Linear Partial Monitoring for Sequential Decision Making Algorithms, Regret Bounds and Applications

Johannes Kirschner JKIRSCHN@UALBERTA.CA
Department of Computer Science
University of Alberta

Tor Lattimore LATTIMORE@GOOGLE.COM
DeepMind
London

Andreas Krause KRAUSEA@ETHZ.CH
Department of Computer Science
ETH Zurich

Editor: Aurelien Garivier

Abstract

Partial monitoring is an expressive framework for sequential decision-making with an abundance of applications, including graph-structured and dueling bandits, dynamic pricing and transductive feedback models. We survey and extend recent results on the linear formulation of partial monitoring that naturally generalizes the standard linear bandit setting. The main result is that a single algorithm, information-directed sampling (IDS), is (nearly) worst-case rate optimal in all finite-action games. We present a simple and unified analysis of stochastic partial monitoring, and further extend the model to the contextual and kernelized setting.

Keywords: Sequential Decision-Making, Linear Partial Monitoring, Information-Directed Sampling, Linear Bandits

1. Introduction

Partial monitoring (Rustichini, 1999) is a flexible framework for stateless sequential decision making. The partial monitoring model captures the standard multi-armed and linear bandit setting, semi- and full information feedback models, dynamic pricing and variants of dueling bandits, to name just a few. Partial monitoring is formalized as a round-based game between a learner and an environment. In each round, the learner chooses an action and the environment provides a feedback. Further, there is an (unknown) reward associated to each action but, unlike in the bandit model, the reward is not necessarily directly observed. Instead, the learner is given a description of how the reward and feedback are related. In order to maximize the cumulative reward, the learner needs to find a careful balance between actions that lead to informative feedback and actions with high reward, which is the essence of the *exploration-exploitation dilemma*. More specifically, the learner needs to collect data that allows it to identify an optimal action while minimizing the cost of playing sub-optimal actions relative to the optimal action (known as regret).

Whether the learner can succeed at this task depends on the structure of feedback and reward. In this work, we focus on linearly correlated feedback models where, akin to the linear bandit model, the reward and observations are both described by linear features. Our formulation of the linear model includes the classical finite stochastic partial monitoring formulation as a special case, where reward and feedback are determined by finite matrices.

Classifying games according to the difficulty of the exploration-exploitation trade-off has been an essential part of previous work on partial monitoring (Bartók et al., 2014; Lattimore and Szepesvári, 2019a; Kirschner et al., 2020). By now it is understood that all finite-action linear partial monitoring games belong to one of four categories: *trivial*, *easy*, *hard* and *hopeless*, corresponding to 0, $\tilde{\mathcal{O}}(\sqrt{n})$, $\tilde{\mathcal{O}}(n^{2/3})$ and $\Omega(1)$ regret in the worst-case over n steps. The categories are defined by precise geometric conditions, which we will introduce in detail later.

Partial monitoring has a moderately well-deserved reputation for being a complicated framework. The complexity is largely a product of the generality and also the discrete nature of the standard setup. The latter leads to some unfortunate practical problems. For example, the finite partial monitoring game that models a finite-armed Bernoulli bandit is exponentially large in the number of actions. Although there now exist unified algorithms for finite partial monitoring, they cannot be practically implemented on large games like this, enormously reducing their applicability.

On the other hand, the *linear* partial monitoring model that we present here is a *natural extension* of the linear bandit setting, that only relaxes the requirement that the observation and reward features for any action are identical. By further introducing (convex) parameter constraints and using a suitable linear embedding, we completely recover the standard *finite* partial monitoring setting. At the same time, many problems are more naturally modelled in the linear framework. The classical k-armed Bernoulli bandit, for example, is modelled by a k-dimensional linear partial monitoring game.

The feedback/reward structure of partial monitoring requires exploration techniques that go beyond what is commonly used in bandit feedback models. In particular, it is known that optimism and Thompson sampling are insufficient to achieve sublinear regret in all types of partial monitoring games. Our presentation focuses on a different design principle, known as *information-directed sampling* (IDS; Russo and Van Roy, 2014; Kirschner et al., 2020). This leads to a universal and practical algorithm for regret minimization in partial monitoring that provably works well in *all* games.

Compared to the previous work, we introduce several innovations, including constrained parameter sets and computationally efficient approximations. Distinct from all previous work, our formulation *unifies* the finite and linear stochastic partial monitoring settings, thereby providing a unified approach that is *(a)* nearly worst-case optimal in all possible games and *(b)* computationally efficient and practical. We present a concise and self-contained analysis, where nearly all geometric constructions only appear in the proof and not the algorithm.

Last but not least, we present two extensions of linear partial monitoring that are of practical importance. The first is a contextual formulation that directly generalizes the standard contextual (stochastic) linear bandit setting. The second is a kernelized formulation, where the reward function is in a reproducing kernel Hilbert space. We provide direct

extensions of IDS to both settings, demonstrating the versatility of the framework. A more detailed historical and contemporary account of related work is deferred to Section 8.

Notation For an integer $i \geq 1$, $[i] \triangleq \{1, \ldots, i\}$. The d-dimensional identity matrix is $\mathbf{1}_d$. For a compact set $\mathcal{X}$, $\mathscr{P}(\mathcal{X})$ is the set of distributions on $\mathcal{X}$ with the Borel σ-algebra. For a measurable function $F : \mathcal{X} \to \mathbb{R}$ and a distribution $\mu \in \mathscr{P}(\mathcal{X})$, we use $F(\mu) = \mathbb{E}_{x \sim \mu}[F(x)]$ to denote the expectation. For a vector $v \in \mathbb{R}^d$, the Euclidean norm is $\|v\|_2$. For a positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$, $\lambda_{\max}(A)$ is the largest eigenvalue, and $\|A\| = \lambda_{\max}(A)$ is the spectral norm. The associated norm to A is $\|b\|_A = \sqrt{b^\top A b}$, where $b \in \mathbb{R}^d$. This notion is generalized for rectangular matrices $B \in \mathbb{R}^{d \times s}$, $\|B\|_A \triangleq \lambda_{\max}(B^\top A B)^{1/2}$. For two positive semi-definite matrices A_1, A_2, $A_1 \preccurlyeq A_2$ means that $A_2 - A_1$ is positive semi-definite. Given a set $X \subset \mathbb{R}^d$, we write $\dim(X)$ for the dimension of the affine hull of X. We use the convention that $1/0 = \infty$, $0/0 = 0$ and $0 \cdot \infty = \infty$.

2. Linear Partial Monitoring

A *linear partial monitoring game* is defined by a quadruple $(\mathcal{A}, (\phi_a)_{a \in \mathcal{A}}, (M_a)_{a \in \mathcal{A}}, \Theta)$. Here, $\mathcal{A}$ is a set of *actions* that is used to index *reward features* $\phi_a \in \mathbb{R}^d$ and *linear observation maps* $M_a \in \mathbb{R}^{m \times d}$ for $a \in \mathcal{A}$. $\Theta \subset \mathbb{R}^d$ is a (convex) parameter set. The actions $\mathcal{A}$, features ϕ_a, observation maps M_a and parameter set Θ are known to the learner.

An *instance* of a partial monitoring game is defined by a fixed parameter $\theta^* \in \Theta$ that is *unknown* to the learner. At time $t \in [n]$, the learner chooses an action $a_t \in \mathcal{A}$ according to a policy π and observes an m-dimensional vector $y_t = M_{a_t} \theta^* + \epsilon_t \in \mathbb{R}^m$, where ϵ_t is a zero-mean noise vector. As usual performance is measured by the *expected regret* at time n,

$$\mathfrak{R}_n(\pi, \theta^*) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^n \langle \phi_a - \phi_{a_t}, \theta^* \rangle \right]$$

where the expectation is over the randomness in the feedback and any randomization used by the policy. For simplicity we write $\mathfrak{R}_n = \mathfrak{R}_n(\pi, \theta^*)$ when the policy and the parameter are clear from the context. We emphasize that the *only* difference to the linear bandit model is that *reward and observation features are decoupled*, whereas in the bandit model $\phi_a = M_a^\top$ for all $a \in \mathcal{A}$. The usefulness of decoupling reward and observation will become apparent in the examples presented in Section 2.1.

The σ-algebra generated by the history up to round t is $\mathcal{F}_t = \sigma(a_1, y_1, \ldots, a_t, y_t)$ and the corresponding filtration over n rounds is $(\mathcal{F}_t)_{t=1}^n$. We define the shorthand $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-1}]$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$. The best action for a parameter $\theta \in \Theta$ is $a^*(\theta) = \arg\max_{a \in \mathcal{A}} \langle \phi_a, \theta \rangle$, chosen arbitrarily if it is not unique. The maximum gap under the true parameter is denoted by $\Delta_{\max} = \max_{a,b \in \mathcal{A}} \langle \phi_a - \phi_b, \theta^* \rangle$.

We further require the following standard boundedness and noise assumptions.

Assumption 1 (Convexity and Boundedness)**.** *The action set $\mathcal{A}$ is compact and the parameter set Θ is convex. Further, we assume that the learner has access to bounds $\|M_a\|_2 \leq L$, $\|\phi\|_2 \leq 1$ and $\|\theta - \theta_0\|_2 \leq B$ for all $\theta \in \Theta$ and a given prior estimate $\theta_0 \in \Theta$.*

Note that assuming $\|\phi\|_2 \leq 1$ is without loss of generality, as we can jointly re-scale ϕ_a, M_a and Θ to satisfy the condition while keeping the rewards and observations the same.

3

Assumption 2 (Sub-Gaussian Noise). *The noise vector $\epsilon_t \in \mathbb{R}^m$ is ρ-sub-Gaussian, i.e.,*

$$\forall u \in \mathbb{R}^m, \quad \mathbb{E}_t\Big[\exp(u^\top \epsilon_t)\Big] \leq \exp\left(\frac{\|u\|^2 \rho^2}{2}\right).$$

Note that we allow the noise distribution to be a function of the action and the parameter, whereas the sub-Gaussian variance ρ^2 is a uniform constant known to the learner.

2.1 Examples

The framework of linear partial monitoring captures a large variety of models for online decision-making, many of which have been studied independently in the literature. We provide some examples below.

Example 1 (Linear Bandits). *As was already mentioned, the linear bandit model is recovered with $a = \phi_a = M_a^\top$ for an action set $\mathcal{A} \subset \mathbb{R}^d$. A direct extension are heteroscedastic linear bandits, where the noise distribution depends on the chosen action. In the setting studied by Kirschner and Krause (2018), the learner observes $y_t = a_t^\top \theta^* + \epsilon_t$, where ϵ_t is $\rho(a_t)$-sub-Gaussian and the noise function $\rho : \mathcal{A} \to \mathbb{R}_{\geq 0}$ governs the signal-to-noise ratio of the associated observation. To formulate the heteroscedastic setting in linear partial monitoring we can define $\phi_a = a$, $M_a = a/\rho(a)$ and $\rho = 1$.*

Example 2 (Graph-Structured Feedback). *Semi-bandit feedback or side-observations refer to models between full information and bandit feedback. Semi-bandit feedback can be specified with a feedback graph defined over actions (Mannor and Shamir, 2011; Caron et al., 2012). When choosing an action, the learner observes the reward of all adjacent actions. Formally, assume that $\mathcal{W} \subset \mathcal{A} \times \mathcal{A}$ is a set of (directed) edges. For each $a \in \mathcal{A}$, the feedback map is defined to reveal the reward of all adjacent actions,*

$$M_a = [\phi_c : c \in \mathcal{A} \text{ s.t. } (a, c) \in \mathcal{W}]^\top,$$

with the minor technicality that the observation dimension now in general depends on the action. This can be resolved by padding the observation matrices with zero vectors, or modifying the model to allow action-dependent observation dimension. If we explicitly require that the edge set contains all self-loops $(a, a) \in \mathcal{W}$ for all $a \in \mathcal{A}$, then the graph feedback structure naturally interpolates between bandit feedback (empty graph, with self-loops) and full information (fully connected graph). We point out that Liu et al. (2018) studied a version of IDS in the corresponding Bayesian regret setting.

Example 3 (Dueling Bandits). *In dueling bandits, the learner chooses pairs of actions and receives noisy feedback about which of the two actions has higher reward (Yue et al., 2012; Sui et al., 2018). The dueling bandit model has many intricate variants and a vast literature on its own (Bengs et al., 2021). Not every dueling bandit is easily modeled as a partial monitoring game. With additional assumptions, Gajane and Urvoy (2015) demonstrated that the utility-based dueling bandit problem can be formulated as a partial monitoring game. Here we focus on a similar setup with quantitative feedback on the reward difference of the chosen actions, opposed to the more common binary signal. The quantitative variant*

has received relatively little attention in the literature and has some interesting applications in robust regret minimization (Kirschner and Krause, 2021).

Formally, let $\mathcal{I}$ be a ground set of actions with an associated feature mapping $\phi : \mathcal{I} \to \mathbb{R}^d, a \mapsto \phi_a$. The action set is $\mathcal{A} = \mathcal{I} \times \mathcal{I}$. For any $(a, b) \in \mathcal{A}$, we define utility-based dueling feedback *by the feedback map $M_{a,b} = \phi_a - \phi_b$. Hence, when choosing the action pair (a_t, b_t), the learner observes the noisy reward difference $y_t = \langle \phi_{a_t} - \phi_{b_t}, \theta \rangle + \epsilon_t$, while collecting (unobserved) reward for both actions with reward features $\phi_{a,b} = \phi_a + \phi_b$. We remark that the sub-Gaussian likelihood combined with appropriate boundedness of the reward includes the standard binary feedback model with $y_t \in \{-1, 1\}$ as a special case.*

Example 4 (Graph-Structured Dueling Bandits). *We propose a novel variant of the dueling bandit that extends Example 2 with a feedback graph structure. Concretely, the learner has access to dueling reward features $\phi_{a,b} = \phi_a + \phi_b$, and dueling feedback $M_{a,b} = \phi_a - \phi_b$, but only for pairs $(a, b) \in \mathcal{A} \subset \mathcal{I} \times \mathcal{I}$. We can think of $\mathcal{A}$ as a set of edges on a graph over the base indices $\mathcal{I}$. In other words, the learner can only observe a dueling action (a, b) if there is a directed edge from a to b. To learn the reward difference between actions that are not neighbors, the learner has to combine dueling observations from a path that connects the two actions. For general action features, the learner can only hope to learn all reward differences if the graph is connected.*

Example 5 (Combinatorial Partial Monitoring). *In the* combinatorial *bandit problem the action set can be exponentially large. Learning algorithms for this scenario are designed to only use the solver for the offline problem, i.e., assuming access to an oracle that solves $\arg\max_{a \in \mathcal{A}} \langle \phi_a, \theta \rangle$ for any $\theta \in \mathcal{M}$. While the IDS algorithm we introduce in Section 3 is not oracle efficient, the theory still applies. The combinatorial setting is the motivation for the linear partial monitoring setting in the work by Lin et al. (2014) and Chaudhuri and Tewari (2016). Both previously proposed methods are oracle efficient, however suffer sub-optimal regret in locally observable games.*

The combinatorial version of the multi-armed bandit setting (Cesa-Bianchi and Lugosi, 2012) makes more specific assumptions on the feedback structure. Let $\mathcal{I}$ be an index set with associated features ϕ_a for $a \in \mathcal{A}$. The action set $\mathcal{A} \subset 2^{\mathcal{I}}$ consists of subsets of $\mathcal{I}$. The reward for choosing an action $a \in \mathcal{A}$ is the sum of rewards $f_\theta(a) = \sum_{i \in a} \langle \phi_i, \theta \rangle$. Equivalently, the features for action a are $\sum_{i \in a} \phi_i$. Two variants for the feedback maps are commonly considered: i) bandit feedback, that is $M_a = \phi_a$, and ii) semi-bandit feedback, $\mathcal{M}_a = [\phi_i : i \in a]^\top$. An important special case is the batch setting where the learner chooses m actions at once, i.e., $\mathcal{A} = \{a \subset \mathcal{I} : |a| = m\}$. Note that the example exhibits an exponential blow-up of the action set in the batch size m.

Example 6 (Transductive Bandits). *In the transductive bandit setting, the learner obtains informative feedback only on a set of actions that is dedicated for exploration. At the same time, the objective is to achieve low regret on a different target set of actions, that when played, do not reveal information. The setting was proposed by Fiez et al. (2019) in the context of best arm identification, and we refer the reader to this work for more examples. A toy example from the partial monitoring literature that fits into this category is that of "apple tasting" (Cesa-Bianchi et al., 2006). In each round, the learner is presented an apple, and decides whether to taste it. Tasting determines if the apple is rotten or not. Apples that have*

been tasted cannot be sold anymore and incur a fixed cost. Not tasting the apple comes with the risk of selling a rotten apple, which also incurs a cost but is not observed. We remark that this setup is closely related to label efficient prediction *Cesa-Bianchi et al. (2005).*

Example 7 (Finite Partial Monitoring). *Most previous work has focused on the* finite partial monitoring *setting that we briefly introduce now. In addition to a finite set of* actions $\mathcal{A} = [k]$, *a finite partial monitoring game consists of a set of* signals $\Sigma = [m]$ *used for the feedback and a finite set of* outcomes $\mathcal{X} = [d]$ *that determines reward and feedback for each action. Reward[1] and feedback are defined by,*

   i) a reward function $R : \mathcal{A} \times \mathcal{X} \to [0,1]$; and

   ii) a signal function $\Phi : \mathcal{A} \times \mathcal{X} \to \Sigma$.

The learner has access to both R and Φ. In each round $t = 1, \ldots, n$ of the game, the learner chooses an action $a_t \in \mathcal{A}$. In the stochastic version of the problem, the outcome x_t is sampled from an unknown and fixed distribution $\vartheta \in \mathscr{P}(\mathcal{X})$. The learner observes a signal $\sigma_t = \Phi(a_t, x_t) \in \Sigma$ and obtains reward $R(a_t, x_t)$. Neither the reward nor the outcome is revealed to the learner.

   As in previous work (e.g., Bartók et al., 2014), we use vector notation to describe the finite setting in the linear framework. Let $e_a \in \mathbb{R}^k$, $e_x \in \mathbb{R}^d$ and $e_\sigma \in \mathbb{R}^m$ be the basis vectors corresponding to action $a \in \mathcal{A}$, outcome $x \in \mathcal{X}$ and signal $\sigma \in \Sigma$. We use $R \in \mathbb{R}^{k \times d}$ as a matrix and function interchangeably, such that $e_a^\top R e_x = R(a, x)$. Further, we introduce reward features $\phi_a = R^\top e_a \in \mathbb{R}^d$, defined as the row of R corresponding to action a. For each action $a \in \mathcal{A}$, the observation matrix $S_a \in \{0, 1\}^{m \times d}$ is such that $e_\sigma^\top S_a e_x = \mathbb{1}(\Phi(a, x) = \sigma)$. We use the symbol S_a instead of M_a to emphasize the particular structure of the feedback map. The distribution $\vartheta \in \Theta = \mathscr{P}(\mathcal{X})$ is identified with a vector in the $(d-1)$-dimensional probability simplex. In particular, $S_a \vartheta$ is the distribution over the observed signals for action $a \in \mathcal{A}$. If the learner chooses action $a_t \in \mathcal{A}$ in round t, and the outcome is $x_t \in \mathcal{X}$, then the corresponding observation vector is $y_t = e_{\sigma_t} = S_{a_t} e_{x_t} \in \mathbb{R}^m$.

