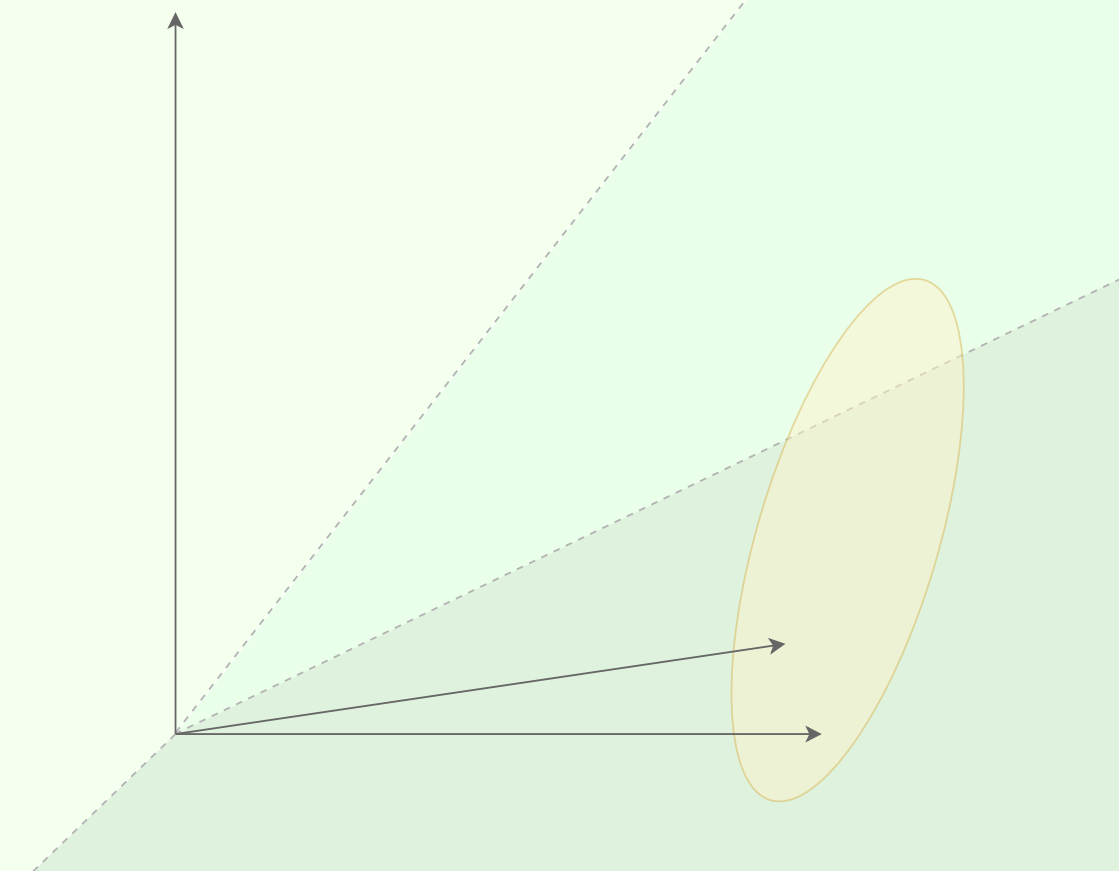
Diss. ETH No. 27627

INFORMATION-DIRECTED SAMPLING FREQUENTIST ANALYSIS AND APPLICATIONS

JOHANNES KIRSCHNER



DISS. ETH NO. 27627

INFORMATION-DIRECTED SAMPLING
FREQUENTIST ANALYSIS AND APPLICATIONS

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

JOHANNES KIRSCHNER
Dipl., Eidgenössisches Polytechnikum

born on 2nd May 1992
citizen of Germany

accepted on the recommendation of

Prof. Dr. Andreas Krause, examiner
Prof. Dr. Benjamin Van Roy, co-examiner
Prof. Dr. Csaba Szepesvári, co-examiner

2021

to my parents

ABSTRACT

Sequential decision-making is an iterative process between a learning agent and an environment. We study the stochastic setting, where the learner chooses an action in each round and the environment returns a noisy feedback signal. The learner's objective is to maximize a reward function that depends on the chosen actions. This basic model has many applications, including adaptive experimental design, product recommendations, dynamic pricing and black-box optimization.

The combination of statistical uncertainty and the objective to maximize reward creates a tension between exploration and exploitation: The learner has to carefully balance between actions that provide informative feedback and actions estimated to yield a high reward. The fields of bandit algorithms and partial monitoring study methods to resolve the exploration-exploitation trade-off optimally, using various regularity assumptions on the feedback-reward structure.

Two of the most widely used methods are optimistic algorithms and Thompson sampling, which have been successfully applied in numerous settings and come with strong theoretical guarantees. More recently, however, an increasing amount of evidence shows that optimism and Thompson sampling are not universal exploration principles. In structured models with correlated feedback, clearly suboptimal actions sometimes provide informative feedback that outweighs their cost. Meanwhile, optimistic approaches and Thompson sampling discard such actions early on, which leads to inefficient exploration.

An alternative and less studied design principle is information-directed sampling (IDS), originally proposed in the Bayesian setting. The main contribution in this thesis is a frequentist interpretation of the IDS framework, complemented with frequentist performance guarantees for several settings with linear reward and feedback structure. Using the IDS approach, we resolve the long-standing challenge to find an asymptotically instance-optimal algorithm for linear bandits that is simultaneously minimax optimal. We further extend the IDS approach to the more general linear partial monitoring setting, making the method applicable to a vast range of previously studied models for sequential decision-making. Along the way, we develop the theory of information-directed sampling, uncover a connection to primal-dual methods and propose computationally faster approximations. Lastly, we discuss extensions of the IDS framework to contextual decision-making and the kernelized setting and highlight example applications.

ZUSAMMENFASSUNG

Sequentielle Entscheidungsfindung ist ein iterativer Prozess zwischen einem Agenten und einer Umgebung. Wir untersuchen den Fall, in dem der Agent in jeder Runde eine Aktion auswählt und die Umgebung ein verrauschtes Feedbacksignal zurückgibt. Das Ziel des Lernenden ist es eine Belohnungsfunktion zu maximieren, die von den gewählten Aktionen abhängt. Dieses Grundmodell hat viele Anwendungen, einschließlich adaptivem Design von Experimenten, Produktempfehlungen, dynamischer Preisgestaltung und Black-Box-Optimierung.

Die Kombination aus statistischer Unsicherheit und dem Ziel die Belohnung zu maximieren erfordert einen Kompromiss zwischen Exploration und Exploitation: Der Lernende muss sorgfältig zwischen Aktionen abwägen, die informatives Feedback liefern, und solchen die voraussichtlich eine hohe Belohnung erzielen. Die Gebiete der Bandit-Algorithmen und Partial Monitoring erforschen Lösungswege, welche optimal zwischen Exploration und Exploitation abwägen, unter Verwendung verschiedener Regelmäßigkeitsannahmen an die Feedback-Belohnungsstruktur.

