linear term comes from checking whether to explore or exploit. This can be improved, by simply computing after each exploration round when the next exploration round will occur.

2.1. Regret Bounds

The regret bounds for Algorithm 1 come in three flavours. In Theorem 2, we show a (nearly) optimal worst-case regret bound of $R_n \le \mathcal{O}(d\sqrt{n}\log(n))$. Second, using a gap-dependent bound on the information ratio, in Theorem 3 we show a gap-dependent regret bound of $R_n \le \mathcal{O}\big(d^3 \Delta_{\min}^{-1} \log(n)^2\big)$. Besides universal constants, the $\mathcal{O}$-notation in the bound only depends on the norm of action features and the parameter. Last, in Theorem 5 we show that the proposed algorithm is asymptotically optimal, that is $R_n \le c^* \log(n) + o(\log(n))$. In contrast to the previous bound, here the lower order terms depend exponentially on problem-dependent quantities such as $\Delta_{\min}^{-1}$.

Theorem 2 (Worst-case regret) *The regret of Algorithm 1 is bounded by*

$$R_n \le \mathcal{O}\big(d\sqrt{n}\log(n)\big)\,.$$

The result matches the best known bound for LinUCB and is optimal up to the logarithmic factor when k is (exponentially) large. On the other hand, when k is small, our bound is worse than basic elimination algorithms that achieve $R_n \leq \mathcal{O}(\sqrt{\log(k)dn})$ (Lattimore and Szepesvari, 2019, §23).

Proof Define $\beta_s = \|\hat{\theta}_s - \theta^*\|_{V_s}^2$ and let $B_s = \mathbb{1}(\beta_s \leq \beta_{s,s^2})$. By Lemma 20 and (6), we have

$$\mathbb{E}[R_n] \leq 2\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(x_s) B_s\right] + \mathcal{O}(\log\log(n))\,,$$

where the $\mathcal{O}$-notation only hides a bound on the largest gap, $\hat{\Delta}(x_s) \leq 1$. Similar to the standard IDS analysis (Russo and Van Roy, 2014), we bound the expected regret,

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(x_s) B_s\right] = \mathbb{E}\left[\sum_{s=1}^{s_n} \sqrt{\Psi_s(\mu_s) I_s(\mu_s) B_s}\right] \leq \sqrt{\mathbb{E}\left[\sum_{s=1}^{s_n} \Psi_s(\mu_s) B_s\right] \mathbb{E}\left[\sum_{s=1}^{s_n} I_s(x_s) B_s\right]}\,,$$

where the equality follows from the tower rule $\mathbb{E}[\hat{\Delta}_s(x_s)B_s] = \mathbb{E}[\mathbb{E}_s[\hat{\Delta}_s(x_s)B_s]] = \mathbb{E}[\hat{\Delta}_s(\mu_s)B_s]$ and the definition of the information ratio. The second inequality follows from Cauchy-Schwarz and another application of the tower rule. To complete the proof, we show that $\Psi_s(\mu_s) \leq 2$ and bound the total information gain, $\gamma_n = \sum_{s=1}^{s_n} I_s(x_s) \leq \mathcal{O}(d^2 \log(n)^2)$. Since μ_s is chosen by IDS to minimize Ψ_s,

$$\Psi_s(\mu_s) = \min_{\mu \in \mathscr{P}(\mathcal{X})} \Psi_s(\mu) \leq \frac{\hat{\Delta}_s(x_s^{\mathrm{UCB}})^2}{I_s(x_s^{\mathrm{UCB}})} \leq 2\,. \tag{11}$$

The last inequality follows from the fact that $\hat{\Delta}_s(x_s^{\mathrm{UCB}}) = \beta_{s,s^2}^{1/2}\|x_s^{\mathrm{UCB}}\|_{V_s^{-1}}$ and bounding

$$I_s(x_s^{\mathrm{UCB}}) = \frac{1}{2}\sum_{z \neq \hat{x}_s} q_s(z)\big(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x_s^{\mathrm{UCB}}\rangle| + \beta_{s,s^2}^{1/2}\|x_s^{\mathrm{UCB}}\|_{V_s^{-1}}\big)^2 \geq \tfrac{1}{2}\beta_{s,s^2}\|x_s^{\mathrm{UCB}}\|_{V_s^{-1}}^2\,,$$

where we used the definition of q_s as a distribution supported on $\mathcal{X} \setminus \{\hat{x}_s\}$. Finally, Lemma 12 provides a worst-case bound on the total information gain, $\gamma_n \leq \mathcal{O}\big(d^2 \log(n)^2\big)$, which is a direct consequence of the elliptic potential bound (Lemma 18) and the soft-min inequality (9). We conclude $R_n \leq \mathcal{O}\big(d\sqrt{n}\log(n)\big)$. ■

Our next result is an instance-dependent logarithmic regret bound. The proof follows along the same lines as the worst-case regret bound, but replaces the worst-case bound on the information ratio with an instance-dependent bound. Interestingly, our bound is attained by a distribution with a close resemblance with Thompson sampling.

Theorem 3 (Gap-dependent regret) *The regret of Algorithm 1 is bounded by*

$$R_n \leq \mathcal{O}\left(\Delta_{\min}^{-1} d^3 \log(n)^2\right)\,.$$

Besides universal constants, the $\mathcal{O}$-notation in the theorem statement hides only the constants required for boundedness of $\mathcal{X}$ and $\mathcal{M}$. The proof makes use of the following lemma, which shows an instance-dependent bound on the information ratio. Recall that $\delta_s = \hat{\Delta}_s(\hat{x}_s)$ is the gap estimate of the empirically best action, and $\hat{\Delta}_s(x) = \delta_s + \langle \hat{x}_s - x, \hat{\theta}_s\rangle$.

Lemma 4 *At any local time s with $\beta_{s,s^2} \geq \beta_s \triangleq \|\hat{\theta}_s - \theta^*\|^2_{V_s}$, the optimal information ratio is bounded as follows,*

$$\min_{\mu \in \mathscr{P}(\mathcal{X})} \Psi(\mu) \leq \frac{4\delta_s(8d+9)}{\Delta_{\min}}\,.$$

Proof Let $a \geq 2$ be a constant to be chosen later. If $2a\delta_s \geq \Delta_{\min}$, then $\min_{\mu\in\mathscr{P}(\mathcal{X})} \Psi_s(\mu) \leq \frac{4a\delta_s}{\Delta_{\min}}$ by (11). Hence we may assume $2a\delta_s \leq \Delta_{\min}$ in the following. By (6), for all s with $\beta_s \leq \beta_{s,s^2}$ and $x \neq x^*$, it holds that $\Delta_{\min} \leq 2\hat{\Delta}_s(x)$, so in particular $\hat{x}_s = x^*$. Define $\tilde{\mu} = \frac{1}{2}e_{\hat{x}_s} + \frac{1}{2}q_s$ to be the uniform mixture[1] of q_s and a Dirac at $\hat{x}_s$. Let $\bar{\Delta}_s(x) = \langle \hat{\theta}_s, \hat{x}_s - x\rangle$ and note that $\bar{\Delta}(\tilde{\mu}) \geq (a-1)\delta_s \geq \delta_s$ by the assumption $a \geq 2$. Therefore, by Lemma 8,

$$\Psi_s(\mu_s) \leq \min_{p\in[0,1]} \frac{(1-p)\delta_s + p\hat{\Delta}_s(\tilde{\mu})}{pI_s(\tilde{\mu})} \leq \frac{4\delta\bar{\Delta}_s(\tilde{\mu})}{I_s(\tilde{\mu})}\,. \tag{12}$$

Note that we can bound the information gain $I_s(\tilde{\mu})$ as follows,

$$I_s(\tilde{\mu}) \geq \frac{1}{2}\sum_{x\in\mathcal{X}} \tilde{\mu}(x) \sum_{z\neq\hat{x}_s} q_s(z)\langle \hat{\nu}_s(z) - \hat{\theta}_s, x\rangle^2 = \tfrac{1}{2}\sum_{z\neq\hat{x}_s} q_s(z) \min_{\nu\in\mathcal{H}_z^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|^2_{V(\tilde{\mu})}\,.$$

On the other hand, we can bound the gap $\bar{\Delta}_s(x) = \langle\hat{\theta}_s, \hat{x}_s - x\rangle$,

$$\langle\hat{\theta}_s, \hat{x}_s - x\rangle = \min_{\nu:\langle\nu, x-\hat{x}_s\rangle\geq 0} \|\nu - \hat{\theta}_s\|_{V(\tilde{\mu})}\|\hat{x}_s - x\|_{V(\tilde{\mu})^{-1}} \leq \min_{\nu\in\mathcal{H}_x^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|_{V(\tilde{\mu})}\|\hat{x}_s - x\|_{V(\tilde{\mu})^{-1}}\,.$$

Combining the last two displays with the definition of $\tilde{\mu}$, the fact that $\hat{x}_s = x^*$ and Cauchy-Schwarz,

$$\begin{aligned}\bar{\Delta}_s(\tilde{\mu})^2 &\leq \frac{1}{4}\sum_{x\neq\hat{x}} q_s(x) \min_{\nu\in\mathcal{C}_x} \|\nu - \hat{\theta}_s\|^2_{V(\tilde{\mu})} \sum_{x\neq\hat{x}} q_s(x)\|\hat{x}_s - x\|^2_{V(\tilde{\mu})^{-1}} \\ &\leq (1+d)\sum_{x\neq\hat{x}} q_s(x) \min_{\nu\in\mathcal{H}_x^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|^2_{V(\tilde{\mu})} \leq 2(1+d)I_s(\tilde{\mu})\,.\end{aligned}$$

The second last step uses $\sum_{x\neq\hat{x}_s} q_s(x)\|x\|^2_{V(\tilde{\mu})^{-1}} \leq 2\sum_{x\neq\hat{x}_s} q_s(x)\|x\|^2_{V(q_s)^{-1}} = 2d$ and $\|\hat{x}_s\|^2_{V(\tilde{\mu})^{-1}} \leq 2$. Next, for $x \neq \hat{x}_s$,

$$\bar{\Delta}_s(x) = \hat{\Delta}_s(x) - \delta_s \geq \frac{1}{2}\Delta_{\min} - \delta_s \geq \frac{1}{2}\left(1 - \frac{1}{a}\right)\Delta_{\min}\,.$$

Hence, by the definition of $\tilde{\mu}$ we have $\bar{\Delta}_s(\tilde{\mu}) \geq \frac{1}{4}(1-1/a)\Delta_{\min}$ and using (12),

$$\Psi_s(\mu_s) \leq \frac{4\delta_s\bar{\Delta}_s(\tilde{\mu})}{I_s(\tilde{\mu})} = \frac{4\delta_s\bar{\Delta}_s(\tilde{\mu})^2}{\bar{\Delta}_s(\tilde{\mu})I_s(\tilde{\mu})} \leq \frac{32\delta_s(1+d)}{\Delta_{\min}\left(1-\frac{1}{a}\right)}\,.$$

The claim follows with $a = 8(1+d)+1$. ■

[1] By a concentration of measure argument (Appendix D), the weights $q_s(x)$ approximate the posterior probability of an action x being preferred over $\hat{x}_s$ by the Bayesian model with Gaussian prior and likelihood. As such, the distribution $\tilde{\mu}$ resembles the *top-two Thompson sampling* approach proposed by Russo (2020).

Proof of Theorem 3 Recall that $B_s = \mathbb{1}(\beta_s \leq \beta_{s,s^2})$ with $\beta_s = \|\hat{\theta}_s - \theta^*\|^2_{V_s}$. As before, by Lemma 20 and using that $\Delta(x_s)B_s \leq 2\hat{\Delta}_s(x_s)B_s$,

$$R_n \leq 2\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(x_s)B_s\right] + \mathcal{O}(\log\log(n))\,.$$

Let $\gamma_n = \sum_{s=1}^{s_n} I_s(x_s)$ be the cumulative information gain. Using Cauchy-Schwarz and the instance-dependent bound on the information ratio from Lemma 4,

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(x_s)B_s\right]^2 \leq \mathbb{E}\left[\sum_{s=1}^{s_n} \Psi_s(\mu_s)B_s\right]\mathbb{E}\left[\sum_{s=1}^{s_n} I_s(x_s)B_s\right] \leq \mathbb{E}\left[\sum_{s=1}^{s_n} \mathcal{O}\left(\frac{\delta_s d B_s}{\Delta_{\min}}\right)\right]\mathbb{E}\left[\gamma_n\right]\,.$$

Further bounding $\delta_s \leq \hat{\Delta}_s(x_s)$ on the right-hand side and re-arranging yields

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(x_s)B_s\right] \leq \mathcal{O}\left(d\Delta_{\min}^{-1}\right)\mathbb{E}\left[\gamma_n\right]\,.$$

The worst-case total information gain is at most $\mathbb{E}[\gamma_n] \leq \mathcal{O}\big(d^2\log(n)^2\big)$ according to Lemma 12, and the claim follows. We remark that the bound can be improved by bounding the term $\mathbb{E}[\log(s_n)]$ (which appears in the upper bound on γ_n) more carefully with the help of Lemma 21. ■

Our next result shows that the proposed version of IDS is asymptotically optimal. The key insight is a connection between information-directed sampling and a primal-dual approach based on online learning to solve the convex program that defines the lower bound. Conceptually, the connection is explained best with an *oracle analysis*, which sets aside the statistical estimation process and highlights the key steps (Appendix C). In particular, Lemma 11 shows that in the asymptotic regime, the information ratio satisfies $\Psi_s(\mu_s) \leq 4\delta_s\big(c^* + \mathcal{O}(\beta_s^{1/2} m_s^{-1/2} + \delta_s)\big)$. Further, Lemma 14 improves the bound on the total information gain to $\gamma_n = \sum_{s=1}^{s_n} I_s(x_s) \leq \log(n) + o(\log(n))$. Lastly, Lemma 21 shows that IDS samples informative actions with large enough probability that $\mathbb{E}[\log(s_n)] \leq \mathcal{O}(\log\log(n))$, which is important to bound lower-order terms in our analysis.

Theorem 5 (Asymptotic regret) *Algorithm 1 is asymptotically optimal,*

$$\lim_{n\to\infty} \frac{R_n}{\log(n)} = c^*\,,$$

where c^* is the solution to the lower bound (2) and we assume that $\|x^*\| > 0$.

We sketch the proof below and defer the complete proof to Appendix B.4. The assumption $\|x^*\| > 0$ is used in Lemma 21 to show that there are not too many exploration steps, which follows from lower bounding the exploration probability. On the other hand, when $\|x^*\| = 0$, the geometry of the lower bound changes, because the optimal action provides no information. Whether the assumption is necessary for Algorithm 1 remains to be determined. As a remedy, we can also replace the gap estimates with thresholded gaps $\hat{\Delta}_s^+(x) = \langle \hat{x}_s - x, \hat{\theta}_s\rangle + \delta_s^+$, where $\delta_s^+ = \max(\delta_s, 1/\sqrt{s})$. Lower bounding the gaps this way ensures that an exploratory action is sampled with probability at least $1/\sqrt{s}$ in each exploration round. We believe that with a thresholded gap estimate, the statement of

Theorem 5 holds without restrictions and Theorems 2 and 3 remain valid. Since it is unclear if the assumptions is required and for simplicity of the proofs, we work with the assumption $\|x^*\| > 0$.

Proof *(Sketch)* The first step is to improve the bound on the information ratio in the asymptotic regime. Recall that $\beta_s = \|\hat{\theta}_s - \theta^*\|_{V_s}^2$ and $m_s = \frac{1}{2}\min_{z\neq x^*} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2$. Then by Lemma 11,

$$\Psi_s(\mu_s) \leq 4\delta_s\big(c^* + \mathcal{O}(\beta_s^{1/2} m_s^{-1/2} + \delta_s)\big)\,,$$

for $\beta_s^{1/2} m_s^{-1/2} \to 0$ and $\delta_s \to 0$. Not surprisingly, the proof bounds the information ratio using a sampling distribution informed from the lower bound (2). Details are given in Appendix B.2.

Second, we improve the bound on the total information gain $\gamma_n = \sum_{s=1}^{s_n} I_s(x_s)$. The key insight is to interpret the information gain as the loss of an online learning algorithm. We adapt the standard regret proof for the exponential weights algorithm Orabona (2019), to bound the total information gain relative to the minimum constraint (Lemma 14). Informally, the result states that

$$\mathbb{E}[\gamma_n] \leq \mathbb{E}\left[\min_{x\neq \hat{x}_{s_n}} \|\hat{\nu}_{s_n}(x) - \hat{\theta}_{s_n}\|_{V_{s_n}}^2 + \mathcal{O}(\log(n)^{1/2}\log(s_n))\right]\,.$$

Exploration rounds are defined by condition (E) to ensure that the minimum remains small,

$$\min_{z\neq \hat{x}_{s_n}} \|\hat{\nu}_{s_n}(z) - \hat{\theta}_{s_n}\|_{V_{s_n}}^2 \leq \beta_{s_n, n\log n} \leq 2\log(n\log(n)) + \mathcal{O}(d\log(s_n))\,.$$

This result improves upon the worst-case bound on the information gain (Lemma 12), as long as the number of exploration rounds s_n is not too large.

Third, the proof hinges on Lemma 21, which shows that $\mathbb{E}[\log(s_n)] \leq \mathcal{O}(\log\log(n))$. Intuitively, IDS samples informative actions with large enough probability to ensure that in expectation, there is only a logarithmic number of exploration rounds, while the exploration probability is small enough to bound the worst-case regret.