   Let $\xi_t = S_{a_t}(e_{x_t} - \vartheta)$ and note that $\mathbb{E}_t[\xi_t] = 0$. The observation can be written as $y_t = S_{a_t}\vartheta + \xi_t$. One directly verifies that for any unit vector $u \in \mathbb{R}^s$, $|u^\top \xi_t| \leq \|u\|_\infty \|\xi_t\|_1 \leq \|S_a e_{x_t} - S_a \vartheta\|_1 \leq 2$. Hence ξ_t is a 4-sub-Gaussian random vector in $\mathbb{R}^m$.

We note that the difficulty of the games depends on the parameter set Θ. In particular, adding constraints to Θ can make the game much easier. The next example shows that a good algorithm for finite partial monitoring has to reason about the parameter set Θ.

Example 8 (Finite vs Linear Partial Monitoring). *Consider the finite game defined by reward and signal matrices*

$$R = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The signal matrix uses only one symbol, therefore the learner cannot distinguish the outcomes. However, when Θ is the probability simplex, the rewards are such that the first

---

1. The finite setting is often formulated with losses instead of rewards. For consistency with our presentation, we use the equivalent setup with rewards. The loss formulation is easily recovered by flipping the sign of the feature vectors.

Algorithm 1: Information-Directed Sampling

Input: Action set $\mathcal{A}$, gap estimate $\hat{\Delta}_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$, information gain $I_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$

1 for $t = 1, 2, 3, \ldots, n$ **do**

2 $\qquad \mu_t \leftarrow \underset{\mu \in \mathscr{P}(\mathcal{A})}{\arg\min} \left\{ \Psi_t(\mu) = \dfrac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \right\}$ $\qquad\qquad$ // IDS distribution

3 $\qquad$ Sample $a_t \sim \mu_t$, observe feedback y_t

---

action is always optimal, so in finite partial monitoring a good algorithm has zero regret. On the other hand, when $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ the learner has to consider the case when $\theta = (-\sqrt{2}, -\sqrt{2})$, and the second action is optimal. Consequently, any learner suffers linear regret on at least one of the two cases.

We conclude the section with a two more examples of in the language of finite partial monitoring.

Example 9 (Multi-Armed Bandits). *In games with bandit information, the learner observes the reward of each action $a \in \mathcal{A}$ by playing it. Since we allow only finitely many signals, the reward of each arm is also one of finitely many values. For Bernoulli bandits with k arms specifically, $\mathcal{A} = [k]$, $\Sigma = \{0, 1\}$ and $\mathcal{X} = \{0, 1\}^k$. The reward and feedback functions are*

$$R(a, x) = \Phi(a, x) = x_a.$$

A consequence of the finite partial monitoring setup is that the parameter dimension $d = 2^k$ is exponentially large in the number of arms. On the other hand, we will see in section Section 3.1 that all relevant information is contained in a $2k$-dimensional subspace of $\mathbb{R}^d$.

Example 10 (Dynamic Pricing). *One of the most notable applications of finite partial monitoring is dynamic pricing. This game is between a seller and a potential customer. The learner takes the role of the seller with the goal to optimally price a product. The action and outcome sets are a (discrete) set of prices corresponding to an offer and the price the customer is willing to pay, e.g. $\mathcal{A} = \mathcal{X} = \{\$1, \$2, \$3\}$. The feedback is whether the customer buys the product ($a \leq x$, $\Phi(a, x) = \text{Y}$), or not ($a > x$, $\Phi(a, x) = \text{N}$). The reward consists of a fixed opportunity cost $c > 0$ and the difference between the offer and the price the customer would have payed, $R(a, x) = (a - x)\mathbb{1}(a \leq x) - c\mathbb{1}(a > x)$. With $c = 2$ and $\mathcal{X}, \mathcal{A}$ as above, the corresponding loss and signal matrices are:*

$$R = \begin{pmatrix} 0 & -1 & -2 \\ -2 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix} \qquad\qquad \Phi = \begin{pmatrix} \text{Y} & \text{Y} & \text{Y} \\ \text{N} & \text{Y} & \text{Y} \\ \text{N} & \text{N} & \text{Y} \end{pmatrix}$$

3. Information-Directed Sampling for Linear Partial Monitoring

Information-directed sampling (IDS) is a *design principle* that requires user choices and leads to different algorithms in different settings. As presented in Algorithm 1, IDS is abstractly defined for a sequence of *gap estimates* $\hat{\Delta}_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ and an *information gain*

functions $I_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$. We present concrete choices for both quantities shortly. One should think of $\hat{\Delta}_t(a)$ as an estimate of the instantaneous regret of playing action a and $I_t(a)$ as some measure of the information gained when playing action a. Naturally, the gap estimates and information gain are computed using observations from previous rounds and are therefore predictable with respect to the filtration $(\mathcal{F}_t)_{t=1}^n$. We also assume that I_t is not zero for at least one action.

Recall that for a distribution $\mu \in \mathscr{P}(\mathcal{A})$, we denote $\Delta_t(\mu) = \mathbb{E}_{a \sim \mu}[\Delta_t(a)]$ and $I_t(\mu) = \mathbb{E}_{a \sim \mu}[I_t(a)]$. The IDS distribution μ_t is defined as the minimizer of the ratio between squared expected regret and expected information gain:

$$\mu_t = \underset{\mu \in \mathscr{P}(\mathcal{A})}{\arg\min} \left\{ \Psi_t(\mu) \triangleq \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \right\}. \tag{1}$$

The objective $\Psi_t(\mu)$ is called the *information ratio* of the sampling distribution $\mu \in \mathscr{P}(\mathcal{A})$. The minimizer always exists for compact $\mathcal{A}$ (Lemma 19). IDS is defined as the policy that samples $a_t \sim \mu_t$ in round t. Intuitively, to achieve a small information ratio, the learner has to sample actions from a distribution with small expected (estimated) regret or large information gain. This intuition will appear formally in the proofs in Section 4.

The information ratio Ψ_t satisfies several important properties, which allow us to solve the optimization problem (1) efficiently (Russo and Van Roy, 2014). First, the function $\mu \mapsto \Psi_t(\mu)$ is convex for any choice of $\hat{\Delta}_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ and $I_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ (Lemma 20). Further, the minimizing distribution μ_t can always be chosen with a support of at most two actions (Lemma 21). Using this property, and provided that $\hat{\Delta}_t(a)$ and $I_t(a)$ have been computed for all $a \in \mathcal{A}$, we can find the exact IDS distribution by enumerating all pairs of actions and solving the trade-off for each pair in closed-form (Lemma 22).

Approximate IDS It is also possible to obtain a $\frac{4}{3}$-approximation of Eq. (1) in $\mathcal{O}(|\mathcal{A}|)$ time with oracle access to $\hat{\Delta}_t(a)$ and $I_t(a)$. To do so, we first find an action that minimizes the gap estimates, $\hat{a}_t = \arg\min_{a \in \mathcal{A}} \hat{\Delta}_t(a)$. We then optimize the trade-off between $\hat{a}_t$ and some other action $b \neq \hat{a}_t$, i.e.

$$p_t, b_t = \underset{p \in [0,1], b \in \mathcal{A}}{\arg\min} \ \Psi_t\big((1-p)e_{\hat{a}_t} + pe_b\big). \tag{2}$$

As before, for fixed $\hat{a}_t, b \in \mathcal{A}$, the optimization problem can be solved in closed form (Lemma 22), and it remains to iterate over the action set to find the best alternative action b_t. The distribution $\tilde{\mu}_t = (1 - p_t)e_{\hat{a}_t} + p_t e_{b_t}$ satisfies $\Psi_t(\tilde{\mu}_t) \leq \frac{4}{3}\Psi_t(\mu_t)$ (Lemma 24).

General Regret Bounds A few more things can be said without committing to specific choices of the gap estimate and information gain. The information ratio Ψ_t appears as a central quantity in the regret analysis. To understand how, consider any adaptive policy $\pi_n = (\mu_t)_{t=1}^n$. We first bound the sum over the gap estimates:

$$\mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(\mu_t)\right] = \mathbb{E}\left[\sum_{t=1}^n \sqrt{\Psi_t(\mu_t)I_t(\mu_t)}\right]$$

$$\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^n \Psi_t(\mu_t)\right]\mathbb{E}\left[\sum_{t=1}^n I_t(a_t)\right]}. \tag{3}$$

8

The first equality uses the tower rule, $\mathbb{E}\big[\hat{\Delta}_t(a_t)\big] = \mathbb{E}\big[\mathbb{E}_t\big[\hat{\Delta}_t(a_t)\big]\big] = \mathbb{E}\big[\hat{\Delta}_t(\mu_t)\big]$. The second equality uses the definition of the information ratio, and Eq. (3) follows from the Cauchy-Schwarz inequality and another application of the tower rule. Note that IDS is the policy that myopically minimizes the first sum in the upper bound. The second sum is the *total information gain*, which we abbreviate with

$$\gamma_n \triangleq \sum_{t=1}^{n} I_t(a_t). \tag{4}$$

For the regret $\mathfrak{R}_n = \mathbb{E}[\sum_{t=1}^{n} \Delta(a_t)]$, Eqs. (3) and (4) imply

$$\mathfrak{R}_n \leq \sqrt{\mathbb{E}\left[\sum_{t=1}^{n} \Psi_t(\mu_t)\right]\mathbb{E}[\gamma_n]} + \mathbb{E}\left[\sum_{t=1}^{n} \Delta(a_t) - \hat{\Delta}_t(a_t)\right]. \tag{5}$$

In the frequentist IDS framework, the gap estimate $\hat{\Delta}_t(a)$ is chosen as a high-probability upper bound on the true gap $\Delta(a)$. This way, the estimation error (the second term in the last display) contributes only negligibly to the overall regret. The total information gain γ_n can be interpreted as a surrogate of the sample complexity of identifying the best action. In the finite-dimensional linear setting γ_n depends only logarithmically on the horizon. Lastly, if $\Psi_t(\mu_t) \leq \alpha$ almost surely for all $1 \leq t \leq n$, then, by construction, the IDS policy has regret at most

$$\mathfrak{R}_n \leq \sqrt{n\alpha\mathbb{E}[\gamma_n]} + \sum_{t=1}^{n} \mathbb{E}\Big[\Delta(a_t) - \hat{\Delta}_t(a_t)\Big].$$

The usefulness of the bound stems from the fact that we can analyze IDS by explicitly designing sampling distributions that achieve a small information ratio, without having to know the exact behavior of the IDS algorithm. Further note that any constant approximation of the IDS distribution directly translates to the regret bound, which allows us to use the approximation in Eq. (2).

In the next theorem, we summarize the result in slightly generalized form. We make use of the *generalized information ratio* introduced by Lattimore and György (2020),

$$\Psi_{\kappa,t}(\mu) \triangleq \frac{\hat{\Delta}_t(\mu)^\kappa}{I_t(\mu)}. \tag{6}$$

Note that the previous definition is recovered with $\kappa = 2$, i.e. $\Psi_{2,t} = \Psi_t$. The next lemma states the regret bound (5) for the generalized information ratio. A similar bound on the Bayesian regret is given by Lattimore and György (2020, Theorem 4).

Theorem 1. *Assume that $\Psi_{\kappa,t}(\mu_t) \leq \alpha_t$ holds almost surely for an $\mathcal{F}_t$-predictable sequence $(\alpha_t)_{t=1}^n$, and let $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^{n} \alpha_t$. Then*

$$\mathfrak{R}_n \leq (\mathbb{E}[\bar{\alpha}_n]\mathbb{E}[\gamma_n])^{\frac{1}{\kappa}} n^{1-\frac{1}{\kappa}} + \sum_{t=1}^{n} \mathbb{E}\Big[\Delta(a_t) - \hat{\Delta}_t(a_t)\Big].$$

Proof We bound the estimated regret similarly as before:

$$\mathbb{E}\left[\sum_{t=1}^{n}\hat{\Delta}(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{n}(\Psi_{\kappa,t}(\mu_t)I_t(\mu_t))^{1/\kappa}\right] \overset{(i)}{\leq} \mathbb{E}\left[\sum_{t=1}^{n}\Psi_{\kappa,t}(\mu_t)^{\frac{1}{\kappa-1}}\right]^{1-\frac{1}{\kappa}} \mathbb{E}\left[\sum_{t=1}^{n}I_t(a_t)\right]^{\frac{1}{\kappa}}$$

$$\overset{(ii)}{\leq} \mathbb{E}\left[\sum_{t=1}^{n}\alpha_t^{\frac{1}{\kappa-1}}\right]^{1-\frac{1}{\kappa}} \mathbb{E}[\gamma_n]^{\frac{1}{\kappa}}$$

$$\overset{(iii)}{\leq} \mathbb{E}\left[\left(\sum_{t=1}^{n}\alpha_t\right)^{\frac{1}{\kappa-1}} n^{\frac{\kappa-2}{\kappa-1}}\right]^{1-\frac{1}{\kappa}} \mathbb{E}[\gamma_n]^{\frac{1}{\kappa}}$$

$$\overset{(iv)}{\leq} n^{\frac{\kappa-1}{\kappa}} \mathbb{E}[\bar{\alpha}_n]^{\frac{1}{\kappa}} \mathbb{E}[\gamma_n]^{\frac{1}{\kappa}} \, .$$

We used (i) and (iii): Hölder's inequality, (ii): definitions of α_t and γ_n, and (iv): Jensen's inequality and the definition of $\bar{\alpha}_n$. The claim follows by introducing the estimation error. ∎

We remark that the generalized information ratio $\Psi_{\kappa,t}$ appears naturally in the analyses of games where the minimax regret is of order $\mathcal{O}(n^{\frac{\kappa-1}{\kappa}})$ (see Section 4). In such cases, it is natural to use an algorithm that optimizes $\Psi_{\kappa,t}$ directly.

The following lemma shows the perhaps surprising result that the IDS distribution obtained as a minimizer of $\Psi_{2,t}(\mu)$ approximately minimizes $\Psi_{\kappa,t}(\mu)$ for any $\kappa \geq 2$. This justifies the use of the $\Psi_{2,t}$ information ratio even in cases when a $\sqrt{n}$ regret rate is not attainable.

Lemma 2 (Lattimore and György (2020, Lemma 21)). *Let $\mu_t = \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_{2,t}(\mu)$ be the IDS distribution computed for $\Psi_{2,t}$. Then for all $\kappa \geq 2$,*

$$\Psi_{\kappa,t}(\mu_t) \leq 2^{\kappa-2} \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_{\kappa,t}(\mu) \, .$$

3.1 IDS for Linear Partial Monitoring

We now propose natural choices for the gap estimate and the information gain functions for linear partial monitoring games $\mathcal{G} = (\mathcal{A}, (\phi_a)_{a \in \mathcal{A}}, (M_a)_{a \in \mathcal{A}}, \Theta)$. The definitions and analysis follow the ideas of Kirschner et al. (2020), with a few innovations to account for constrained parameter sets Θ. We note that parameter constraints are required to recover optimal bounds in the finite partial monitoring setting. The complete approach is summarized in Algorithm 2. Various improvements and extensions are discussed later in Section 5.

Parameter Estimate We assume prior knowledge of some $\theta_0 \in \Theta$ such that $\|\theta_0 - \theta^*\|_2 \leq B$, where both θ_0 and B are known to the learner (Assumption 1). As a main tool for estimating θ^*, we rely on regularized linear least-squares on Θ,

$$\hat{\theta}_t = \arg\min_{\theta \in \Theta} \sum_{s=1}^{t-1} \|M_{a_s}\theta - y_s\|^2 + \lambda\|\theta - \theta_0\|^2 \, , \tag{7}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ is a regularizer. For the case where $\Theta = \mathbb{R}^d$, the usual closed-form is available, $\hat{\theta}_t = V_t^{-1}(\lambda\theta_0 + \sum_{s=1}^{t-1} M_{a_s}^\top y_s)$ with inverse regularized covariance $V_t = \lambda I +$

Algorithm 2: IDS for Linear Partial Monitoring

Input: Action set $\mathcal{A}$, parameter set Θ, feature maps ϕ_a, feedback maps M_a, basis W, regularizer $\lambda > 0$, prior estimate θ_0, norm bound $B > 0$, noise variance ρ^2.

1 **for** $t = 1, 2, 3, \ldots, n$ **do**

2 $\quad \hat{\theta}_t \leftarrow \arg\min_{\theta \in \Theta} \sum_{s=1}^{t-1} \|M_{a_s}\theta - y_s\|^2 + \lambda\|\theta - \theta_0\|^2$ $\qquad$ `// solve least-squares`

3 $\quad V_t \leftarrow \sum_{s=1}^{t-1} M_{a_s} M_{a_s}^\top + \lambda \mathbf{1}_d \,, \ W_t \leftarrow W^\top V_t W$

4 $\quad \beta_{t,\delta}^{1/2} \leftarrow \rho\sqrt{\log\det(W_t) - \log\det(\lambda\mathbf{1}_d) + 2\log(1/\delta)} + \sqrt{\lambda}B$

5 $\quad \mathcal{E}_t \leftarrow \{\theta \in \Theta : \|\theta - \hat{\theta}_t\|_{V_t}^2 \leq \beta_{t,1/t^2}\}$ $\qquad\qquad$ `// confidence set`

6 $\quad \hat{\Delta}_t(a) \leftarrow \max_{\theta \in \mathcal{E}_t} \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \theta \rangle$ $\qquad\qquad$ `// gap estimates`

7 $\quad I_t(a) \leftarrow \frac{1}{2}\log\det\left(\mathbf{1}_m + M_a V_t^{-1} M_a^\top\right)$ $\qquad\qquad$ `// information gain`

8 $\quad \mu_t \leftarrow \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \dfrac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)}$ $\qquad\qquad\qquad$ `// IDS distribution`

9 $\quad a_t \sim \mu_t$

10 $\quad$ Choose a_t, observe $y_t = \langle M_{a_t}, \theta \rangle + \epsilon_t$

$\sum_{s=1}^{t-1} M_{a_s}^\top M_{a_s}$. In general form, $\hat{\theta}_t$ is the projection of the unconstrained least-square estimate onto Θ with respect to the $\|\cdot\|_{V_t}$ norm.

In settings where Θ is contained in a lower-dimensional affine subspace of $\mathbb{R}^d$, or the dimension of the observation subspace $\mathrm{span}(\mathrm{im}(M_c^\top) : c \in \mathcal{A})$ is significantly smaller than d, it can be useful to introduce a basis $W \in \mathbb{R}^{d \times r}$ to ease computation and improve the dependence of the regret bounds on the dimension d. To be a valid basis, W needs to satisfy $W^\top W = \mathbf{1}_r$ and

$$\forall a \in \mathcal{A}, \ \theta, \nu \in \Theta \quad M_a(\theta - \nu) = M_a W W^\top (\theta - \nu) \,. \tag{8}$$

We remark that $W \in \mathbb{R}^{d \times r}$ can be chosen to satisfy

$$r \leq \min\{\dim(\Theta), \dim(\mathrm{im}(M_c^\top) : c \in \mathcal{A})\} \leq \min\{d, m|\mathcal{A}|\} \,. \tag{9}$$

$W = \mathbf{1}_d$ is a perfectly valid (and sometimes the only possible) choice. The next lemma provides an elliptical confidence set for $\hat{\theta}_t$.

Lemma 3. *Let* $\beta_{t,\delta}^{1/2} \triangleq \rho\sqrt{2\log\frac{1}{\delta} + \log\det(W_t) - \log\det(\lambda\mathbf{1}_r)} + \sqrt{\lambda}B$ *be a confidence coefficient where* $W_t = W^\top V_t W \in \mathbb{R}^{r \times r}$. *Then*

$$\mathbb{P}\big[\forall t \geq 1, \ \theta^* \in \mathcal{E}_{t,\delta} \triangleq \{\theta \in \Theta : \|\theta - \hat{\theta}_t\|_{V_t}^2 \leq \beta_{t,\delta}\}\big] \geq 1 - \delta \,.$$

The proof generalizes the standard ellipsoidal confidence set (cf. Abbasi-Yadkori et al., 2011) and is deferred to Appendix B.1.