Zwei der am häufigsten verwendeten Methoden sind optimistische Algorithmen und Thompson-Sampling, die in zahlreichen Fällen erfolgreich angewendet wurden und starke theoretische Garantien bieten. In letzer Zeit zeigen jedoch immer mehr Resultate, dass Optimismus und Thompson-Sampling keine universellen Explorationsprinzipien sind. In strukturierten Modellen mit korreliertem Feedback können eindeutig suboptimale Aktionen manchmal informatives Feedback liefern, das die Kosten überwiegt. Optimistische Ansätze und Thompson-Sampling verwerfen solche Aktionen frühzeitig, was zu einer ineffizienten Exploration führt.

Ein alternatives und weniger untersuchtes Entwurfsprinzip ist Information-Directed Sampling (IDS), das ursprünglich im Bayes'schen Model vorgeschlagen wurde. Der Hauptbeitrag in dieser Arbeit ist eine frequentistische Interpretation des IDS-Frameworks, sowie frequentistische Performance-Guaranteen für verschiedenen Umgebungen mit linearer Belohnungs- und Feedbackstruktur. Mit dem IDS-Ansatz lösen wir das schon lange bestehende Problem einen asymptotisch optimalen Algorithmus für lineare Bandits zu finden, der gleichzeitig minimax-optimal ist. Wir erweitern den IDS-Ansatz zudem auf Partial Monitoring, wodurch die Methode auf eine Vielzahl zuvor untersuchter Modelle für die sequentielle Entscheidungsfindung anwendbar ist. Desweiteren entwickeln wir die Theorie für Information-Directed Sampling, decken einen Zusammenhang mit Primal-Dual-Methoden auf und schlagen rechnerisch schnellere Approximationen vor. Zuletzt diskutieren wir Erweiterungen des IDS-Frameworks für die kontextbezogene Entscheidungsfindung sowie Kernel-Methoden und heben Beispielanwendungen hervor.

ACKNOWLEDGEMENTS

I am truly grateful to all my friends and colleagues, who made the last five years of my life outstanding in every respect and significantly shaped the thesis in one way or another. My sincere thank-you goes to:

1. Andreas Krause, for always outstanding mentorship, academic advice, valuable feedback and the freedom to pursue the ideas that lead to this thesis,
2. Rita Klute, for her invaluable backstage support and making everything possible,
3. Ben Van Roy and Csaba Szepesvári for kindly agreeing to be in the PhD committee and providing feedback on this manuscript,
4. Carl-Johann Simon-Gabriel, Felix Berkenkamp, Ilija Bogunovic, Jonas Rothfuss, Kfir Levy, Mohammad-Reza Karimi, Mojmir Mutný, Sebastian Curi, Tobias Sutter, Tor Lattimore, Ya-Ping Hsieh and Zalán Borsos for detailed feedback and careful proof-reading of paper drafts and thesis chapters (in alphabetical order),
5. Nicole Hiller, Jochem Snuverink, Jaime Coello de Portugal, Rasmus Ischebeck and the crew at PSI for infamous day & night shifts,
6. Claire Vernade, Tor Lattimore and the foundation team at DeepMind for hosting my internship in extraordinary times,
7. my family, for their endless support and dedication,
8. and Laura, for her love.

The research in this thesis was funded by SNSF grant 200020 159557, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 815943.

CONTENTS

NOTATION

$\mathcal{A}$	action set
a, b, c	actions
ϕ_a	action features
M_a	linear feedback map
k	size of finite action set
$\mathcal{M}$	parameter set
θ, ν, ω	parameters
$\hat{\theta}_t$	least-squares estimate
V_t	precision matrix
d	parameter / feature dimension
f, f_θ	reward function
y_t	feedback in round t
ϵ_t	noise
n, t, s	horizon, time step
π_n	policy
μ, μ_t	sampling distribution
$\mathfrak{R}_n$	regret
$\mathcal{Z}$	context set
z	context
$\mathcal{X}$	outcome set in finite partial monitoring
Σ	symbol set in finite partial monitoring
ρ	sub-Gaussian variance
$\Psi_t(\mu)$	information ratio
$\Delta(a)$	suboptimality gap
$\hat{\Delta}_t(a)$	gap estimate
$I_t(a)$	information gain
γ_n	total information gain
$\beta_{t,1/\delta}$	confidence coefficient

1

INTRODUCTION

The fields of bandit algorithms and partial monitoring study strategies for *sequential decision-making* under uncertainty. The setting is formalized as a round-based game between a learner and an environment. At the beginning of each round, the learner chooses an action, and the environment reveals a stochastic feedback signal. The learner then receives a reward that depends on the action and is correlated with the feedback. To maximize the total reward, the learner has to carefully balance actions that lead to informative feedback and actions estimated to yield a high reward. This trade-off is known as the *exploration-exploitation dilemma*. The topic of this thesis is about understanding the exploration-exploitation trade-off under specific structural assumptions on how the feedback and the reward are generated.

The basic bandit setting dates back to Thompson [159], who proposed it for designing adaptive allocation schemes in clinical trials. Since then, bandit algorithms have found many more applications, and the field has grown rapidly in recent years. For a broader introduction, we refer the reader to the excellent books by Lattimore & Szepesvari [103] and Slivkins [149].

In this work, we focus on the *stochastic* version of the bandit problem, where the feedback for each action is sampled from a fixed distribution. There is also the *adversarial* bandit setting, where the data is adversarially generated [15]. We do not further expand in this direction here. A common characteristic of most bandit problems is that the environment is *stateless*. This means the feedback for a specific action is sampled from the same distribution in all rounds. Hence, unlike in reinforcement learning [157], the learner does not face a planning problem.

Let us first give some applications. Bandit algorithms are an attractive tool to optimize high-throughput internet services such as *recommender systems* and *online advertising* [6, 66, 107, 158]. In this scenario, the system presents the user a set of options, for example, a list of movies that the user might like. The goal is to learn the user's preferences through the interaction and, at the same time, maximize the success rate that the user chooses one of the presented options. The important twist is that the learner faces an *online learning problem*: The data acquired by the policy early on is used later to optimize the decisions, and care is required to

avoid an unfavorable selection bias. On the other hand, probing too many unsuccessful options comes at the cost of eventually turning the user away.

Another application is *dynamic pricing* [94]. Here, the learner represents a salesperson who wants to price a product optimally. The price the customer is willing to pay is unknown initially, but for each offer, the learner observes whether the customer buys the product or not. The learner then adapts the price based on this feedback in a way to maximize the revenue.