In the remaining proof sketch we only discuss the case where $\Psi_s(\mu_s) \leq 4\delta_s(c^* + o(1))$ and $\mathbb{E}[\gamma_n] \leq \log(n) + o(\log(n))$ holds (the actual proof requires to also bound the regret in early rounds, when the asymptotic statements do not hold). Asymptotically, the mean gap estimate $\bar{\Delta}_s(x) \triangleq \langle x_s - x, \hat{\theta}_s\rangle$ is a good estimate of the actual regret. Therefore, we get

$$R_n \leq \mathbb{E}\left[\sum_{t=1}^{s_n} \bar{\Delta}_s(x_s)\right] + o(\log(n))$$

Next using that $4ab \leq (a+b)^2$ and Cauchy-Schwarz combined with a few applications of the tower rule, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{s=1}^{s_n} \bar{\Delta}_s(x_s)\right] = \mathbb{E}\left[\sum_{s=1}^{s_n} \bar{\Delta}_s(\mu_s)\right] &\leq \frac{1}{4}\mathbb{E}\left[\sum_{s=1}^{s_n}\delta_s\right]^{-1}\mathbb{E}\left[\sum_{s=1}^{s_n}\hat{\Delta}_s(\mu_s)\right]^2 \\ &\leq \frac{1}{4}\mathbb{E}\left[\sum_{s=1}^{s_n}\delta_s\right]^{-1}\mathbb{E}\left[\sum_{s=1}^{s_n}\Psi_s(\mu_s)\right]\mathbb{E}\left[\sum_{s=1}^{s_n} I_s(x_s)\right].\end{aligned}$$

The bound on the information ratio yields

$$\frac{1}{4}\mathbb{E}\left[\sum_{s=1}^{s_n}\delta_s\right]^{-1}\mathbb{E}\left[\sum_{s=1}^{s_n}\Psi_s(\mu_s)\right] \leq \frac{1}{4}\mathbb{E}\left[\sum_{s=1}^{s_n}\delta_s\right]^{-1}\mathbb{E}\left[\sum_{s=1}^{s_n}4\delta_s(c^*+o(1))\right] \leq c^* + o(1)\,.$$

Combined with the bound on the information gain, asymptotic optimality follows. ∎

2.2. Alternative Definitions of the Information Gain

Our definition of the information gain ensures that $I_s(x) \approx \frac{1}{2}\sum_{z\neq\hat{x}_s} q_s(z)\langle\hat{\nu}_s(z)-\hat{\theta}_s, x\rangle^2$ asymptotically. In finite time, however, the mean estimates can be inaccurate. Therefore, we add an optimistic term in the definition of the information gain (7), which is an essential ingredient in the proof of Theorem 2. At the same time, the optimistic term corresponds to an information gain which was analyzed in earlier work (Kirschner and Krause, 2018; Kirschner et al., 2020). Since this choice is motivated from a worst-case perspective, empirically it sometimes leads to over-exploration in the finite-time regime. A closer inspection of the worst-case regret proof (in particular, Eq. 11) reveals that the optimistic term is only needed for the UCB action. This motivates the following definition:

$$I_s^{\mathcal{H}\text{-UCB}}(x) = \frac{1}{2}\sum_{z\neq\hat{x}_s} q_s(z)\left(|\langle\hat{\nu}_s(z)-\hat{\theta}_s, x\rangle| + \mathbb{1}(x = x_s^{\text{UCB}})\beta_{s,s^2}^{1/2}\|x\|_{V_s^{-1}}\right)^2 . \tag{13}$$

With a few additional steps in the proof of Lemma 11 and Theorem 5, the resulting algorithm is shown to satisfy the same regret bounds as presented in Theorems 2, 3 and 5. Since the proofs are very similar, we omit the details. We compare both information gain functions in our experiments. Another variant is to set the alternative parameters to

$$\tilde{\nu}_s(x) = \arg\min_{\nu\in\mathcal{C}_x} \|\nu - \hat{\theta}_s\|_{V_s}^2 , \quad \text{where} \quad \mathcal{C}_x = \{\nu \in \mathcal{M} : \max_{z\in\mathcal{X}}\langle\nu, z - x\rangle = 0\} .$$

Note that $\mathcal{C}_x$ is the set of parameters where x is optimal and is sometimes called the *cell* of x. Let $\tilde{q}(z) \propto \exp(-\eta\|\tilde{\nu}_s(z)-\hat{\theta}_s\|_{V_s}^2)$ and define

$$I_s^{\mathcal{C}}(x) \triangleq \frac{1}{2}\sum_{z\neq\hat{x}_s}\tilde{q}_s(z)\left(|\langle\tilde{\nu}_s(z)-\hat{\theta}_s, x\rangle| + \beta_{s,s^2}^{1/2}\|x\|_{V_s^{-1}}\right)^2 . \tag{14}$$

Note that all bounds that we obtain hold true for IDS defined with $I_s^{\mathcal{C}}$ as well, by replacing $\mathcal{H}_x^{\hat{x}_s}$ with $\mathcal{C}_x$ in the proof. The key insight is that $\mathcal{C}^* = \cup_{x\neq x^*}\mathcal{C}_x = \cup_{x\neq x^*}\mathcal{H}_x^{x^*}$, hence the change is simply a different decomposition of the set of alternative parameters $\mathcal{C}^*$ into convex regions. One might expect faster convergence from the fact that $\tilde{q}_s$ is more concentrated, but empirically we find little difference compared to I_s. On the other hand, for unconstrained parameter sets $\mathcal{M}$, we can compute $\hat{\nu}_s(z)$ in closed form (Eq. 10), whereas $\tilde{\nu}_s(z)$ can only be computed by solving a positive definite quadratic program with k linear constraints for each action $z \neq \hat{x}_s$. Interestingly, however, the information gain (14) relates to the Bayesian mutual information $\mathbb{I}_s(y_s; x^*|x_s = x)$. The argument uses concentration of measure to show that $\tilde{q}_s(x)$ approximates the posterior probability that an action $x \neq x^*$ is optimal in the Bayesian model. We refer to Appendix D for details.

2.3. Information-Directed Sampling as a Primal-Dual Approach

Lemma 9 shows that the IDS distribution μ_s is supported on actions x that minimize the function

$$g_s(x) = \hat{\Delta}_s(x) - \frac{\Psi_s(\mu_s)}{2\hat{\Delta}_s(\mu_s)} I_s(x) \overset{n\to\infty}{\approx} \hat{\Delta}_s(x) - c^* I_s(x) .$$

The approximation holds because asymptotically, $\Psi_s(\mu_s) \approx 4c^*\delta_s$ and $\hat{\Delta}_s(\mu_s) \approx 2\delta_s$. The weight c^* appears from normalizing the Lagrange multipliers as discussed in Appendix C. Therefore, the

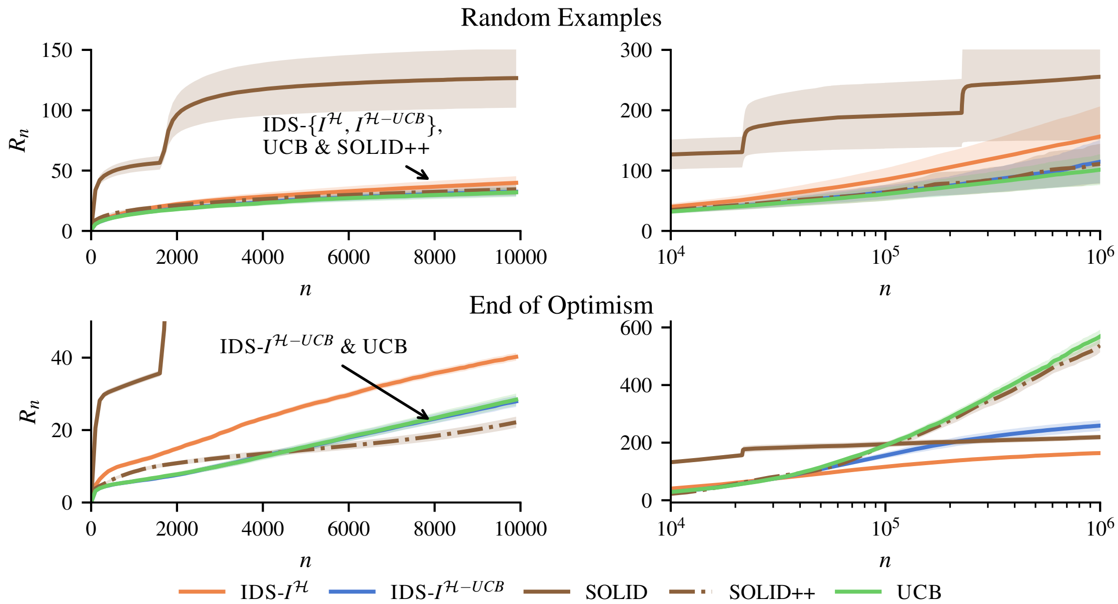

Figure 1: Top: Worst-case regret on randomly drawn action sets. Bottom: Counter-example problem from Lattimore and Szepesvári (2017). Early stages are shown in linear scale, asymptotics in log scale. Results are averaged over 100 repetitions and the confidence region shows 2×standard error.

IDS distribution can be understood as a type of best-response on the primal-dual game defined by the Lagrangian of the lower bound, where the dual variables correspond to the q-weights of the information gain. Note that the best response on g_s is not unique, and IDS chooses a particular, randomized trade-off, which is imposed by the IDS objective (3).

The first work which exploits the primal-dual formulation for regret minimization is by Degenne et al. (2020). In our notation, their algorithm corresponds to choosing the action with the best information-regret trade-off $z_s = \arg\min_{x \in \mathcal{X}} \hat{\Delta}_s(x)/I_s(x)$. IDS instead asymptotically randomizes between x^* and z_s, which allows it to maintain the worst-case regret bound. Another more recent primal-dual approach is the SOLID algorithm by Tirinzoni et al. (2020). This approach uses a different Lagrangian, which is defined by keeping the minimum over $\mathcal{C}^*$ in (2). Accordingly, the dual variable is one-dimensional, but the constraints appear non-smooth. SOLID is defined by alternating (optimistic) sub-gradient steps on the allocation and the dual variable. This leads to a randomized strategy over actions with exponential weights that are only updated when an exploration condition is satisfied.

3. Experiments

We compare IDS with LINUCB (Abbasi-Yadkori et al., 2011) and SOLID (Tirinzoni et al., 2020), the latter being our closest competitor. Note that SOLID was shown to outperform OAM (Hao et al., 2019) and LINTS in a variety of settings. To the best of our knowledge, SOLID is the current state-of-the-art for asymptotically optimal algorithms.

To enable a fair comparison, we use the same confidence coefficient $\beta_{t,1/\delta}$ (4) for all algorithms. We also run the same experiment with the (tighter) confidence coefficient derived by Tirinzoni et al. (2020), but we found no significant difference in the results, see Appendix E. For SOLID, we use the default hyper-parameters suggested by Tirinzoni et al. (2020, Appendix K). Finally, as recommended by the authors, we implement a variant of the SOLID algorithm, which is (heuristically) optimized for better performance in finite time and does not reset the sampling vector ω_t at the beginning of each phase. We display that improved version as SOLID++.

IDS is implemented as in Algorithm 1 with the computational improvements described at the end of Section 2. In particular, we use an unconstrained parameter set ($\mathcal{M} = \mathbb{R}^d$), which allows us to compute the parameter $\hat{\nu}_s(x)$ in closed form. We further compute the IDS distribution randomizing only between $\hat{x}_t$ and one other action (Lemma 9) to reduce the per-round computational complexity from $\mathcal{O}(k^2)$ to $\mathcal{O}(k)$. All variants of IDS used in the experiments satisfy the theoretical guarantees presented in this paper with minor proof modifications. We also compare to IDS-$I^{\mathcal{H}\text{-UCB}}$ defined with information gain (13). In Appendix E, we present further empirical evidence, including a benchmark with Thompson Sampling and Bayesian IDS, a comparison of information gain functions, and an evaluation of the tuning sensitivity of the β_s and η_s parameters.

Average performance on random problems. For each repetition, we sample an action set with 6 actions drawn uniformly from the unit sphere. We set $d = 2$ and the variance of the noise to $\sigma^2 = 0.1$, which is chosen so that the asymptotic regime is observed after fewer rounds relative to $\sigma^2 = 1$. The results are shown in the first row of Figure 1. We display the average over 100 runs and 95% confidence intervals. All policies except for SOLID have comparable averaged performances, but the latter is not designed to optimize for worst-case regret in principle. IDS-$I^{\mathcal{H}\text{-UCB}}$ is similar to LINUCB, followed by IDS-$I^{\mathcal{H}}$.

The End of Optimism? This example of a 2-dimensional linear bandit dates back to Soare et al. (2014, Appendix A), and was used by Lattimore and Szepesvári (2017) to show that algorithms based on optimism and Thompson sampling are not asymptotically optimal in the linear setting. There are three arms $x_1 = (1, 0)$, $x_2 = (1 - \epsilon, 2\epsilon)$ and $x_3 = (0, 1)$ with a tuning variable $\epsilon > 0$. The true parameter is $\theta = (1, 0)$ which makes action x_1 optimal. The situation is illustrated in Figure 2. The colored regions $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ are the corresponding *cells*, i.e. the subset of parameters in $\mathbb{R}^2$ for which x_1, x_2 or x_3 is optimal respectively. When the confidence ellipsoid $\mathcal{E}_t = \{\theta : \|\theta - \hat{\theta}_t\|^2_{V_t} \leq c\log(n)\}$ for the least squares estimator $\hat{\theta}_t$ is contained in the cell $\mathcal{C}_1$, the learner has identified the best action with high probability.

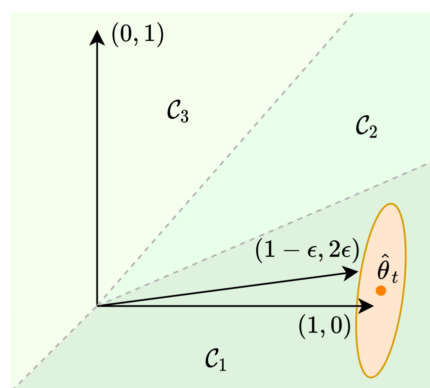

Figure 2: The 'end of optimism' example.

Algorithms based on optimism and Thompson sampling quickly rule out the suboptimal arm x_3 and just play either x_1 or x_2. The twist is that the third arm is still informative for determining a^*, and in fact an asymptotically optimal algorithm plays only on $\{x_1, x_3\}$. To see why, note that any no-regret learner plays $x^* = x_1$ a lot, therefore the parameter is well-estimated along the direction x_1. It remains to shrink the confidence ellipsoid approximately along the direction x_3, which means increasing the V_t-norm of x_3. Choosing arm x_2 incurs a small cost ϵ, but the increase of the confidence ellipsoid in direction x_3 is only small, $\langle x_3, (V_{t+1} - V_t)x_3 \rangle = \langle x_3, x_2 \rangle^2 = \epsilon^2$. On the other hand, choosing x_3 implies a higher regret cost of 1, but the confidence set is increased by 1 along direction x_3, which allows to identify the optimal action at a much smaller cost. An optimistic algorithm has an asymptotic regret that scales with $R_n \approx \log(n)/\epsilon$, while for an optimal algorithm, $R_n \approx 1 \cdot \log(n)$. In fact, for some small ϵ, the lower bound constant (2) is $c^* = 64$ and does not depend on ϵ, so optimistic algorithms cannot be asymptotically optimal.

For the experiments, we use noise variance $\sigma^2 = 0.1$, and $\epsilon = 0.01$, which is sufficiently large to reach the asymptotic regime within $n = 10^6$ rounds, and small enough to highlight the difference between UCB and IDS. Results in this setting are shown in the bottom row of Figure 1. As expected, LINUCB's asymptotics show a suboptimal log-slope, but it is surprisingly followed by SOLID++. Despite our attempts, we are presently not able to provide a good explanation for this result and it might require a more involved analysis of the SOLID++ heuristic. However, both versions of IDS and the theoretical SOLID reach the optimal asymptotic around $t = 10^5$ (10^4 for SOLID) and significantly outperform LINUCB on that problem. An interesting observation is that IDS-I_s^{UCB} performs better in finite time, whereas IDS-I_s reaches the asymptotic regime earlier.

4. Conclusion

We introduced a simple and efficient algorithm for linear bandits that is (nearly) worst-case optimal and matches the asymptotic lower bound exactly. Note that the algorithm is essentially hyper-parameter free with the usual boundedness assumptions. Nonetheless, the confidence parameter $\beta_{s,1/\delta}$ and the learning rate η_s used in the definition of I_s provide some tuning knobs to improve performance in practice.

Our theoretical results still rely on some restrictive assumptions, such as the boundedness requirement for the parameter set, uniqueness of x^* and $\|x^*\| > 0$ for the asymptotic regret, and the need to discard data in exploitation rounds. Also, the dependence on d and k is sub-optimal in some regimes, in particular for the worst-case regret bound and small k. On the upside, our analysis is relatively simple, and raises the hope that there exists a *really* simple proof. Finding an information gain which preserves the guarantees and telescopes more easily could be a first step towards this end.

Finally, it appears likely that our framework generalizes in several directions. The contextual case is already covered in previous work on asymptotic algorithms (Hao et al., 2019; Tirinzoni et al., 2020). We point out that IDS can be defined to optimize the marginals of the joint distribution between context and action (Kirschner et al., 2020). Decoupling the reward from the observation features leads to the linear partial monitoring framework, where IDS is known to achieve the optimal worst-case rate in all possible games (Kirschner et al., 2020). The structured bandit setting and information gain functions for a non-Gaussian likelihood are yet other promising directions.

Acknowledgements

Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

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

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

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, March 2003. ISSN 1532-4435.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved Second-Order Bounds for Prediction with Expert Advice. In *Learning Theory*, pages 217–232, Berlin, Heidelberg, 2005.

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

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, pages 355–366. Omnipress, 2008.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*, pages 14492–14501, 2019.

Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. 07 2020.

Naveen Garg and Jochen Koenemann. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. *SIAM Journal on Computing*, 37(2):630–652, 2007.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. *arXiv preprint arXiv:1910.06996*, 2019.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. 06 2020.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. *arXiv preprint arXiv:1801.09667*, 2018.

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. PMLR, 09–12 Jul 2020.

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

Tor Lattimore and Czsaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Philippe Rigollet. 18. s997: High dimensional statistics. 2015.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 2020.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.

Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications, 2007.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pages 828–836, 2014.

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Bart PG Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *arXiv preprint arXiv:2007.07302*, 2020.

Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Andrew Wagenmaker, Julian Katz-Samuels, and Kevin Jamieson. Experimental design for regret minimization in linear bandits. *arXiv preprint arXiv:2011.00576*, 2020.