In the following, we set $\beta_t = \beta_{t,1/t^2}$ and $\mathcal{E}_t = \mathcal{E}_{t,1/t^2}$, which allows us to derive bounds on the expected regret. It is also possible to fix the confidence level δ to obtain high-probability bounds, or tune the confidence coefficient empirically.

Gap Estimates   The gap estimates are defined as conservative estimates of the true gaps:

$$\hat{\Delta}_t(a) = \max_{\theta \in \mathcal{E}_t,\, b \in \mathcal{A}} \langle \phi_b - \phi_a, \theta \rangle . \tag{10}$$

When $\Theta = \mathbb{R}^d$, the maximum over θ can be computed in closed-form: $\hat{\Delta}_t(a) = \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \hat{\theta}_t \rangle + \beta_t^{1/2} \|\phi_b - \phi_a\|_{V_t^{-1}}$. Note that $\hat{\Delta}_t(a)$ is chosen as a high-probability upper bound on the true gap $\Delta(a)$, which allows us to control the error term in Eq. (5). Specifically, denoting $\Delta_{\max} = \max_{a \in \mathcal{A}} \Delta(a)$, we get

$$\sum_{t=1}^n \mathbb{E}\Big[\Delta(a_t) - \hat{\Delta}_t(a_t)\Big] \leq \sum_{t=1}^n \Delta(a_t) \mathbb{P}[\theta^* \notin \mathcal{E}_{t,1/t^2}] \leq \sum_{t=1}^n \Delta_{\max} t^{-2} \leq \mathcal{O}(\Delta_{\max}) . \tag{11}$$

Information Gain   The next step is to choose an information gain function $I_t(a)$. We define the information gain as the increase of the log-determinant, given by

$$\begin{aligned} I_t(a) &= \frac{1}{2} \log \det(W_t + W^\top M_a^\top M_a W) - \frac{1}{2} \log \det(W_t) \\ &= \frac{1}{2} \log \det\left(\mathbf{1}_m + (M_a W) W_t^{-1} (M_a W)^\top \right) . \end{aligned} \tag{12}$$

The log-determinant of the covariance matrix is a common progress measure in linear experimental design that captures the log-volume of the confidence ellipsoid (D-optimal design). This choice was primarily analyzed in the frequentist IDS framework by Kirschner et al. (2020), with the difference that here we introduced the basis W. The definition further has a natural interpretation in the Bayesian setting (Russo and Van Roy, 2014). Assume for a moment that inference is done with Gaussian prior $\vartheta \sim \mathcal{N}(W^\top \theta_0, \lambda^{-1} \mathbf{1}_r)$ and an observation likelihood $y_t \sim \mathcal{N}(M_{a_t}(W\vartheta + (\mathbf{1}_d - WW^\top)\theta_0), \mathbf{1}_m)$. The posterior distribution corresponds to the least-squares estimate, $\mathcal{N}(W^\top \hat{\theta}_t, W_t^{-1})$ and the entropy of the Gaussian posterior distribution on the subspace defined by W is $\mathbb{H}(\theta) = \frac{1}{2} \log((2\pi e)^d \det(W_t^{-1}))$. Therefore Eq. (12) corresponds to the entropy reduction when choosing $a_t = a$ and observing $y_t = M_a \theta + \epsilon_t$, known as the *mutual information*,

$$\mathbb{I}_t(\theta; y_t | a_t = a) = \frac{1}{2} \log \det(W_{t+1}) - \frac{1}{2} \log \det(W_t) = I_t(a) .$$

In light of Theorem 1, an important quantity in our analysis is the total information gain $\gamma_n = \sum_{t=1}^n I_t(a_t)$. The next lemma is a standard result closely related to the *elliptical potential lemma*. It provides a worst-case bound on γ_n that is independent of the sequence of actions.

Lemma 4 (Total Information Gain). *The total information gain $\gamma_n = \sum_{t=1}^n I_t(a_t)$ is bounded as follows,*

$$\gamma_n = \frac{1}{2} \log \det(W_n) - \frac{1}{2} \log \det(\lambda \mathbf{1}_r) \leq \frac{r}{2} \log\left(1 + \frac{nL}{\lambda r}\right) .$$

For a proof, see, e.g., (Lattimore and Szepesvari, 2020, Lemma 19.4). The lemma motivates the use of a lower-dimensional basis W (where possible) as it makes the upper bound independent of the ambient dimension d. Further note that the lemma implies an upper bound on the confidence coefficient since $\beta_{t,\delta}^{1/2} \leq \beta_{n,\delta}^{1/2} = \rho(\gamma_n + 2\log\frac{1}{\delta})^{1/2} + \lambda^{1/2} B$.

Computation Note that the least-squares estimate and inverse of the covariance matrix can be computed incrementally in $\mathcal{O}(d^2)$ steps per round. Computing gap estimates requires $\mathcal{O}(|\mathcal{A}|^2 d^2)$ operations and computing the information gain can be done in $\mathcal{O}(|\mathcal{A}|d^2)$. Computing the IDS distribution requires $\mathcal{O}(|\mathcal{A}|^2)$, and an approximate distribution that minimizes the information ratio up to constant factor can be computed in linear time (see Eq. (2)). In Section 5.1 we show how to improve the per-step computation complexity to $\mathcal{O}(|\mathcal{A}|d^2)$ by introducing a gap estimate that can be computed in $\mathcal{O}(|\mathcal{A}|d^2)$. By carefully computing quantities using the basis W, the computation complexity can be reduced to $\mathcal{O}(|\mathcal{A}|r^2)$ per round and $\mathcal{O}(d)$ once at the beginning of the game.

4. Regret Bounds

How fast the learner can determine an optimal action in a linear partial monitoring game depends on the geometric structure of feedback and reward. Some terminology is necessary to state the results. The main distinction is between *locally* and *globally* observable games. Informally, in globally observable games, the learner has access to actions with which it can estimate $\langle \phi_a - \phi_b, \theta^* \rangle$ for all actions a and b that are Pareto optimal (defined formally below). The Pareto optimal actions have the property that, for any θ^*, one of them is always optimal. However, acquiring sufficient information might incur a constant regret cost per round, and appropriately trading off exploration and exploitation leads to $\mathcal{O}(n^{2/3})$ regret in the worst-case. In locally observable games, the cost payed for information is at most proportional to the statistical estimation error of the optimal action, in which case the learner can achieve $\mathcal{O}(\sqrt{n})$ regret. Note that local observability is a stronger requirement than global observability. Any locally observable game is also globally observable.

4.1 Local and Global Observability

Following the standard terminology (Bartók et al., 2014), the cell of $a \in \mathcal{A}$ is defined as the set of parameters for which $a \in \mathcal{A}$ is optimal,

$$\mathcal{C}_a \triangleq \{\theta \in \Theta : \langle \phi_a, \theta \rangle = \max_{b \in \mathcal{A}} \langle \phi_b, \theta \rangle\}. \tag{13}$$

Two actions $a, b \in \mathcal{A}$ are called *duplicates* if $\langle \phi_a, \theta \rangle = \langle \phi_b, \theta \rangle$ for all $\theta \in \Theta$. Note that duplicate actions can still differ on the feedback maps M_a and M_b. An action is called *Pareto optimal* if $\dim(\mathcal{C}_a) = \dim(\Theta)$. If a is Pareto optimal, then the only actions that are optimal on the relative interior of $\mathcal{C}_a$ are a and its duplicates (which are also Pareto optimal). The set of all Pareto optimal actions is $\mathcal{P}$. An action is called *degenerate* if $0 \leq \dim(\mathcal{C}_a) < \dim(\Theta)$, and *dominated* if $\mathcal{C}_a = \emptyset$. Degenerate actions can be optimal but not uniquely so, whereas dominated actions are never optimal. We denote the linear span of parameter differences by $\mathcal{V} = \text{span}(\{\theta - \nu : \theta, \nu \in \Theta\})$ and introduce the orthogonal projection $\text{Proj}_{\mathcal{V}} : \mathbb{R}^d \to \mathcal{V}$.

Global Observability A linear partial monitoring game is called *globally observable* if

$$\forall a, b \in \mathcal{P}, \quad \text{Proj}_{\mathcal{V}}(\phi_a - \phi_b) \in \text{span}(\text{im}(\text{Proj}_{\mathcal{V}} M_c^\top) : c \in \mathcal{A}). \tag{14}$$

Intuitively, the requirement is that the learner can estimate the difference in reward $\langle \phi_a - \phi_b, \nu - \theta \rangle$ for Pareto optimal actions $a, b \in \mathcal{P}$ and parameters $\nu, \theta \in \Theta$ by combining the

feedback from all actions. The projection onto $\mathcal{V}$ appears naturally as the set of possible directions in which two parameters can differ. The worst-case cost-to-signal ratio is captured by the *global alignment constant*:

$$\alpha \triangleq \max_{\nu \in \mathcal{V}} \max_{a,b \in \mathcal{P}} \min_{c \in \mathcal{A}} \frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\|M_c \nu\|^2} \,. \tag{15}$$

An immediate consequence is that $\alpha < \infty$ if and only if the condition in Eq. (14) is satisfied.

Local Observability The definition of *local observability* strengthens the previous definition by requiring that the learner can estimate the reward difference between Pareto optimal actions that are plausibly optimal and that it can do so by playing actions with small regret. Formally, given a set $\mathcal{E} \subset \Theta$, let $\mathcal{P}(\mathcal{E}) = \cup_{\theta \in \mathcal{E}} \{a \in \mathcal{P} : \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \theta \rangle = 0\}$ be the set of Pareto optimal actions that are optimal for some $\nu \in \mathcal{E}$. Later we will take $\mathcal{E}$ to be the confidence set constructed by the algorithm in a given round. In this case $\mathcal{P}(\mathcal{E})$ is the set of plausibly optimal actions in $\mathcal{P}$. Denote by $\Delta(a|\theta) = \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \theta \rangle$ the gap of action $a \in \mathcal{A}$ for parameter $\theta \in \Theta$. For any $\eta > 0$, the *extended plausible Pareto* set is defined as follows:

$$\bar{\mathcal{P}}_\eta(\mathcal{E}) = \{a \in \mathcal{A} : \max_{\theta \in \mathcal{E}} \Delta(a|\theta) \le \eta \cdot \max_{b \in \mathcal{P}(\mathcal{E})} \max_{\theta \in \mathcal{E}} \Delta(b|\theta)\} \,.$$

Actions in $\bar{\mathcal{P}}(\mathcal{E})$ may be dominated or degenerate in general. The point is that for any action $a \in \bar{\mathcal{P}}(\mathcal{E})$, there exists a Pareto optimal action $b \in \mathcal{P}(\mathcal{E})$ with larger regret under some plausible parameter (up to a constant factor). The local alignment constant for $\mathcal{E}$ is

$$\alpha_\eta(\mathcal{E}) \triangleq \max_{\nu \in \mathcal{V}} \max_{a,b \in \mathcal{P}(\mathcal{E})} \min_{c \in \bar{\mathcal{P}}_\eta(\mathcal{E})} \frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\|M_c \nu\|^2} \,. \tag{16}$$

We let $\alpha(\mathcal{E}) \triangleq \inf_{\eta > 0} \eta^2 \alpha_\eta(\mathcal{E})$. A game is called *locally observable* if

$$\sup_{\mathcal{E} \subset \Theta} \alpha(\mathcal{E}) < \infty \,. \tag{17}$$

The next lemma is helpful to to bound the alignment constant (Kirschner et al., 2020, c.f. Lemma 13).

Lemma 5. *Let $a, b \in \mathcal{A}$ with $\phi_a \ne \phi_b$, and $\mathcal{B} \subset \mathcal{A}$ be a subset of actions such that*

$$Proj_\mathcal{V}(\phi_a - \phi_b) \in \text{span}(\text{im}(Proj_\mathcal{V} M_c^\top) : c \in \mathcal{B}) \,.$$

Then there exist weights $w_{ab}^c \in \mathbb{R}^m$ for each $c \in \mathcal{B}$, such that $\langle \phi_a - \phi_b, \nu \rangle = \sum_{c \in \mathcal{B}} \langle M_c^\top w_{ab}^c, \nu \rangle$ for all $\nu \in \mathcal{V}$. Further, for any such weights,

$$\max_{\nu \in \mathcal{V}} \min_{c \in \mathcal{B}} \frac{\langle \phi_a - \phi_b, v \rangle^2}{\|M_c^\top v\|^2} \le \left(\sum_{c \in \mathcal{B}} \|w_{ab}^c\| \right)^2 \,.$$

Proof The existence of the weights $w_{ab}^c \in \mathbb{R}^m$ is immediate by assumption. Therefore, we can write

$$\langle \phi_a - \phi_b, \nu \rangle^2 = \langle \sum_{c \in \mathcal{B}} M_c^\top w_{ab}^c, \nu \rangle^2 = \left(\sum_{c \in \mathcal{B}} \langle w_{ab}^c, M_c \nu \rangle \right)^2 ,$$

14

Example	Local / Global	Alignment Constant	IDS Regret $\mathfrak{R}_n$
Linear Bandit (Ex. 1)	Local	$\alpha(\mathcal{E}_t) \leq 4$	$\mathcal{O}(\sqrt{n}d\log(n))$
Dueling Bandit (Ex. 3)	Local	$\alpha(\mathcal{E}_t) \leq 4$	$\mathcal{O}(\sqrt{n}d\log(n))$
Graph Dueling Bandit (Ex. 4)	Global	$\alpha(\mathcal{E}_t) \leq 4W_{\max}^2$	$\mathcal{O}((nd\log(n)W_{\max})^{2/3})$
Non-Degenerate Finite PM	Local	$\alpha(\mathcal{E}_t) \leq 4k^2m^2$	$\mathcal{O}(km^{3/2}n^{1/2}\,r\log(nr))$
Finite PM	Local	$\alpha(\mathcal{E}_t) \leq 4mdk^{d+2}$	$\mathcal{O}((nmdk^{d+2})^{1/2}r\log(rn))$
Finite PM	Global	$\alpha \leq mdk^{d+2}$	$\mathcal{O}\big((mdk^{d+2})^{1/3}(rn\log(rn))^{2/3}\big)$

Table 1: The regret bound follows from Theorems 7 and 8, and $\beta_n, \gamma_n \leq \mathcal{O}(r\log(rn))$.

An application of the Cauchy-Schwarz inequality shows:

$$\frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\max_{c\in\mathcal{B}} \|M_c\nu\|^2} \leq \frac{\left(\sum_{c\in\mathcal{B}} \|w_{ab}^c\|\|M_c\nu\|\right)^2}{\max_{c\in\mathcal{B}} \|M_c\nu\|^2} \leq \left(\sum_{c\in\mathcal{B}} \|w_{ab}^c\|\right)^2.$$

∎

The lemma can be used to bound the alignment constant for various games that were introduced as examples. We refer to Table 1 for an overview.

Linear Bandits are locally observable. This follows from the observation that for any two actions a, b, we have $\phi_a - \phi_b \in \text{span}\{\phi_a, \phi_b\}$. We can choose the estimation vector $w_{a,b}^a = -w_{a,b}^b = 1$ and $w_{a,b}^c = 0$ for all $c \notin \{a, b\}$. Therefore $\alpha(\mathcal{E}_t) \leq \alpha_1(\mathcal{E}_t) \leq \max_{a,b\in\mathcal{A}}(|w_{a,b}^a| + |w_{a,b}^b|)^2 = 4$.

Dueling Bandits are also locally observable. Recall that we defined dueling bandits on a ground set $\mathcal{I}$ with features ϕ_a for all $a \in \mathcal{I}$ by letting $\mathcal{A} = \mathcal{I} \times \mathcal{I}$, $\phi_{a,b} = \phi_a + \phi_b$ and $M_{a,b} = \phi_a - \phi_b$. Note that dueling actions (a, b) with $\phi_a \neq \phi_b$ cannot be Pareto optimal. Let $(a, a), (b, b) \in \mathcal{A}$ be a pair of Pareto optimal dueling actions. A simple calculation reveals that $\phi_{a,b} \in \text{conv}(\{\phi_{a,a}, \phi_{b,b}\})$, and $\phi_{a,a} - \phi_{b,b} = 2M_{a,b}$. It follows from Lemma 5 that $\alpha(\mathcal{E}_t) \leq \alpha_1(\mathcal{E}_t) \leq 4$.

Graph Dueling Bandits use the same reward features and feedback maps as dueling bandits, but restrict the action set to a subset $\mathcal{A} \subset \mathcal{I} \times \mathcal{I}$. Define $\text{dist}(a, b)$ as the shortest undirected path from a to b in the graph defined by $(\mathcal{I}, \mathcal{A})$, or infinity if no such path exists. Let $W_{\max} = \max_{a,b\in\mathcal{P}} \text{dist}(a, b)$ be the maximum length of a shortest path between any two Pareto optimal actions. This game is globally observable if and only if $W_{\max} < \infty$.

To see this, note that for any pair of Pareto optimal dueling actions (a, a), (b, b) we can find a sequence $c_0, c_1, \ldots, c_l$ with $c_0 = a$ and $c_l = b$ and $l \leq W_{\max}$, such that $(c_i, c_{i+1}) \in \mathcal{A}$ for all $i = 0, \ldots, l-1$. In particular, we can write $\phi_{a,a} - \phi_{b,b} = 2\sum_{i=1}^{l-1} M_{c_i,c_{i+1}}$ and Lemma 5 implies a bound on the alignment constant, $\alpha \leq 2W_{\max}$.

Finite Partial Monitoring Bounds for finite partial monitoring games are derived in the next lemma. A finite partial monitoring game is called non-degenerate if every action is either Pareto optimal or dominated and there are no duplicate actions.

Lemma 6. Let $\mathcal{E} \subset \Theta$ be convex. For finite partial monitoring games the following bounds on the alignment constant hold:

15

(a) *For globally observable games $\alpha \leq mdk^{d+2}$; and*

(b) *for locally observable games $\alpha(\mathcal{E}) \leq 4mdk^{d+2}$; and*

(c) *for non-degenerate locally observable games $\alpha(\mathcal{E}) \leq 4k^2m^3$.*

The proof is given in Appendix B.2.

4.2 Regret Bounds

Using the notion of local and global observability, we can now state the main results. To interpret the results, note that $\gamma_n, \beta_n \leq \mathcal{O}(r\log(1+n))$ by Lemma 4. For an overview on how the regret bounds behave on standard examples, see Table 1.

Theorem 7 (Globally Observable Games). *Let $\mathcal{G}$ be a globally observable game that saties-fies Assumption 1 with worst-case alignment α according to Eq. (15). Then the regret of IDS (Algorithm 2) with $\lambda \geq L$ satisfies*

$$\mathfrak{R}_n \leq n^{2/3}(54\alpha\mathbb{E}[\beta_n]\mathbb{E}[\gamma_n(\Delta_{\max} + 4B)])^{1/3} + \mathcal{O}(\Delta_{\max}).$$

Consequently, $\mathfrak{R}_n \leq \mathcal{O}(\alpha^{1/3}r\log(1+n)n^{2/3})$.