More generally, the bandit setting is related to the field of adaptive experimental design [37, 41, 57]. For instance, bandit algorithms can be used to design customer surveys and for A/B testing. With a continuous action set, the bandit problem essentially becomes that of zero-order noisy optimization. In this context, the field of *Bayesian optimization* is understood to solve a similar problem, [119, 145], and the algorithms are based on similar ideas [152]. Techniques from bandit algorithms are also used in the tree-search algorithm that is behind the celebrated success of AlphaGo [95, 148], and are used to develop reinforcement learning algorithms [157].

A longer list of applications and further references have been conveniently compiled by Bouneffouf, Rish & Aggarwal [26] and in the book by Lattimore & Szepesvari [103, §1.2]. We will mention more examples as we go, and Chapter 8 is dedicated to applications of the algorithmic ideas developed in this thesis.

Among many algorithms designed for the stochastic bandit setting, two approaches stand out. The first is the *upper confidence bound* (UCB) algorithm [13, 99]. The idea is to compute a high-probability upper bound on the reward of each action based on the accumulated feedback. UCB algorithms choose the action with the largest plausible reward and, in that sense, are *optimistic* about the outcome.

The second popular approach is *Thompson sampling* [138, 159], which is a randomized strategy that samples actions according to their posterior probability of being optimal in a Bayesian model of the rewards. UCB and Thompson sampling are known to satisfy strong theoretical guarantees and are widely successful in practice. They have also been adapted beyond the basic bandit setting, for example, in reinforcement learning [77, 125, 153].

More recently, however, there is evidence that the *optimism principle* and Thompson sampling are *not* universal recipes for exploration [103, 105]. The reason is that both methods are designed to choose only actions that might be optimal. In settings where the feedback is correlated, actions that are known to be suboptimal sometimes provide a substantial amount of information that outweighs the cost. Such actions are never chosen by UCB

or Thompson sampling, which is why these approaches can be inefficient in models with structured feedback.

In this thesis we focus on an alternative *design principle* proposed by Russo & Van Roy [135] called *information-directed sampling* (IDS). The idea is to myopically optimize a trade-off between *information* and the *estimated cost* of choosing a suboptimal action. Without getting too formal yet, we denote by $\mathcal{A}$ the set of actions and the set of sampling distribution over actions by $\mathscr{P}(\mathcal{A})$. Let $\hat{\Delta}_t(\mu)$ be an estimate of the expected loss of sampling actions from a distribution μ in round t, compared to the best action in hindsight. Further, assume that we quantify the expected information from the feedback with a function $I_t(\mu)$. Information-directed sampling is defined to optimize the following trade-off:

$$\mu_t^{\text{IDS}} = \operatorname*{arg\,min}_{\mu \in \mathscr{P}(\mathcal{A})} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \tag{1.1}$$

Previous work [111, 135] and the theoretical analysis we present in this thesis show that IDS satisfies strong guarantees for regret minimization in an exceptionally wide range of settings. Our analysis goes well beyond the previous work and provides new insights into how IDS solves the exploration-exploitation trade-off. An interesting and previously little explored feature of IDS is that it adapts to the hardness of the problem. This is an important property, because not every instance encountered in practice is as difficult as the theoretical analysis suggests. We will also be specific about the computational complexity required to compute the trade-off defined by Eq. (1.1). In most cases, we provide pseudo code to efficiently compute or approximate the IDS algorithm, and the implementation often requires only a few lines of code. Last but not least, IDS is often competitive with state of the art methods on numerical benchmarks.

BASIC NOTATION Before we introduce the setting formally, we settle on the basic mathematical notation that is used throughout. The real numbers are $\mathbb{R}$, and $\mathbb{R}_{\geq 0}$ denotes the non-negative reals. The standard Euclidean norm is $\|\cdot\|$ and the Euclidean inner product is $\langle \cdot, \cdot \rangle$. The Euclidean basis in $\mathbb{R}^d$ is $e_1, \ldots, e_d$. The identity matrix in $\mathbb{R}^{d \times d}$ is $\mathbf{1}_d$. For a positive (semi-)definite, symmetric matrix $A \in \mathbb{R}^{d \times d}$ and a vector $v \in \mathbb{R}^d$, the associated matrix (semi-)norm is $\|v\|_A^2 = \langle v, Av \rangle$. The smallest and largest eigenvalues of a matrix A are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. For two square matrices A, B, $A \preceq B$ means that $B - A$ is positive semi-definite. The operator norm of a matrix C is $\|C\|_2 = \sqrt{\lambda_{\max}(C^\top C)}$.

For a set $\mathcal{C} \subset \mathbb{R}^d$ we define $\dim(\mathcal{C})$ as the dimension of the affine hull of $\mathcal{C}$. The diameter is $\operatorname{diam}(\mathcal{C}) = \sup_{x,y\in\mathcal{X}} \|x - y\|$. The relative interior of a convex set $\mathcal{C}$ is $\operatorname{relint}(\mathcal{C}) = \{x \in \mathcal{C} : \forall y \in \mathcal{C}\ \exists \lambda > 1 : \lambda x + (1-\lambda)y \in \mathcal{C}\}$, and the set of extreme points is $\operatorname{ext}(\mathcal{C})$.

The space of Borel probability measures on a topological set $\mathcal{S}$ is $\mathscr{P}(\mathcal{S})$. The support of a distribution is $\operatorname{supp}(\mu)$. For measurable functions $F : \mathcal{S} \to \mathbb{R}$ and a distribution $\mu \in \mathscr{P}(\mathcal{S})$, we extend the argument linearly to denote the Lebesgue integral $F(\mu) = \int_{\mathcal{X}} F(s) d\mu(s)$. For finite $\mathcal{S}$ and where more convenient, we use vector notation, including inner products to denote evaluation of functions, for example $F(s) = \langle e_s, F\rangle$. Occasionally, we use the Dirac measure on $s \in \mathcal{S}$, denoted by e_s to match the interpretation as a standard basis vector.

1.1 REGRET MINIMIZATION

Let $\mathcal{A}$ be a compact set of *actions*. Sometimes, we additionally require that $\mathcal{A}$ is finite, which is then specified in the relevant context. The *reward function* $f_\theta : \mathcal{A} \to \mathbb{R}$ is parameterized by $\theta \in \mathcal{M}$, where $\mathcal{M}$ is the *model class*. At the beginning of the game, the learner knows both $\mathcal{A}$ and $\mathcal{M}$, but the reward function is unknown. The interaction between the learner and the environment is on time steps $t = 1, \dots, n$. In practice, it is often useful to have strategies that do not rely on the knowledge of the *horizon* n.

The learner's actions are defined by a *policy* $\pi_n = (\mu_t)_{t=1}^n$ that consists of an adaptive sequence of *sampling distributions* $\mu_t \in \mathscr{P}(\mathcal{A})$. At time t, the learner samples an action a_t from μ_t, and obtains a *feedback* y_t. The filtration $\mathcal{F}_t$ is the σ-algebra generated by the observation history $\{a_s, y_s\}_{s=1}^{t-1}$. We abbreviate the conditional probability and expectation with $\mathbb{P}_t[\cdot] \triangleq \mathbb{P}[\cdot|\mathcal{F}_t]$ and $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot|\mathcal{F}_t]$.