Appendix A. Notation

Linear Bandit Setting

d		feature dimension
$\mathcal{X}$	$\subset$	$\mathbb{R}^d$, action (feature) set
$\mathcal{M}$	$\subset$	$\mathbb{R}^d$, parameter set
θ^*	$\in$	$\mathcal{M}$, unknown, true parameter
k	$\triangleq$	$\lvert\mathcal{X}\rvert$, number of actions
x^*	$\triangleq$	$\arg\max_{x\in\mathcal{X}}\langle x, \theta^*\rangle$, best action
$\Delta(x)$	$\triangleq$	$\langle x^* - x, \theta^*\rangle$, suboptimality gap
$\mathcal{C}_x$	$\triangleq$	$\{\nu \in \mathcal{M} : \langle x, \nu\rangle \geq \max_{y\in\mathcal{X}}\langle y, \nu\rangle\}$, cell of action x
$\mathcal{H}_x^{x^*}$	$\triangleq$	$\{\nu \in \mathcal{M} : \langle x, \nu\rangle \geq \langle x^*, \nu\rangle\}$
n	$\triangleq$	horizon
R_n	$\triangleq$	$\sum_{t=1}^n \langle x^* - x_t, \theta^*\rangle$, regret
s_n	$\triangleq$	effective horizon / exploration step counter
c^*	$\triangleq$	asymptotic regret, see (2)
α^*	$\triangleq$	asymptotically optimal allocation
x_s	$\triangleq$	x_{t_s} action choice at local time s
y_s	$\triangleq$	$\langle x_s, \theta^*\rangle + \epsilon_s$, observation with (sub-)Gaussian noise ϵ_s

Least-Squares Estimate

$V(\alpha)$	$\triangleq$	$\sum_{x\in\mathcal{X}} \alpha(x) x x^\top$, covariance matrix for allocation α
V_s	$\triangleq$	$\sum_{i=1}^s x_i x_i^\top + \mathbf{1}_d$, (regularized) empirical covariance matrix
$\hat{\theta}_s$	$\triangleq$	$V_s^{-1}\sum_{i=1}^s x_i y_i$, least squares estimate
$\beta_{s,1/\delta}$	$\triangleq$	$\left(\sqrt{2\log\frac{1}{\delta} + \log\det V_s} + 1\right)^2$ concentration coefficient
$\hat{x}_s$	$\triangleq$	$\arg\max_{x\in\mathcal{X}}\langle x, \hat{\theta}_s\rangle$, empirically best action for the estimate $\hat{\theta}_s$
x_s^{UCB}	$\triangleq$	$\arg\max_{x\in\mathcal{X}}\langle x, \hat{\theta}_s\rangle + \beta_{s,s^2}^{1/2}\lVert x\rVert_{V_s^{-1}}$, UCB action
$\hat{\nu}_s(x)$	$\triangleq$	$\arg\min_{\nu\in\mathcal{C}_x}\lVert\nu - \hat{\theta}_s\rVert_{V_s}^2$, alternative parameter in $\mathcal{C}_x$
m_s	$\triangleq$	$\frac{1}{2}\min_{x\neq\hat{x}_s}\lVert\hat{\nu}_s(x) - \hat{\theta}_s\rVert_{V_s}^2$, minimum constraint value

Information-Directed Sampling

$\hat{\Delta}_s(x)$	$\triangleq$	$\delta_s + \langle \hat{x}_s - x, \hat{\theta}_s\rangle$ gap estimate with estimation error δ_s
$I_s(x)$	$\triangleq$	information gain
γ_n	$\triangleq$	$\sum_{s=1}^{s_n} I_s(x_s)$, total information gain
$\Psi_s(\mu)$	$\triangleq$	$\frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)}$, information ratio
μ_s	$\triangleq$	$\arg\min_{\mu\in\mathscr{P}(\mathcal{X})}\Psi_s(\mu)$, IDS distribution

Appendix B. Additional Proofs and Technical Lemmas

B.1. Properties of the IDS Distribution

The results in this section are mostly known or refine previous results. We start with a lemma by Kirschner et al. (2020, Lemma 5), which shows that IDS plays close to greedy.

Lemma 6 (Almost greedy) *The IDS distribution is almost greedy,* $\hat{\Delta}_s(\mu_s) \leq 2\delta_s$.

The next result is by Russo and Van Roy (2014, Proposition 6).

Lemma 7 (Convexity & support on two actions) *The information ratio as a function of the distribution,* $\mu \mapsto \Psi_s(\mu)$ *is convex. Further, the IDS distribution* $\mu_s = \arg\min_\mu \Psi_s(\mu)$ *can always be chosen with a support of at most two actions.*

In light of this lemma, the IDS distribution can be understood and computed by optimizing the information ratio between pairs of actions. We provide a closed-form solution for the IDS distribution over two actions in the next lemma.

Lemma 8 *Let* $0 < \Delta_1 \leq \Delta_2$ *denote the gaps of two actions and* $0 \leq I_1, I_2$ *the corresponding information gain. Define the ratio*

$$\Psi(p) = \frac{((1-p)\Delta_1 + p\Delta_2)^2}{(1-p)I_1 + pI_2} \,.$$

Then the optimal trade-off probability $p^* = \arg\min_{0 \leq p \leq 1} \Psi(p)$ *is*

$$p^* = \begin{cases} 0 & \textit{if } I_1 \geq I_2 \\ \text{clip}_{[0,1]}\left(\frac{\Delta_1}{\Delta_2 - \Delta_1} - \frac{2I_1}{I_2 - I_1}\right) & \textit{else,} \end{cases}$$

where we use the convention that $\Delta_1/0 = \infty$ *and* $\text{clip}_{[0,1]}(a) = \max(\min(a,1),0)$.

Proof The case $I_1 \geq I_2$ is immediate, because $p > 0$ increases the numerator and decreases the denominator. Hence we can assume $I_1 < I_2$. Recall that $\Psi(p)$ is convex on the domain $[0,1]$ (Lemma 7). The derivative is

$$\Psi'(p) = -\frac{\big(\Delta_1 + p(\Delta_2 - \Delta_1)\big)\big(\Delta_1(I_2 - I_1) - (\Delta_2 - \Delta_1)(2I_1 + p(I_2 - I_1))\big)}{(I_1 + p(I_2 - I_1))^2} \,.$$

Note that $(\Delta_2 - \Delta_1) \geq 0$ and $(I_2 - I_1) > 0$. Solving for the first order condition $\Psi'(p) = 0$ gives $p = \frac{\Delta_1}{\Delta_2 - \Delta_1} - \frac{2I_1}{I_2 - I_1}$. We can also read off that $p < 0$ implies $\Psi'(0) > 0$, and $p > 1$ implies $\Psi'(1) < 0$. Hence clipping p to $[0,1]$ leads to the correct solution. ■

The next lemma characterizes the support of the IDS distribution.

Lemma 9 (IDS support) *Let* $\Psi_s^* = \min_\mu \Psi_s(\mu)$ *and define*

$$g_s(x) = \hat{\Delta}_s(x) - \frac{\Psi_s^*}{2\hat{\Delta}_s(\mu_s)} I_s(x) \,.$$

For any $x \in supp(\mu_s)$*, it holds* $g_s(x) = \min_{z \neq \mathcal{X}} g_s(z)$*, and further* $g_s(x) = g_s(\mu) = \frac{1}{2}\hat{\Delta}_s(\mu_s)$.

Proof The proof is similar to the proof of (Russo and Van Roy, 2014, Proposition 6). It is easy to see that the solution sets to the following objectives are equal:

$$\min_\mu \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)} \quad \text{and} \quad \min_\mu \left\{ S(\mu) = \hat{\Delta}_s(\mu)^2 - \Psi_s^* I_s(\mu) \right\} ,$$

where $\Psi_s^* = \min_\mu \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)}$ is the optimal information ratio. Thinking of μ_s as a vector in $\mathbb{R}^k$, we compute the gradient of $S(\mu)$ at μ_s,

$$\nabla_\mu S(\mu)|_{\mu=\mu_s} = 2\hat{\Delta}_s \hat{\Delta}_s(\mu_s) - \Psi_s^* I_s = h_s \in \mathbb{R}^k$$

It must be that for each $x \in \text{supp}(\mu_s)$, $h_s(x) = \min_x h_s(x)$. Suppose otherwise, that the optimal solution is supported on some x and there exists a $z \neq x$ with $h_s(x) > h_s(z)$. Then $(e_x - e_z)^\top h_s > 0$, hence moving probability mass from x to z would decrease the objective. In other words, the IDS distribution must be minimizing h_s,

$$h_s(\mu_s) = \min_\mu h_s(\mu) .$$

Now, simply dividing h_s by $2\hat{\Delta}(\mu_s)$ and taking expectation over the support of the IDS distribution yields the second claim. ■

B.2. Bounds on the Information Ratio

For the asymptotic bound on the information ratio, we define $\alpha^* \in (\mathbb{R}_{\geq 0} \cup \{\infty\})^k$ as the solution to the lower bound (2), which is obtained as the appropriate limit. Further, let $\tilde{\alpha}^* = \alpha^* \mathbb{1}(x \neq x^*)$ be the optimal allocation on the sub-optimal actions, which is always finite.

Lemma 10 (Truncated optimal allocation) *Let $\alpha_\lambda^*(x) = \tilde{\alpha}^* + \lambda \mathbb{1}(x = x^*)$ be the optimal allocation truncated on x^* such that $\alpha_\lambda^*(x^*) = \lambda$. There exists a constant $C(\theta, \mathcal{X})$ depending only on the instance and the action set, such that for all $\nu \in \mathcal{C}^*$,*

$$\tfrac{1}{2}\|\nu - \theta^*\|_{V(\alpha_\lambda^*)}^2 \geq 1 - 2C(\theta, \mathcal{X})\|\tilde{\alpha}\|_1 \lambda^{-1} .$$

Proof Assume $2C(\theta, \mathcal{X})\|\tilde{\alpha}\|_1 \leq \lambda$, otherwise the claim is immediate. Let $\tilde{\alpha}^*(x) = \alpha^* \mathbb{1}(x \neq x^*)$ be the optimal allocation on sub-optimal actions. We have

$$\tfrac{1}{2}\|\nu - \theta^*\|_{V(\alpha_\lambda^*)}^2 = \tfrac{1}{2}\|\nu - \theta^*\|_{V(\tilde{\alpha}^*)}^2 + \tfrac{\lambda}{2}\langle \nu - \theta^*, x^* \rangle^2 .$$

If $\lambda \langle \nu - \theta^*, x^* \rangle^2 \geq 2$ the claim follows. Hence we may assume $\langle \nu - \theta^*, x^* \rangle^2 \leq 2\lambda^{-1}$. In other words, ν is in a $(2/\lambda)^{1/2}$-neighbourhood of the affine subspace, which is defined by x^* and offset θ^*. Now we fix any $x \neq x^*$, such that $\nu \in \mathcal{H}_x^{x^*}$ and define $\mathcal{H}_x^* = \mathcal{H}_x^{x^*} \cap \{\nu : \langle \nu - \theta^*, x^* \rangle = 0\}$ as the intersection of the affine subspace with $\mathcal{H}_x^{x^*}$. This is the set of parameters in $\mathcal{H}_x^{x^*}$, which is indistinguishable from observations of x^*. By definition, $\nu^* \in \mathcal{H}_x^*$ satisfies $\langle \nu^* - \theta^*, x^* \rangle = 0$, hence by definition of the optimal allocation,

$$\tfrac{1}{2}\|\nu^* - \theta^*\|_{V(\tilde{\alpha}^*)}^2 = \tfrac{1}{2}\|\nu^* - \theta^*\|_{V(\alpha^*)}^2 \geq 1 .$$

We expect the same holds approximately for ν with $\langle \nu - \theta^*, x^* \rangle^2 \leq 2\lambda^{-1}$. Lemma 23 with an appropriate shift of the parameter space and $\lambda_{\max}(V(\tilde{\alpha}^*)) \leq \|\tilde{\alpha}^*\|_1$ imply

$$\min_{\nu^* \in \mathcal{H}_x^*} \|\nu - \nu^*\|_{V(\tilde{\alpha}^*)}^2 \leq C(\theta, \mathcal{X}) \|\tilde{\alpha}^*\|_1 \langle \nu - \theta^*, x^* \rangle^2 \leq 2\lambda^{-1} C(\theta, \mathcal{X}) \|\tilde{\alpha}\|_1 \leq 1\,,$$

where the last inequality is our case assumption. Considering that $\|\nu^* - \theta^*\|_{V(\alpha^*)} \geq \|\nu - \nu^*\|_{V(\tilde{\alpha}^*)}$, we further get

$$\begin{aligned}
\tfrac{1}{2}\|\nu - \theta^*\|_{V(\alpha_\lambda^*)}^2 &= \tfrac{1}{2}\|\nu - \theta^*\|_{V(\tilde{\alpha}^*)}^2 + \tfrac{\lambda}{2}\langle \nu - \theta^*, x^* \rangle^2 \\
&\geq \tfrac{1}{2}(\|\nu^* - \theta^*\|_{V(\tilde{\alpha}^*)} - \|\nu - \nu^*\|_{V(\tilde{\alpha}^*)})^2 + \tfrac{\lambda}{2}\langle \nu - \theta^*, x^* \rangle^2\,.
\end{aligned}$$

The case $\|\nu^* - \theta^*\|_{V(\tilde{\alpha}^*)} \geq 2$ is again immediate, so we may assume $\sqrt{2} \leq \|\nu^* - \theta^*\|_{V(\tilde{\alpha}^*)} \leq 2$, which leaves us with

$$\begin{aligned}
\tfrac{1}{2}\|\nu - \theta^*\|_{V(\alpha_\lambda^*)}^2 &\geq \tfrac{1}{2}\|\nu^* - \theta^*\|_{V(\tilde{\alpha}^*)}^2 - \|\nu - \nu^*\|_{V(\tilde{\alpha}^*)}\|\nu^* - \theta^*\|_{V(\tilde{\alpha}^*)} + \tfrac{\lambda}{2}\langle \nu - \theta^*, x^* \rangle^2 \\
&\geq 1 - 2\|\nu - \nu^*\|_{V(\tilde{\alpha}^*)} + \tfrac{\lambda}{2}\langle \nu - \theta^*, x^* \rangle^2 \\
&\overset{(i)}{\geq} 1 - 2(C(\theta, \mathcal{X})\|\tilde{\alpha}\|_1 \langle \nu - \theta^*, x^* \rangle^2)^{1/2} + \tfrac{\lambda}{2}\langle \nu - \theta^*, x^* \rangle^2 \\
&\overset{(ii)}{\geq} 1 - 2C(\theta, \mathcal{X})\|\tilde{\alpha}\|_1 \lambda^{-1}\,.
\end{aligned}$$

For (i) we choose ν^* with $\|\nu - \nu^*\|_{V(\tilde{\alpha}^*)}^2 \leq C(\theta, \mathcal{X})\|\tilde{\alpha}\|_1 \langle \nu - \theta^*, x^* \rangle^2$ and for (ii) we minimize over $\langle \nu - \theta^*, x^* \rangle$. This completes the proof. ■

Lemma 11 (Asymptotic information ratio) *Recall that* $\beta_s = \|\hat{\theta}_s - \theta^*\|_{V_s}^2$ *and* $m_s = \frac{1}{2}\min_{z \neq x^*} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2$. *Assume that* $4\beta_s \leq m_s$ *and* $\beta_s \leq \beta_{s,s^2}$. *Then,*

$$\Psi_s(\mu_s) \leq 4\delta_s\big(c^* + \mathcal{O}(\beta_s^{1/2} m_s^{-1/2} + \delta_s)\big)\,,$$

for $\beta_s^{1/2} m_s^{-1/2} \to 0$ *and* $\delta_s \to 0$.

Proof First note that the assumption $m_s \geq 4\beta_s$ implies $\hat{x}_s = x^*$ by Lemma 19. Introduce the shorthand $\bar{\Delta}_s(x) = \langle \hat{\theta}_s, \hat{x}_s - x \rangle$ for the estimated mean gap and let $\tilde{\mu} \in \mathscr{P}(\mathcal{X})$ be a distribution with $2\delta_s \leq \hat{\Delta}(\tilde{\mu}) = \delta_s + \bar{\Delta}_s(\tilde{\mu})$. Then, by Lemma 8,

$$\min_{\mu \in \mathscr{P}(\mathcal{X})} \Psi_s(\mu) \leq \min_{0 \leq p \leq 1} \frac{\big((1-p)\hat{\Delta}_s(x^*) + p\hat{\Delta}(\tilde{\mu})\big)^2}{pI_s(\tilde{\mu})} = \frac{4\delta_s(\hat{\Delta}_s(\tilde{\mu}) - \delta_s)}{I_s(\tilde{\mu})} = \frac{4\delta_s \bar{\Delta}_s(\tilde{\mu})}{I_s(\tilde{\mu})}\,.$$

Note that the last ratio is invariant in constant rescaling $\tilde{\mu}$, so we may plug in non-normalized allocations. Recall that $\tilde{\alpha}^*$ is the optimal allocation over suboptimal actions, as defined at the beginning of Appendix B.2. We let $\alpha_\lambda^* = \tilde{\alpha}^* + \lambda \mathbb{1}(x = x^*)$ be the truncated optimal allocation and $\tilde{\mu}_\lambda = \alpha_\lambda^*/(\|\tilde{\alpha}^*\|_1 + \lambda)$ be the corresponding normalized distribution. With Lemma 19 , we get

$$\Delta(\tilde{\mu}_\lambda) - \bar{\Delta}_s(\tilde{\mu}_\lambda) \leq \|\hat{\theta}_s - \theta^*\|_{V_s} \max_{x \neq x^*} \|x^* - x\|_{V_s^{-1}} \leq \frac{\beta_s^{1/2}}{(2m_s)^{1/2} - \beta_s^{1/2}} \leq \frac{2\beta_s^{1/2}}{m_s^{1/2}}\,.$$

The last inequality simplifies the expression with $4\beta_s \leq m_s$. Note that $\Delta(\tilde{\mu}_\lambda) = \frac{c^*}{\|\tilde{\alpha}^*\|_1 + \lambda}$. Hence, to satisfy $\delta_s \leq \bar{\Delta}_s(\tilde{\mu}_\lambda)$, it is sufficient to satisfy the constraint,

$$\delta_s \leq \frac{c^*}{\|\tilde{\alpha}^*\|_1 + \lambda} - \frac{2\beta_s^{1/2}}{m_s^{1/2}}.$$

At equality, we get

$$\lambda = \frac{c^*}{\delta_s + \frac{2\beta_s^{1/2}}{m_s^{1/2}}} - \|\tilde{\alpha}^*\|_1 \,.$$

Note that as $\delta_s \to 0$ and $m_s \to \infty$, we get $\lambda \to \infty$ as expected. Next we compute the approximation errors. Using again Lemma 19,

$$\begin{aligned}\bar{\Delta}(\alpha^*_\lambda) &= \Delta(\tilde{\alpha}^*) + \sum_{x \neq x^*} \tilde{\alpha}^*(x) \langle \hat{\theta}_s - \theta^*, x^* - x\rangle \\ &\leq c^* + \frac{\|\tilde{\alpha}^*\|_1 \beta_s^{1/2}}{(2m_s)^{1/2} + \beta_s^{1/2}} \leq c^* + 2\|\tilde{\alpha}^*\|_1 \beta_s^{1/2} m_s^{-1/2} \,.\end{aligned}$$

To bound the approximation error of $I_s(\alpha^*_\lambda)$, note that $\beta_s = \|\hat{\theta}_s - \theta^*\|^2_{V_s} \leq \beta_{s,s^2}$ implies

$$\begin{aligned}I_s(\alpha^*_\lambda) &= \tfrac{1}{2} \sum_{z \in \mathcal{X}} \alpha^*_\lambda(z) \sum_{x \neq x^*} q_s(x) \left(|\langle \hat{\nu}_s(x) - \hat{\theta}_s, z\rangle| + \beta_{s,s^2}^{1/2} \|z\|_{V_s^{-1}} \right)^2 \\ &\geq \tfrac{1}{2} \sum_{z \in \mathcal{X}} \alpha^*_\lambda(z) \sum_{x \neq x^*} q_s(x) \langle \hat{\nu}_s(x) - \theta^*, z\rangle^2 \\ &= \tfrac{1}{2} \sum_{x \neq x^*} q_s(x) \|\hat{\nu}_s(x) - \theta^*\|^2_{V(\alpha^*_\lambda)} \\ &\geq 1 - 2C(\mathcal{X}, \theta)\|\tilde{\alpha}^*\|_1 \lambda^{-1} \,.\end{aligned}$$

The last step is by Lemma 10. Finally, the proof is completed by using $\frac{c^* + A}{1 - B} \leq c^* + A + c^* B$, which yields

$$\Psi_s(\mu_s) \leq \frac{4\delta_s \bar{\Delta}_s(\alpha^*_\lambda)}{I_s(\alpha^*_\lambda)} \leq 4\delta_s \big(c^* + 2\|\tilde{\alpha}^*\|_1 \beta_s^{1/2} m_s^{-1/2} + 2c^* C(\mathcal{X}, \theta)\|\tilde{\alpha}^*\|_1 \lambda^{-1}\big) \,.$$

Since $\lambda^{-1} = \mathcal{O}\Big(c^{*-1}\big(\delta_s + 2\beta_s^{1/2} m_s^{-1/2}\big)\Big)$ for $\beta_s^{1/2} m_s^{-1/2} \to 0$ and $\delta_s \to 0$, we get

$$\Psi_s(\mu_s) \leq 4\delta_s \big(c^* + \mathcal{O}(\beta_s^{1/2} m_s^{-1/2} + \delta_s)\big) \,.$$

■

B.3. Bounds on the Information Gain

We start by proving a worst-case bound on the total information gain $\gamma_n = \sum_{s=1}^{s_n} I_s(x_s)$.