Proof Combining the general IDS regret bound in Theorem 1 with $\kappa = 3$ and using Eq. (11) to bound the estimation error, we find

$$\mathfrak{R}_n \leq n^{2/3}\mathbb{E}[\gamma_n]^{1/3}\left(\frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n}\Psi_{3,t}\right]\right)^{1/3} + \mathcal{O}(\Delta_{\max}).$$

The next step is to bound the information ratio. The main idea is to optimize the trade-off between playing a greedy action $\hat{a}_t = \arg\max_{a\in\mathcal{A}}\langle\phi_a, \hat{\theta}_t\rangle$ and the action that maximizes information gain. As it turns out, this is sufficient to bound the information ratio in globally observable games. We emphasize that this particular choice of actions appears only in the analysis. The actual IDS distribution may be supported on actions that achieve an even smaller information ratio.

Note that we may always choose $\hat{a}_t \in \mathcal{P}$ as a Pareto optimal action. Hence

$$\hat{\Delta}_t(\hat{a}_t) = \max_{\theta\in\mathcal{E}_t}\max_{b\in\mathcal{A}}\langle\phi_b - \phi_{\hat{a}_t}, \theta\rangle = \max_{\theta\in\mathcal{E}_t}\max_{b\in\mathcal{P}}\langle\phi_b - \phi_{\hat{a}_t}, \theta\rangle \leq \max_{\theta\in\mathcal{E}_t}\max_{b\in\mathcal{P}}\langle\phi_b - \phi_{\hat{a}_t}, \theta - \hat{\theta}_t\rangle.$$

The equality uses that we can choose the maximizer in $\mathcal{P}$. Using the definition of the global alignment constant α in Eq. (15), we find

$$\hat{\Delta}_t(\hat{a}_t)^2 \leq \max_{\theta\in\mathcal{E}_t}\max_{b\in\mathcal{P}}\langle\phi_b - \phi_{\hat{a}_t}, \theta - \hat{\theta}_t\rangle^2 \leq \alpha\max_{\theta\in\mathcal{E}_t}\max_{c\in\mathcal{A}}\|M_c(\theta - \hat{\theta}_t)\|^2. \tag{18}$$

Using Eq. (8) and Cauchy-Schwarz, we can further bound for all $c \in \mathcal{A}$,

$$\max_{\theta\in\mathcal{E}_t}\|M_c(\theta - \hat{\theta}_t)\|^2 = \max_{\theta\in\mathcal{E}_t}\|M_c WW^\top(\theta - \hat{\theta}_t)\|^2$$

$$\leq \max_{\theta\in\mathcal{E}_t}\|W^\top(\theta - \hat{\theta}_t)\|_{W_t}^2\|M_c W\|_{W_t^{-1}}^2 \leq \beta_t\|M_c W\|_{W_t^{-1}}^2 \tag{19}$$

16

The last inequality follows from the definition of the confidence set $\mathcal{E}_t$. Moreover, note that $\|M_c W\|^2_{W_t^{-1}} \leq \|M_c W\|^2_{W_0^{-1}} \leq L\lambda^{-1} \leq 1$ by our assumption on λ. Hence, using further that $x \leq 2\log(1+x)$ for $x \in [0,1]$,

$$
\begin{aligned}
\|M_c W\|^2_{W_t^{-1}} &= \lambda_{\max}(M_c W W_t^{-1}(M_c W)^\top) \\
&\leq 2\log(1 + \lambda_{\max}(M_c W W_t^{-1}(M_c W)^\top)) \\
&\leq 2\log\det(\mathbf{1}_m + M_c W W_t^{-1}(M_c W)^\top) = 4I_t(c)
\end{aligned}
\tag{20}
$$

Taking the previous two displays together, we get $\max_{\theta \in \mathcal{E}_t} \|M_c(\theta - \hat{\theta}_t)\|^2 \leq 4\beta_t I_t(c)$. Combined with Eq. (18), we get

$$
\hat{\Delta}_t(\hat{a}_t)^2 \leq 4\alpha\beta_t \max_{c \in \mathcal{A}} I_t(c) \,.
$$

To bound the generalized information ratio $\Psi_{3,t}$ let $c_t = \arg\max_{a \in \mathcal{A}} I_t(a)$. Using Lemma 2, we find

$$
\begin{aligned}
\Psi_{3,t}(\mu_t) \leq 2\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_{3,t}(\mu) &\overset{(i)}{\leq} 2\min_{p \in [0,1]} \frac{\left((1-p)\hat{\Delta}_t(\hat{a}_t) + p\hat{\Delta}_t(c_t)\right)^3}{pI_t(c_t)} \\
&\overset{(ii)}{\leq} \frac{27\hat{\Delta}_t(\hat{a}_t)^2 \hat{\Delta}_t(c_t)}{2I_t(c_t)} \\
&\overset{(iii)}{\leq} 54\alpha\beta_t(\Delta_{\max} + 4B)
\end{aligned}
\tag{21}
$$

Step (i) uses that $I_t(\hat{a}_t) \geq 0$, (ii) follows with $p = \frac{\hat{\Delta}_t(\hat{a}_t)}{2\hat{\Delta}_t(c_t)}$ and (iii) the inequality in the previous display and a direct consequence of the boundedness Assumption 1, $\hat{\Delta}_t(c_t) \leq \Delta_{\max} + \max_{a,b} \max_{\theta \in \mathcal{E}_t} \langle \theta - \theta^*, \phi_a - \phi_b \rangle \leq \Delta_{\max} + 4B$.

The final bound follows using Lemma 4 to bound β_n and γ_n. ∎

The locally observable case is summarized in the next theorem.

Theorem 8. *Let $\mathcal{G}$ be a locally observable partial monitoring game that satisfies Assumption 1. Then the regret of IDS (Algorithm 2) with regularizer $\lambda \geq L$ satisfies*

$$
\mathfrak{R}_n \leq \sqrt{8\mathbb{E}[\bar{\alpha}_n\beta_n]\mathbb{E}[\gamma_n]n} + \mathcal{O}(\Delta_{\max}) \,,
$$

where $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha(\mathcal{E}_t)$ is the average realized local alignment constant. In particular, $\mathfrak{R}_n \leq \mathcal{O}(r\sqrt{n\mathbb{E}[\bar{\alpha}_n]}\log(1+n))$.

Proof We start once more with the general IDS regret bound in Theorem 1 for $\kappa = 2$ and use Eq. (11) to bound the estimation error, which gives

$$
\mathfrak{R}_n \leq \sqrt{\mathbb{E}[\gamma_n]n}\sqrt{\frac{1}{n}\sum_{t=1}^n \Psi_{2,t}(\mu_t)} + \mathcal{O}(\Delta_{\max}) \,.
$$

To bound the information ratio, we make use of the plausible Pareto action set $\mathcal{P}(\mathcal{E}_t)$ and its convex relaxation $\bar{\mathcal{P}}(\mathcal{E}_t)$. For any $\eta > 0$, let $c_t = c_t(\eta) = \arg\max_{a \in \bar{\mathcal{P}}_\eta(\mathcal{E}_t)} I_t(a)$ be the most informative action in $\bar{\mathcal{P}}_\eta(\mathcal{E}_t)$. By the definition of $\bar{\mathcal{P}}_\eta(\mathcal{E}_t)$, we bound the gap estimate as follows,

$$\eta^{-1}\hat{\Delta}_t(c_t) \leq \max_{a,b \in \mathcal{P}(\mathcal{E}_t)} \max_{\theta \in \mathcal{E}_t} \langle \phi_a - \phi_b, \theta \rangle \leq \max_{a,b \in \mathcal{P}(\mathcal{E}_t)} \max_{\theta, \nu \in \mathcal{E}_t} \langle \phi_a - \phi_b, \theta - \nu \rangle. \tag{22}$$

The second inequality follows since $\nu \in \mathcal{E}_t$ can be chosen such that b is optimal for ν. Consequently the definition of the local alignment constant in Eq. (16) implies

$$\hat{\Delta}_t(c_t)^2 \leq \eta^2 \alpha_\eta(\mathcal{E}_t) \max_{\nu,\omega \in \mathcal{E}_t} \max_{c \in \bar{\mathcal{P}}_\eta(\mathcal{E}_t)} \|M_c(\nu - \omega)\|^2$$

To relate the norm to the information gain, note that for any $c \in \mathcal{A}$,

$$\max_{a,b \in \mathcal{P}(\mathcal{E}_t)} \|M_c(\nu - \omega)\|^2 = \max_{a,b \in \mathcal{P}(\mathcal{E}_t)} \|M_c W W^\top (\nu - \omega)\|^2$$
$$\leq \max_{a,b \in \mathcal{P}(\mathcal{E}_t)} \|W^\top(\nu - \omega)\|_{W_t}^2 \|M_c W\|_{W_t^{-1}}^2 \leq 2\beta_t \|M_c W\|_{W_t^{-1}}^2.$$

From the argument in Eq. (20), it follows that $\|M_c W\|_{W_t^{-1}}^2 \leq 4I_t(c)$. Therefore we bound the information ratio for the action c_t as follows,

$$\hat{\Delta}_t(c_t)^2 \leq \max_{c \in \bar{\mathcal{P}}_\eta} 2\eta^2 \alpha_\eta(\mathcal{E}_t)\beta_t \|M_c W\|_{W_t^{-1}}^2 \leq 8\eta^2 \alpha_\eta(\mathcal{E}_t)\beta_t I_t(c_t).$$

This shows that to bound the information ratio it suffices to play c_t,

$$\Psi_t(\mu_t) = \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) \leq \inf_{\eta > 0} \min_{c \in \bar{\mathcal{P}}_\eta(\mathcal{E}_t)} \frac{\hat{\Delta}_t(c)^2}{I_t(c)} \leq \inf_{\eta > 0} 8\eta^2 \alpha_\eta(\mathcal{E}_t)\beta_t = 8\alpha(\mathcal{E}_t)\beta_t,$$

and the regret bound follows because (β_t) is increasing, and using Lemma 4 to bound β_n and γ_n. ∎

4.3 Classification of Finite Action Games

For the upper bounds, so far we have encountered two cases: globally and locally observable games that satisfy the conditions in Eqs. (14) and (17) respectively. There are two other types of games with a less interesting structure. A game is called *trivial* if there exists an action a such that $\mathcal{C}_a = \Theta$. By definition, a is optimal for all possible values of θ_*. IDS achieves zero regret on these games because $\hat{\Delta}_t(a) = 0$ and the information ratio is minimized by a Dirac on a.

Games that are locally observable but not trivial are called *easy*. Games that are globally observable but not locally observable are called *hard*. The last category of games are those that are not globally observable, which are called *hopeless*. The reason is that in these games there are necessarily multiple (non-duplicate) Pareto optimal actions that cannot be distinguished, which means the learner suffers linear regret in the worst case.

The *minimax regret* is $\mathfrak{R}_n^* = \inf_\pi \sup_{\theta \in \Theta} \mathbb{E}[R_n(\pi, \theta)]$, where the infimum is over all possible policies π that map observation histories to actions. The classification theorem provides the minimax regret rate for each type of game.

Theorem 9 (Classification). *For any finite linear partial monitoring game, the minimax regret $\mathfrak{R}_n^*$ satisfies*

$$\mathfrak{R}_n^* = \begin{cases} 0 & \text{for trival games,} \\ \tilde{\Theta}(n^{1/2}) & \text{for easy games,} \\ \tilde{\Theta}(n^{2/3}) & \text{for hard games,} \\ \Omega(n) & \text{otherwise, for hopeless games.} \end{cases}$$

The classification theorem recovers the known result for finite stochastic partial monitoring up to logarithmic factors (Bartók et al., 2014; Lattimore and Szepesvári, 2019c). It further generalizes the results for linear partial monitoring without explicit constraints on the parameter sets by Kirschner et al. (2020).

Proof As we have argued, IDS provides the upper bounds in each case. Therefore it remains to provide lower bounds for each category. These follow essentially from the results in (Kirschner et al., 2020, Appendix F), by restricting all constructions to the subspace $\mathcal{V}$. ∎

We conclude the section remarking that the situation in continuous action games is much more delicate. As shown in (Kirschner et al., 2020, Section 2.4) the achievable rates are not only determined by the observability conditions, but further depend on the curvature of the action set. A full classification of continuous action games is still a fascinating open question.

5. Extensions

5.1 Faster Gap Estimation

The version of IDS presented in Algorithm 2 requires $\mathcal{O}(|\mathcal{A}|^2)$ computation steps per round in general, which quickly becomes prohibitive for large action sets. The main bottleneck is the computation of the gap estimates as defined in Eq. (10). We propose a novel gap estimate that can be computed in linear time. To this end, let $\hat{a}_t = \arg\max_{a \in \mathcal{P}}\langle \hat{\theta}, a \rangle$ be the empirically best action, chosen as a Pareto optimal action if not unique. We relax the gap estimate using $\hat{a}_t$ as an intermediate action:

$$\hat{\Delta}_t(a; \hat{a}_t) \triangleq \delta_t + \max_{\theta \in \mathcal{E}_t}\langle \phi_{\hat{a}_t} - \phi_a, \theta \rangle, \quad \text{where} \quad \delta_t \triangleq \max_{\theta \in \mathcal{E}_t}\max_{b \in \mathcal{A}}\langle \phi_b - \phi_{\hat{a}_t}, \theta \rangle. \tag{23}$$

It is immediate that $\hat{\Delta}_t(a) \leq \Delta_t(a; \hat{a}_t)$ holds for all $a \in \mathcal{A}$. Note that $\hat{a}_t$ can be found by enumerating the Pareto optimal actions, and the gap $\Delta_t(a; \hat{a}_t)$ estimate can be computed for all $a \in \mathcal{A}$ by solving $2k$ second-order cone programs over Θ with positive semi-definite quadratic constraints. Combined with the $\frac{4}{3}$-approximation of the IDS distribution using Eq. (2), this allows us to reduce the computational complexity of Algorithm 2 to $\mathcal{O}(|A|)$ per round. The next lemma confirms that the new algorithm satisfies essentially the same bounds as before.

Lemma 10. *Algorithm 2 with the gap estimate $\hat{\Delta}_t(a; \hat{a}_t)$ and approximate IDS sampling according to Eq. (23) satisfies the same bounds up to constants as in Theorems 7 and 8.*

Proof The proofs of Theorems 7 and 8 go through almost unchanged. We only need to derive an upper bound on the new gap estimate, as we explain below.

In the globally observable case, we show that the analog of Eq. (18) holds. By definition of $\hat{\Delta}_t(a; \hat{a}_t)$ we get

$$\hat{\Delta}_t(\hat{a}_t; \hat{a}_t)^2 \leq \max_{\theta \in \mathcal{E}_t} \max_{b \in \mathcal{P}} \langle \phi_b - \phi_{\hat{a}_t}, \theta - \hat{\theta}_t \rangle^2$$

The remaining proof remains unchanged.

In the locally observable case, we show that Eq. (22) continues holds up to a factor of two. For all $c \in \mathcal{P}(\mathcal{E}_t)$, we get

$$\hat{\Delta}_t(c; \hat{a}_t) \leq 2 \max_{a, b \in \mathcal{P}(\mathcal{E}_t)} \max_{\theta \in \mathcal{E}_t} \langle \phi_a - \phi_b, \theta \rangle$$

The remaining proof remains unchanged with an additional factor of two in the leading term. ∎

Lastly, we remark that when $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$, then we can also use the following estimate based on truncation:

$$\tilde{\Delta}_t(a; \hat{a}_t) \triangleq \min\{\delta_t + \langle \phi_{\hat{a}_t} - \phi_a, \hat{\theta}_t \rangle, B\}, \quad \text{where} \quad \delta_t \triangleq \max_{\theta \in \mathcal{E}_t} \max_{b \in \mathcal{A}} \langle \phi_b - \phi_{\hat{a}_t}, \theta \rangle. \tag{24}$$

The main difference to $\hat{\Delta}_t(a; \hat{a}_t)$ is that the mean-gap $\langle \phi_{\hat{a}_t} - \phi_a, \hat{\theta}_t \rangle$ appears directly in the estimation. One easily confirms that $\hat{\Delta}_t(a, \hat{a}_t) \leq 2\tilde{\Delta}_t(\hat{a}_t, a_t)$ and the proof can be adjusted to accommodate this change. The advantage of $\tilde{\Delta}_t(a; \hat{a}_t)$ is that it is less dependent on the tightness of the confidence bound thereby can be more accurate in practice.

5.2 Directed Information Gain

The information gain function Eq. (12) based on the log determinant potential is a convenient choice for the worst-case analysis, but can be conservative in practice. The main reason is that Eq. (12) is agnostic to the current parameter estimate and the geometric structure of the game. Therefore, IDS acquires information that leads to uncertainty reduction on the true parameter θ^* overall, independent of the true maximizer a^*.

To introduce a form of *directed* information gain, we define for any $\omega \in \mathbb{R}^d$,

$$J_t(a; \omega) \triangleq \frac{1}{2} \|M_a \omega\|^2. \tag{25}$$

Note that $J_t(a; \omega)$ measures the sensitivity of the feedback obtained from action $a \in \mathcal{A}$ along the direction ω. Next we define optimistic and pessimistic parameters

$$\theta_t^+, \theta_t^- = \arg\max_{\theta_1, \theta_2 \in \mathcal{E}_t} \max_{a, b \in \mathcal{P}(\mathcal{E}_t)} \langle \phi_b - \phi_a, \theta_1 - \theta_2 \rangle.$$

Both parameters can be computed by solving $|\mathcal{P}|^2$ second-order cone programs over $\mathcal{E}_t$. The computation complexity can be reduced to $2|\mathcal{P}|$ using a relaxation on $\hat{a}_t$, similarly to the previous section. We define the directed information gain

$$J_t(a) = \beta_t^{-1} J_t(a; \theta_t^+ - \hat{\theta}_t^-).$$

Lemma 11. *Algorithm 2 with the information gain $J_t(a)$ instead of $I_t(a)$ satisfies the same bounds as in Theorems 7 and 8.*

Proof First note that the new information gain is upper bounded by the information gain I_t (Eq. (12)) as follows,

$$J_t(c) = \beta_t^{-1}\|M_c(\theta_t^+ - \theta_t^-)\|^2 \leq 4I_t(c).$$

The inequality follows along the same lines as Eq. (19). In particular, the total information gain remains bounded, $\gamma_{n,J} = \sum_{t=1}^n J_t(a_t) \leq 4\sum_{t=1}^n I_t(a_t) = 4\gamma_{n,I}$.

Therefore it remains to show that the information ratio is bounded. For the globally observable case, note that

$$\hat{\Delta}_t(\hat{a}_t)^2 \leq \max_{b\in\mathcal{P}}\max_{\nu\in\mathcal{E}_t}\langle\phi_b - \phi_{\hat{a}_t}, \nu - \hat{\theta}_t\rangle^2 \leq \max_{b\in\mathcal{P}}\langle\phi_b - \phi_{\hat{a}_t}, \theta^+ - \theta^-\rangle^2$$

$$\leq \alpha\max_{c\in\mathcal{P}}\|M_c(\theta^+ - \theta^-)\| = \alpha\beta_t\max_{c\in\mathcal{A}}J_t(c).$$

This follows along the same lines as Eq. (18), and the remaining proof in the globally observable case remains unchanged.

In the locally observable case, we reproduce Eq. (22) as follows

$$\hat{\Delta}_t(c_t)^2 \leq \max_{a,b\in\mathcal{P}(\mathcal{E}_t)}\langle\phi_b - \phi_{\hat{a}}, \theta^+ - \theta^-\rangle^2$$

$$\leq \alpha(\mathcal{E}_t)\max_{c\in\bar{\mathcal{P}}(\mathcal{E}_t)}\|M_c(\theta^+ - \theta^-)\|^2 = \beta_t\alpha(\mathcal{E}_t)\max_{c\in\mathcal{A}^+(\mathcal{E}_t)}J_t(c).$$

The remaining proof follows unchanged. ∎

6. Contextual Partial Monitoring

In the contextual bandit problem, the environment provides a context in each round *before* the learner chooses an action (Langford and Zhang, 2008). Depending on our assumptions, the context may be either sampled from a fixed distribution or chosen adversarially by the environment. The reward depends on the context and the chosen action, and the learner competes with the best context-dependent action. In applications, the context can represent additional information available to the learner, such as temperature measurements, time of day, or the profile of a user visiting a website. We now introduce a *contextual version of linear partial monitoring*. This setting directly generalizes the linear contextual bandit model.