So far, we have not specified the feedback. *Bandit feedback* is defined as the reward of the chosen action, $y_t = f_\theta(a_t) + \epsilon_t$, subject to zero-mean observation noise ϵ_t. This type of feedback is the topic of Chapters 3 to 5. Throughout, we assume that the noise is light-tailed, which allows us to make use of standard concentration results for estimation. Concretely, we require that ϵ_t is conditionally independent of a_t and ρ-sub-Gaussian,

$$\forall \eta \in \mathbb{R}, \quad \mathbb{E}_t[\exp(\eta \epsilon_t)|a_t] \le \exp(\eta^2 \rho^2/2)\,. \tag{1.2}$$

Partial monitoring uses a more general feedback structure, which we preview below in Section 1.2.3 and then study in detail in Chapters 6 to 8.

The generalized feedback model includes most stateless online learning problems, including *full information* and semi-bandit models.

The learner's objective is to maximize the sum of rewards $\sum_{t=1}^{n} f_\theta(a_t)$ collected in n rounds. Equivalently, the learner minimizes the loss relative to the best solution in hindsight, which leads us to the *expected regret*,

$$\mathfrak{R}_n(\pi, \theta) \triangleq \mathbb{E}\left[\sum_{t=1}^{n} f_\theta(a^*) - f_\theta(a_t)\right]. \tag{1.3}$$

The expectation is over the randomness of the policy and the noise. When the policy and the instance is clear from the context, we omit the dependence of the regret on π_n and θ, and write $\mathfrak{R}_n = \mathfrak{R}_n(\pi_n, \theta)$. We denote by $\Delta(a) \triangleq \max_{b \in \mathcal{A}} f_\theta(b) - f_\theta(a)$ for the *suboptimality gap* of an action $a \in \mathcal{A}$. The regret $\Delta(a_t)$ suffered at time t is also called the *instantaneous regret*, and the cumulative regret is written as $\mathfrak{R}_n(\pi_n, \theta) = \mathbb{E}[\sum_{t=1}^{n} \Delta(a_t)]$. We are primarily interested in two ways of bounding the regret: *worst-case regret* and *instance-dependent regret*, which we formally define next.

1.1.1 *Worst-case Regret*

The worst-case regret is defined over a fixed model class $\mathcal{M}$,

$$\mathfrak{R}_n(\pi, \mathcal{M}) \triangleq \sup_{\theta \in \mathcal{M}} \mathfrak{R}_n(\pi, f_\theta).$$

Besides the horizon n, an upper bound on the worst-case regret depends on quantities that are known to the learner at the beginning of the game: the model class $\mathcal{M}$, the action set $\mathcal{A}$ and properties of the observation likelihood like the sub-Gaussian noise variance. However, the supremum makes the bound independent of the parameter $\theta \in \mathcal{M}$ that defines the instance the learner is facing. We say a policy π has *no regret* on the model class $\mathcal{M}$ if $\limsup_{n \to \infty} \frac{1}{n} \mathfrak{R}_n(\pi, \mathcal{M}) \to 0$, which requires the learner to query eventually (near-) optimal actions most of the time.

1.1.2 *Instance-dependent and Asymptotic Regret*

The difficulty of the learning problem also varies with the instance f. Some instances can be significantly easier compared to the worst-case realization. It is therefore of interest to bound the *instance-dependent* regret, $\mathfrak{R}_n(\pi, \theta)$, where we allow the bound to additionally depend on the instance defined by $\theta \in\in \mathcal{M}$. Achieving near-optimal bounds on the instance-dependent

regret is often a challenging endeavor, and a less demanding starting point is to study the appropriately normalized *asymptotic regret*,

$$\limsup_{n\to\infty} \frac{\mathfrak{R}_n(\pi,\theta)}{\log(n)}\,.$$

1.2 MULTI-ARMED BANDITS AND BEYOND

The simplest and most extensively studied case is the *multi-armed bandit* setting, where the action set is finite, and the model class contains all functions that map actions to a bounded reward, e.g. $\mathcal{M} = [0,1]^{\mathcal{A}}$. Traditionally, the literature refers to actions as *arms*, which pictorially stands for levers of a gambling slot machine that the player can pull to obtain a random reward [32]. The arms are *independent*, in the sense that reward observations from one action do not allow to infer the reward of any other action.

A widely successful idea for exploration is the *optimism principle*, which refers to choosing actions that promise the potentially largest reward. The reasoning is that when the optimism is justified, the learner obtains a high reward. Otherwise, when the reward is significantly smaller than expected, the estimates are updated, and consequently, the action is excluded from future experimentation. The strategy is designed to rule out sub-optimal actions systematically and is often highly effective in practice. Adaptation of the optimism principle to the bandit setting dates back to Lai & Robbins [99], who formalized the idea with upper confidence bound (UCB) strategies. The first finite-time analysis of UCB in the multi-armed bandit setting is by Auer, Cesa-Bianchi & Fischer [14].

A different approach is by Gittins [65], who derived score functions for each arm from a dynamical programming solution in a Bayesian model with discounted reward. The agent that maximizes the *Gittins indices* is a Bayes-optimal policy, that satisfies a distinct relation to the UCB approach [134].

For action sets of size $k = |\mathcal{A}|$, a carefully balanced algorithm based on the upper confidence bound approach achieves worst-case regret at most

$$\mathfrak{R}_n \leq \mathcal{O}(\sqrt{nk})\,,$$

and, at the same time, instance-dependent regret that satisfies

$$\lim_{n\to\infty} \frac{\mathfrak{R}_n}{\log(n)} \leq \sum_{a\in\mathcal{A}:\Delta(a)>0} \frac{2}{\Delta(a)}\,.$$

The result and variants are by Audibert & Bubeck [12], Degenne & Perchet [49], and Ménard & Garivier [117]. The bound on the worst-case regret

matches the order of the lower bound by Auer *et al.* [15] and the asymptotic bound matches the lower bound by Lai & Robbins [99] exactly.

1.2.1 *Structured Bandits*

The general case with arbitrary model class $\mathcal{M}$ is known as the *structured bandit setting* [43] and provides flexibility to impose structural constraints on the reward function such as linearity, Lipschitz continuity or unimodality. The general structured setting is used in Chapter 4, whereas all other chapters assume a linear reward and feedback model. In Chapter 7 we discuss parameter sets with convex constraints.

INFORMATION-DIRECTED SAMPLING Similar to the optimism principle, the IDS framework describes a way to design algorithms for different settings. As briefly described in Eq. (1.1), IDS is defined to optimize the sampling distributions on a trade-off between squared expected regret and information gain. The original formulation by Russo & Van Roy [135] is in the structured bandit setting with a Bayesian prior on the parameter. This includes linear feedback as a special case, which was explicitly analyzed in the same paper. We defer a formal and more detailed introduction of information-directed sampling to Chapter 2.