Lemma 12 (Total information gain) *For any sequence $x_1, \dots, x_n$, the total information gain $\gamma_n = \sum_{s=1}^{s_n} I_s(x_s)$ is bounded as follows,*

$$\gamma_n \leq 2\left(\beta_{s_n, n\log(n)} + \beta_{s_n, n\log(n)}^{1/2} + \beta_{s_n, s_n^2}\right) d \log(s_n) \leq \mathcal{O}\big(d^2 \log(n)^2\big)\,.$$

Proof Note that

$$\begin{aligned}
\gamma_n = \sum_{s=1}^{s_n} I_s(x_s) &= \tfrac{1}{2}\sum_{s=1}^{s_n}\left(\sum_{x \neq \hat{x}_s} q_s(x)|\langle \hat{\nu}_s(x) - \hat{\theta}_s, x_s\rangle| + \beta_{s,s^2}^{1/2}\|x_s\|_{V_s^{-1}}\right)^2 \\
&\leq \sum_{s=1}^{s_n}\sum_{x\neq\hat{x}_s} q_s(x)\langle \hat{\nu}_s(x) - \hat{\theta}_s, x_s\rangle^2 + \beta_{s,s^2}\|x_s\|_{V_s^{-1}}^2 \\
&\overset{(i)}{\leq} \sum_{s=1}^{s_n}\sum_{x\neq\hat{x}_s} q_s(x)\big(\|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2 + \beta_{s,s^2}\big)\|x_s\|_{V_s^{-1}}^2 \\
&\overset{(ii)}{\leq} \sum_{s=1}^{s_n}\left(\min_{x\neq\hat{x}_s}\|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2 + \frac{2\log(k)}{\eta_s} + \beta_{s,s^2}\right)\|x_s\|_{V_s^{-1}}^2 \\
&\overset{(iii)}{\leq} \left(\beta_{s_n, n\log(n)} + \beta_{s_n, n\log(n)}^{1/2} + \beta_{s_n, s_n^2}\right)\sum_{s=1}^{s_n}\|x_s\|_{V_s^{-1}}^2 \\
&\overset{(iv)}{\leq} 2\left(\beta_{s_n, n\log(n)} + \beta_{s_n, n\log(n)}^{1/2} + \beta_{s_n, s_n^2}\right) d\log(s_n)
\end{aligned}$$

Step (i) uses Cauchy-Schwarz, (ii) the soft-min bound for the q-weights (see Lemma 22). For (iii) we used that $m_s = \frac{1}{2}\min_{x\neq\hat{x}_s}\|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2 \leq \frac{1}{2}\beta_{s_n, n\log(n)}$ holds in all exploration rounds and the choice $\eta_s = \min_{l\leq s} m_l^{-1/2}\log(k)$ and lastly, (iv) bounds the elliptic potential (Lemma 18). Considering that $\beta_{s,1/\delta} = 2\log\frac{1}{\delta} + \mathcal{O}(d\log s)$ completes the proof. ■

Lemma 13 (Constant information gain) *Assume that $\hat{x}_s = x^*$ and $2\delta_s \leq \hat{\Delta}_s(x)$ for all $x \neq \hat{x}_s$. If $z_s \neq x^*$ is contained in the support of the IDS distribution, supp(μ_s), then the information gain of z_s is at least a constant,*

$$I_s(z_s) \geq \frac{\Delta_{\min}^2}{8(8d+9)}\,.$$

Proof Note that by $z_s \in \operatorname{supp}(\mu_s)$ and Lemma 9,

$$I_s(z_s) = \left(\hat{\Delta}_s(z_s) - \frac{\hat{\Delta}_s(\mu_s)}{2}\right)\frac{2\hat{\Delta}_s(\mu_s)}{\Psi_s} \geq \left(\hat{\Delta}_s(z_s) - \delta_s\right)\frac{2\delta_s}{\Psi_s} \geq \frac{\hat{\Delta}_s(z_s)\delta_s}{\Psi_s}\,.$$

We first used that $\delta_s \le \hat{\Delta}_s(\mu) \le 2\delta_s$ (Lemma 6) and then the assumption that $2\delta_s \le \hat{\Delta}_s(z_s)$. Further, $2\hat{\Delta}_s(z_s) \ge \Delta_{\min}$, and by Lemma 4,

$$\Psi_s(\mu_s) \le \frac{4\delta_s(8d+9)}{\Delta_{\min}} .$$

Combining the inequalities yields the result. ■

Lemma 14 *Let $q_s^*(z) \propto \exp(-\eta_s\|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2)$ be mixing weights defined on $\mathcal{X} \setminus x^*$ (also when $\hat{x}_s \neq x^*$), where $\hat{\nu}_s(z) = \arg\min_{\nu \in \mathcal{H}_z^{x^*}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ for all $z \neq x^*$. Define $l_s(q_s) = \sum_{z\neq x^*} q_s^*(z)\langle\hat{\nu}_s(z) - \theta_s\rangle^2$ and let $J_s = \mathbb{1}(24^2\eta_s\beta_s\|x_s\|_{V_s^{-1}}^2 \le 1; \beta_s\|x_s\|_{V_s^{-1}}^2 \le 1)$. Then*

$$\mathbb{E}\Big[\textstyle\sum_{s=1}^{s_n} J_s l_s(q_s^*) - \min_{x\neq x^*}\|\hat{\nu}_{s_n}(x) - \hat{\theta}_{s_n}\|_{V_{s_n}}^2\Big] \le \mathcal{O}\big(\log(n)^{1/2}\mathbb{E}[\log(s_n)^2]\big) .$$

Proof The statement is a regret bound for the exponential weights learner that defines the q_s^*-weights, excluding steps where $J_s = 0$. The difference to standard online learning bounds is that the cumulative loss $L_s(x) = \frac{1}{2}\|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2$, which defines the mixing weights and the baseline, does not exactly equal the sum of instantaneous loss $\sum_{s=1}^{s_n} l_s(x)$. For the analysis we make use of well-known connections between the exponential weights algorithm and the mirror descent framework, in particular the *follow the regularized leader* (FTRL) algorithm (Shalev-Shwartz and Singer, 2007). To this end, let $\psi(q) = \sum_{x\neq x^*} q(x)\log(q(x))$ be the entropy function defined for $q \in \mathscr{P}(\mathcal{X} \setminus \hat{x}_s)$. For learning rate $\eta > 0$, we define

$$\psi_\eta(q) = \frac{1}{\eta}\left(\psi(q) - \min_{q' \in \mathscr{P}(\mathcal{X}\setminus x^*)} \psi(q')\right) .$$

We denote $\psi_s = \psi_{\eta_s}$. The choice of mixing weights q_s^* can be equivalently written as

$$q_s^* = \mathop{\arg\min}_{q \in \mathscr{P}(\mathcal{X}\setminus x^*)} L_s(q) + \psi_s(q) .$$

Denote $\Lambda_n = \sum_{s=1}^{s_n} J_s l_s(q_s^*) - \min_{x\neq x^*}\|\hat{\nu}_{s_n}(x) - \hat{\theta}_{s_n}\|_{V_{s_n}}^2$. The following inequality is easily verified by telescoping (c.f. Lemma 7.1 Orabona, 2019),

$$\Lambda_n \le -\frac{1}{\eta_{s_n}}\min_{q'}\psi(q') + \sum_{s=1}^{s_n}\big([L_s + J_s l_s + \psi_s](q_s) - [L_{s+1} + \psi_{s+1}](q_{s+1})\big) .$$

For the first term, we immediately get $-\frac{1}{\eta_s}\min_{q'}\psi(q') \le \frac{\log(k)}{\eta_{s_n}}$. The second term is often referred to as stability term. We first address steps s where $J_s = 1$. Define $\tilde{q}_{s+1} = \arg\min_{q\in\mathscr{P}(\mathcal{X}\setminus x^*)}[L_{s+1} + \psi_s](q) \propto \exp(-\eta_s L_{s+1})$. Using that the learning rate is decreasing, we get

$$\begin{aligned} &[L_s + l_s + \psi_s](q_s) - [L_{s+1} + \psi_{s+1}](q_{s+1}) \\ &\le [L_{s+1} + \psi_s](q_s) - [L_{s+1} + \psi_s](\tilde{q}_{s+1}) + [L_s + l_s - L_{s+1}](q_s) . \end{aligned} \tag{15}$$

Note that L_{s+1} exhibits an intricate dependence on the outcome y_s, whereas all other quantities appearing in the last display are $\mathcal{F}_s$-predictable. Using that $\tilde{q}_{s+1}$ is a minimizer of $L_{s+1} + \psi_s$ and the definition of the Bregman divergence $D_\psi(p\|q) = \psi(p) - \psi(q) - \langle \nabla\psi(q), p - q\rangle$, we find

$$[L_{s+1} + \psi_s](q_s^*) - [L_{s+1} + \psi_s](\tilde{q}_{s+1}) = \frac{1}{\eta_s} D_{\psi_s}(q_s^*, \tilde{q}_{s+1}) = \frac{1}{\eta_s} \sum_{x \neq x^*} q_s^*(x) \log \frac{q_s^*(x)}{\tilde{q}_{s+1}(x)}$$

Using that $\log(x) \leq x - 1$ for all $x > 0$, we find

$$\begin{aligned}
\sum_{x\neq x^*} q_s^* \log \frac{q_s^*}{\tilde{q}_{s+1}} &= \eta_s [L_{s+1} - L_s](q_s^*) + \log\bigg(\sum_{x\neq x^*} q_s^* \exp\big(-\eta_s(L_{s+1} - L_s)\big)\bigg) \\
&\leq -1 + \eta_s [L_{s+1} - L_s](q_s^*) + \sum_{x\neq x^*} q_s^* \exp\big(-\eta_s(L_{s+1} - L_s)\big) \\
&= \sum_{x\neq x^*} q_s^*(x) \sum_{i=2}^{\infty} \frac{(-\eta_s(L_{s+1} - L_s))^i}{i!}
\end{aligned}$$

A technical calculation which directly bounds the moments of the subgaussian noise under the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_s]$ with the condition $J_s = 1$, is summarized in Lemma 17. This yields

$$\begin{aligned}
&\sum_{s=1}^{s_n} J_s \mathbb{E}\Bigg[[L_{s+1} + \psi_s](q_s^*) - [L_{s+1} + \psi_s](\tilde{q}_{s+1})\Bigg| \mathcal{F}_s\Bigg] \\
&\leq \sum_{s=1}^{s_n} \frac{J_s}{\eta_s} \sum_{x\neq x^*} q_s^*(x) \mathbb{E}\Bigg[\sum_{i=2}^{\infty} \frac{(-\eta_s(L_{s+1}(x) - L_s(x)))^i}{i!}\Bigg| \mathcal{F}_s\Bigg] \\
&\leq \sum_{s=1}^{s_n} \sum_{x\neq x^*} q_s^*(x) \mathcal{O}\Big(\eta_s\big(\beta_s \|x_s\|_{V_s^{-1}}^2 + \|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2 \|x_s\|_{V_s^{-1}}^2\big)\Big) \\
&\leq \mathcal{O}\Big(\log(n)^{1/2} \log(s_n)^2\Big)
\end{aligned}$$

The last step makes use of Lemma 22, $\eta_s m_s \leq \beta_{s_n, n\log n}^{1/2} \leq \mathcal{O}(\log(n)^{1/2} + \log(s_n)^{1/2})$ and Lemma 22. Going back to (15), still for the case where $J_s = 1$, it remains to bound the shift term $S_s = L_s + l_s - L_{s+1}$. We have

$$\begin{aligned}
\mathbb{E}[S_s(q_s^*)|\mathcal{F}_s] &\overset{(i)}{\leq} 2\|x_s\|_{V_s^{-1}}^2 \big(\textstyle\sum_{x\neq x^*} q_s \|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s} \beta_s^{1/2} + \beta_s + 1\big) \\
&\overset{(ii)}{\leq} 2\|x_s\|_{V_s^{-1}}^2 \Big(\sqrt{\textstyle\sum_{x\neq x^*} q_s \|\hat{\nu}_s(x) - \hat{\theta}_s\|_{V_s}^2} \beta_s^{1/2} + \beta_s + 1\Big) \\
&\overset{(iii)}{\leq} 2\|x_s\|_{V_s^{-1}}^2 \Big(\big((m_s + \log(k)/\eta_s)\beta_s\big)^{1/2} + \beta_s + 1\Big)
\end{aligned}$$

Here, (i) follows from the Lemma 15, Cauchy-Schwarz and taking the expectation; (ii) is Jensen's inequality and (iii) is the softmin inequality (Lemma 22). Hence, using that $m_s \leq \beta_{s_n, n\log(n)} \leq \mathcal{O}(\log(n) + \log(s_n))$ and the elliptic potential lemma (Lemma 18), we find

$$\sum_{s=1}^{s_n} \mathbb{E}[S_s(x)|\mathcal{F}_s] \leq \mathcal{O}\big(\log(s_n)^2 \log(n)^{1/2}\big)\,.$$

Lastly, we address (15) for the case $J_s = 0$, which then reads

$$[L_s + \psi_s](q_s) - [L_{s+1} + \psi_{s+1}](q_{s+1}) \leq [L_s - L_{s+1}](q_{s+1})\,. \tag{16}$$

We can reuse Lemma 16 to find

$$\begin{aligned}\mathbb{E}_s[[L_s - L_{s+1}](x)] &\leq \mathcal{O}\big(\beta_s \|x_s\|^2_{V_s^{-1}} + |\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle| + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2\big)\\ &\leq \mathcal{O}\big(\beta_s \|x_s\|^2_{V_s^{-1}} + 1)\big)\end{aligned}$$

Using that when $J_s = 0$ we have $1 \leq \beta_s \|x_s\|^2_{V_s}$, or $1 \leq 24^2 \eta_s \beta_s \|x_s\|^2_{V_s^{-1}}$, we can sum up these terms to

$$\sum_{s=1}^{s_n} \mathbb{E}_s[L_s - L_{s+1}](x) \leq \sum_{s=1}^{s_n} \mathcal{O}(\beta_s \|x_s\|^2_{V_s^{-1}}) \leq \mathcal{O}(\log(s_n)^2)$$

The claim follows. ■

Lemma 15 *Let $L_s(x) = \|\hat{\nu}_s(x) - \hat{\theta}_s\|^2_{V_s}$ defined for $x \neq x^*$ and assume that $\langle \nu - \theta^*, x\rangle \leq 1$ for all $\nu \in \mathcal{M}$ and $x \in \mathcal{X}$. Then*

$$[L_s + l_s - L_{s+1}](x) \leq 2\langle \hat{\nu}_s(x) - \hat{\theta}_s, x_s\rangle \frac{\epsilon_s + \langle x_s, \theta^* - \hat{\theta}_s\rangle}{1 + \|x_s\|^2_{V_s^{-1}}} + 2\|x_s\|^2_{V_s^{-1}}(1 + \beta_s)$$

Proof For the proof we adopt the notation $\omega_s(x) = \hat{\nu}_s(x) - \hat{\theta}_s$.

$$\begin{aligned}L_s + l_s - L_{s+1} &= \|\omega_s\|^2_{V_{s+1}} - \|\omega_{s+1}\|^2_{V_{s+1}}\\ &= \|\omega_s\|^2_{V_{s+1}} - \|\omega_s + \omega_{s+1} - \omega_s\|^2_{V_{s+1}}\\ &= 2\langle \omega_s - \omega_{s+1}, V_{s+1}\omega_s\rangle - \|\omega_{s+1} - \omega_s\|^2_{V_{s+1}}\\ &= \underbrace{2\langle \omega_s - \omega_{s+1}, V_s\omega_s\rangle}_{(A)} + \underbrace{2\langle \omega_s - \omega_{s+1}, x_s\rangle\langle x_s, \omega_s\rangle - \|\omega_{s+1} - \omega_s\|^2_{V_{s+1}}}_{(B)}\end{aligned}$$

To avoid clutter, the dependence on x is implicit below. Note that because $\hat{\nu}_s$ is a projection of $\hat{\theta}_s$ V_s-norm onto the convex set $\mathcal{H}_x^{x^*}$, we have $\langle \hat{\nu}_s - \hat{\nu}_{s+1}, V_s(\hat{\nu}_s - \hat{\theta}_s)\rangle \leq 0$. Therefore

$$(A) \leq 2\langle \hat{\theta}_{s+1} - \hat{\theta}_s, V_s(\hat{\nu}_s - \hat{\theta}_s)\rangle = 2\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle \frac{\epsilon_s + \langle x_s, \theta^* - \hat{\theta}_s\rangle}{1 + \|x_s\|^2_{V_s^{-1}}}$$

The equality follows from Lemma 24. Next, we derive an upper bound to the term (B).