Let $\mathcal{Z}$ be a compact context set and $z_t \in \mathcal{Z}$ the context presented at time t. Reward features $\phi_a^z \in \mathbb{R}^d$ and feedback map $M_a^z \in \mathbb{R}^{d\times m}$ depend on the context $z \in \mathcal{Z}$. For each context $z \in \mathcal{Z}$ a subset of actions $\mathcal{A}(z) \subset \mathcal{A}$ is available for playing. The reward function is parameterized by a single $\theta^* \in \mathbb{R}^d$ shared among all contexts. The feedback in round t for action a_t and context z_t is $y_t = M_{a_t}^{z_t}\theta^* + \epsilon_t$, where $\epsilon_t \in \mathbb{R}^m$ is conditionally independent ρ-sub-Gaussian noise. The regret is defined so that the learner competes with the best action $a^*(z) = \arg\max_{a\in\mathcal{A}(z)} f(a, z)$ chosen in hindsight for each context $z \in \mathcal{Z}$. For a sequence of contexts $(z_t)_{t=1}^n$, the *contextual regret* is

$$\mathfrak{R}_n(\pi, \theta^*, (z_t)_{t=1}^n) = \mathbb{E}\left[\sum_{t=1}^n\langle\phi_{a^*(z_t)}^{z_t} - \phi_{a_t}^{z_t}, \theta^*\rangle\right].$$

Over the next two sections, we develop IDS policies for the contextual partial monitoring setting. The first variant directly extends Algorithm 1 by conditioning the information ratio on the observed context z_t. We refer to this variant as *conditional IDS*. The second variant assumes that the context is sampled from a fixed and known distribution. This allows the learner to optimize the information ratio directly over the joint action-context distribution, which we refer to *contextual IDS*. We show that this allows for much weaker conditions where sublinear regret is possible.

For simplicity, our presentation focuses on the case with unconstrained parameter set, $\Theta = \mathbb{R}^d$. As before, we assume boundedness of the observation and reward features $\|M_a^z\|_2 \leq L$ and $\mathrm{diam}(\phi(a,z) : a \in \mathcal{A}, z \in \mathcal{Z}) \leq 1$ and parameter $\|\theta^*\|_2 \leq B$. The case with constrained parameter set can be handled as explained in Section 3 by introducing appropriate projections. The product action space over the contexts with some fixed ordering is $\mathcal{A}(\mathcal{Z}) \triangleq \times_{z \in \mathcal{Z}} \mathcal{A}(z)$.

6.1 Conditional IDS

The definitions of the gap estimate and the information gain directly extend to the contextual setting. As before $\hat{\theta}_t$ denotes the least squares estimate defined in Eq. (7), and $\mathcal{E}_t = \mathcal{E}_{t,1/t^2}$ the calibrated confidence set defined in Lemma 3. The conditional gap estimate for a context $z \in \mathcal{Z}$ and action $a \in \mathcal{A}(z)$ is

$$\hat{\Delta}_t(a,z) = \max_{b \in \mathcal{A}(z)} \max_{\theta \in \mathcal{E}_t} \langle \phi_b^z - \phi_a^z, \theta \rangle . \tag{26}$$

The (undirected) information gain for action $a \in \mathcal{A}$ and context $z \in \mathcal{Z}$ is

$$I_t(a,z) = \frac{1}{2} \log \det(\mathbf{1} + M_a^{z\top} V_t^{-1} M_a^z) . \tag{27}$$

Conditional IDS is the policy that optimizes the information ratio conditioned on the observed context z_t at time t,

$$\mu_t(z_t) = \underset{\mu \in \mathscr{P}(\mathcal{A}(z_t))}{\arg\min} \left\{ \Psi(\mu, z_t) = \frac{\hat{\Delta}_t(\mu, z_t)^2}{I_t(\mu, z_t)} \right\} , \tag{28}$$

and samples the action according to $a_t \sim \mu_t(z_t)$. The computational complexity required to find the minimizer of the information ratio is the same as in the non-contextual case.

Regret Bounds For the analysis, we extend the notion of the alignment constant with the contextual argument. We let $\mathcal{P}(z)$ be the set of Pareto optimal actions for context $z \in \mathcal{Z}$ defined as actions that are uniquely optimal for some parameter θ and context z. The conditional global alignment constant is

$$\alpha(z) \triangleq \max_{\omega \in \mathbb{R}^d} \max_{a,b \in \mathcal{P}(z)} \min_{c \in \mathcal{A}(z)} \frac{\langle \phi_a^z - \phi_b^z, \omega \rangle^2}{\|M_c^z \omega\|^2} . \tag{29}$$

Further, let $\Delta(a,z|\theta) = \max_{b \in \mathcal{A}(z)} \langle \phi_b^z - \phi_a^z, \theta \rangle$ be the gap of action $a \in \mathcal{A}(z)$ under parameter $\theta \in \Theta$ and context $z \in \mathcal{Z}$ and define contextual extensions of the (extended) plausible Pareto

optimal actions as follows:

$$\mathcal{P}(\mathcal{E}, z) \triangleq \cup_{\theta \in \mathcal{E}} \{a \in \mathcal{P}(z) : \langle \phi_a^z, \theta \rangle = \max_{b \in \mathcal{A}(z)} \langle \phi_b^z, \theta \rangle \},$$

$$\bar{\mathcal{P}}_\eta(\mathcal{E}, z) \triangleq \{a \in \mathcal{A}(z) : \Delta(a, z|\theta) \leq \eta \cdot \max_{b \in \mathcal{P}(\mathcal{E}, z)} \Delta(a, z|\theta) \}.$$

The local alignment constant for $\eta > 0$ under context $z \in \mathcal{Z}$ is then

$$\alpha_\eta(\mathcal{E}, z) \triangleq \max_{\omega \in \mathbb{R}^d} \max_{a, b \in \mathcal{P}(\mathcal{E}, z)} \min_{c \in \bar{\mathcal{P}}_\eta(\mathcal{E}, z)} \frac{\langle \phi_a^z - \phi_b^z, \omega \rangle^2}{\|M_c^z \omega\|^2}, \tag{30}$$

and we let $\alpha(\mathcal{E}, z) \triangleq \inf_{\eta > 0} \eta^2 \alpha_\eta(\mathcal{E}, z)$.

The next two results are immediate extensions of the upper bound for globally and locally observable games that replace the bounds on the information ratio with their conditional counterpart. These bounds are only meaningful if the game is globally or locally observable for each observed context $z_t \in \mathcal{Z}$.

Corollary 12. *For any $\mathcal{F}_t$-predictable sequence $(z_t)_{t=1}^n$ in $\mathcal{Z}$, the regret of conditional IDS satisfies,*

$$\mathfrak{R}_n \leq \mathcal{O}\left(n^{2/3}(\mathbb{E}[\bar{\alpha}_n \beta_n]\mathbb{E}[\gamma_n])^{1/3}\right),$$

where $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha(z_t)$ is the average global alignment on the observed sequence of contexts.

Proof The claim follows along the same lines of as Theorem 7. ∎

The result for the locally observable case is stated similarly.

Corollary 13. *For any $\mathcal{F}_t$-predictable sequence $(z_t)_{t=1}^n$ in $\mathcal{Z}$, the regret of conditional IDS satisfies*

$$\mathfrak{R}_n \leq \sqrt{8\mathbb{E}[\bar{\alpha}_n \beta_n]\mathbb{E}[\gamma_n]n} + \mathcal{O}(1),$$

where $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha(\mathcal{E}_t, z_t)$ is the average local alignment constant for the sequence of confidence sets $(\mathcal{E}_t)_{t=1}^n$ realized by the algorithm.

Proof Along the lines of Theorem 8. ∎

In the contextual linear bandit setting, $\alpha(\mathcal{E}_t, z) \leq 4$ holds for all $z \in \mathcal{Z}$ (cf. Table 1). Therefore, Corollary 13 recovers the same bounds as the optimistic approach in the linear contextual bandit setting (cf., Abbasi-Yadkori et al., 2011). Moreover, we immediately get a regret bound for the contextual extension of the dueling bandit setting in Theorem 8.

6.2 Using the Context Distribution for Exploration

Perhaps surprisingly, the contextual case allows for much weaker conditions under which the learner can achieve sublinear regret. This is possible if the learner exploits the distribution of contexts. Here we study the case where the context follows a fixed and known distribution $\chi \in \mathscr{P}(\mathcal{Z})$. We only address the case where χ is known exactly. If the distribution is unknown, it is natural to replace χ with an online estimate of the context distribution (cf. Tirinzoni et al., 2020). Extending the analysis to unknown context distribution is left as an important question for future work.

Example 11 (Non-Informative Context). *Consider the case where for some $z \in \mathcal{Z}$ the learner obtains no information, i.e. $M_a^z = 0$ for all $a \in \mathcal{A}(z)$. In this case, the only sensible choice is the greedy action $a^*(\hat{\theta}_t)$. The learner has to explore in rounds where information is available and the sampling distribution needs to be sufficiently diverse to account for rounds where the learner is forced to play greedily. Note that while there can be vanishing information gain in* some *rounds, the* expected *information gain with respect to the context distribution χ is non-zero. A natural application is in customer surveys: Clients who agree to provide feedback can be asked specifically targeted questions, whereas feedback from other customers is never observed.*

Example 12 (Greedy Exploration). *Another interesting case is when feedback from the optimal action in each context is sufficiently diverse to allow estimation of the parameter without* further *exploration. In such cases, the greedy algorithm can be highly effective. This effect has been studied in the bandit literature before (Bastani et al., 2017; Hao et al., 2019). Conceptually, one can think of the context as part of the action space, where the sampling distribution is imposed by the environment.*

The conditional IDS distribution (Eq. (28)) is independent of the context distribution. It is easy to see that it behaves sub-optimally in both examples, and the bounds in Corollaries 12 and 13 become vacuous when a context occurs where the information gain is zero for all actions. To exploit the randomness of the context in the regret bounds, we include the contextual distribution in the information ratio. By optimizing the joint distribution over action and context, we obtain an IDS algorithm that makes use of the contextual distribution for exploration.

Denote by $\mathscr{P}_\times(\mathcal{A}(\mathcal{Z})) = \times_{z \in \mathcal{Z}} \mathscr{P}(\mathcal{A}(z))$ the set of probability kernels that assign each context z a sampling distribution over $\mathcal{A}$. Suggestively we write $\xi(a|z) = \xi(a, z)$ for elements in $\mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))$. For context distribution $\chi \in \mathscr{P}(\mathcal{Z})$ and kernel $\xi \in \mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))$, we extend the definition of the gap estimates in Eq. (26) to

$$\hat{\Delta}_t(\xi, \chi) \triangleq \int_{\mathcal{Z}} \int_{\mathcal{A}(z)} \hat{\Delta}_t(a, z) d\xi(a|z) d\chi(z) \,. \tag{31}$$

As information gain we use Eq. (27) with the same convention that

$$I_t(\xi, \chi) = \int_{\mathcal{Z}} \int_{\mathcal{A}(z)} I_t(a, z) d\xi(a|z) d\chi(z) \,.$$

For $a \in \mathcal{A}(\mathcal{Z})$ we also write $\hat{\Delta}_t(a, \chi) = \int_{\mathcal{Z}} \hat{\Delta}_t(a(z), z) d\chi(z)$ and $I_t(a, \chi) = \int_{\mathcal{Z}} I_t(a(z), z) d\chi(z)$. *Contextual IDS* is defined to optimize the conditional distribution $\xi \in \mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))$,

$$\xi_t \triangleq \underset{\xi \in \mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))}{\arg \min} \left\{ \Psi_t(\xi, \chi) = \frac{\hat{\Delta}_t(\xi, \chi)^2}{I_t(\xi, \chi)} \right\} \,. \tag{32}$$

The action $a_t \sim \xi_t^{\mathrm{IDS}}(\cdot|z_t)$ is sampled from the conditional distribution corresponding to the observed context z_t. In the joint minimization of the information ratio, the contextual distribution contributes to exploration and a smaller information ratio. By Jensen's inequality

and Lemma 20, the contextual information ratio is never worse then the conditional information ratio:

$$\min_{\xi \in \mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))} \Psi_t(\xi, \chi) \leq \min_{\xi \in \mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))} \int_{\mathcal{Z}} \Psi_t(\xi(z), z) d\chi(z) = \int_{\mathcal{Z}} \min_{\mu \in \mathscr{P}(\mathcal{A}(z))} \Psi_t(\mu, z) d\chi(z)$$

Regret Bounds Before presenting the regret bounds, we extend the definition of the alignment constant. Let $\mathcal{P}(\mathcal{Z}) \triangleq \times_{z \in \mathcal{Z}} \mathcal{P}(z)$ be the set of functions that map a context z to a Pareto optimal action $a(z) \in \mathcal{P}(z)$. For context distribution $\chi \in \mathscr{P}(\mathcal{Z})$, the global alignment constant is

$$\alpha(\chi) \triangleq \max_{\omega: \mathcal{Z} \to \mathbb{R}^d} \max_{a,b \in \mathcal{P}(\mathcal{Z})} \min_{c \in \mathcal{A}(\mathcal{Z})} \frac{\left(\int_{\mathcal{Z}} \langle \phi^z_{a(z)} - \phi^z_{b(z)}, \omega^z \rangle d\chi(z)\right)^2}{\int_{\mathcal{Z}} \int_{\mathcal{Z}} \|M^{z'}_{c(z')} \omega^z\|^2 d\chi(z) d\chi(z')} . \tag{33}$$

For the locally observable case, the definition of plausible maximizers is extended to the product space $\mathcal{P}(\mathcal{E}, \mathcal{Z}) \triangleq \times_{z \in \mathcal{Z}} \mathcal{P}(\mathcal{E}, z)$ and $\bar{\mathcal{P}}_\eta(\mathcal{E}, \mathcal{Z}) \triangleq \times_{z \in \mathcal{Z}} \bar{\mathcal{P}}_\eta(\mathcal{E}, z)$. Correspondingly, the local alignment constant for $\mathcal{E} \subset \Theta$ is

$$\alpha_\eta(\mathcal{E}, \chi) \triangleq \max_{\omega \in \mathbb{R}^{d \times \mathcal{Z}}} \max_{a,b \in \mathcal{P}(\mathcal{E}, \mathcal{Z})} \min_{c \in \bar{\mathcal{P}}_\eta(\mathcal{E}, \mathcal{Z})} \frac{\left(\int_{\mathcal{Z}} \langle \phi^z_{a(z)} - \phi^z_{b(z)}, \omega^z \rangle d\chi(z)\right)^2}{\int_{\mathcal{Z}} \int_{\mathcal{Z}} \|M^{z'}_{c(z')} \omega^z\|^2 d\chi(z) d\chi(z')} . \tag{34}$$

Finally, $\alpha(\mathcal{E}, \chi) \triangleq \inf_{\eta > 0} \eta^2 \alpha_\eta(\mathcal{E}, \chi)$.

Note that, reassuringly, the alignment constant $\alpha(\mathcal{E}, \chi)$ is finite as long as any direction $\phi^z_a - \phi^z_b$ for context z and Pareto optimal actions $a, b \in \mathcal{P}(z)$ can be estimated under *some* context $z' \in \mathcal{Z}$ that occurs with positive probability $\chi(z') > 0$. Specifically,

$$\alpha_\eta(\mathcal{E}, \chi) \leq \max_{z \in \mathcal{Z}, \omega \in \mathbb{R}^d} \max_{a,b \in \mathcal{P}(\mathcal{E}, \mathcal{Z})} \min_{z' \in \mathcal{Z}, c \in \bar{\mathcal{P}}_\eta(\mathcal{E}, \mathcal{Z})} \frac{\langle \phi^z_{a(z)} - \phi^z_{b(z)}, \omega \rangle^2}{\chi(z') \|M^{z'}_{c(z')} \omega\|^2} .$$

Theorem 14. *For fixed $\chi \in \mathscr{P}(\mathcal{Z})$ and context sequence $(z_t)^n_{t=1}$ sampled independently from χ, the regret of contextual IDS (Eq. (32)) satisfies*

$$\mathbb{E}[\mathfrak{R}_n] \leq n^{2/3} (54\alpha(\chi) \mathbb{E}[\beta_n] \mathbb{E}[\gamma_n])^{1/3} + \mathcal{O}(\Delta_{\max}) ,$$

where the expectation is over the random context.

Proof The proof follows along the lines of Theorem 7. Define $\hat{a}_t \in \mathcal{P}(\mathcal{Z})$ by $\hat{a}_t(z) = \arg\max_{a \in \mathcal{P}(z)} \langle \phi^z_a, \theta_t \rangle$. It follows from the same steps in the aforementioned theorem and the definition of $\alpha(\chi)$,

$$\hat{\Delta}_t(\hat{a}_t, \chi)^2 \leq \max_{b \in \mathcal{P}(\mathcal{Z})} \left(\int_{\mathcal{Z}} \max_{\theta \in \mathcal{E}_t} \langle \phi^z_{b(z)} - \phi^z_{\hat{a}_t(z)}, \theta - \hat{\theta}_t \rangle d\chi(z) \right)^2$$

$$\leq \alpha(\chi) \max_{\theta \in \mathcal{E}_t} \max_{c \in \mathcal{A}(\mathcal{Z})} \int_{\mathcal{Z}} \int_{\mathcal{Z}} \|M^{z'}_{c(z')}(\theta^z - \hat{\theta}_t)\|^2 d\chi(z') d\chi(z)$$

$$\leq 4\beta_t \max_{c \in \mathcal{A}(\mathcal{Z})} \int_{\mathcal{Z}} \log \det(\mathbf{1}_m + M^z_c V^{-1}_t M^{z\top}_c) d\chi(z)$$

$$\leq 8 \max_{c \in \mathcal{A}(\mathcal{Z})} \beta_t I_t(c, \chi) .$$

The information ratio is bounded by optimizing the trade-off between the greedy action $\hat{a}_t$ and $c_t = \arg\max_{c \in \mathcal{A}(\mathcal{Z})} I_t(c, \chi)$. Similar to Eq. (21), we obtain

$$\Psi_{3,t}(\xi_t, \chi) \triangleq \frac{\hat{\Delta}_t(\xi_t, \chi)^3}{I_t(\xi_t, \chi)} \leq \alpha(\chi)\beta_t(\Delta_{\max} + 4B)$$

The proof is concluded with Theorem 1 and bounding the estimation error. ∎

The result for the locally observable case is stated in the next theorem.

Theorem 15. *For fixed $\chi \in \mathscr{P}(\mathcal{Z})$ and context sequence $(z_t)_{t=1}^n$ sampled independently from χ, the regret of contextual IDS (Eq. (32)) satisfies,*

$$\mathbb{E}[\mathfrak{R}_n] \leq \sqrt{8\mathbb{E}[\bar{\alpha}_n \beta_n]\mathbb{E}[\gamma_n]n} + \mathcal{O}(\Delta_{\max}),$$

where $\bar{\alpha}_n = \frac{1}{n}\alpha(\mathcal{E}_t, \chi)$ is the average local alignment constant for the sequence of confidence sets $(\mathcal{E}_t)_{t=1}^n$ realized by the algorithm.