1.2.2 *Stochastic Linear Bandits*

In the linear bandit model, $\mathcal{M} \subset \mathbb{R}^d$ is a compact set of d-dimensional parameters, and the actions are represented by features $\mathcal{A} \subset \mathbb{R}^d$. The reward function is defined by a parameter $\theta \in \mathcal{M}$ such that $f_\theta(a) = \langle a, \theta \rangle$. Linear bandits date back to Abe & Long [4] and have been extensively studied since then. An overview of the existing work is given below.

There are many reasons for studying the linear case. Unlike in the multi-armed bandit setting, the linear structure allows the learner to estimate the reward of an action without directly observing it. This also means that regret bounds do not necessarily scale with the number of actions, and we can get meaningful bounds even for large or continuous action sets. Linear least-squares estimation conveniently offers analytically tractable inference. The linear structure also implies that observations are correlated in a way that makes the exploration-exploitation trade-off much more subtle, in particular for asymptotically optimal exploration. Lastly, the choice of action features provides considerable flexibility; ultimately with

kernel methods, where the basis is chosen in a Hilbert space that represents continuous sets of non-linear functions. We roughly categorize previous work in the following.

OPTIMISTIC APPROACHES Upper confidence bound algorithms were first adapted to the linear bandit setting by Auer [13]. The analysis was later refined by Dani, Hayes & Kakade [46], Rusmevichientong & Tsitsiklis [132] and Abbasi-Yadkori, Pál & Szepesvári [3]. The latter work introduces the self-normalized confidence sets for online least-squares regression, which we repeatedly use in this thesis. Note that in the linear bandit setting, the upper confidence bound algorithm is sometimes abbreviated as LinUCB to distinguish it from the algorithm for the independent arm case, or OFUL for "optimism in the face of uncertainty linear bandit algorithm" [3]. Here we write UCB for all upper confidence type algorithms and explicitly specify the way the confidence bounds are defined in the relevant context.

THOMPSON SAMPLING As mentioned in the introduction, the first bandit algorithm is by Thompson [159], and the method is known today as *Thompson sampling*. The strategy is defined to sample actions according to their probability of being optimal in a Bayesian model of the reward function. A connection between Thompson sampling and UCB algorithms is established by Russo & Van Roy [136], who provided the first bound on the Bayesian regret in the linear setting. An elegant information-theoretic analysis is by the same authors Russo & Van Roy [137], which then lead to the information-directed sampling framework. The first frequentist analysis of Thompson sampling for linear bandits is by Agrawal & Goyal [7], and an alternative proof with an optimistic interpretation is by Abeille & Lazaric [5].

Thompson sampling is a widely popular method due to its simplicity and wide applicability when a Bayesian posterior is available [138]. The method is often accredited superior empirical performance compared to UCB algorithms [38]. Arguably, this is mainly due to tighter concentration of the Bayesian credible sets, which make explicit use of the Bayesian realizability assumption. There is evidence that carefully tuned UCB algorithms are competitive with Thompson sampling [138, §7.4] and [103, §36.2].

EXPLORE-THEN-COMMIT AND PHASED EXPLORATION A simple and more explicit way to implement the exploration-exploitation trade-off is by introducing phases that alternate between exploration and exploitation. In the exploration phase, the learner uses a fixed sampling strategy to reduce

the uncertainty about the unknown maximizer. In exploitation rounds, the learner chooses the empirically best action. The effectiveness of this approach largely depends on the design of the exploration distribution, which is directly related to experimental design. This approach has been successfully applied to the linear bandit setting by Abbasi-Yadkori, Antos & Szepesvári [2] and Rusmevichientong & Tsitsiklis [132]. Interestingly, a carefully designed algorithm that sequentially eliminates sub-optimal actions with an i.i.d. sampling scheme achieves better bounds than the UCB approach in some cases, see Lattimore & Szepesvari [103, §22]. Some asymptotic algorithms also rely on a *forced exploration* to initialize the estimates [72, 104]. More recently, Wagenmaker, Katz-Samuels & Jamieson [169] combine optimal experimental design with a phased elimination-style algorithm to derive finite-time guarantees that scale with the Gaussian width of the action set.

ASYMPTOTICALLY OPTIMAL ALGORITHMS Lattimore & Szepesvári [104] showed that algorithms based on optimism or Thompson sampling are not asymptotically optimal in the linear setting. They proposed an approach based on the explore-then-commit framework that computes an estimate of the optimal allocation and updates the allocation to match the predicted target. Combes, Magureanu & Proutiere [43] follow a similar plan for the more general structured bandit setting. This idea was subsequently extended to the contextual setting by Hao, Lattimore & Szepesvari [72]. Unfortunately, these algorithms are not really practical and do not enjoy reasonable minimax regret. More recently, Jun & Zhang [79] refined this technique in the structured setting with a finite model class to avoid forced exploration and the knowledge of the horizon. Similarly, Van Parys & Golrezaei [166] use a dual formulation of the lower bound to devise an algorithm that achieves the optimal asymptotic regret up to a constant and avoids resolving for the predicted optimal allocation at every round.

A different route is taken by Degenne, Shao & Koolen [50], who translate the Lagrangian of the lower bound into a fictitious two-player game, where the saddle point corresponds to the optimal asymptotic regret. Using tools from online convex optimization [74, 124], this leads to a family of asymptotically optimal algorithms, which incrementally update the allocation in each round based on primal-dual updates on the Lagrangian of the lower bound. The concurrent work by Tirinzoni *et al.* [160] is also a primal-dual method, and unlike previous methods both worst-case and asymptotically optimal. The approach further applies to the contextual

case. We explore a connection between information-directed sampling and primal-dual methods in Chapter 4 and a contextual version of IDS in Chapter 8.

KERNEL METHODS Algorithms for the *kernelized* linear bandit setting make use of the *kernel trick* to avoid a direct representation of the feature vectors, and instead only rely on computation of inner products [130, 143]. The advantage is that one can represent the reward function in a potentially infinite-dimensional reproducing kernel Hilbert space (RKHS). The choice of kernel function and the norm of the parameter correspond to a smoothness prior on the reward function. Cumulative regret of kernelized versions of UCB is studied by Abbasi-Yadkori [1], Chowdhury & Gopalan [42], Srinivas *et al.* [152], and Valko *et al.* [164]. A kernelized analysis of Thompson sampling is by Chowdhury & Gopalan [42]. The main contribution of this line of work is to replace the (potentially infinite) dimension with an appropriate notion of an *effective dimension*. Closely related in this context is also the field of Bayesian optimization [119] with a large body of work and applications on its own. For further details, the reader is referred to the surveys by Shahriari *et al.* [145] and Frazier [60].