$$\begin{aligned}(B) &\leq 2\langle \omega_s - \omega_{s+1}, x_s\rangle\langle x_s, \omega_s\rangle - \|\omega_{s+1} - \omega_s\|^2_{V_{s+1}}\\ &\leq 2\|\omega_s - \omega_{s+1}\|_{V_s}\|x_s\|_{V_s^{-1}}\langle x_s, \omega_s\rangle - \|\omega_{s+1} - \omega_s\|^2_{V_{s+1}}\\ &\leq 2\|\omega_s - \omega_{s+1}\|_{V_{s+1}}\|x_s\|_{V_s^{-1}}\langle x_s, \omega_s\rangle - \|\omega_{s+1} - \omega_s\|^2_{V_{s+1}}\\ &\leq \|x_s\|^2_{V_s^{-1}}\langle x_s, \omega_s\rangle^2 \leq 2\|x_s\|^2_{V_s^{-1}}(1 + \beta_s)\end{aligned}$$

We used Cauchy-Schwarz and $\|\cdot\|^2_{V_s} \leq \|\cdot\|^2_{V_{s+1}}$ in the first and second inequality. Then we use $2ab - b^2 \leq a^2$, and in the last step boundedness, $|\langle \omega_s(x), x_s\rangle| \leq \langle \hat{\nu}_s(x) - \theta^*, x_s\rangle| + \beta_s^{1/2}\|x_s\|_{V_s^{-1}} \leq 1 + \beta_s^{1/2}$. The claim follows from combining the bounds. ■

Lemma 16 *Let $L_s(x) = \|\hat{\nu}_s(x) - \hat{\theta}_s\|^2_{V_s}$ defined for $x \neq x^*$ and assume that $\langle \nu - \theta^*, x\rangle \leq 1$ for all $\nu \in \mathcal{M}$ and $x \in \mathcal{X}$. Then*

$$|[L_s - L_{s+1}](x)| \leq 4|\epsilon|^2\|x_s\|^2_{V_s^{-1}} + 2|\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle||\epsilon_s| + 8\beta_s\|x_s\|^2_{V_s^{-1}} + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2$$

Proof For one direction, we can reuse Lemma 15,

$$\begin{aligned}[L_s - L_{s+1}](x) &\leq [L_s + l_s - L_{s+1}](x) \\ &\leq 2|\epsilon_s||\langle \hat{\nu}_s(x) - \hat{\theta}_s, x_s\rangle| + 2\|x_s\|_{V_s^{-1}}\beta_s^{1/2} + 2\|x_s\|^2_{V_s^{-1}}(1+\beta_s)\,.\end{aligned}$$

For the other direction, we have

$$\begin{aligned}[L_{s+1} - L_s](x) &= \|\hat{\nu}_{s+1} - \hat{\theta}_{s+1}\|^2_{V_{s+1}} - \|\hat{\nu}_s - \hat{\theta}_s\|^2_{V_s} \\ &\leq \|\hat{\nu}_s - \hat{\theta}_{s+1}\|^2_{V_{s+1}} - \|\hat{\nu}_s - \hat{\theta}_s\|^2_{V_s} \\ &= \|\hat{\nu}_s - \hat{\theta}_s + V_s^{-1}x_s u_s\|^2_{V_{s+1}} - \|\hat{\nu}_s - \hat{\theta}_s\|^2_{V_s}\,,\end{aligned}$$

where for the last step we denote $u_s = \frac{\epsilon_s + \langle x_s, \theta^* - \hat{\theta}_s\rangle}{1 + \|x_s\|^2_{V_s^{-1}}}$ and use Lemma 24. Further unwrapping the square gives

$$\begin{aligned}&\|\hat{\nu}_s - \hat{\theta}_s - V_s^{-1}x_s u_s\|^2_{V_{s+1}} - \|\hat{\nu}_s - \hat{\theta}_s\|^2_{V_s} \\ &= \|\hat{\nu}_s - \hat{\theta}_s - V_s^{-1}x_s u_s\|^2_{V_s} + \langle \hat{\nu}_s - \hat{\theta}_s - V_s^{-1}x_s u_s, x_s\rangle^2 - \|\hat{\nu}_s - \hat{\theta}_s\|^2_{V_s} \\ &= -2\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle u_s + u_s^2\|x_s\|^2_{V_s^{-1}} + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2 - 2\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle\|x_s\|^2_{V_s^{-1}} + \|x_s\|^4_{V_s^{-1}}u_s^2 \\ &\leq -2\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle u_s(1 + \|x_s\|^2_{V_s^{-1}}) + 2u_s^2\|x_s\|^2_{V_s^{-1}} + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2 \\ &\leq 2|\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle|(|\epsilon_s + \beta_s^{1/2}\|x_s\|_{V_s^{-1}}) + 4(|\epsilon|^2 + \beta_s\|x_s\|^2_{V_s^{-1}})\|x_s\|^2_{V_s^{-1}} + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2 \\ &\leq 2|\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle||\epsilon_s| + 2\beta_s^{1/2}\|x_s\|_{V_s^{-1}} + 4|\epsilon|^2\|x_s\|^2_{V_s^{-1}} + 6\beta_s\|x_s\|^2_{V_s^{-1}} + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2\end{aligned}$$

Combining both directions yields the claim. ■

Lemma 17 *Let s such that $24^2\eta_s\beta_s\|x_s\|^2_{V_s^{-1}} \leq 1$ and $\beta_s\|x_s\|^2_{V_s^{-1}} \leq 1$. Then*

$$\mathbb{E}\left[\sum_{i=2}^{\infty} \frac{|\eta_s(L_{s+1}(x) - L_s(x))|^i}{i!}\middle|\mathcal{F}_s\right] \leq \mathcal{O}\Big(\eta_s^2\big(\beta_s\|x_s\|^2_{V_s^{-1}} + \|\hat{\nu}_s(x) - \hat{\theta}_s\|^2_{V_s}\|x_s\|^2_{V_s^{-1}}\big)\Big)\,.$$

Proof

$$\begin{aligned}&|(L_{s+1}(x) - L_s(x))|^i| \\ &\leq \big(4|\epsilon|^2\|x_s\|^2_{V_s^{-1}} + 2|\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle||\epsilon_s| + 8\beta_s\|x_s\|^2_{V_s^{-1}} + \langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2\big)^i \\ &\leq \big(12|\epsilon|^2\|x_s\|^2_{V_s^{-1}}\big)^i + \big(6|\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle||\epsilon_s|\big)^i + \big(24\beta_s\|x_s\|^2_{V_s^{-1}} + 3\langle \hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2\big)^i\end{aligned}$$

For the last step we used that for $a, b, c, \geq 0$, $(a+b+c)^i \leq (3a)^i + (3b)^i + (3c)^i$. Further note that for the σ-subgaussian noise ϵ_s, it holds that for all $i \in \mathbb{N}$, $\mathbb{E}[|\epsilon|^i] \leq (2\sigma^2)^{i/2} i\Gamma(i/2) \leq (2\sigma^2)^i i!$ and $\mathbb{E}[|\epsilon|^2 i] \leq (2\sigma^2)^i 2i!$ (c.f. Lemma 1.4, Rigollet, 2015). Hence we get

$$\mathbb{E}_s\left[\frac{|\eta_s(L_{s+1}(x) - L_s(x))|^i}{i!}\right] \leq \mathbb{E}_s\left[\frac{(12\eta_s|\epsilon|^2\|x_s\|^2_{V_s^{-1}})^i}{i!}\right] + \mathbb{E}_s\left[\frac{(6\eta_s|\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle||\epsilon_s|)^i}{i!}\right] + \frac{(24\eta_s\beta_s\|x_s\|^2_{V_s^{-1}} + 3\eta_s\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2)^i}{i!}$$

We address each term individually, also using that $24^2\eta_s\beta_s\|x_s\|^2_{V_s^{-1}} \leq 1$.

$$\begin{aligned}
\mathbb{E}_s\left[\frac{(12\eta_s|\epsilon|^2\|x_s\|^2_{V_s^{-1}})^i}{i!}\right] &\leq (24\eta_s\sigma^2\|x_s\|^2_{V_s^{-1}})^i \\
&\leq (24\eta_s\sigma^2\|x_s\|^2_{V_s^{-1}})^2 \cdot 2^{-i+2} \\
\mathbb{E}_s\left[\frac{(6\eta_s|\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle||\epsilon_s|)^i}{i!}\right] &\leq (12\eta_s|\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle|\sigma^2)^i \\
&\leq (12\eta_s|\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle|\sigma^2)^2 \cdot 2^{-i+2} \\
\frac{(24\eta_s\beta_s\|x_s\|^2_{V_s^{-1}} + 3\eta_s\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2)^i}{i!} &\leq (24\eta_s\beta_s\|x_s\|^2_{V_s^{-1}} + 3\eta_s\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2)^{i-2}\frac{2^{i-2}}{i!}
\end{aligned}$$

Summing over $i = 2, \ldots, \infty$ gives

$$\begin{aligned}
&\sum_{i=2}^{\infty}\mathbb{E}_s\left[\frac{|\eta_s(L_{s+1}(x) - L_s(x))|^i}{i!}\right] \\
&\leq \mathcal{O}\Big(\big(\eta_s\|x_s\|^2_{V_s^{-1}}\big)^2 + \big(\eta_s|\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle|\big)^2 + \big(\eta_s\beta_s\|x_s\|^2_{V_s^{-1}} + \eta_s\langle\hat{\nu}_s - \hat{\theta}_s, x_s\rangle^2\big)^2\Big) \\
&\leq \mathcal{O}\Big(\eta_s^2\big(\beta_s\|x_s\|^2_{V_s^{-1}} + \|\hat{\nu}_s(x) - \hat{\theta}_s\|^2_{V_s}\|x_s\|^2_{V_s^{-1}}\big)\Big)
\end{aligned}$$

For the last step we summarize the terms using also that for $J_s = 1$, we have $\beta_s\|x_s\|^2_{V_s^{-1}} \leq 1$. ■

B.4. Asymptotic Regret: Proof of Theorem 5

Proof of Theorem 5 As before, we let $\beta_s = \|\hat{\theta}_s - \theta^*\|^2_{V_s}$ and $B_s = \mathbb{1}(\beta_s \leq \beta_{s,s^2})$. With Lemma 20 we get

$$\mathbb{E}[R_n] \leq \mathbb{E}\Big[\textstyle\sum_{s=1}^{s_n}\Delta(x_s)B_s\Big] + \mathcal{O}\big(\log\log(n)\big)$$

Recall that $m_s = \frac{1}{2}\min_{x\neq\hat{x}_s}\|\hat{\nu}_s(x) - \hat{\theta}_s\|^2_{V_s}$. Let λ be a trade-off parameter, which in hindsight is chosen as $\lambda = \log(n)^{-2/3} \leq \frac{1}{4}$ for n large enough. We decompose the exploration rounds into three disjoint sets, which capture different regimes as $\beta_{s,s^2}/m_s \to 0$ and $\delta_s \to 0$:

$$S_1 = \left\{s \in [s_n] : \tfrac{\beta_{s,s^2}}{m_s} > \lambda, \beta_s \leq \beta_{s,s^2}\right\}$$
$$S_2 = \left\{s \in [s_n] : \tfrac{\beta_{s,s^2}}{m_s} \leq \lambda, \tfrac{\delta_s^2}{16} > \tfrac{\beta_{s,s^2}}{m_s}, \beta_s \leq \beta_{s,s^2}\right\}$$
$$S_3 = \left\{s \in [s_n] : \tfrac{\delta_s^2}{16} \leq \tfrac{\beta_{s,s^2}}{m_s} \leq \lambda, \beta_s \leq \beta_{s,s^2}\right\}$$

In particular, we can write

$$\mathbb{E}\Big[\textstyle\sum_{s=1}^{s_n} \Delta(x_s) B_s\Big] = \mathbb{E}\Big[\textstyle\sum_{s\in S_1} \Delta(x_s)\Big] + \mathbb{E}\Big[\textstyle\sum_{s\in S_2} \Delta(x_s)\Big] + \mathbb{E}\Big[\textstyle\sum_{s\in S_3} \Delta(x_s)\Big]\,.$$

We address the three terms in order.

Sum over S_1: Cauchy-Schwarz and a few applications of the tower rule as before show that

$$\mathbb{E}\big[\textstyle\sum_{s\in S_1} \hat{\Delta}(x_s)\big]^2 \leq \mathbb{E}\left[\textstyle\sum_{s\in S_1} \Psi_s\right] \mathbb{E}\left[\textstyle\sum_{s\in S_1} I_s(x_s)\right]\,.$$

To bound the information-ratio, the definition of S_1 implies the conditions of Lemma 4, which combined with $\delta_s \leq \hat{\Delta}_s(x_s)$ yields

$$\sum_{s\in S_1} \mathbb{E}\big[\Psi_s\big] \leq \mathcal{O}\left(\tfrac{d}{\Delta_{\min}}\right) \sum_{s\in S_1} \mathbb{E}\big[\hat{\Delta}_s(x_s)\big]\,.$$

The total information gain on S_1 is bounded using the same steps as in the proof of Lemma 12,

$$\begin{aligned}
\sum_{s\in S_1} I_s(x_s) &\leq \sum_{s\in S_1} \big(m_s + \tfrac{\log(k)}{\eta_s} + \beta_{s,s^2}\big)\|x_s\|^2_{V_s^{-1}} \\
&\overset{(i)}{\leq} \sum_{s\in S_1} \big(\beta_{s,s^2}(\lambda^{-1}+1) + \tfrac{\log(k)}{\eta_s}\big)\|x_s\|^2_{V_s^{-1}} \\
&\overset{(ii)}{\leq} \mathcal{O}\big(\lambda^{-1} d^2 \log(s_n)^2 + d^{3/2}\log(n)^{1/2}\log(s_n)\big)\,,
\end{aligned}$$

where (i) follows because $m_s < \beta_{s,s^2}\lambda^{-1}$ for $s \in S_1$ and (ii) from the elliptic potential (Lemma 18) and using that $\log(k)\eta_s^{-1} \leq \beta^{1/2}_{s_n,n\log(n)}$. Combining and rearranging the last three displays and using $\Delta(x_s)B_s \leq 2\hat{\Delta}_s(x_s)B_s$ with $B_s = 1$ for $s \in S_1$ yields

$$\mathbb{E}\left[\sum_{s\in S_1} \Delta(x_s)\right] \leq \mathcal{O}\Big(\lambda^{-1}\Delta_{\min}^{-1} d^3 \mathbb{E}[\log(s_n)^2] + \Delta_{\min}^{-1} d^{5/2}\log(n)^{1/2}\mathbb{E}[\log(s_n)]\Big)\,.$$

Sum over S_2: First note that $\beta_s \leq \beta_{s,s^2} < m_s$ implies $\hat{x}_s = \hat{x}^{\mathrm{UCB}} = x^*$. For any $x \in \mathcal{X}$,

$$\begin{aligned}
\beta_{s,s^2}^{-1/2}\delta_s - \|x\|_{V_s^{-1}} &\overset{(i)}{=} \|x^*\|_{V_s^{-1}} - \|x\|_{V_s^{-1}} \overset{(ii)}{\leq} \|x^* - x\|_{V_s^{-1}} \\
&\overset{(iii)}{\leq} \frac{1}{(2m_s)^{1/2} - \beta_s^{1/2}} \overset{(iv)}{\leq} \frac{2}{m_s^{1/2}} \overset{(v)}{<} \frac{\delta_s}{2\beta_{s,s^2}^{1/2}}\,,
\end{aligned} \tag{17}$$

where (i) follows because $\hat{x}_s = x_s^{\text{UCB}} = x^*$, implying that $\delta_s = \beta_{s,s^2}^{1/2}\|x^*\|_{V_s^{-1}}$. (ii) follows from the triangle inequality, (iii) from Lemma 19 and (iv) because $\beta_s \leq m_s/4$. Finally, (v) holds since $\delta_s^2/16 > \beta_{s,s^2}/m_s$. With $x = x_s$ and rearranging yields $\delta_s \leq 2\beta_{s,s^2}^{1/2}\|x_s\|_{V_s^{-1}}$ and hence

$$\sum_{s\in S_2} \mathbb{E}[\hat{\Delta}_s(x_s)] = \sum_{s\in S_2} \mathbb{E}[\hat{\Delta}_s(\mu_s)] \overset{(i)}{\leq} 2\sum_{s\in S_2} \mathbb{E}[\delta_s] \leq 4\sum_{s\in S_2} \mathbb{E}[\beta_{s,s^2}^{1/2}\|x_s\|_{V_s^{-1}}]\,,$$

where (i) uses $\hat{\Delta}_s(\mu_s) \leq 2\delta_s$ (Lemma 6). From here, we can apply Cauchy-Schwarz in a similiar manner as before, to get

$$\begin{aligned}\mathbb{E}\left[\sum_{s\in S_2} \beta_{s,s^2}^{1/2}\|x_s\|_{V_s^{-1}}\right]^2 &\leq \mathbb{E}\left[\sum_{s\in S_2} \beta_{s,s^2}^{1/2} m_s^{-1/2}\right]\mathbb{E}\left[\sum_{s\in S_2} m_s^{1/2}\beta_{s,s^2}^{1/2}\|x_s\|_{V_s^{-1}}^2\right] \\ &\leq \mathbb{E}\left[\sum_{s\in S_2} \hat{\Delta}_s(x_s)\right]\mathcal{O}\big(d^2\log(n)^{1/2}\mathbb{E}[\log(s_n)^2]\big)\,.\end{aligned}$$

For the last inequality, we used that $4\beta_{s,s^2}^{1/2}m_s^{-1/2} \leq \delta_s \leq \hat{\Delta}_s(x_s)$, the elliptic potential (Lemma 18) and $m_s \leq \beta_{s_n,n\log(n)} \leq \mathcal{O}(\log(n) + d\log(s_n))$. Hence, combining the last two displays and $\Delta_s(x_s) \leq 2\hat{\Delta}_s(x_s)$, we get

$$\mathbb{E}\left[\sum_{s\in S_2} \Delta(x_s)\right] \leq \mathcal{O}\big(d^2\log(n)^{1/2}\mathbb{E}[\log(s_n)^2]\big)\,.$$

Sum over S_3: Denote $\bar{\Delta}_s(x) = \langle\hat{\theta}_s, \hat{x}_s - x\rangle$. Note that $\hat{x}_s = x^*$ continues to hold, and hence

$$\mathbb{E}\Big[\textstyle\sum_{s\in S_3} \Delta(x_s)\Big] \leq \mathbb{E}\Big[\textstyle\sum_{s\in S_3} \bar{\Delta}_s(x_s)\Big] + \mathbb{E}\Big[\textstyle\sum_{s\in S_3} \beta_s^{1/2}\|x^* - x_s\|_{V_s^{-1}}\Big]\,. \tag{18}$$

For the second sum, note that by Lemma 13 the information gain of $x_s \neq x^*$ is lower bounded by a constant, $I_s(x_s) \geq \Omega\left(\frac{\Delta_{\min}^2}{d}\right)$. As in (17), Lemma 19 implies

$$\beta_s^{1/2}\|x^* - x_s\|_{V_s^{-1}} \leq 2\beta_s^{1/2}m_s^{-1/2}\mathbb{1}(x_s \neq x^*) \leq \mathcal{O}\big(\lambda^{1/2}d\Delta_{\min}^{-2}I_s(x_s)\big)\,.$$