Proof Again, we generalize the proof of Theorem 8. Denote $\mathcal{E}_t = \times_{z \in \mathcal{Z}}\mathcal{E}_t$. Let $c_t = \arg\max_{a \in \bar{\mathcal{P}}(\mathcal{E}_t, \mathcal{Z})} I_t(a)$ be the most informative action in $\bar{\mathcal{P}}_\eta(\mathcal{E}_t, \mathcal{Z})$ for some $\eta > 0$. For the gap estimate of c_t we find

$$\eta^{-1}\hat{\Delta}_t(c_t) \leq \max_{a,b \in \mathcal{P}(\mathcal{E}_t, \mathcal{Z})} \int_{\mathcal{Z}} \max_{\theta \in \mathcal{E}_t} \langle \phi_a^z - \phi_b^z, \theta \rangle d\chi(z) \leq \max_{a,b \in \mathcal{P}(\mathcal{E}_t)} \max_{\theta, \nu \in \mathcal{E}_t(\mathcal{Z})} \int_{\mathcal{Z}} \langle \phi_a^z - \phi_b^z, \theta - \nu \rangle d\chi(z).$$

Consequently we can bound the information ratio for the action c_t as follows,

$$\hat{\Delta}_t(c_t)^2 \leq \eta^2 \alpha_\eta(\mathcal{E}_t, \chi) \max_{\theta, \nu \in \mathcal{E}_t(\mathcal{Z})} \int_{\mathcal{Z}} \int_{\mathcal{Z}} \|M_{c_t}^{z'}(\theta^z - \nu^z)\|^2 d\chi(z')d\chi(z)$$

$$\leq 2\eta^2 \alpha_\eta(\mathcal{E}_t, \chi)\beta_t \int_{\mathcal{Z}} \lambda_{\max}(M_{c_t(z)}^z V_t^{-1} M_{c_t(z)}^{z\top}) d\chi(z) \leq 8\eta^2 \alpha_\eta(\mathcal{E}_t, \chi)\beta_t I_t(c_t, \chi).$$

We conclude that $\Psi_t(c_t, \chi) \leq 8 \inf_{\eta > 0} \eta^2 \alpha_\eta(\mathcal{E}_t, \chi)\beta_t$ and the result follows. ∎

Computation Note that optimizing the conditional distribution is computationally more demanding than optimizing the conditional information ratio. Since the information ratio is a convex function of the distribution (Lemma 20), we can optimize the conditional distribution using standard convex solvers. A particularly simple implementation uses the Frank-Wolfe algorithm (Frank et al., 1956), that only relies on solving linear functions over $\mathscr{P}_\times(\mathcal{A}(\mathcal{Z}))$. The contextual IDS algorithm with Frank-Wolfe is summarized in Algorithm 3. The gradient is

$$\nabla_\xi \Psi(\xi, \chi) = \Psi(\xi, \chi)\left(\frac{2\hat{\Delta}(\cdot, \cdot)}{\hat{\Delta}_t(\xi, \chi)} - \frac{\hat{I}_t(\cdot, \cdot)}{I_t(\xi, \chi)}\right) \in \mathbb{R}^{\mathcal{A} \times \mathcal{Z}}$$

Convergence of Frank-Wolfe is guaranteed assuming that the gradient is Lipschitz (Jaggi, 2013, Theorem 1). In this case, the iterates $\xi_t^{(k)}$ approach the exact IDS distribution ξ_t at rate

$$\Psi(\xi_t^{(k)}, \chi) - \Psi(\xi_t, \chi) \leq \mathcal{O}\left(\frac{1}{k+2}\right).$$

Algorithm 3: Contextual IDS with Frank-Wolfe

Input: Action set $\mathcal{A}$, context set $\mathcal{Z}$, context distribution $\chi \in \mathscr{P}(\mathcal{Z})$, action-context features ϕ_a^z, feedback maps M_a^z.

1 for $t = 1, 2, 3, \ldots, n$ **do**

2 $\hat{\Delta}_t(a, c) \leftarrow \max_{\theta \in \mathcal{E}_t} \max_{b \in \mathcal{A}(z)} \langle \phi_b^z - \phi_a^z, \theta \rangle$ `// gap-estimates`

3 $I_t(a, z) \leftarrow \frac{1}{2} \log \det(\mathbf{1}_m + M_a^z V_t^{-1} M_a^{z, \top})$ `// information gain`

4 $\xi_t^{(1)}(a, z) \leftarrow 1/|\mathcal{A}|, \forall a \in \mathcal{A}, z \in \mathcal{Z}$

5   **for** $k = 1, \ldots, t^2$ **do**

6 $\bar{\Delta}^{(k)} \leftarrow \sum_{z \in \mathcal{Z}, a \in \mathcal{A}} \chi(z) \xi_t^{(k-1)}(a, z) \hat{\Delta}_t(a, z)$

7 $\bar{I}^{(k)} \leftarrow \sum_{z \in \mathcal{Z}, a \in \mathcal{A}} \chi(z) \xi_t^{(k-1)}(a, z) I_t(a, z)$

       `// Gradient` $\nabla_\xi \Psi_t(\xi, \chi)|_{\xi = \xi_t^{(k-1)}}$`, up to a positive factor:`

8 $G^{(k)}(a, z) \leftarrow 2\chi(z) \hat{\Delta}_t(a, z) \bar{\Delta}^{(k)} \bar{I}^{(k)} - \chi(z) I_t(a, z) (\bar{\Delta}^{(k)})^2$

       `// Frank-Wolfe step`

9 $a^*(z) \leftarrow \arg\min_{a \in \mathcal{A}} G^{(k)}(a, z), \quad \forall z \in \mathcal{Z}$

10 $\xi_t^{(k)}(a, z) \leftarrow (1 - \frac{2}{k+2}) \xi_t^{(k-1)}(a, z) + \frac{2}{k+2} \mathbb{1}(a = a^*(z)), \quad \forall a \in \mathcal{A}, z \in \mathcal{Z}$

11   `Observe context:` $z_t \sim \chi$

12   `Sample action:` $a_t \sim \xi_t^{(k)}(\cdot, z_t)$

13   Choose a_t, observe $y_t = \langle M_{a_t}^{z_t}, \theta \rangle + \epsilon_t$

Unfortunately, smoothness of the gradient $\nabla_\xi \Psi_t(\xi, \chi)$ is not guaranteed in general, in particular if for some ξ the information gain $I_t(\xi, \chi) \approx 0$ is vanishing. On the other hand, for our choice of information gain and gap estimate, the gradient is smooth around the IDS distribution ξ_t. This follows from $\hat{\Delta}_t(a) \geq \Omega(t^{-1/2})$ and the bound on the information ratio, which implies that the information gain is $I_t(\xi_t) \geq \Omega(1/t)$. More explicitly, we can ensure smoothness by bounding the information gain away from zero. Define $I_t^\epsilon(a, z) = I_t(a, z) + \epsilon$. Using $\epsilon = 1/t$ ensures that the gradient is $\mathcal{O}(1/t^2)$-Lipschitz while only marginally increasing the total information gain. This suggests that $K = t^2$ iterations of Frank-Wolfe suffice to obtain a good approximation of the contextual IDS distribution.

Tighter Gap Estimates and Alignment Constants We can obtain a tighter definition of the gap estimates using integrated reward features $\phi_\xi^\chi \triangleq \int_{\mathcal{Z}} \int_{\mathcal{A}(z)} \phi_a^z d\xi(a|z) d\chi(z)$ and defining the gap estimates $\bar{\Delta}_t(\xi, \chi) \triangleq \max_{\theta \in \mathcal{E}_t} \max_{b \in \mathcal{A}(\mathcal{Z})} \langle \phi_b^\chi - \phi_\xi^\chi, \theta \rangle$. By Jensen's inequality we get $\bar{\Delta}_t(\xi, \chi) \leq \hat{\Delta}_t(\xi, \chi)$. It is easy to verify that the definition of the alignment constant can be simplified and tightened for this choice of gap estimate:

$$\bar{\alpha}_\eta(\mathcal{E}, \chi) \triangleq \max_{\omega \in \mathbb{R}^d} \max_{a, b \in \mathcal{P}(\mathcal{E}, \mathcal{Z})} \min_{c \in \bar{\mathcal{P}}_\eta(\mathcal{E}, \mathcal{Z})} \frac{\int_{\mathcal{Z}} \langle \phi_{a(z)}^z - \phi_{b(z)}^z, \omega \rangle^2 d\chi(z)}{\int_{\mathcal{Z}} \|M_{c(z)}^z \omega\|^2 d\chi(z)}. \tag{35}$$

Jensen's inequality implies the natural property $\bar{\alpha}_\eta(\mathcal{E}, \chi) \leq \int_{\mathcal{Z}} \alpha_\eta(\mathcal{E}, z) d\chi(z)$ where $\alpha(\mathcal{E}, z)$ is the conditional alignment constant in Eq. (29). Moreover, the definition is never worse than the contextual alignment constant in Eq. (34), i.e. $\bar{\alpha}(\mathcal{E}, \chi) \leq \alpha(\mathcal{E}, \chi)$ where $\alpha(\mathcal{E}, \chi)$.

The downside is that computing the tighter gap estimate $\bar{\Delta}_t(\xi, \chi)$ appears to require a search over $\mathcal{A}(\mathcal{Z})$, which leads computationally intractable methods in general.

7. Kernelized Partial Monitoring

Linear partial monitoring captures the relationship between reward and observation through linear reward features and linear feedback maps. Whereas we have focused on the finite-dimensional setting so far, we now consider the infinite-dimensional setting. In particular, let $\mathcal{H}$ be a Hilbert space over $\mathbb{R}$ with norm $\|\cdot\|_{\mathcal{H}}$ and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We identify the unknown true reward function with a vector in $f^\star \in \mathcal{H}$ that satisfies the known bound $\|f^\star\|_{\mathcal{H}} \leq B$. The reward features $\phi_a \in \mathcal{H}$ correspond to linear functionals and we adopt the notation $f^\star(a) = \langle \phi_a, f^\star \rangle_{\mathcal{H}}$. Similarly, the feedback maps are modeled as linear operators $M_a : \mathcal{H} \to \mathbb{R}^m$ that observe the parameter on the subspace $\text{im}(M_a^*) \subset \mathcal{H}$, where M_a^* is the adjoint mapping. While this adds a great amount of flexibility, it poses two additional challenges.

First, computing the least-square estimate in the feature space requires $\Omega(d^2)$ memory and computation, which becomes prohibitive if d is large or infinite. Kernelized methods circumvent this limitation using a *representer theorem* (Kimeldorf and Wahba, 1970; Girosi, 1998; Schölkopf et al., 2001). Such results express all quantities of interest as inner products that are specified by a known kernel function. Kernel methods are widely used in machine learning (Schölkopf et al., 2002), and several kernelized bandit algorithms have been analyzed (Srinivas et al., 2010; Abbasi-Yadkori, 2012; Valko et al., 2013, 2014; Chaudhuri and Tewari, 2016). More broadly, by interpreting kernel regression as a Gaussian process (Rasmussen, 2004; Kanagawa et al., 2018), the field of Bayesian optimization is understood to solve a closely related problem (Mockus, 1982; Srinivas et al., 2010).

A second imminent issue that arises when the feature dimension is large or infinite, is that the dimension renders the previous results vacuous. In the literature on kernelized bandits, this challenge is circumvented by replacing the dimension by an appropriate notion of an *effective dimension* (Valko et al., 2013), or bounding the log-determinant using the eigendecay of the covariance matrix (Srinivas et al., 2010; Vakili et al., 2020).

7.1 Kernel Regression for Partial Monitoring Feedback

The least-squares estimate is defined over the Hilbert space $\mathcal{H}$, using the observations $y_t = M_{a_t} f^\star + \epsilon_t$ and regularizer $\lambda > 0$,

$$\hat{f}_t \triangleq \underset{f \in \mathcal{H}}{\arg\min} \sum_{s=1}^{t-1} \|M_{a_s} f - y_s\|^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{36}$$

For simplicity of the exposition, we present the version without additional parameter constraints. We remark that certain types of affine linear constraints can be handled by solving a finite-dimensional SOC-constrained problem (e.g. Bagnell and Farahmand, 2015; Aubin-Frankowski and Szabo, 2020) As usual, the regularized least-squares solution is always contained in a finite-dimensional subspace spanned by the data,

$$\hat{f}_t \in \text{span}(\text{im}(M_{a_s}^\top) : s \in [t-1]).$$

In other words, the least-squares solution can be parameterized by $\alpha_1, \ldots, \alpha_{t-1} \in \mathbb{R}^m$ such that $\hat{f}_t = \sum_{s=1}^{t-1} M_{a_s}^* \alpha_i$. We are interested in sufficient conditions to ensure that the coefficients and the evaluations maps $\langle \phi_a, \hat{f}_t \rangle$ can be computed efficiently. Define the *joint evaluation map* for $a, b \in \mathcal{A}$,

$$E_{a,b} : \mathcal{H} \to \mathbb{R}^{m+1}, \quad f \mapsto [\phi_a f, (M_b f)^\top]^\top. \tag{37}$$

To enable efficient computation in the potentially infinite-dimensional Hilbert space $\mathcal{H}$, we now assume that $E_{a,b}$ is an evaluation functional of a vector-valued reproducing kernel Hilbert space (RKHS; Aronszajn, 1950; Pedrick, 1957). For a modern introduction to RKHS theory see (Carmeli et al., 2006).

Assumption 3 (RKHS). *The subspace* $\mathrm{span}(\mathrm{im}(E_{a,b}^*) : a, b \in \mathcal{A}) \subset \mathcal{H}$ *is a* $\mathbb{R}^{m+1}$-*valued RKHS over* $\mathcal{A} \times \mathcal{A}$ *with evaluation functionals* $E_{a,b}$ *defined in Eq. (37) and a known corresponding kernel*

$$k : \mathcal{A}^2 \times \mathcal{A}^2 \to \mathbb{R}^{(m+1) \times (m+1)}, \quad k(a, b, a', b') = E_{a,b} E_{a',b'}^* .$$

The assumption gives access to the covariance of actions $k_\phi(a, b) \triangleq \langle \phi_a, \phi_b \rangle \in \mathbb{R}$, feedback $k_M(a, b) \triangleq M_a M_b^* \in \mathbb{R}^{m \times m}$ and action-feedback $k_{\phi,M}(a, b) \triangleq \phi_a M_b^* \in \mathbb{R}^{1 \times m}$. The next lemma shows that the reward estimate $\hat{f}_t(a) \triangleq \langle \phi_a, \hat{f}_t \rangle$ can be computed efficiently from finite-dimensional quantities.

Theorem 16 (Partial Monitoring RepresPresenter Theorem). *Under Assumption 3 the kernel least-squares estimate Eq. (36) is for any* $a \in \mathcal{A}$:

$$\hat{f}_t(a) = k_t(a)^\top (K_t + \lambda \mathbf{1}_{m(t-1)})^{-1} \boldsymbol{y}_t ,$$

where we define the following finite-dimensional expressions:

$$\boldsymbol{y}_t \triangleq [y_1^\top, \ldots, y_{t-1}^\top]^\top \in \mathbb{R}^{m(t-1)} \qquad \text{(the observation vector)}$$
$$K_t \triangleq [M_{a_r} M_{a_s}^*]_{r,s=1,\ldots,t-1} \in \mathbb{R}^{m(t-1) \times m(t-1)} \qquad \text{(kernel matrix)}$$
$$k_t(a) \triangleq [\phi_a M_{a_s}^*]_{s=1,\ldots,t-1}^\top \in \mathbb{R}^{m(t-1)} \qquad \text{(evaluation weights)}$$

In particular, the quantities above are defined by the kernel:

$$k(a, b, a', b') \triangleq \begin{bmatrix} k_\phi(a, a') & k_{\phi,M}(a, b') \\ k_{\phi,M}(a', b)^\top & k_M(b, b') \end{bmatrix} = \begin{bmatrix} \phi_a \phi_{a'}^* & \phi_a M_{b'}^* \\ M_b \phi_{a'}^* & M_b M_{b'}^* \end{bmatrix} .$$

Proof We define the map

$$\Phi_t : \mathcal{H} \to \mathbb{R}^{m(t-1)}, \quad \theta \mapsto [(M_{a_1} \theta)^\top, \ldots, (M_{a_{t-1}} \theta)^\top]^\top \tag{38}$$

as the stack of evaluation maps in the observation history. The regularized least-square solution of Eq. (36) is $\hat{\theta}_t = V_t^{-1} \Phi_t^* \boldsymbol{y}_t$, where $V_t \theta \triangleq (\Phi_t^* \Phi_t + \lambda \mathbf{1}_\mathcal{H}) \theta$ is an invertable linear map $\mathcal{H} \to \mathcal{H}$ and $\mathbf{1}_\mathcal{H}$ is the identity operator. The claim follows with the identity $(\Phi_t^* \Phi_t + \lambda \mathbf{1}_\mathcal{H})^{-1} \Phi_t^* = \Phi_t^* (\Phi_t \Phi_t^* + \lambda \mathbf{1}_t)^{-1}$ and replacing the inner products with the kernel expressions. ∎

29

In order to make use of the estimator in the IDS algorithm, we also need a kernelized statement of the confidence bounds. The next lemma directly extends the confidence bounds by Abbasi-Yadkori (2012, Corollary 3.15) to linear partial monitoring feedback.

Lemma 17. *Let $V_t = \sum_{s=1}^{t-1} M_{a_s}^* M_{a_s} + \lambda \mathbf{1}_{\mathcal{H}}$ and $\mathcal{E}_{t,\delta} = \{f \in \mathcal{H} : \|f - \hat{f}_t\|_{V_t}^2 \leq \beta_{t,\delta}\}$ where $\beta_{t,\delta}^{1/2} = \rho\sqrt{2 \log \frac{1}{\delta} + \log \det(\mathbf{1} + \lambda^{-1} K_t)} + \lambda^{1/2} B$. Let $(a_t)_{t=1}^\infty$ be a $\mathcal{F}_t$-adapted sequence of actions and corresponding observations $y_t = M_{a_t}\theta + \epsilon_t \in \mathbb{R}^m$ with conditionally independent ρ-sub-Gaussian vector ϵ_t. If $\|f\|_{\mathcal{H}} \leq B$, then*

$$\mathbb{P}[\forall t \geq 1, f^\star \in \mathcal{E}_t] \geq 1 - \delta.$$

Further, with probability at least $1 - \delta$, for all $t \geq 1$,

$$|\hat{f}_t(a) - \hat{f}_t(b) - (f^\star(a) - f^\star(b))| = |\langle \phi_a - \phi_b, \hat{f}_t - f^\star \rangle| \leq \sqrt{\beta_{t,\delta}\psi_t(a,b)},$$

where $\psi_t(a,b) \triangleq \frac{1}{\lambda}\left(\psi(a,b) - (k_t(a) - k_t(b))^\top (K_t + \lambda\mathbf{1})^{-1}(k_t(a) - k_t(b))\right)$ and the kernel metric is $\psi(a,b) \triangleq k_\phi(a,a) + k_\phi(b,b) - 2k_\phi(a,b)$. The evaluation weights $k_t(a)$ and kernel matrix K_t are defined in Theorem 16.