1.2.3 *Linear Stochastic Partial Monitoring*

Partial monitoring generalizes the linear bandit framework by decoupling the feedback signal from the reward. This makes it an extremely flexible framework that models numerous stateless online decision-making problems studied in the literature. A simple example is dynamic pricing, which we mentioned in the introduction. Partial monitoring encompasses many other common settings such as dueling bandits [176], bandits with graph feedback [115] and cascading bandits [98]. We present more detailed examples in Chapters 6 to 8.

In the linear variant of the setting, each action $a \in \mathcal{A}$ is associated with a d-dimensional feature vector $\phi_a \in \mathbb{R}^d$. The reward $f_\theta(a) = \langle \phi_a, \theta \rangle$ is defined by a parameter $\theta \in \mathbb{R}^d$ in the same way as in the linear bandit model. The difference is that in partial monitoring, the learner does *not* observe the reward directly. Instead, each action $a \in \mathcal{A}$ has an associated linear feedback operator $M_a : \mathbb{R}^d \to \mathbb{R}^m$, and the observation in round t is $y_t = M_{a_t}\theta + \epsilon_t$ where ϵ_t is a sub-Gaussian random vector in $\mathbb{R}^m$.

FINITE PARTIAL MONITORING The partial monitoring framework was first introduced by Rustichini [139]. The general class of partial monitoring games has a rich geometric structure [16, 106] where the achievable worst-case regret rate depends on precise observability conditions. Classifying games according to the difficulty of the exploration-exploitation trade-off is an essential part of previous work. The complete classification of all finite games is achieved in a line of work by Antos *et al.* [9], Bartók *et al.* [16], Cesa-Bianchi, Lugosi & Stoltz [35], Lattimore & Szepesvári [106], and Piccolboni & Schindelhauer [128], with a focus on the stochastic version of the problem in the work by Bartók, Pál & Szepesvári [17] and Bartók, Zolghadr & Szepesvári [18]. Asymptotics for finite games are known as well [96]. Partial monitoring with prior information was studied by Vanchinathan, Bartók & Krause [167], and with side-information by Bartók & Szepesvári [19]. Thompson sampling has been analyzed for games that satisfy a strong local observability condition by Tsuchiya, Honda & Sugiyama [162].

LINEAR PARTIAL MONITORING The linear version of the problem and a phased exploration scheme is due to Lin *et al.* [109]. Another phased exploration method was proposed by Chaudhuri & Tewari [40]. Both algorithms rely on oracle solvers for the offline problem and are therefore suitable for potentially exponential action sets. However, the analysis and algorithm design is tailored to a *global* observability assumption (formally defined in Section 6.2), and is not adaptive towards more benign cases where faster rates are possible. The analysis of Thompson sampling by Tsuchiya, Honda & Sugiyama [162] also applies in the linear setting but requires a strong local observability assumption.

The use of information-directed sampling for partial monitoring has been suggested already by Russo & Van Roy [135], and in fact, some of the examples in this work capture the spirit of partial monitoring. However, to the best of our knowledge, there is no prior work that provides an explicit formulation of IDS for the partial monitoring setting.

1.3 CONTRIBUTIONS

In this thesis, we develop a *frequentist version* of information-directed sampling. Using the IDS design principle, we derive novel algorithms for regret minimization in linear bandits and partial monitoring, each complemented with strong theoretical performance guarantees. A reading guide and overview of the contributions is given below:

- In Chapter 2, we formally introduce information-directed sampling and present general results that are useful throughout. This includes a new *approximation* of IDS that is computationally more efficient.
- In Chapter 3, we develop an algorithm for heteroscedastic linear bandits based on the frequentist IDS principle. Using the tools from the second chapter, we derive bounds on the worst-case regret that match the best-known bounds for UCB in the homoscedastic case, but can be better by an arbitrarily large factor when the noise is heteroscedastic.
- In Chapter 4, we uncover a fundamental connection between IDS and primal-dual methods for regret minimization on structured bandits. Our analysis sheds light on how IDS balances the trade-off between regret and information in the asymptotic regime.
- In Chapter 5, we build on the primal-dual analysis and introduce an information gain function, for which IDS is asymptotically instance-optimal and (nearly) minimax optimal on linear bandits, resolving a long-standing challenge. The resulting algorithm is remarkably simple, anytime, and has the same computational complexity as UCB.
- In Chapter 6, we generalize the IDS algorithm to the linear partial monitoring framework. We show that the same IDS algorithm is worst-case optimal in all linear games with finitely many actions. The result is complemented with a classification of linear partial monitoring with finitely many actions, showing that the regret rate is either 0, $\tilde{\Theta}(\sqrt{n})$, $\tilde{\Theta}(n^{2/3})$ or $\Omega(n)$.
- In Chapter 7, we refine the analysis for linear partial monitoring for parameter sets with convex constraints. We then apply these results to the classical finite partial monitoring setting, where the parameter is in the simplex, and show that IDS matches the established classification result for finite partial monitoring.
- In Chapter 8, we conclude the technical part with two extensions: A contextual formulation of linear partial monitoring and a kernelized variant of IDS. We further supply a set of example applications, including contextual customer surveys, dueling bandits, robust regret minimization, and Bayesian optimization with gradient observations.
- Lastly, in Chapter 9 briefly compare the Bayesian and frequentist IDS frameworks and conclude with a list of open problems.

1.3.1 *Publications and Collaborations*

This thesis would not have been possible without my advisor, Andreas Krause, and many of the ideas presented here have been shaped in our meetings. To a large extend, the results on partial monitoring are joint work with Tor Lattimore. I further enjoyed collaborating with my colleagues on numerous ideas. Individual contributions are accredited in more detail at the end of each chapter. Results presented and not otherwise cited are by the author and collaborators. The thesis is based on the following conference papers:

- Kirschner, J. & Krause, A. *Information Directed Sampling and Bandits with Heteroscedastic Noise* in *Proc. International Conference on Learning Theory (COLT)* (July 2018)
- Kirschner, J., Lattimore, T. & Krause, A. *Information Directed Sampling for Linear Partial Monitoring* in *Proc. International Conference on Learning Theory (COLT)* (July 2020)
- Kirschner, J., Lattimore, T., Vernade, C. & Szepesvári, C. *Asymptotically Optimal Information-Directed Sampling* in *Proc. International Conference on Learning Theory (COLT)* (Aug. 2021)
- Kirschner, J. & Krause, A. *Bias-Robust Bayesian Optimization via Dueling Bandits* in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)* (July 2021)

FURTHER PUBLICATIONS The following publications of the author and collaborators are more broadly relevant to the topic of this thesis, but have not been directly included. The first is an heuristic extension of IDS to reinforcement learning, using the ideas developed in Chapter 3. The numerical results show that the IDS approach leads to improvements on the Atari benchmark suite [21].

- Nikolov, N., Kirschner, J., Berkenkamp, F. & Krause, A. *Information-Directed Exploration for Deep Reinforcement Learning* in *Proc. International Conference on Learning Representations (ICLR)* (May 2019)

The second set of publications studies variants of the bandit model, such as linear bandits with stochastic context, distributionally robust optimization and best arm identification. The last publication in the list below makes use of the primal-dual approach that we explore in Chapter 4.