Summing the last display inside the expectation and using Lemma 12 yields

$$\mathbb{E}\Big[\textstyle\sum_{s\in S_3} \beta_s^{1/2}\|x^* - x_s\|_{V_s^{-1}}\Big] \leq \mathcal{O}\big(\lambda^{1/2}d\log(n)\mathbb{E}[\log(s_n)]\big)\,.$$

For the first sum in (18), we use $4ab \leq (a+b)^2$ and Cauchy-Schwarz combined with a few applications of the towering rule, to get

$$\begin{aligned}\mathbb{E}\left[\sum_{s\in S_3} \bar{\Delta}_s(\mu_s)\right] &\leq \frac{1}{4}\mathbb{E}\left[\sum_{s\in S_3} \delta_s\right]^{-1}\mathbb{E}\left[\sum_{s\in S_3} \hat{\Delta}_s(\mu_s)\right]^2 \\ &\leq \frac{1}{4}\mathbb{E}\left[\sum_{s\in S_3} \delta_s\right]^{-1}\mathbb{E}\left[\sum_{s\in S_3} \Psi_s(\mu_s)\right]\mathbb{E}\left[\sum_{s\in S_3} I_s(x_s)\right]\end{aligned} \tag{19}$$

Lemma 11 bounds the information ratio, $\Psi_s(\mu_s) \leq 4\delta_s(c^* + \mathcal{O}(\delta_s + \beta_s^{1/2} m_s^{-1/2}) \leq 4\delta_s(c^* + \mathcal{O}(\lambda))$, making use of $\delta_s/4 \leq \beta_{s,s^2}^{1/2} m_s^{-1/2} \leq \lambda^{1/2}$. In particular,

$$\frac{1}{4}\mathbb{E}\left[\sum_{s \in S_3} \delta_s\right]^{-1} \mathbb{E}\left[\sum_{s \in S_3} \Psi_s(\mu_s)\right] \leq c^* + \mathcal{O}(\lambda^{1/2})$$

To bound the information gain on S_3, denote $l_s(q_s) = \sum_{x \neq x^*} q_s(x)\langle \hat{\nu}_s(x) - \hat{\theta}_s, x_s\rangle^2$. Note that since $\hat{x}_s = x^*$ on S_3, $l_s(q_s) = I_s(x_s)$. Further, let $J_s = \mathbb{1}(24^2 \eta_s \beta_s \|x_s\|^2_{V_s^{-1}} \leq 1; \beta_s \|x_s\|^2_{V_s^{-1}} \leq 1)$. It is easy to verify that for small enough λ, $J_s = 1$ for all $s \in S_3$. Hence, by Lemma 14 and $m_s \leq \log(n) + \log\log(n) + \mathcal{O}(d \log(s_n))$,

$$\mathbb{E}\left[\sum_{s \in S_3} I_s(x_s)\right] = \mathbb{E}\left[\sum_{s \in S_3} l_s(q_s)\right] \leq \mathbb{E}\left[\sum_{s=1}^{s_n} J_s l_s(q_s)\right] \leq \log(n) + \mathcal{O}\big(\log(n)^{1/2}\mathbb{E}[\log(s_n)^2]\big)$$

Combing the bounds on the information ratio and the information gain, we get

$$\mathbb{E}\left[\sum_{s \in S_3} \bar{\Delta}_s(\mu_s)\right] \leq \big(c^* + \mathcal{O}(\lambda^{1/2})\big)\big(\log(n) + \mathcal{O}(\log(n)^{1/2}\mathbb{E}[\log(s_n)^2])\big)$$

Hence we conclude

$$\mathbb{E}\left[\sum_{s \in S_3} \Delta(s)\right] \leq c^* \log(n) + \mathcal{O}\big(\lambda^{1/2} \log(n)\big)\,.$$

Finally, with Lemma 21, we get that $\mathbb{E}[\log(s_n)^b] \leq \mathcal{O}(\log\log(n))$. Therefore, with $\lambda = \log(n)^{-2/3}$ all terms except for $c^* \log(n)$ are of lower order and the claim follows. ■

B.5. Technical Lemmas

Lemma 18 (Elliptic potential lemma) *Assume that* $\|x_s\|^2_{V_s^{-1}} \leq 1$ *and* $\|x_s\|_2 \leq 1$. *Then*

$$\sum_{s=1}^{s_n} \|x_s\|^2_{V_s^{-1}} \leq 2 \log\det(V_{s_n}) \leq 2d \log\left(\frac{s_n + d}{d}\right)$$

A proof can be found in (Abbasi-Yadkori et al., 2011, Lemma 11). Note that by $\mathrm{diam}(\mathcal{X}) \leq 1$ and the choice $V_0 = \mathbf{1}_d$, the assumptions of the lemma are always satisfied for our setting.

Lemma 19 *Let* $\beta_s = \|\hat{\theta}_s - \theta^*\|^2_{V_s}$ *and* $m_s = \frac{1}{2}\min_{x \neq \hat{x}_s} \|\hat{\nu}(x) - \hat{\theta}_s\|^2_{V_s}$. *Assume that* $\beta_s < 2m_s$ *and* $\max_{x \in \mathcal{X}} \Delta(x) \leq 1$. *Then* $\hat{x}_s = x^*$ *and further, for all* $x \in \mathcal{X}$,

$$\big((2m_s)^{1/2} - \beta_s^{1/2}\big)\|x^* - x\|_{V_s^{-1}} \leq 1\,.$$

Proof Since $m_s = \frac{1}{2}\min_{x\neq\hat{x}_s}\min_{\nu\in\mathcal{C}_x}\|\nu-\hat{\theta}_s\|_{V_s}^2$, the assumption that $\beta_s = \|\hat{\theta}_s-\theta^*\|_{V_s}^2 < 2m_s$ implies that $\theta^*\in\mathcal{C}_{\hat{x}_s}$, and therefore $\hat{x}_s = x^*$. Further, for any $x\in\mathcal{X}$,

$$0 = \min_{\nu:\|\nu-\hat{\theta}_s\|_{V_s}^2\leq 2m_s}\langle\nu, x^*-x\rangle = \langle\hat{\theta}, x^*-x\rangle - (2m_s)^{1/2}\|x^*-x\|_{V_s^{-1}}$$
$$\leq \langle\theta^*, x^*-x\rangle + (\|\hat{\theta}_s-\theta^*\|_{V_s} - (2m_s)^{1/2})\|x^*-x\|_{V_s^{-1}}\,.$$

Using $\Delta(x) = \langle\theta^*, x^*-x\rangle \leq 1$ and rearranging completes the proof. ■

Lemma 20 *Let $\beta_s = \|\hat{\theta}_s-\theta^*\|_{V_s}^2$ and define the indicator $B_s = \mathbb{1}(\beta_{s,s^2}\geq\beta_s)$ for rounds s where the confidence bounds at level β_{s,s^2} are valid. Assume that $\max_{x\in\mathcal{X}}\Delta(x)\leq 1$. Then*

$$R_n \leq \mathbb{E}\left[\sum_{s=1}^{s_n}\Delta_s(x_s)B_s\right] + \mathcal{O}\big(\log\log(n)\big)\,.$$

Proof Naturally, the regret decomposes into exploration and exploitation rounds. When $\beta_s > \beta_{s,s^2}$ (in exploration rounds, indexed by local time s) or $\beta_{s_t} > \beta_{s_t,t\log t}$ (in exploitation rounds, indexed by global time t), the parameter estimate is too inaccurate to bound the regret, and we simply bound $\Delta(x)\leq 1$. On the other hand, in exploitation rounds where $\beta_{s_t}\leq\beta_{s_t,t\log t}$, by the definition of an exploitation round, it holds that $m_{s_t}\geq\beta_{s_t,t\log t}\geq\beta_{s_t}$ and by Lemma 19 this implies that $\hat{x}_s = x^*$ and the regret vanishes. Hence,

$$R_n = \sum_{t=1}^n\Delta(x_t) \leq \sum_{s=1}^{s_n}\Delta(x_s)B_s + \sum_{s=1}^{s_n}\mathbb{1}(\beta_{s,s^2}<\beta_s) + \sum_{t=1}^n\mathbb{1}(\beta_{s_t,t\log t}<\beta_{s_t})$$

Note that by (4), we have $\mathbb{P}[\beta_{s,s^2}<\beta_s]\leq 1/s^2$ and $\mathbb{P}[\beta_{s_t,t\log t}<\beta_{s_t}] < \frac{1}{t\log t}$. Hence, in expectation we get

$$\mathbb{E}[R_n] \leq \mathbb{E}\left[\sum_{s=1}^{s_n}\Delta(x_s)B_s + \sum_{s=1}^{s_n}\frac{1}{s^2} + \sum_{t=1}^n\frac{1}{t\log t}\right]$$
$$\leq \mathbb{E}\left[\sum_{s=1}^{s_n}\Delta(x_s)B_s\right] + \mathcal{O}(\log\log(n))\,.$$

■

Lemma 21 *Assume that $\|x^*\|_2 > 0$. Then the number of exploration steps s_n in Algorithm 1 is bounded in expectation,*

$$\mathbb{E}[s_n^{1/2}] \leq \mathcal{O}\left(d^2\Delta_{\min}^{-1}\log(n)^2\|x^*\|_2^{-1}\right)\,.$$

In particular, for any $b\geq 1$, we have $\mathbb{E}[\log(s_n)^b]\leq\mathcal{O}(\log\log(n))$.

Proof By Theorem 3,

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \delta_s\right] \le \mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(x_s)\right] \le \mathcal{O}\left(d^2 \Delta_{\min}^{-1} \|x^*\|_2^{-1} \log(n)^2\right) .$$

We can assume that $2\delta_s < \Delta_{\min}$, since there can be at most $\mathcal{O}\left(d^2 \Delta_{\min}^{-2} \log(n)^2\right)$ steps where this condition is not satisfied. In particular, the assumption implies that $x^* = \hat{x}_s$, since for all $x \neq x^*$, $2\hat{\Delta}_s(x) \ge \Delta_{\min}$. Therefore,

$$\delta_s = \max_{z \in \mathcal{X}} \langle z - x^*, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|z\|_{V_s^{-1}} \ge \beta_{s,s^2}^{1/2} \|x^*\|_{V_s^{-1}} \ge \|x^*\|_2 s^{-1/2} .$$

The last inequality follows from since $\lambda_{\max}(V_s) \le s$. Hence further

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \delta_s\right] \ge \|x^*\| (s_n^{1/2} - \mathcal{O}\left(d \Delta_{\min}^{-1} \log(n)^1\right)) .$$

This proves the first claim. For the second part, note that $\log(s)^b$ is concave for $s \ge \exp(b-1)$. Hence

$$\begin{aligned} \mathbb{E}[\log(s_n)^b] = 2^b \mathbb{E}[\log(s_n^{1/2})^b] &\le 2^b \mathbb{E}[\log\left(\max(s_n^{1/2}, \exp(b-1))\right)^b] \\ &\le 2^b \log(\mathbb{E}[s_n^{1/2}] + \exp(b-1)) \\ &\le \mathcal{O}(\log\log(n)) \end{aligned}$$

■

Lemma 22 (Softmin approximation) $A_1, \ldots A_k \ge 0$ *be a sequence of positive numbers and* $a = \min_{i \in [k]} A_i$. *Let* $q_i(x) \propto \exp(-\eta A_i)$ *be exponential mixing weights with* $\eta > 0$. *Then*

$$\sum_{i \in [k]} q_i A_i \le a + \frac{\log(k)}{\eta} .$$

Further, the mixing weights q_i *are bounded as follows,*

$$\frac{1}{k} \exp\left(-\eta(A_i - a)\right) \le q_i \le \exp\left(-\eta(A_i - a)\right) .$$

Proof Let $\psi_\eta^*(A) = \frac{1}{\eta} \log\left(\sum_{i \in [k]} \exp(\eta A_i))\right)$ be the Fenchel conjugate of the normalized entropy function. A direct calculation confirms that

$$q = \nabla_A \psi_\eta^*(-A) .$$

By convexity of ψ_η^*,

$$\sum_i q_i A_i = \langle \nabla \psi_\eta^*(-A), A \rangle \le \psi_\eta^*(0) - \psi_\eta^*(-A) \le \tfrac{1}{\eta} \log(k) + \min_i A_i .$$

The last inequality follows from

$$\psi_\eta^*(-A) = \eta^{-1} \log \Big(\sum_i \exp(-\eta A_i) \Big) \geq \eta^{-1} \log \Big(\exp(-\eta \min_i A_i) \Big) = -\min_i A_i \,.$$

For the bound on the mixing weights, note that the claim is equivalent to the following bound on the normalization constant,

$$\exp(-\eta a) \leq \sum_i \exp(-\eta A_i) \leq k \exp(-\eta a) \,.$$

■

Lemma 23 (Convex Polytopes) *Let K be a convex polytope. For unit vector $\eta \in \mathbb{R}^d$, let $K_0 = \{x \in K : \langle x, \eta \rangle = 0\}$ be the intersection of k with a $(d-1)$-dimensional hyperplane, which is assumed to be non-empty. Then there exists a constant $c > 0$ such that for all $z \in K$,*

$$\min_{x \in K_0} \|x - z\|_2 \leq c \langle z, \eta \rangle \,.$$

Proof Let $A = \{x \in K : \langle x, \eta \rangle \geq 0\}$, which is also a convex polytope. We first show there exists a $c > 0$ such that for all $z \in A$,

$$\min_{x \in K_0} \|x - z\|_2 \leq c \langle z, \eta \rangle \,. \tag{20}$$

The result follows by making a symmetric argument for $\{x \in K : \langle x, \eta \rangle \leq 0\}$. To establish (20), let $V \subset \mathbb{R}^d$ be the vertices of A, which is a finite set. Define $h : A \setminus K_0 \to \mathbb{R}$ by

$$h(z) = \max_{x \in K_0} \frac{\langle \eta, z - x \rangle}{\|z - x\|} \,.$$

Clearly, $1/c \triangleq \min_{v \in V : \langle v, \eta \rangle > 0} h(v) > 0$. Hence, the mapping $\varphi : V \to K_0$ such that $\varphi(v) = v$ for $v \in K_0$ and $\varphi(v) = \arg\max_{x \in \mathcal{K}_0} \frac{\langle \eta, v - x \rangle}{\|v - x\|}$ satisfies $\|v - \varphi(v)\|_2 \leq c \langle \eta, v - \varphi(v) \rangle$. Given any $z \in A$, let α be a probability distribution on V such that $z = \sum_{v \in V} \alpha(v) v$ and let $x = \sum_{v \in V} \alpha(v) \varphi(v) \in K_0$. Then,

$$\begin{aligned}
\|z - x\|_2 &= \left\| \sum_{v \in V} \alpha(v) v - \sum_{v \in V} \alpha(v) \varphi(v) \right\|_2 \\
&\leq \sum_{v \in V} \alpha(v) \, \|v - \varphi(v)\|_2 \\
&\leq c \sum_{v \in V} \alpha(v) \langle \eta, v - \varphi(v) \rangle \\
&= c \langle \eta, z \rangle \,.
\end{aligned}$$

■

Lemma 24 *The one-step update to the least-squares estimator with data $y_s = \langle x_s, \theta^* \rangle + \epsilon_s$ is*

$$\hat{\theta}_{s+1} - \hat{\theta}_s = V_s^{-1} x_s \left(\frac{\epsilon_s + x_s^\top (\theta^* - \hat{\theta}_s)}{1 + \|x_s\|^2_{V_s^{-1}}} \right) .$$

Proof The difference can be computed with the Sherman-Morrison-Woodbury formula,

$$\begin{aligned}
\hat{\theta}_{s+1} - \hat{\theta}_s &= V_{s+1}^{-1} \sum_{i=1}^{s} x_i y_i - \hat{\theta}_s \\
&= V_s^{-1} \sum_{i=1}^{s-1} x_i y_i + V_s^{-1} x_s y_s - \frac{V_s^{-1} x_s x_s^\top V_s^{-1}}{1 + \|x_s\|^2_{V_s^{-1}}} \sum_{i=1}^{s} x_i y_i - \hat{\theta}_s \\
&= V_s^{-1} x_s y_s - \frac{V_s^{-1} x_s \|x_s\|^2_{V_s^{-1}} y_s}{1 + \|x_s\|^2_{V_s^{-1}}} - \frac{V_s^{-1} x_s x_s^\top \hat{\theta}_s}{1 + \|x_s\|^2_{V_s^{-1}}} \\
&= V_s^{-1} x_s \left(y_s - \frac{\|x_s\|^2_{V_s^{-1}} y_s}{1 + \|x_s\|^2_{V_s^{-1}}} - \frac{x_s^\top \hat{\theta}_s}{1 + \|x_s\|^2_{V_s^{-1}}} \right) \\
&= V_s^{-1} x_s \left(\frac{y_s - x_s^\top \hat{\theta}_s}{1 + \|x_s\|^2_{V_s^{-1}}} \right) \\
&= V_s^{-1} x_s \left(\frac{\epsilon_s + x_s^\top (\theta^* - \hat{\theta}_s)}{1 + \|x_s\|^2_{V_s^{-1}}} \right) .
\end{aligned}$$

■

Appendix C. Information-Directed Sampling as a Primal-Dual Method

This section serves as self-contained exposition to establish the link between information-directed sampling and primal-dual approaches used to solve the lower bound (2). *Note that in this section, quantities such as $\hat{\Delta}_t$, δ_t and I_t are re-defined independently of the main text.*

For simplicity, for the remainder of this section we fix finitely many alternative parameters $\nu_1, \dots, \nu_l \in \mathcal{M}$ for which $x^*(\nu) \neq x^*(\theta^*)$. Define constraint vectors $h_j \in \mathbb{R}^{\mathcal{X}}$ as $h_j(x) = \frac{1}{2}\langle \nu_j - \theta^*, x \rangle^2$ for each $x \in \mathcal{X}$ and $j = 1, \dots, l$. Our boundedness assumptions imply $\|h_j\|_2 \leq 1$. With this notation, the lower bound (2) can be written as a *linear covering program*,

$$c^* = \inf_{\alpha \in \mathbb{R}^k_{\geq 0}} \sum_{x \in \mathcal{X}} \alpha(x) \langle x^* - x, \theta^* \rangle \qquad \text{s.t.} \qquad \forall j = 1, \dots, l, \quad h_j(\alpha) \geq 1 . \tag{21}$$

It is immediate from the assumption that $\mathcal{X}$ spans $\mathbb{R}^d$ that the program is feasible. Further, there is no cost for playing the optimal action x^* since the corresponding gap is zero, $\Delta(x^*) = 0$. Following the terminology of Jun and Zhang (2020), we refer to a constraint h_j with $h_j(x^*) > 0$ as *docile*. Such constraints are trivially satisfied by letting $\alpha(x^*) \to \infty$, while the regret from allocating x^* remains zero in the limit. To simplify our exposition further, here we assume that there are *no docile constraints*, i.e. $h_j(x^*) = 0$ holds for all $j = 1, \dots, l$.