Proof The confidence set is the same as (Abbasi-Yadkori, 2012, Corollary 3.15) applied to the observation maps. For the second claim, note that $\psi_t(a,b) = \|\phi_a - \phi_b\|_{V_t^{-1}}^2$. The statement in the lemma follows using Cauchy-Schwarz and computing the feature uncertainty with the Sherman-Morrison identity,

$$\lambda V_t^{-1} = \mathbf{1}_{\mathcal{H}} - \Phi_t^*(\Phi_t\Phi_t^* + \lambda\mathbf{1})^{-1}\Phi_t,$$

where Φ_t is defined as in Eq. (38). ∎

7.2 Kernelized Information-Directed Sampling

Equipped with the representer theorem and the kernelized confidence bound, we can define kernelized gap estimates and kernelized information gain functions for information-directed sampling. We use the relaxed gap estimate from Eq. (24), which is computationally simpler. Let $\hat{f}_t(a)$ as defined in Theorem 16, and $\beta_t \triangleq \beta_{t,1/t^2}$ and $\psi_t(a,b)$ as defined in Lemma 17. The kernelized gap estimate is

$$\hat{\Delta}_t(a) = \min\left\{ \max_{b \in \mathcal{A}} \hat{f}_t(\hat{a}_t) + (\beta_t \psi_t(\hat{a}_t, b))^{1/2} - \hat{f}_t(a), B \right\}, \tag{39}$$

where $\hat{a}_t = \arg\max_{a \in \mathcal{A}} \hat{f}_t(a)$ is the empirical maximizer. Other variants of the gap estimate are derived similarly. The (undirected) information gain corresponding to Eq. (12) is

$$I_t(a) = \frac{1}{2} \log \det\left(\mathbf{1}_m + \frac{1}{\lambda}\left(k_M(a,a) - L_t(a)K_t^{-1}L_t(a)^\top\right)\right), \tag{40}$$

where $L_t(a) = M_a\Phi_t^* \in \mathbb{R}^{m \times (t-1)m}$ and Φ_t is the kernel design matrix defined in Eq. (38). The total information gain is

$$\gamma_n = \frac{1}{2} \log \det(\mathbf{1} + \lambda^{-1} K_{n+1}).$$

Corollary 18. *The theoretical guarantees for IDS Theorems 7 and 8 stated in terms of the confidence coefficient β_n and the total information gain $\gamma_n = \sum_{t=1}^{n} I_t(a_t)$ continue to hold up to constant factors.*

The log-determinant is understood as a complexity measure of adaptive exploration in $\mathcal{H}$ and is closely related to the notion of the Eluder dimension (Russo and Van Roy, 2013; Huang et al., 2021). In particular, γ_n is often bounded independently of the dimension of $\mathcal{H}$. For the large class of Mercer kernels, the literature has produced bounds depending on the decay of the eigenvalues in the Mercer decomposition (Srinivas et al., 2010; Mutný and Krause, 2018; Vakili et al., 2020). A kernelized version of the directed information gain Eq. (25) can be derived similarly.

Example 13 (Kernelized Dueling Bandits). *We present a kernelized version of the linear dueling bandit setting (Example 3). As before, let $\mathcal{I}$ be a ground set of actions. The action space $\mathcal{A} = \mathcal{I} \times \mathcal{I}$ consists of pairs of elements in the ground set. In the* utility-based dueling feedback *model, the reward and feedback is determined by a* utility function $g : \mathcal{I} \to \mathbb{R}$. *Upon choosing the pair $a_t = (a_t^1, a_t^2) \in \mathcal{A}$ in round t, and the learner observes the reward difference*

$$y_t = g(a_t^1) - g(a_t^2) + \epsilon_t, \tag{41}$$

and suffers instantaneous regret $f^\star(a_t^1, a_t^2) = g(a_t^1) + g(a_t^2)$ for both actions.

Let $\mathcal{H}(\mathcal{I})$ be an RKHS with kernel function $k : \mathcal{I} \times \mathcal{I}$ and assume that the utility function $g \in \mathcal{H}(\mathcal{I})$ satisfies $\|g\|_{\mathcal{H}} \leq \frac{B}{2}$, therefore $\|f^\star\|_{\mathcal{H}} \leq B$. For an action $a \in \mathcal{I}$, denote by $k_a \in \mathcal{H}(\mathcal{I})$ the kernel features of the evaluation functionals, which satisfy $k(a, b) = \langle k_a, k_b \rangle_{\mathcal{H}}$. The features and evaluation maps corresponding to our reward and feedback model are $\phi_{a,b} = k_a + k_b$ and $M_{a,b} = k_a - k_b$. The covariance between reward and feedback for actions $a = (a^1, a^2)$ and $b = (b^1, b^2)$ is

$$k_M(a, b) = k(a^1, b^1) - k(a^2, b^1) - k(a^1, b^2) + k(a^2, b^2)$$
$$k_{\phi,M}(a, b) = k(a^1, b^1) + k(a^2, b^1) - k(a^1, b^2) - k(a^2, b^2).$$

Hence, the kernel matrix and evaluation weights at time t are

$$K_t = [k(a_r^1, a_s^1) - k(a_r^2, a_s^1) - k(a_r^1, a_s^2) + k(a_r^2, a_s^2)]_{r,s=1,\ldots,t-1},$$
$$k_t(a) = [k(a^1, a_s^1) + k(a^2, a_s^1) - k(a^1, a_s^2) - k(a^2, a_s^2)]_{s=1,\ldots,t-1}.$$

With the above, we can directly apply Algorithm 2 and the corresponding results. A caveat is that the size of action space $|\mathcal{A}|$ scales quadratically in the size of the ground set $|\mathcal{I}|$. This leads to $\mathcal{O}(|\mathcal{I}|^2)$ computation complexity per round, even with the relaxed gap estimate Eq. (39) and the approximate IDS distribution. This can be improved to $\mathcal{O}(|\mathcal{I}|)$, by directly estimating the utility function g and using the structure of the dueling feedback (Kirschner and Krause, 2021). As usual, computing the kernel estimate in the data space requires $\mathcal{O}(t^2)$ steps per round with incremental updates, or $\mathcal{O}(n^3)$ overall on a horizon of length n.

8. Related Work

Partial monitoring dates back to the work by Rustichini (1999) and generalizes a considerable number of models for (stateless) sequential decision making that have been studied separately in the literature, most prominently the bandit setting (Lattimore and Szepesvari, 2020). Like in the standard bandit model, one can consider the *adversarial setting*, where the data is generated adversarially and the learner is compared to a fixed baseline. For a long time, the primary focus was on understanding the relationship between the minimax regret and the structure of the loss and feedback functions. The dependence on the horizon is now completely understood as proven in a long line of work (Piccolboni and Schindelhauer, 2001; Cesa-Bianchi et al., 2006; Antos et al., 2013; Lattimore and Szepesvári, 2019a,c). An algorithm with rate-optimal worst-case regret in all classes of games is by Lattimore and Szepesvári (2019b).

In the *stochastic setting* the hidden outcomes are independent and identically distributed according to some unknown distribution (Bartók et al., 2011). An algorithm for stochastic feedback that adapts to the game structure is by Bartók et al. (2012). Asymptotically optimal instance-dependent bounds were studied by Komiyama et al. (2015). For games that satisfy a local observability condition, Vanchinathan et al. (2014b) analyze an algorithm that exploits prior knowledge on the loss distribution. They also propose a computationally more efficient variant based on Thompson sampling, which often has outstanding performance, while also suffering linear regret in certain games. More recently, Tsuchiya et al. (2020) derive logarithmic regret bounds for Thompson sampling on partial monitoring games that satisfy a strong local observability condition. Note, however, that for general partial monitoring games, Thompson sampling and algorithms based on optimism are insufficient to resolve the exploration-exploitation trade-off and might suffer linear regret (Lattimore and Szepesvári, 2019c, Appendix G). The setting with linear reward and feedback structure was first introduced by Lin et al. (2014) who provide an elimination-style algorithm that achieves $\tilde{\mathcal{O}}(n^{2/3})$ regret under a global observability condition. A similar approach that achieves the same regret scaling is by Chaudhuri and Tewari (2016).

The information-directed sampling framework was first proposed by Russo and Van Roy (2014), and later extended to the frequentist framework by Kirschner and Krause (2018); Kirschner et al. (2020); extending instance optimality to other model classes is still an open problem. The latter work provides the basis of the current work. IDS was applied to the contextual linear bandit setting by Hao et al. (2022). In the adversarial setting, the IDS framework was used by Lattimore and Szepesvári (2019c) to classify minimax rates and to derive an algorithm that applies to all finite games (Lattimore and Szepesvári, 2019b). For the linear bandit setting, Kirschner et al. (2021) show that for a specific choice of information gain function, IDS achieves the asymptotic lower bound while also being near worst-case optimal (Graves and Lai, 1997; Lattimore and Szepesvári, 2017; Combes et al., 2017). Beyond the bandit setting, information-directed sampling was also applied to reinforcement learning (Nikolov et al., 2019; Lu et al., 2021; Hao and Lattimore, 2022; Zanette and Sarkar, 2017). Various numerical results on the IDS approach can be found in (Russo and Van Roy, 2014; Kirschner et al., 2021; Kirschner and Krause, 2021; Kirschner, 2021).

Tunable IDS and Estimation-To-Decisions   The closely related estimation-to-decisions (E2D) framework was recently proposed by Foster et al. (2021). The relation can be un-

Algorithm	Reference	Theory	Compute
FeedExp	Piccolboni and Schindelhauer (2001)	global, local, adversarial	
CBP	Bartók et al. (2012)	global, local, frequentist	LP
NeighborhoodWatch	Foster and Rakhlin (2012)	local, adversarial	
BPM-TS	Vanchinathan et al. (2014a)	–	Gaussian sampling
BPM-Least	Vanchinathan et al. (2014a)	local, frequentist	SOCP
PM-DMED	Komiyama et al. (2015)	–	LSIP+ECP†
PM-DMED-Hinge	Komiyama et al. (2015)	local, global, asymptotic	LSIP+ECP†
Mario sampling	Lattimore and Szepesvári (2019c)	local, Bayesian	posterior sampling
ExpByOpt	Lattimore and Szepesvári (2019b)	global, local, adversarial	ECP
TSPM	Tsuchiya et al. (2020)	strongly local, frequentist	
IDS	this work	global, local, frequentist	SOCP

LP = linear programming, LSIP = linear semi-infinite program, SOCP = second-order cone programming, ECP = exponential cone programming
† Alternatively, a convex/concave saddle-point problem that requires solving an ECP to evaluate.

Table 2: All algorithms need basic linear programming at initialization to determine estimation vectors and/or the cell decomposition. Algorithms with a blank compute entry can be computed using elementary matrix calculations only. Most algorithms can be sped up, at least heuristically, by re-computing various quantities only intermittently.

derstood by introducing a *gap-based decision-making coefficient* (g-dec):

$$\mu_t^{\text{E2D}} = \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \left\{ \text{g-dec}_\lambda(\mu) \triangleq \hat{\Delta}_t(\mu) - \lambda I_t(\mu) \right\}$$

We emphasize that our gap-based formulation of the decision-estimation coefficient is a relaxation of the formulation by Foster et al. (2021). Observe that the minimization is solved by a Dirac on $a_t = \arg\min_{a \in \mathcal{A}} \hat{\Delta}_t(a) - \lambda I_t(a)$. For a fixed λ, the gap-based E2D objective can be essentially solved by an offline oracle once the gap estimates and information gain function have been computed. The algebraic inequality $2ab \leq a^2 + b^2$ implies

$$\hat{\Delta}_t(\mu) - \lambda I_t(\mu) = \frac{\hat{\Delta}_t(\mu)\sqrt{\lambda I_t(\mu)}}{\sqrt{\lambda I_t(\mu)}} - \lambda I_t(\mu) \leq \frac{\hat{\Delta}_t(\mu)^2}{4\lambda I_t(\mu)} = \frac{\Psi_t(\mu)}{4\lambda}$$

Using this to bound the regret yields

$$\sum_{t=1}^{n} \hat{\Delta}_t(a_t) = \sum_{t=1}^{n} \left(\hat{\Delta}_t(a_t) - \lambda I_t(a_t) \right) + \lambda \sum_{t=1}^{n} I_t(a_t)$$

$$= \sum_{t=1}^{n} \min_{\mu \in \mathscr{P}(\mathcal{A})} \left(\hat{\Delta}_t(\mu) - \lambda I_t(\mu) \right) + \lambda \gamma_n \leq \frac{1}{4\lambda} \sum_{t=1}^{n} \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) + \lambda \gamma_n$$

By optimizing λ and bounding the estimation error, we recover the IDS regret bound

$$\mathfrak{R}_n \leq \sqrt{\sum_{t=1}^{n} \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) \gamma_n} + \mathcal{O}(1)$$

In other words, the gap-based E2D algorithm with appropriately chosen λ achieves the same worst-case bound as IDS. A similar argument can be made for the globally observable case.

A clear advantage of the gap-based E2D algorithm is that the optimal sampling distribution is realized as a Dirac. This is a significant simplification in the contextual setting, where E2D obtains the same bounds as contextual IDS without optimizing over the space of marginal distributions. The price for this simplification is that an optimal choice of λ requires access to a bound on the information ratio and the information gain, and the algorithm becomes more dependent on choices that are informed by a worst-case analysis.

9. Conclusion

We presented *linear partial monitoring*, a versatile framework for interactive decision making. Building upon and extending earlier work, we show that a single algorithm, information-directed sampling, achieves near-optimal regret rates in various settings, including parameter-constrained, kernelized and contextual decision-making problems. The framework includes the classical finite partial monitoring setting as a special case, and unlike at least some of the prior work, the proposed algorithm is simple practical to implement. Promising directions for future work includes broadening the scope of information-directed sampling beyond the linear setting, deriving information gain functions for non-Gaussian observation likelihood functions and characterizing the exact minimax rate for continuous action sets.

Acknowledgments

Appendix A. Information-Directed Sampling: General Results

A.1 Properties of the Information-Ratio

Lemma 19 (Existence). *Let $\mathcal{A}$ be compact, $\hat{\Delta}_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ continuous and $I_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ continuous and not zero everywhere. Then there exists a $\mu^* \in \mathscr{P}(\mathcal{A})$ such that $\Psi_t(\mu^*) = \inf_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$.*

Proof The claim essentially follows from the fact that $\mathscr{P}(\mathcal{A})$ is compact in the weak*-topology, which is also the topology that makes the maps $\mu \mapsto \hat{\Delta}_t(\mu)$ and $\mu \mapsto I_t(\mu)$ continuous. More specifically, pick a sequence $(\mu_j)_{j=1}^{\infty}$ in $\mathscr{P}(\mathcal{A})$ such that $\Psi_t(\mu_j) \to \inf_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$ as $j \to \infty$. Note that μ_j is a tight sequence of probability distributions because $\mathcal{A}$ is compact. Prokhorov's theorem (Prokhorov, 1956) guarantees the existence of a subsequence μ_{j_i} converging weakly to some $\mu^* \in \mathscr{P}(\mathcal{A})$. By definition of weak convergence of probability measures, $\hat{\Delta}_t(\mu_{j_i}) \to \hat{\Delta}_t(\mu^*)$ and $I_t(\mu_{j_i}) \to I_t(\mu^*)$. By the assumption that $I_t(\cdot)$ is not zero everywhere, we have $I_t(\mu^*) > 0$. Continuity of the map $(v, w) \mapsto v^2/w$ on $[0, \infty) \times (0, \infty)$ completes the proof. ∎

Lemma 20 (Convexity (Russo and Van Roy, 2014, Proposition 6)). *$\Psi_t(\mu)$ is convex in μ.*

Proof Note that $(v, w) \mapsto v^2/w$ is convex on the domain $\mathbb{R} \times (0, \infty)$ as shown in (Boyd et al., 2004, Chapter 3). Further, $\mu \mapsto (\hat{\Delta}_t(\mu), I_t(\mu))$ is an affine function on $\mathscr{P}(\mu)$. Since $\Psi_t(\mu) = \hat{\Delta}_t(\mu)^2/I_t(\mu)$ can be written as a composition of a convex and an affine function, the result follows. ∎

The next lemma extends Russo and Van Roy (2014, Prop. 6) to compact $\mathcal{A}$.

Lemma 21 (Support). *The IDS distribution $\mu_t \in \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$ can always be chosen such that $|\operatorname{supp}(\mu_t)| \leq 2$. Further, for $a \in \mathcal{A}$ define*

$$h_t(a) \triangleq 2\, \hat{\Delta}_t(\mu_t)\hat{\Delta}_t(a) - \Psi_t(\mu_t)I_t(a)\,.$$

Then any $a \in \operatorname{supp}(\mu_t)$ satisfies $h_t(a) = \min_{b \in \mathcal{A}} h_t(b) = \hat{\Delta}_t(\mu_t)^2$.

Proof We claim that

$$h_t(a) = \min_{b \in \mathcal{A}} h_t(b) \quad \text{for all } a \in \operatorname{supp}(\mu_t)\,. \tag{42}$$

We first show how this implies all other claims. Choose any minimizing distribution $\mu^* \in \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$, not necessarily supported on two actions. Taking the expectation of $h_t(a)$ on μ^* gives $h_t(\mu^*) = \hat{\Delta}_t(\mu^*)^2$. Let $a_{\min} = \arg\min_{a \in \operatorname{supp}(\mu^*)} \hat{\Delta}_t(a)$ and $a_{\max} = \arg\max_{a \in \operatorname{supp}(\mu^*)} \hat{\Delta}_t(a)$. Then we can define $\mu^{\mathrm{IDS}}(p) = (1 - p)e_{a_{\min}} + pe_{a_{\max}}$, where e_a is a Dirac on $a \in \mathcal{A}$ and $p \in [0, 1]$ is a trade-off probability. We can choose p^* such that $\hat{\Delta}_t(\mu^{\mathrm{IDS}}(p^*)) = \hat{\Delta}_t(\mu^*)$ and let $\mu_t = \mu_t(p^*)$. By Eq. (42) we get $I_t(\mu_t) = I_t(\mu^*)$. Therefore $\Psi_t(\mu^*) = \Psi_t(\mu_t)$ and μ_t is a minimizing distribution with support size at most 2.

To show Eq. (42), let $\Psi_t^* = \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$ and define for $\mu \in \mathscr{P}(\mathcal{A})$,

$$H_t(\mu) \triangleq \hat{\Delta}_t(\mu)^2 - I_t(\mu)\Psi_t^*\,.$$

Note that H_t has the same minimizers as Ψ_t. To see this, observe that $H(\mu) \geq 0$ and $H_t(\mu^*) = 0$, which shows one direction. For the converse, assume that μ' minimizes $H(\mu')$, i.e. $H_t(\mu') = 0$, which immediately gives $\Psi_t(\mu') = \Psi_t(\mu^*)$. Let $a = \arg\min_{b \in \text{supp}(\mathcal{A})} h_t(b)$ which exists by compactness and continuity of h. Define the measure $\mu_\lambda = (1-\lambda)\mu^* + \lambda e_a$ obtained from shifting mass to a. Since μ^* is a minimizer of H_t, we must have that

$$0 \leq \frac{d}{d\lambda}H_t(\mu_\lambda)|_{\lambda=0} = 2\hat{\Delta}_t(\mu^*)(\hat{\Delta}_t(e_a) - \hat{\Delta}_t(\mu^*)) - (I_t(e_a) - I_t(\mu^*))$$

$$= h_t(a) - h_t(\mu^*).$$

The claim follows after rearranging. ∎

Lemma 22 (Closed form). *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

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

where we define $\Delta_1/0 = \infty$ and $\text{clip}_{[0,1]}(p) = \max(\min(p,1),0)$.