- Kirschner, J. & Krause, A. *Stochastic Bandits with Context Distributions* in *Proc. Neural Information Processing Systems (NeurIPS)* (Dec. 2019)
- Mutný, M., Kirschner, J. & Krause, A. *Experimental Design for Optimization of Orthogonal Projection Pursuit Models* in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)* (Feb. 2020)
- Kirschner, J., Bogunovic, I., Jegelka, S. & Krause, A. *Distributionally Robust Bayesian Optimization* in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)* (Aug. 2020)
- Jourdan, M., Mutný, M., Kirschner, J. & Krause, A. *Efficient Pure Exploration for Combinatorial Bandits with Semi-Bandit Feedback* in *Algorithmic Learning Theory* (2021), 805

Lastly, in collaboration with Nicole Hiller, Jochem Snuverink, Jaime Coello and Rasmus Ischebeck at the Paul Scherrer Institute, we developed data driven tuning methods for particle accelerators, using techniques from kernelized bandits. These works focus on practical aspects of Bayesian optimization, including safety constraints and high-dimensional settings. We successfully deployed our methods on the Swiss Free Electron Laser [SwissFEL, 118] and the High-Intensity Proton Accelerator [HIPA, 144]. While these projects did not fit the scope of the thesis, the application has shaped a few examples presented in Chapter 8.

- Kirschner, J., Mutný, M., Hiller, N., Ischebeck, R. & Krause, A. *Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional Subspaces* in *Proc. International Conference for Machine Learning (ICML)* (June 2019)
- Kirschner, J., Nonnenmacher, M., Mutný, M., Hiller, N., Adelmann, A., Ischebeck, R. & Krause, A. *Bayesian Optimization for Fast and Safe Parameter Tuning of SwissFEL* in *Proc. International Free-Electron Laser Conference (FEL2019)* (June 2019)

2

INFORMATION-DIRECTED SAMPLING

Information-directed sampling (IDS) is a *design principle*, that, like the optimism principle, leads to different algorithms in different settings. The framework was introduced with a Bayesian analysis by Russo & Van Roy [135]. We briefly review the Bayesian formulation of IDS and related results in Section 2.3. All other results in this chapter are purely algorithmic, and do not rely on a modeling assumption.

We assume that the learner is provided a *gap estimate* $\hat{\Delta}_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ and an *information gain* $I_t : \mathcal{A} \to \mathbb{R}_{\geq 0}$ at the beginning of each round t. We assume that $\hat{\Delta}_t$ and I_t are continuous (which only really is a restriction if $\mathcal{A}$ is not finite). For the purpose of this chapter, there is no need to commit to specific choices yet, but we will develop concrete choices for both $\hat{\Delta}_t$ and I_t in the subsequent chapters.

Naturally, the gap estimates and information gain are computed using observations from previous rounds, and are therefore predictable on the filtration $\mathcal{F}_t$. We also require that I_t is not zero everywhere. The IDS distribution μ_t^{IDS} is defined to optimize the trade-off between squared expected regret and expected information gain:

$$\mu_t^{\mathrm{IDS}} = \operatorname*{arg\,min}_{\mu \in \mathscr{P}(\mathcal{A})} \left\{ \Psi_t(\mu) \triangleq \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \right\}. \tag{2.1}$$

IDS is defined as the policy $\pi_n^{\mathrm{IDS}} = (\mu_t^{\mathrm{IDS}})_{t=1}^n$ that samples $a_t \sim \mu_t^{\mathrm{IDS}}$ in round t. A minimizer always exists on compact $\mathcal{A}$ and can be chosen arbitrarily if the IDS distribution μ_t^{IDS} is not unique. The objective $\Psi_t(\mu)$ is called the *information ratio* of the sampling distribution $\mu \in \mathscr{P}(\mathcal{A})$. Sometimes, we overload the notation and write $\Psi_t(a) = \Psi_t(e_a)$ were e_a is a Dirac on action $a \in \mathcal{A}$. Intuitively, a small information ratio requires the learner to sample actions from a distribution with small expected regret and large information gain. Since there is only a certain amount of information, eventually the information gain is vanishing and the learner has to play actions with small regret. The complete approach is summarized in Algorithm 1.

Algorithm 1: Information-Directed Sampling

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

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

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

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

A REGRET INEQUALITY The information ratio 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 estimated cumulative regret:

$$\mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(\mu_t)\right] = \mathbb{E}\left[\sum_{t=1}^n \sqrt{\Psi_t(\mu_t) I_t(\mu_t)}\right]$$
$$\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^n \Psi_t(\mu_t)\right] \mathbb{E}\left[\sum_{t=1}^n I_t(a_t)\right]} \tag{2.2}$$

The first step 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 the (2.2) follows from the Cauchy-Schwarz inequality and another application of the tower rule. Note that IDS is the policy that myopically minimized the first sum in the upper bound. The second sum is the *total information gain*, which we abbreviate with

$$\gamma_n \triangleq \sum_{t=1}^n I_t(a_t)\,. \tag{2.3}$$

For the regret $\mathfrak{R}_n = \mathbb{E}[\sum_{t=1}^n \Delta(a_t)]$, Eqs. (2.2) and (2.3) imply

$$\mathfrak{R}_n = \mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(a_t)\right] + \mathbb{E}\left[\sum_{t=1}^n \Delta(a_t) - \hat{\Delta}_t(a_t)\right]$$
$$\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^n \Psi_t\right] \mathbb{E}[\gamma_n]} + \mathbb{E}\left[\sum_{t=1}^n \Delta(a_t) - \hat{\Delta}_t(a_t)\right]. \tag{2.4}$$

This is as far as we can go without a more concrete setting, and specific choices for the gap estimate and information gain. We summarize the re-occurring arguments used in the IDS analysis:

i) 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 $\mathbb{E}\left[\sum_{t=1}^n \Delta(a_t) - \hat{\Delta}_t(a_t)\right]$ contributes only negligibly to the overall regret.

ii) The total information gain $\gamma_n = \sum_{t=1}^n I_t(a_t)$ can be interpreted as the sample complexity of identifying the best action and typically has a logarithmic dependence on the horizon.

iii) The information ratio is bounded in a way that for any sampling path up to time t, there exists a sampling distribution μ_t with bounded information ratio, $\Psi_t(\mu_t) \leq \alpha$.

By Eq. (2.4), the policy $\pi_n = (\mu_t)_{t=1}^n$ constructed from iii) has regret at most

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

Moreover, the bound is non-trivial under the premises i-ii). The usefulness of the bound stems from the fact that we can explicitly design sampling distributions that achieve a small information ratio as in iii). The same bound automatically applies by definition to the IDS algorithm, and we do not have to know exactly which action is realized as a sample from the IDS distribution. It has already been demonstrated by Russo & Van Roy [135] that bounds derived with this way are (near) optimal in many settings, and we will see further examples in this thesis. In the next section we study properties of the information ratio and the IDS distribution. The regret bounds are formalized in Section 2.2.