The objective of this section is to derive sequential strategies to solve (21) in the *oracle setting*, where the exact cost and constraint vectors are known. Thereby, we set aside all complications that arise in the statistical setting. Specifically, we seek to incrementally determine a sequence of distributions $\mu_1, \dots \mu_n \in \mathscr{P}(\mathcal{X})$ over actions, which define a cumulative allocation $\alpha_n = \sum_{t=1}^n \mu_t$. We say an allocation is *asymptotically optimal* at rate β_n if

$$\lim_{n\to\infty} \frac{\Delta(\alpha_n)}{\beta_n} = c^*, \quad \text{and} \quad \forall j = 1, \dots, l, \lim_{n\to\infty} \frac{h_j(\alpha_n)}{\beta_n} \geq 1\,. \tag{22}$$

The lower bound suggests a choice which satisfies $\lim_{n\to\infty} \beta_n = \log(n)$.

Online Convex Optimization We review an approach due to Garg and Koenemann (2007); Arora et al. (2012), which solves covering LPs – such as the oracle lower bound – using *online convex optimization* (OCO). The same idea has recently inspired bandit algorithms for best arm identification (Degenne et al., 2019) and regret minimization (Degenne et al., 2020). The approach sets up a fictitious two-player game that converges to the saddle point of the Lagrangian,

$$\max_{\lambda \geq \mathbb{R}^l_{\geq 0}} \min_{\alpha \in \mathbb{R}^k_{\geq 0}} \left\{ \mathcal{L}(\alpha, \lambda) = \Delta(\alpha) - \sum_{j=1}^{l} \lambda_j (h_j(\alpha) - 1) \right\}\,.$$

It is easy to verify that strong duality holds, and we can interchange the maximum and minimum. Note that the dual variables are on an unbounded space, but it turns out that we can normalize them. The KKT conditions are

$$\begin{aligned}
\Delta(x) - \sum_{j=1}^{l} \lambda_j h_j(x) &= 0 && \text{(stationarity)}\\
\lambda_j (h_j(\alpha) - 1) &= 0 && \text{(complementary slackness)}
\end{aligned}$$

Combining both, we find that $c^* = \sum_{j=1}^l \lambda_j$. This implies that the optimal cost c^* normalizes the dual variables $q_j = \lambda_j / c^*$. The normalized Lagrangian is

$$\mathcal{L}(\alpha, q) = \Delta(\alpha) - c^* \sum_{j=1}^{l} q_j (h_j(\alpha) - 1)\,, \tag{23}$$

where $q \in \mathscr{P}([l])$ is a distribution over the constraints. Recall that the allocation $\alpha_n = \sum_{t=1}^n \mu_t$ is chosen sequentially. In each iteration of the game, the *first player*, or q-learner, chooses a distribution $q_t \in \mathscr{P}([l])$ over the constraints. Then, the response of the *second player* is a distribution $\mu_t \in \mathscr{P}(\mathcal{X})$ over actions, which defines the allocation $\alpha_n = \sum_{t=1}^n \mu_t$. The loss of the q-learner is defined by the second player's response μ_t,

$$l_t(q) = \sum_{j=1}^{l} q_t(j) h_j(\mu_t)\,, \tag{24}$$

which is linear in the dual variable q_t. The loss sequence defines the q-learner regret Λ_n (not to be confused with R_n), which is

$$\Lambda_n = \sum_{t=1}^{n} l_t(q_t) - \min_{q \in \mathscr{P}([l])} \sum_{t=1}^{n} l_t(q)\,. \tag{25}$$

For concreteness, we choose the exponential weights learner (Vovk, 1990; Littlestone and Warmuth, 1994),

$$q_t(j) \propto \exp\left(-\eta_t \sum_{s=1}^{t-1} l_t(j)\right)$$

with learning rate η_t. Standard regret bounds for online convex optimization guarantee $\Lambda_n \leq \mathcal{O}(\sqrt{n})$ for suitably chosen learning rate schedules. More refined techniques lead to first-order regret bounds, which scale with the best loss in hindsight $\Lambda_n \leq \mathcal{O}(\sqrt{\min_i \sum_{t=1}^n l_t(i)})$, see for example (Cesa-Bianchi et al., 2005). Given the choice q_t of the q-learner, we define the combined constraint $I_t = \sum_{j=1}^l q_t(j) h_j$. The policy response is defined as

$$\mu_t = e_{x_t}, \quad \text{where} \quad x_t = \begin{cases} \arg\min_{x \in \mathcal{X} \setminus x^*} \frac{\Delta(x)}{I_t(x)} & \text{if } \min_j \alpha_{t-1}^\top h_j < \beta_n \\ x_t = e_{x^*} & \text{else.} \end{cases} \tag{26}$$

The second case corresponds to *exploitation*, which happens as soon as the constraints are satisfied:

$$\min_j h_j^\top \alpha_{t-1} = \min_j \sum_x \alpha_{t-1}(x) \langle \nu_j - \theta^*, x \rangle^2 \geq \beta_n \,.$$

Note that x^* is the only action which does not incur cost. On the contrary, when $\min_j h_j(\alpha_t) < \beta_n$, the policy allocates on the suboptimal action $x_t \neq x^*$ with the best cost/constraint ratio, $\min_{x \neq x^*} \Delta(x)/I_t(x)$. Since there are no docile constraints, we have $I_t(x^*) = 0$ and $\mu_t = e_{x_t}$ corresponds to the *optimal* allocation for the rescaled linear program with the single combined constraint $I_t = \sum_{j=1}^l q_t(j) h_j$,

$$\min_{\alpha \in \mathbb{R}_{\geq 0}^{\mathcal{X}}} \Delta(\alpha) \quad \text{s.t.} \quad I_t(\alpha) \geq I_t(x_t) \,.$$

Rescaling the optimal solution α^* to the original covering program (21), we obtain an upper bound to the cost of choosing $\mu_t = e_{x_t}$,

$$\Delta(\mu_t) \leq \Delta(I_t(\mu_t)\alpha^*) = c^* I_t(\mu_t) \,.$$

Since $I_t(\mu_t) = l_t(q_t)$, we can make use of the regret bound for the q-learner,

$$\sum_{t=1}^n I_t(\mu_t) = \sum_{t=1}^n l_t(q_t) \leq \min_j \alpha_n^\top h_j + \Lambda_n \leq \beta_n + \mathcal{O}\big(\beta_n^{1/2}\big) \,. \tag{27}$$

For the last inequality, we used that $\min_j \alpha_n^\top h_j \leq \beta_n + 1$ is guaranteed by the definition (26) and boundedness, $I_t(x) \leq 1$. Further, we assume a first-order regret bound $\Lambda_n \leq \mathcal{O}(\beta_n^{1/2})$ for the q-learner. From here, we easily bound the regret R_n of the allocation α_n,

$$R_n = \langle \alpha_n, \Delta \rangle = \sum_{t=1}^n \Delta(\mu_t) \leq c^* \sum_{t=1}^n I_t(x_t) \leq c^* \beta_n + \mathcal{O}\big(c^* \beta_n^{1/2}\big) \,.$$

With some care, this approach can be translated to a bandit algorithm, by replacing all unknown quantities with statistical estimates, see Degenne et al. (2020). The formulation presented here differs from previous work in that it avoids a re-parametrization of the allocation, and the argument to bound the regret is more direct. We are now in the position to establish a link between information-directed sampling and the two-payer minimax game setup.

Oracle Information-Directed Sampling For reasons that become apparent soon, we refer to the combined constraints $I_t = \sum_{j=1}^{l} q_t(j)h_j \in \mathbb{R}^{\mathcal{X}}_{\geq 0}$ as the *information gain*, where q_t is the output of the same q-learner as before. We also introduce a positive *error* term $\delta_t > 0$ that is added to the gaps, to obtain *approximate gaps* $\hat{\Delta}_t(x) = \Delta(x) + \delta_t$. This choice anticipates the definition for the gap estimate, which we use later in the bandit setting. Moreover, $\delta_t > 0$ avoids a degenerate regret-information trade-off and allows us to treat all actions in a unified manner.

Information-directed sampling approaches the regret minimization problem by sampling actions from a distribution that minimizes the *information ratio*,

$$\mu_t = \underset{\mu \in \mathscr{P}(\mathcal{X})}{\arg\min} \left\{ \Psi_t(\mu) = \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \right\} .$$

We follow this strategy as long as $\min_j h_j(\alpha_t) < \beta_n$. Once the constraints are satisfied, we resort to playing the optimal action x^* as before. This allows to bound the q-learner regret Λ_n, and therefore the total information gain as in (27). We make the assumption that the estimation gap δ_t is small compared to the minimum gap $\Delta_{\min} = \min_{x \neq x^*} \Delta(x)$,

$$2\delta_t \leq \min_{x \neq x^*} \hat{\Delta}_t(x) \tag{28}$$

or equivalently, $\delta_t \leq \Delta_{\min}$. At first sight, the IDS distribution does not relate to the previous analysis, since the ratio appears with the cost squared. However, a closer inspection reveals a strong connection, which is summarized in the following lemma.

Lemma 25 *Let $\mu_t = \arg\min_{\mu \in \mathscr{P}(\mathcal{X})} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)}$ be the IDS distribution. If $2\delta_t \leq \min_{x \neq x^*} \hat{\Delta}_t(x)$ and $I_t(x^*) = 0$, then $\mu_t = (1 - p_t)e_{x^*} + p_t e_{z_t}$ with alternative action $z_t = \arg\min_{z \in \mathcal{X}} \frac{\Delta(z)}{I_t(z)}$ and trade-off probability $p_t = \frac{\delta_t}{\hat{\Delta}_t(z_t) - \delta_t} = \frac{\delta_t}{\Delta(z_t)}$.*

Proof Let $\psi(p) = \frac{((1-p)\hat{\Delta}_t(\mu_t) + p\delta_t)^2}{(1-p)I_t(\mu_t)}$ be the ratio obtained from shifting probability mass to x^*. By definition of the IDS distribution, we must have

$$0 \leq \frac{d}{dp}\psi(p)|_{p=0} = \frac{2\hat{\Delta}_t(\mu_t)\delta_t - \hat{\Delta}_t(\mu_t)^2}{I_t(\mu_t)} .$$

Re-arranging shows that $\hat{\Delta}_t(\mu_t) \leq 2\delta_t$. The IDS distribution can always be chosen with a support of at most two actions, which is a result by Russo and Van Roy (2014, Proposition 6). With the condition $2\delta_t \leq \min_{x \neq x^*} \hat{\Delta}_t(x)$, it therefore suffices to optimize over distributions $\mu(p, z) = (1 - p)e_{x^*} + pe_z$. A simple calculation reveals that $\arg\min_{p \in [0,1]} \Psi_t(\mu(p, z)) = \frac{\delta_t}{\hat{\Delta}_t(z) - \delta_t}$, and

$$\min_{\mu} \Psi_t(\mu) = \min_{z \neq x^*} \min_{p \in [0,1]} \Psi(\mu(p, z)) = \min_{z \neq x^*} \frac{4\delta_t(\hat{\Delta}_t(z) - \delta_t)}{I_t(z)} .$$

Therefore the alternative action is $z_t = \arg\min_{z \neq x^*} \left\{ \frac{\hat{\Delta}_t(z) - \delta_t}{I_t(z)} = \frac{\Delta(z)}{I_t(z)} \right\}$. ■

The lemma shows that $\text{supp}(\mu_t) = \{x^*, z_t\}$ and the alternative action $z_t \neq x^*$ minimizes the same cost-to-constraint ratio as before. Hence, almost the same argument implies a regret bound $R_n \leq c^* \beta_n + \mathcal{O}(c^* \beta_n^{1/2})$.

Unlike the approach presented before, the distribution μ_t allocates mass to the zero-cost action, even before the constraint threshold β_n is reached. Importantly, the randomization also allows to bound the regret in a worst-case manner. In the statistical setting, we expect that the estimation error roughly decreases at a rate $\delta_t \approx t^{-1/2}$. With the trade-off probability $p_t = \frac{\delta_t}{\Delta(z_t)}$ the expected cost per round is $\Delta(\mu_t) = \delta_t$. In other words, we get a finite-time, problem-independent bound on the regret R_n,

$$R_n = \Delta(\alpha_n) \leq \sum_{t=1}^{n} \delta_t \leq \mathcal{O}(\sqrt{n}) \, .$$

Lastly, we link our analysis to the standard Cauchy-Schwarz argument that appears in all previous regret bounds for IDS (c.f. Russo and Van Roy (2014); Kirschner and Krause (2018)). A direct calculation using the trade-off probability p_t reveals that the expected approximate cost of the IDS distribution is exactly two times the actual cost and equals the estimation gap,

$$\frac{1}{2}\hat{\Delta}_t(\mu_t) = \delta_t = \Delta(\mu_t) \, .$$

Note that exact equality only holds because we have $I_t(x^*) = 0$ (no docile constraints). We continue to bound the regret with the Cauchy-Schwarz inequality and using the definition of the information ratio $\Psi_t = \frac{\hat{\Delta}_t(\mu_t)^2}{I_t(\mu_t)}$,

$$R_n = \sum_{t=1}^{n} \Delta(\mu_t) = \frac{1}{2}\sum_{t=1}^{n} \hat{\Delta}_t(\mu_t) \leq \frac{1}{2}\sqrt{\sum_{t=1}^{n} \Psi_t \sum_{t=1}^{n} I_t(\mu_t)} \, .$$

Let $\tilde{\alpha}^* = \alpha^* \mathbb{1}(x \neq x^*)$ be the optimal allocation (21) restricted to suboptimal actions. Note that by definition and the fact that we excluded docile constraints, $I_t(\tilde{\alpha}^*) \geq 1$ and $\Delta(\tilde{\alpha}^*) = c^*$. Define a distribution $\mu(p) = (1-p)e_{x^*} + p\tilde{\alpha}^*/\|\tilde{\alpha}^*\|_1$, which randomizes between the best action and optimal allocation with trade-off probability p. A simple calculation reveals that,

$$\Psi_t = \min_{\mu} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \leq \min_{p} \frac{\hat{\Delta}_t(\mu(p))^2}{I_t(\mu(p))} \leq \frac{4\delta_t \Delta(\alpha^*)}{I_t(\tilde{\alpha}^*)} \leq 4c^*\delta_t \, .$$

We combine the inequality and the regret bound for the q-learner to get

$$R_n = \sum_{t=1}^{n} \Delta(\mu_t) \leq \frac{1}{2}\sqrt{\sum_{t=1}^{n} 4c^*\delta_t \left(\beta_n + \mathcal{O}(\beta_n^{1/2})\right)} \, .$$

Squaring both sides, using again that $\delta_t = \Delta(\mu_t)$ and solving for the regret yields the desired bound,

$$R_n = \sum_{t=1}^{n} \Delta(\mu_t) \leq c^*\beta_n + \mathcal{O}(c^*\beta_n^{1/2}) \, .$$

Appendix D. Approximating Mutual Information

The information gain function that was primarily analyzed in the Bayesian framework by Russo and Van Roy (2014) is the mutual information

$$I_t^{\mathrm{MI}}(x) = \mathbb{I}_t(y_t; x^*|x_t = x) = \mathbb{H}_t(x^*) - \mathbb{H}_t(x^*|y_t, x_t = x)\,.$$

The second equality rewrites the mutual information as the entropy reduction on x^*, which is a random variable in the Bayesian setting. Computation of the posterior distribution is tractable with a Gaussian prior $\mathcal{N}(0, \lambda^{-1})$ on the parameter and Gaussian observation likelihood $y_t \sim \mathcal{N}(\langle x_t, \theta\rangle, 1)$. In this case the posterior distribution is $\mathcal{N}(\hat{\theta}_t, V_t^{-1})$. However, computing the mutual information requires further evaluations of d-dimensional integrals which is challenging even with Gaussian distributions.

As a remedy, Russo and Van Roy (2014) proposed a *variance-based* information gain

$$I_t^{\mathrm{VAR}}(x) \stackrel{\mathrm{def}}{=} \mathbb{E}_t\big[\big(\mathbb{E}_t[\langle x, \theta\rangle|x^*] - \mathbb{E}_t[\langle x, \theta\rangle]\big)^2\big] = \mathbb{E}_t[\langle \bar{\nu}_t(x^*) - \hat{\theta}_t, x\rangle^2]\,. \tag{29}$$

The last step uses that $\mathbb{E}_t[\theta] = \hat{\theta}_t$ and we defined $\bar{\nu}_t(x) = \mathbb{E}_t[\theta|x^* = x]$. They further showed that the variance-based information gain lower-bounds the mutual information, $I_t^{\mathrm{MI}}(x) \geq 2I_t^{\mathrm{VAR}}(x)$, while, at the same time, the information ratio is still bounded in the Bayesian setting with linear reward (Russo and Van Roy, 2014, Proposotion 7). Importantly, (29) can be approximated for a moderate number of actions using samples from the posterior distribution.

We compute the posterior probability $\bar{q}_t(c) \stackrel{\mathrm{def}}{=} \mathbb{P}_t[x^* = z]$ with a Laplace approximation of the integral over the cell $\mathcal{C}_z = \{\theta \in \mathcal{M} : x^*(\theta) = z\}$,

$$\bar{q}_t(z) = \frac{1}{\sqrt{(2\pi)^d \det(V_t)}} \int_{\mathcal{C}_z} \exp\left(-\tfrac{1}{2}\|\nu - \hat{\theta}_t\|_{V_t}^2\right) d\nu \approx Q_z^{-1} \exp\left(-\tfrac{1}{2}\|\tilde{\nu}_t(z) - \hat{\theta}_t\|_{V_t}^2\right)\,,$$

where $\tilde{\nu}_t(x) = \arg\min_{\nu \in \mathcal{C}_x} \|\nu - \hat{\theta}_s\|_{V_s}^2$. Similarly, in the Laplace limit, the conditional distribution $\mathbb{P}_t[\theta|x^* = x]$ concentrates on $\tilde{\nu}_t(x)$, which allows us to approximate $\bar{\nu}_t(x) \approx \tilde{\nu}_t(x)$. This leads to

$$I_t^{\mathrm{VAR}}(x) \approx \sum_{z \neq x^*} \bar{q}_t(z)\langle \tilde{\nu}_t(x) - \hat{\theta}_t, x\rangle^2\,,$$

which resembles the definition of the cell-based information gain in (14).