Proof The case $I_1 \geq I_2$ is immediate, because any $p > 0$ increases the numerator and decreases the denominator. For the remaining part we assume $I_1 < I_2$. The derivative is

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

Lemma 20 implies that $\Psi(p)$ is convex on the domain $[0,1]$. Solving for the first order condition $\Psi'(p) = 0$ gives $p_0 \triangleq \frac{\Delta_1}{\Delta_2 - \Delta_1} - \frac{2I_1}{I_2 - I_1}$. If $p_0 \in [0,1]$ we are done. Otherwise, note that $p_0 < 0$ implies $\Psi'(0) > 0$ and $p_0 > 1$ implies $\Psi'(1) < 0$, which follows from calculating the sign of both factors in the nominator. Convexity on $[0,1]$ implies that clipping p_0 to $[0,1]$ leads to the correct solution. ∎

We frequently use this lemma in the following way. Assume that $\tilde{\mu} \in \mathscr{P}(\mathcal{A})$ is a sampling distribution, possibly chosen as a Dirac on some action $a \in \mathcal{A}$. Let $\hat{a}_t = \arg\min_{a \in \mathcal{A}} \hat{\Delta}_t(a)$ be the action with the smallest estimated gap and denote $\delta_t = \hat{\Delta}_t(\hat{a}_t)$. Then

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) \leq \min_{p \in [0,1]} \frac{\big((1-p)\delta_t + p\hat{\Delta}_t(\tilde{\mu})\big)^2}{(1-p)I_t(\hat{a}_t) + pI_t(\tilde{\mu})}$$

$$\leq \min_{p \in [0,1]} \frac{\big((1-p)\delta_t + p\hat{\Delta}_t(\tilde{\mu})\big)^2}{pI_t(\tilde{\mu})}.$$

The first inequality is by restricting the sampling distribution as a mixture between a Dirac on $\hat{a}_t$ and $\tilde{\mu}_t$. The second inequality uses that $I_t(\hat{a}_t) \geq 0$. If we minimize the right-hand side using Lemma 22, we get

$$\Psi_t(\mu_t) \leq \begin{cases} \frac{4\delta_t(\hat{\Delta}_t(\tilde{\mu}) - \delta_t)}{I_t(\tilde{\mu}_t)} & \text{if } 2\delta_t \leq \hat{\Delta}_t(\tilde{\mu}) \\ \frac{\hat{\Delta}_t(\tilde{\mu})^2}{I_t(\tilde{\mu})} & \text{else.} \end{cases} \tag{43}$$

Lemma 23 (Almost greedy). *Let $\hat{a}_t = \arg\min_{a \in \mathcal{A}} \hat{\Delta}_t(a)$ be the greedy action, chosen arbitrarily if not unique. The IDS distribution μ_t satisfies*

$$\hat{\Delta}_t(\mu_t) \leq 2\hat{\Delta}_t(\hat{a}_t) \,.$$

Proof Note that by definition, the information ratio cannot be improved by shifting mass to $\hat{a}_t$ and discarding the information $I_t(\hat{a}_t)$,

$$\Psi_t(\mu_t) \leq \min_{p \in [0,1]} \left\{ \frac{\left((1-p)\hat{\Delta}_t(\mu_t) + p\hat{\Delta}_t(\hat{a}_t)\right)^2}{(1-p)I_t(\mu_t)} \triangleq \psi(p) \right\} \,.$$

Note that the gradient of $\psi(p)$ cannot be negative at $p = 0$. Hence

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

Rearranging yields the claim. ∎

Lemma 24 (Approximate IDS). *Define the restricted set of sampling distributions $\mathscr{P}_a = \{e_a(1-p) + e_b p : b \in \mathcal{A}, p \in [0,1]\}$ that randomize between a fixed $a \in \mathcal{A}$ and a second action $b \in \mathcal{A}$. Let $\hat{a}_t = \arg\min_{a \in \mathcal{A}} \hat{\Delta}_t(a)$ be the greedy action. Define $\tilde{\mu}_t = \arg\min_{\mu \in \mathscr{P}_{\hat{a}_t}} \Psi_t(\mu)$ as the distribution that minimizes the information ratio among distribution in $\mathscr{P}_{\hat{a}_t}$. Then*

$$\Psi_t(\tilde{\mu}_t) \leq \frac{4}{3} \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) \,,$$

and the bound is tight for general Δ_t and I_t. Further, if $2\hat{\Delta}_t(\hat{a}_t) \leq \hat{\Delta}_t(b)$ for all $b \in \mathcal{A}$ with $\hat{\Delta}_t(b) > \hat{\Delta}_t(\hat{a}_t)$, then $\Psi_t(\tilde{\mu}_t) = \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$.

Proof By Lemma 21 it suffices to consider three actions with gaps $\Delta_1 < \Delta_2 < \Delta_3$ and information gain I_1, I_2, I_3. Let Ψ_{12}, Ψ_{13} and Ψ_{23} denote the ratio obtained by minimizing the trade-off only between the actions indicated in the subscript. Assume that $\Psi^* \triangleq \Psi_{23} \overset{!}{=} \min\{\Psi_{12}, \Psi_{13}, \Psi_{23}\}$ and let $\tilde{\Psi} = \min\{\Psi_{12}, \Psi_{13}\}$. The claim follows if we show $\tilde{\Psi} \leq \frac{4}{3}\Psi^*$.

Note that we can assume that $I_1 = 0$, since this choice does not affect Ψ_{23} and can only make $\tilde{\Psi}$ larger. Further, the minimizer of the information ratio is invariant to rescaling of the gap and information gain functions. Therefore without loss of generality, we can assume that $\Delta_1 = 1$ and $\tilde{\Psi} = 1$.

37

We show that $\Psi^{*-1} \leq \frac{4}{3}$. First, we make some calculations with the help of Lemma 22. The trade-off probability between actions 2 and 3 is

$$p_{23} = \frac{\Delta_2}{\Delta_3 - \Delta_2} - \frac{2I_2}{I_3 - I_2},$$

and we require that the trade-off is non-trivial, $0 < p_{23} < 1$. The ratio Ψ_{23} is

$$\Psi_{23} = \frac{4(\Delta_2 I_3 - \Delta_3 I_2)(\Delta_3 - \Delta_2)}{(I_3 - I_2)^2}.$$

We complete the proof with two cases. For the first case, we assume $1 < \Delta_2 < \Delta_3 \leq 2$. We again use Lemma 22 to compute $\Psi_{12} = \Delta_2^2/I_2$ and $\Psi_{13} = \Delta_3^2/I_3$. In fact, we can assume that $\Psi_{12} = \Psi_{13}$ since that does not affect $\tilde{\Psi}$ and only makes Ψ^* smaller. The normalization $\tilde{\Psi} = 1$ implies that $I_2 = \Delta_2^2$ and $I_3 = \Delta_3^2$. Hence,

$$\begin{aligned}
\Psi_{23}^{-1} &= \frac{(I_3 - I_2)^2}{4(\Delta_2 I_3 - \Delta_3 I_2)(\Delta_3 - \Delta_2)} \\
&= \frac{(\Delta_3^2 - \Delta_2^2)^2}{4(\Delta_2 \Delta_3^2 - \Delta_3 \Delta_2^2)(\Delta_3 - \Delta_2)} \\
&= \frac{(\Delta_3 + \Delta_2)^2}{4\Delta_2 \Delta_3} \leq \frac{9}{8}.
\end{aligned}$$

The last inequality holds for $1 \leq \Delta_2, \Delta_3 \leq 2$, and note that the constraint on p_{23} is satisfied.

For the second case, assume that $1 < \Delta_2 \leq 2 < \Delta_3$. In this case $\Psi_{13} = 4(\Delta_3 - 1)$. The same normalization argument implies $I_2 = \Delta_2^2$ and $I_3 = 4(\Delta_3 - 1)$. With this, the ratio Ψ_{23} is

$$\begin{aligned}
\Psi_{23}^{-1} &= \frac{(I_3 - I_2)^2}{4(\Delta_2 I_3 - \Delta_3 I_2)(\Delta_3 - \Delta_2)} \\
&= \frac{(4(\Delta_3 - 1) - \Delta_2^2)^2}{4(4\Delta_2(\Delta_3 - 1) - \Delta_3 \Delta_2^2)(\Delta_3 - \Delta_2)} \triangleq \varphi(\Delta_2, \Delta_3).
\end{aligned}$$

To eliminate Δ_3 we compute the derivative

$$\frac{d}{d\Delta_3}\varphi(\Delta_2, \Delta_3) = \frac{(\Delta_2 - 2)^3(\Delta_2 - 2\Delta_3 + 2)(-\Delta_2^2 + 4\Delta_3 - 4)}{4\Delta_2(\Delta_2 - \Delta_3)^2((\Delta_2 - 4)\Delta_3 + 4)^2} > 0$$

The inequality holds for all $1 \leq \Delta_2 \leq 2 < \Delta_3$. Hence it suffices to consider the limit

$$\lim_{\Delta_3 \to \infty} \varphi(\Delta_2, \Delta_3) = \frac{4^2}{4(4\Delta_2 - \Delta_2^2)} = \frac{4}{\Delta_2(4 - \Delta_2)} \leq \frac{4}{3}.$$

The last inequality holds for $1 \leq \Delta_2 \leq 2$ and the constraint on p_{23} is satisfied. By Lemma 23, Ψ_{23} cannot be optimal if $\Delta_2 > 2 = 2\Delta_1$. Finally, note that the bound is tight in the same limit. ∎

Appendix B. Additional Proofs

B.1 Proof of Lemma 3

Proof The proof leverages the method of mixtures confidence set for the unconstrained least squares estimate (c.f. Abbasi-Yadkori et al., 2011), where we make the necessary adjustments to express the bound in terms of the basis W. Let $\hat{\vartheta}_t$ be the unconstrained least-square estimate,

$$\hat{\vartheta}_t = \arg\min_{\vartheta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \|M_{a_s}\vartheta - y_s\|^2 + \lambda\|\vartheta - \theta_0\|^2,$$

Note that $\|M_{a_s}\vartheta - y_s\|^2 = \|M_{a_s}WW^\top(\vartheta - \theta^*) - \epsilon_s\|^2$. Hence the minimizer can be parametrized as $\hat{\vartheta}_t = (1_d - WW^\top)\theta_0 + Ww$ for some $w \in \mathbb{R}^d$. Further, $\hat{\theta}_t$ is the projection of $\hat{\vartheta}_t$ onto the convex parameter set Θ with respect to the $\|\cdot\|_{V_t}$ norm. Therefore,

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \|\hat{\vartheta}_t - \theta^*\|_{V_t} = \|W^\top(\hat{\vartheta}_t - \theta^*)\|_{W_t},$$

where we used Eq. (8) in the second step to introduce the basis W. Define $\hat{w}_t = W^\top(\hat{\vartheta}_t - \theta^*) \in \mathbb{R}^r$. Then, using that $y_s = M_{a_s}\theta^* + \epsilon_s$ and writing the least-squares objective directly in $\mathbb{R}^r$, we find

$$\hat{w}_t = \arg\min_{w \in \mathbb{R}^r} \sum_{s=1}^{t-1} \|M_{a_s}Ww - \epsilon_s\|^2 + \lambda\|Ww + \theta^* - \theta_0\|^2$$

$$= (WV_tW)^{-1}\left(\sum_{s=1}^{t-1} WM_{a_s}^\top\epsilon_t - \lambda W(\theta^* - \theta_0)\right)$$

Recall that $W_t = W^\top V_t W$. Using the closed-form to write the quantity of interest, we find

$$\|W^\top(\hat{\vartheta}_t - \theta^*)\|_{W_t} = \|\hat{w}_t\|_{W_t} = \left\|\sum_{s=1}^{t-1} WM_{a_s}^\top\epsilon_t - \lambda W(\theta^* - \theta_0)\right\|_{W_t^{-1}}$$

$$\leq \left\|\sum_{s=1}^{t-1} WM_{a_s}^\top\epsilon_t\right\|_{W_t^{-1}} + \lambda\|W(\theta^* - \theta_0)\|_{W_t^{-1}}$$

$$\leq \left\|\sum_{s=1}^{t-1} WM_{a_s}^\top\epsilon_t\right\|_{W_t^{-1}} + \sqrt{\lambda}\|\theta^* - \theta_0\|$$

The claim follows now directly form Abbasi-Yadkori et al. (2011, Theorem 1), and using that $\lambda\|W(\theta^* - \theta_0)\|_{W_t^{-1}} \leq \lambda\|W(\theta^* - \theta_0)\|_{W_0^{-1}} \leq \sqrt{\lambda}B$. ∎

B.2 Proof of Lemma 6

Proof [Sketch] Recall the definition of the observation matrix S_a in Example 7. By the definition of global observability, for any pair of Pareto optimal actions a and b there

exist vectors $(w_{ab}^c)_{c \in [k]}$ in $\mathbb{R}^m$ such that $\phi_a - \phi_b = \sum_{c \in [k]} S_c^\top w_{ab}^c$. By Proposition 37.18 (Lattimore and Szepesvari, 2020), w_{ab}^c can be chosen so that $\|w_{ab}^c\|_\infty \leq d^{1/2} k^{d/2}$. The bound in *(a)* follows from Cauchy-Schwarz to bound $\|w_{ab}^c\| \leq \sqrt{m} \|w_{ab}^c\|_\infty$ and the second part of Lemma 5. Part *(b)* follows in an identical fashion. For *(c)* we need the concept of neighbours. Pareto actions a and b are neighbours if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d-2$. Given neighbours a and b let $\theta_\circ$ be in the relative interior of $\mathcal{C}_a \cap \mathcal{C}_b$ and $\mathcal{E}' = \{\theta : \|\theta - \theta_\circ\| \leq \epsilon\}$ where $\epsilon > 0$ is small enough that $\mathcal{E}' \subset \mathcal{C}_a \cup \mathcal{C}_b$. By the definition of local observability applied to $\mathcal{E}'$ there exist a vectors $w_{ab}^a, w_{ab}^b \in \mathbb{R}^m$ such that $\phi_a - \phi_b = S_a^\top w_{ab}^a + S_b^\top w_{ab}^b$ and by Proposition 37.18 (Lattimore and Szepesvari, 2020), these vectors can be chosen so that $\|w_{ab}^a\|_\infty \leq m$ and $\|w_{ab}^b\|_\infty \leq m$. Finally, let a and b be arbitrary actions in $\bar{\mathcal{P}}(\mathcal{E})$ that need not be neighbours. Then there exists a sequence $c_1, \ldots, c_j \in \bar{\mathcal{P}}(\mathcal{E})$ with $j \leq k$ and $c_1 = a$ and $c_j = b$ such that c_i and c_{i+1} are neighbours. Therefore,

$$\phi_a - \phi_b = \sum_{i=1}^{j-1} \Big(S_{c_i} w_{c_i c_{i+1}}^{c_i} + S_{c_{i+1}} w_{c_i c_{i+1}}^{c_{i+1}} \Big).$$

By Lemma 5,

$$\alpha(\mathcal{E}) \leq \left(\sum_{i=1}^{j-1} \left\| w_{c_i c_{i+1}}^{c_i} \right\| + \left\| w_{c_i c_{i+1}}^{c_{i+1}} \right\| \right)^2 \leq 4k^2 m^3 \,,$$

which establishes *(c)*. ∎

References

Yasin Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, 2012.

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.

A. Antos, G. Bartók, D. Pál, and Cs. Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Pierre-Cyril Aubin-Frankowski and Zoltan Szabo. Hard shape-constrained kernel machines. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 384–395. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/03fa2f7502f5f6b9169e67d17cbf51bb-Paper.pdf.

J Andrew Bagnell and Amir-massoud Farahmand. Learning positive functions in a hilbert space. In *NIPS Workshop on Optimization (OPT2015)*, volume 20, pages 3240–3255, 2015.

G. Bartók, D. Pál, and Cs. Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 133–154, 2011.

G. Bartók, N. Zolghadr, and Cs. Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML, pages 1779–1786, USA, 2012. Omnipress.

G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and Cs. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39 (4):967–997, 2014.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*, 2017.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4 (04):377–408, 2006.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 142–151, 2012.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31:562–580, 2006.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Sougata Chaudhuri and Ambuj Tewari. Phased exploration with greedy exploitation in stochastic combinatorial partial monitoring games. In *Advances in Neural Information Processing Systems*, pages 2433–2441, 2016.

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

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems 32*, pages 10666–10676. Curran Associates, Inc., 2019.

Dean Foster and Alexander Rakhlin. No internal regret via neighborhood watch. In *Artificial Intelligence and Statistics*, pages 382–390. PMLR, 2012.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Pratik Gajane and Tanguy Urvoy. Utility-based dueling bandits as a partial monitoring game. *arXiv preprint arXiv:1507.02750*, 2015.

Federico Girosi. An equivalence between sparse approximation and support vector machines. *Neural computation*, 10(6):1455–1480, 1998.

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 and Tor Lattimore. Regret bounds for information-directed reinforcement learning. *arXiv preprint arXiv:2206.04640*, 2022.

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

Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8446–8464. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/hao22b.html`.

Kaixuan Huang, Sham M Kakade, Jason D Lee, and Qi Lei. A short note on the relationship of information gain and eluder dimension. *arXiv preprint arXiv:2107.02377*, 2021.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

George S Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

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

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proc. International Conference on Learning Theory (COLT)*, July 2018. URL `https://arxiv.org/abs/1801.09667`.

Johannes Kirschner and Andreas Krause. Bias-robust bayesian optimization via dueling bandits. In *Proc. International Conference on Artificial Intelligence and Statistics (AIS-TATS)*, July 2021.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Proc. International Conference on Learning Theory (COLT)*, July 2020. URL `https://arxiv.org/abs/2002.11182`.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Proc. International Conference on Learning Theory (COLT)*, August 2021.

J. Komiyama, J. Honda, and H. Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 1792–1800. Curran Associates, Inc., 2015.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, NIPS, pages 817–824. Curran Associates, Inc., 2008.

Tor Lattimore and András György. Mirror descent and the information ratio. *arXiv preprint arXiv:2009.12228*, 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 Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory*, pages 529–556, 2019a.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. *arXiv preprint arXiv:1907.05772*, 2019b.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. *arXiv preprint arXiv:1902.00470*, 2019c.

Tor Lattimore and Czsaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2020.

Tian Lin, Bruno Abrahao, Robert Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pages 901–909, 2014.

Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Jonas Mockus. The bayesian approach to global optimization. *System Modeling and Optimization*, pages 473–481, 1982.

Mojmir Mutný and Andreas Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Neural and Information Processing Systems (NeurIPS)*, December 2018.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *Proc. International Conference on Learning Representations (ICLR)*, May 2019. URL `https://arxiv.org/abs/1812.07544`.

George Pedrick. *Theory of reproducing kernels for Hilbert spaces of vector valued functions*. PhD thesis, University of Kansas, 1957.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *International Conference on Computational Learning Theory*, pages 208–223. Springer, 2001.

Yu V Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

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

A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1): 224–243, 1999.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* 2002.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *International Conference on Machine Learning*, 2010.

Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. Advancements in dueling bandits. In *IJCAI*, pages 5502–5510, 2018.

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.

Taira Tsuchiya, Junya Honda, and Masashi Sugiyama. Analysis and design of thompson sampling for stochastic partial monitoring. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8861–8871. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/649d45bf179296e31731adfd4df25588-Paper.pdf`.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. *arXiv preprint arXiv:2009.06966*, 2020.

Michal Valko, Nathaniel Korda, Remi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-Time Analysis of Kernelised Contextual Bandits. *arXiv:1309.6869 [cs, stat]*, September 2013.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54. PMLR, 2014.

Hastagiri Vanchinathan, Gábor Bartók, and Andreas Krause. Efficient partial monitoring with prior information. In *Neural Information Processing Systems (NIPS)*, 2014a.

Hastagiri Vanchinathan, Gábor Bartók, and Andreas Krause. Efficient partial monitoring with prior information. In *Neural Information Processing Systems (NIPS)*, 2014b.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Andrea Zanette and Rahul Sarkar. Information directed reinforcement learning. Technical report, Technical report, Technical report, 2017.