Using the Laplace argument, we can also compute the mutual information more directly. Assuming that the posterior is well-concentrated, there exists an action $\bar{x}_t^*$ with $\bar{q}_t(\bar{x}_t^*) \approx 1$. For all $z \neq \bar{x}_t^*$ and interpolation variable $\tau \in [0, 1]$, we define the conditional weights

$$\bar{q}_t^\tau(z|x) \stackrel{\mathrm{def}}{=} \bar{q}_t(z) \exp\left(-\tfrac{\tau}{2}\langle \tilde{\nu}_t(z) - \hat{\theta}_t, x\rangle^2\right)\,,$$

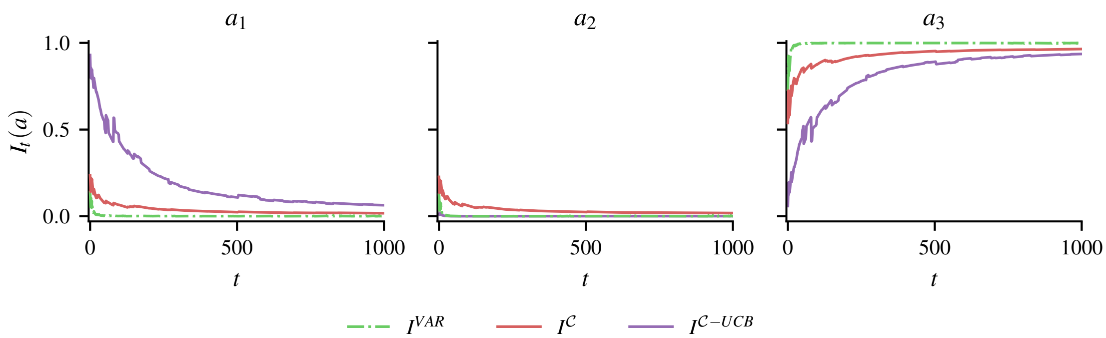

Figure 3: Comparison of information gain functions on the 'end of optimism' example with $\epsilon = 0.01$. The information gain functions are evaluated on the same trajectory generated by IDS-$I_t^{\mathcal{H}\text{-UCB}}$ and normalized such that $\sum_{x \in A} I_t(x) = 1$. On this instance, x_1 is optimal, x_2 is ϵ-suboptimal, and x_3 is 1-suboptimal, but asymptotically more informative than action x_2. Clearly visible is that the lower-order terms of the $I_t^{\mathcal{H}}$ and $I_t^{\mathcal{H}\text{-UCB}}$ are increasingly dominated by the asymptotic term where x_3 is the most informative action. I_t^{VAR} is approximated using 10^4 samples from the posterior distribution, and converges much faster than the information gain functions based on the q-learner, which uses a more conservative learning rate. Note that the approximation with posterior samples is unstable on a larger horizon without increasing the number of samples accordingly.

and $q_t^\tau(\bar{x}_t^*|x) \overset{\text{def}}{=} 1 - \sum_{z \neq \bar{x}_t^*} q_t^\tau(z|x)$. Using the approximate posterior probabilities, the entropy reduction up to first order is

$$
\begin{aligned}
\mathbb{I}_t(y_t; x^*|x_t = x) &\approx -\sum_{z \in \mathcal{A}} \bar{q}_t(z) \log \bar{q}_t(z) + \sum_{z \in \mathcal{A}} \left(\bar{q}_t^\tau(z|x) \log(\bar{q}_t^\tau(z|x))\right)\big|_{\tau=1} \\
&\approx \sum_{z \in \mathcal{A}} \frac{d}{d\tau} \left(\bar{q}_t^\tau(z|x) \log(\bar{q}_t^\tau(z|x))\right)\big|_{\tau=1} \\
&= -\frac{1}{2} \sum_{z \neq \bar{x}_t^*} \bar{q}_t(z) \langle \nu_z - \theta, x\rangle^2 \log\left(\frac{\bar{q}_t(z)}{1 - \sum_{z' \neq \bar{x}_t^*} \bar{q}_t(z')}\right).
\end{aligned}
$$

Using that $-x \log x \geq x$ for $x \ll 1$, the last expression can be lower bounded to arrive at a form similar to the cell-based information gain (14).

While our reasoning here is rather informal, we think that it warrants a more formal investigation in the future. Such results could be fruitful in two directions. First, interpreting the mutual information as an approximation of a dual loss could lead to an instance-dependent analysis for the Bayesian IDS algorithm, either on the frequentist or Bayesian regret. Second, the Bayesian information gain might serve as a starting point to design more effective information gain functions in the frequentist framework, for example adapted to other likelihood functions and regularizers.

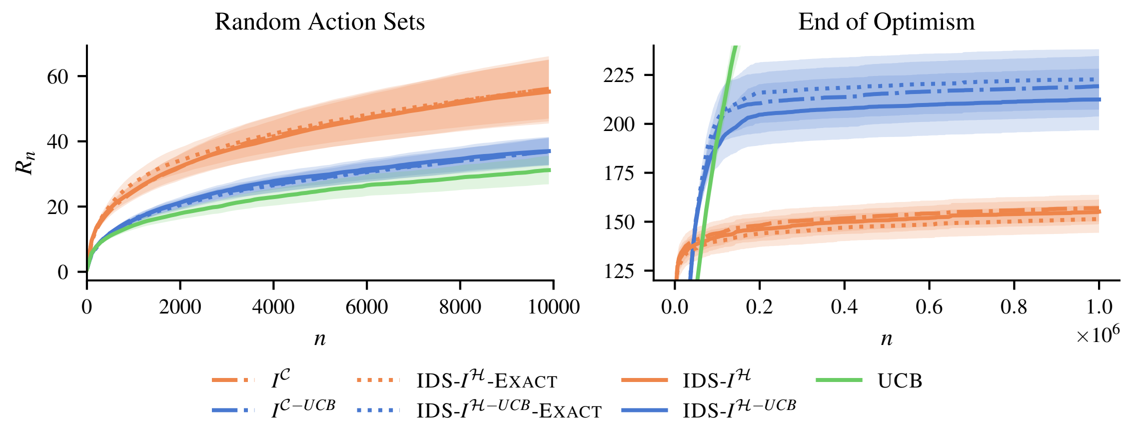

Figure 4: Comparison of information gain functions defined on cells and halfspaces respectively, as well as exact and approximate sampling from the IDS distribution. All variants achieve similar performance within the standard error, however the correction term has a larger impact on the regret. In the right plot, the y-axis is scaled to make the difference visible.

Appendix E. Additional Experiments

In this section we summarize further numerical results. In Section E.1 we compare different information gain functions and show evidence that the cell based information gain IDS-$I_s^{\mathcal{C}}$ variant from Eq.(14) behaves similarly to IDS despite much longer runtimes. In Section E.2 we show the effect of the choice of confidence coefficient β_t and the learning rate η_s on the performance, and also evaluate the confidence coefficient derived by Tirinzoni et al. (2020). In Section E.3, we provide a benchmark with Bayesian methods including Bayesian IDS and a runtime evaluation.

E.1. Comparison of Information Gain Functions

The information gain functions used in the experiments are summarized below.

- The information gain defined in the main text with halfspaces-based alternatives (7):

$$I_s^{\mathcal{H}}(x) = \frac{1}{2} \sum_{z \neq \hat{x}_s} q_s(z) \left(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}} \right)^2$$

- As before, but with correction only for the UCB action (13):

$$I_s^{\mathcal{H}\text{-UCB}}(x) = \frac{1}{2} \sum_{z \neq \hat{x}_s} q_s(z) \left(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \mathbb{1}(x = x_s^{\text{UCB}}) \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}} \right)^2$$

- The information gain defined with cell-based alternatives (14):

$$I_s^{\mathcal{C}}(x) \triangleq \frac{1}{2} \sum_{z \neq \hat{x}_s} \tilde{q}_s(z) \left(|\langle \tilde{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}} \right)^2$$

- The information gain defined on cells and UCB correction:

$$I_s^{\text{C-UCB}}(x) = \frac{1}{2} \sum_{z \neq \hat{x}_s} \tilde{q}_s(z) \left(|\langle \tilde{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \mathbb{1}(x = x_s^{\text{UCB}}) \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}} \right)^2$$

- The variance-based information gain defined in (29) and used for Bayesian IDS:

$$I_t^{\text{VAR}}(x) = \mathbb{E}_t[\big(\mathbb{E}_t[\langle x, \theta \rangle | x^*] - \mathbb{E}_t[\langle x, \theta \rangle]\big)^2]$$

Alternative definitions of the information gain function based on the log-determinant potential are proposed by Kirschner et al. (2020). The resulting IDS algorithm satisfies similar worst-case guarantees but does not achieve asymptotic optimality, e.g. on the end of optimism example.

Figure 3 shows a quantitative comparison of the information gain functions evaluated on the same trajectory on the end of optimism example. The asymptotic information gain based on half-spaces is not shown since it was empirically indistinguishable from the cell-based variant (which might be also due to the fact that there are only three cells in this example). This finding is confirmed by the regret plot in Figure 4, where compare information gain functions, as well as the approximate IDS distribution (optimized directly on $\hat{x}_s$ and one other action) and the exact IDS distribution. The results show that, at least on our examples, there is almost no difference between the information gain defined with $\hat{\nu}_s$ and $\tilde{\nu}_s$, and the approximate and exact IDS sampling.

E.2. Choice of Confidence Coefficient and Learning Rate

We run all our experiments with the simplified rate $\beta_t = \sigma^2(2\log(t) + d\log\log(t))$ instead, as suggested in Tirinzoni et al. (2020). These result are shown on Figure 5 and confirm the statement in Section 3 that there is no significant difference in the conclusions. However tuning β_t to minimize regret *significantly* improves the performance as shown in Figures 6 and 7. On the other hand, tuning the learning rate η_s has much less effect on the regret. The choice $\eta_s = 1/\sqrt{\beta_s}$ as suggested by the theory leads to good results and can be used to reduce the number of tuning parameters.

E.3. Comparison with Bayesian Methods

In our last empirical benchmark, we include Bayesian methods, specifically Thompson sampling (TS) and an approximation of Bayesian IDS. Our implementation of Bayesian IDS uses the variance-based information gain defined in (29), and we approximate the Bayesian gap estimates and information gain using 10^4 posterior samples per round as suggested in (Russo and Van Roy, 2014, Algorithm 6). The performance plots are in Figure 8. Thompson sampling significantly outperforms UCB and the frequentist IDS variants, unless we set $\beta_s = 1$, which, as noted before, improves performance of the frequentist methods. The approximation of Bayesian IDS is the most effective on our benchmark, outperforming the best frequentist method on the 'end of optimism' example roughly by a factor two. Lastly, we show runtime of all methods on a horizon $n = 10^6$ in Table 2. Note that despite the approximation, Bayesian IDS is computationally much more demanding, whereas the frequentist IDS is only about a factor of 5 slower than Thompson sampling on instances in $\mathbb{R}^5$ with $k = 50$ actions.

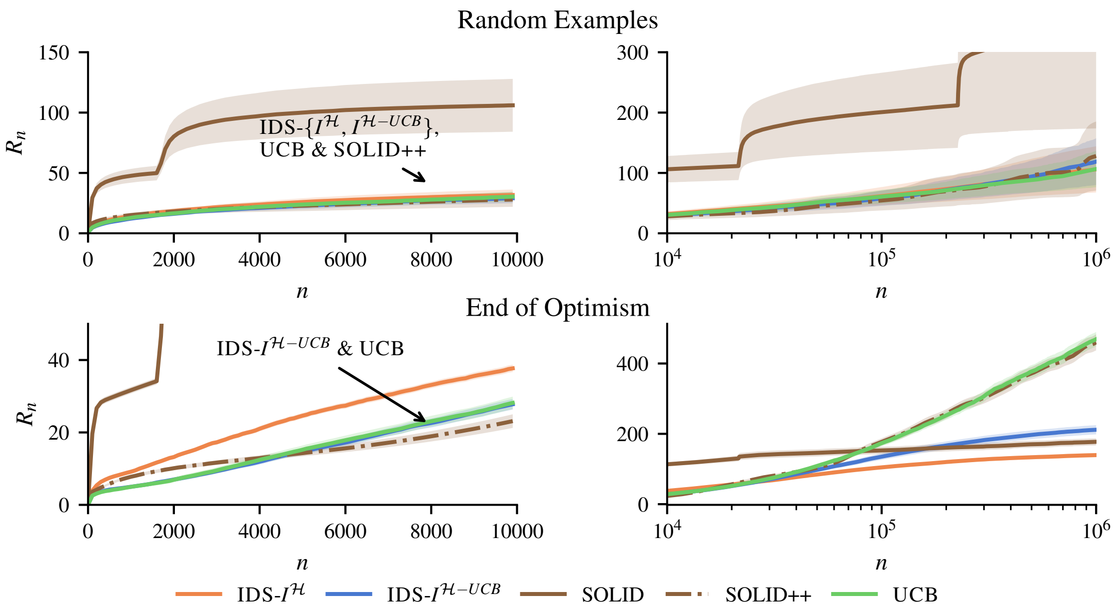

Figure 5: Experiments with $\beta_t = \sigma^2(2\log(t) + d\log\log(t))$. The numerical performance is comparable to the log-determinant confidence coefficient used in the main paper.

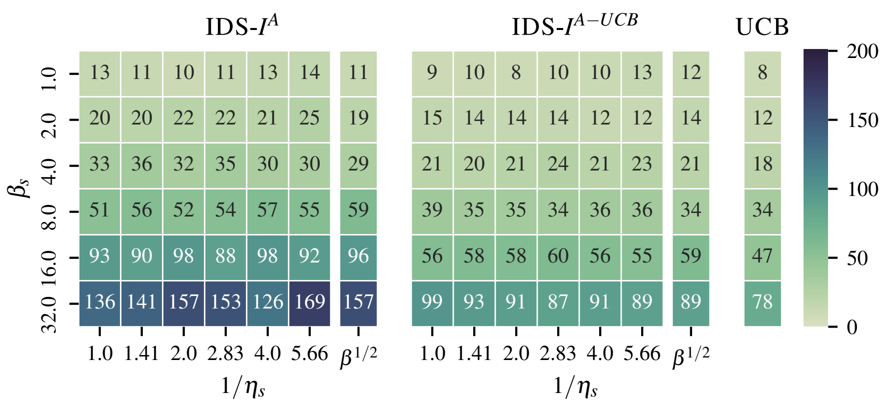

Figure 6: The matrix shows the regret on randomly generated action sets after $n = 10^4$ steps for different values of β_s and η_s. The first observation is that the regret can be *significantly* reduced by choosing a smaller value for β_s. On the other hand, tuning the q-learning rate η_s affects performance marginally. Tuning *only* β_s and setting $\eta_s = 1/\sqrt{\beta_s}$ as suggested by the theory leads to near optimal results.

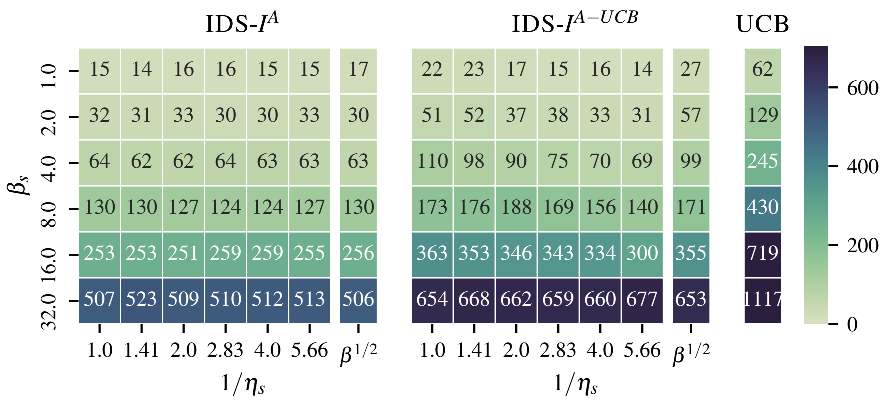

Figure 7: The matrix shows the regret on the 'end of optimism' example after $n = 10^6$ steps for different values of β_s and η_s. The observations are similar as in Figure 6. Note that IDS is consistently better than UCB for any value of β_s.

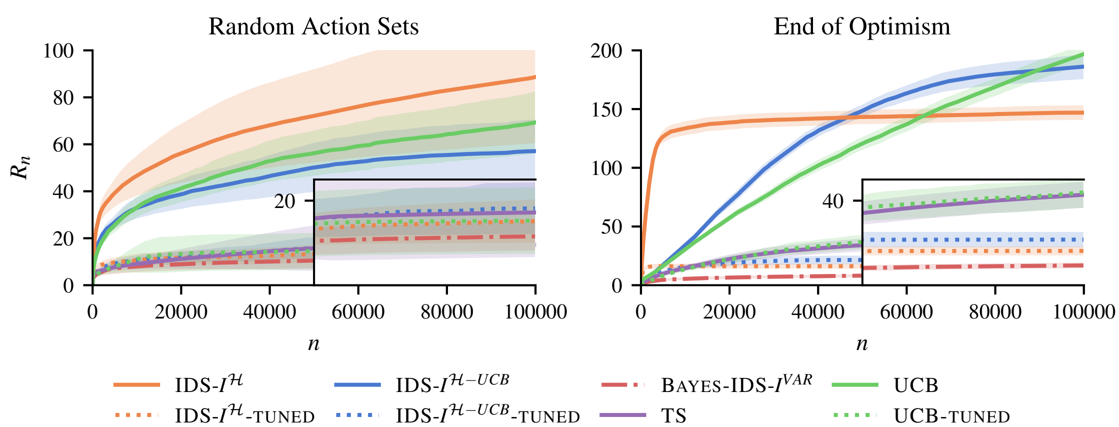

Figure 8: Comparison with Bayesian methods. On these examples, Bayesian IDS outperforms the frequentist methods, even when tuning the frequentist counterpart ($\beta_s = 1$).

Algorithm	$d = 2, k = 6$	$d = 5, k = 50$
BAYES-IDS-I^{VAR}-EXACT	561.7 ± 58.8	2560.0 ± 78.4
BAYES-IDS-I^{VAR}	544.4 ± 69.7	1771.9 ± 40.5
IDS-$I^{\mathcal{H}\text{-UCB}}$-EXACT	50.5 ± 22.6	798.5 ± 233.5
IDS-$I^{\mathcal{H}\text{-UCB}}$	45.7 ± 18.8	106.8 ± 28.6
UCB	26.9 ± 7.7	23.9 ± 5.7
TS	21.6 ± 5.9	22.2 ± 6.9

Table 2: Runtime comparison on random action sets with horizon $n = 10^5$. The table shows mean and standard-deviation of the runtime in seconds on 50 runs, computed on a single core at 2.30GHz. The EXACT-suffix indicates that the IDS distribution is computed exactly, whereas no suffix means that we minimize the tradeoff directly between $\hat{x}_s$ and an informative action as discussed at the end of Section 2.