2.1 PROPERTIES OF THE INFORMATION RATIO

The information ratio $\Psi_t(\mu)$ has many remarkable properties that make it a far more tractable object than it suggests at first sight. The existence of the IDS distribution $\mu_t^{\text{IDS}} = \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$ is immediate for finite action sets, and is formally proven for compact $\mathcal{A}$ in Lemma 2.1. The information ratio and the IDS distribution further satisfy the following properties:

i) The IDS distribution is invariant under constant re-scaling of the gap estimate or the information gain (Lemma 2.2).

ii) The function $\mu \mapsto \Psi_t(\mu)$ is convex on $\mathscr{P}(\mathcal{A})$ (Lemma 2.3).

iii) The support of IDS distribution μ_t^{IDS} can always be chosen on at most two actions that satisfy an affine-linear relationship (Lemma 2.4).

iv) For two actions, the IDS distribution μ_t^{IDS} has an analytic closed form (Lemma 2.5).

v) The expected regret of the IDS distribution is not too far from greedy, specifically $\hat{\Delta}(\mu_t^{\text{IDS}}) \leq 2 \min_{a \in \mathcal{A}} \hat{\Delta}(a)$ (Lemma 2.6).

vi) The information ratio obtained by a distribution that randomizes only between the greedy action $\hat{a}_t = \arg\min_{a \in \mathcal{A}} \hat{\Delta}_t(a)$ and one other action in $\mathcal{A} \setminus \{\hat{a}_t\}$ is within a constant factor $\frac{4}{3}$ of the optimal ratio (Lemma 2.7).

Properties ii) and iii) were established by Russo & Van Roy [135] for finite action sets, and we only provide the technical extension to compact $\mathcal{A}$. The fact that the IDS distribution can always be chosen with a support on at most two actions is frequently used in the following, and has implications for computation as discussed in Section 2.1.1. In Chapter 4 we show that the two actions chosen by IDS correspond to exploration and exploitation in a precise mathematical sense.

Lemma 2.1 (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 [129] 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 2.2 (Invariance)**.** *The information ratio is invariant under re-scaling of the gap estimate and the information gain.*

Proof. Immediate. □

Lemma 2.3 (Convexity [135, 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 [27, 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 & Van Roy [135, Prop. 6] to compact $\mathcal{A}$.

Lemma 2.4 (Support)**.** *The IDS distribution $\mu_t^{\mathrm{IDS}} \in \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu)$ can always be chosen such that $|\operatorname{supp}(\mu_t^{\mathrm{IDS}})| \leq 2$. Further, for $a \in \mathcal{A}$ define*

$$h_t(a) \triangleq 2\, \hat{\Delta}_t(\mu^*)\hat{\Delta}_t(a) - \Psi_t(\mu_t^{\mathrm{IDS}}) I_t(a)\,.$$

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

Proof. We claim that

$$h_t(a) = \min_{b \in \mathcal{A}} h_t(b) \quad \text{for all } a \in \operatorname{supp}(\mu_t^{\mathrm{IDS}})\,. \tag{2.5}$$

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 $\mu^{\mathrm{IDS}}(p) = (1-p)e_{a_{\min}} + p e_{a_{\max}}$, 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 $\mu_t^{\mathrm{IDS}} = \mu_t(p^*)$. By Eq. (2.5) we get $I_t(\mu_t^{\mathrm{IDS}}) = I_t(\mu^*)$. Therefore $\Psi_t(\mu^*) = \Psi_t(\mu_t^{\mathrm{IDS}})$ and μ_t^{IDS} is a minimizing distribution with support size at most 2.

To show Eq. (2.5), 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 $a = \arg\min_{b \in \operatorname{supp}(\mathcal{A})} h_t(b)$ 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

$$\begin{aligned} 0 \leq \frac{d}{d\lambda} H_t(\mu_\lambda)|_{\lambda=0} &= 2\hat{\Delta}_t(\mu^*)(\hat{\Delta}_t(e_a) - \hat{\Delta}_t(\mu^*)) - (I_t(e_a) - I_t(\mu^*)) \\ &= h_t(a) - h_t(\mu^*)\,. \end{aligned}$$

The claim follows after rearranging. □

Lemma 2.5 (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*

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

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

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

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

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 2.3 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

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

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 2.5, we get

$$\Psi_t(\mu_t^{\text{IDS}}) \leq \begin{cases} \frac{4\delta_t(\hat{\Delta}_t(\tilde{\mu})-\delta_t)}{I_t(\tilde{\mu}_t)} & \text{if } 2\delta_t \leq \hat{\Delta}_t(\tilde{\mu}) \\ \frac{\hat{\Delta}_t(\tilde{\mu})^2}{I_t(\tilde{\mu})} & \text{else.} \end{cases} \tag{2.6}$$

Lemma 2.6 (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^{IDS} satisfies*

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

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

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

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

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

Rearranging yields the claim. □

Lemma 2.7 (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 2.4 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 $\Psi^* = \Psi_{23} = \min\{\Psi_{12}, \Psi_{13}, \Psi_{23}\}$ 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, by Lemma 2.2, we can normalize the gaps and information gain such that $\Delta_1 = 1$ and $\tilde{\Psi} = 1$.

We show that $\Psi^{*-1} \leq \frac{4}{3}$. First, we make some calculations with the help of Lemma 2.5. 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 2.5 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 2.6, Ψ_{23} cannot be optimal if $\Delta_2 > 2 = 2\Delta_1$. Finally, note that the bound is tight in the same limit. □

2.1.1 *Computation of the IDS distribution*

Given access to the gap estimates $\hat{\Delta}_t(a)$ and information gain $I_t(a)$ for each action $a \in \mathcal{A}$, we can compute the IDS distribution efficiently on finite action sets of size $k = \|\mathcal{A}|$ [135, §6.2]. The idea is to make use of Lemma 2.4, which shows that the IDS distribution μ_t^{IDS} can always be chosen with a support on two actions. In particular, we can compute the optimal trade-off for all $k(k+1)/2$ pairs of actions using the closed form provided in Lemma 2.5, which leads to an overall computation complexity of $\mathcal{O}(k^2)$. However, the quadratic runtime in the number of actions is a limiting factor for larger action sets. Standard algorithm often use score functions over actions, that can be optimized in $\mathcal{O}(k)$ steps, which is significantly faster.

To improve the sample complexity, we compute the action $\hat{a}_t \in \mathcal{A}$ that minimizes the estimated gaps in $\mathcal{O}(k)$. Then we find the minimal information ratio among distribution that randomize only between $\hat{a}_t$ and some other action $b \in \mathcal{A}$, using $\mathcal{O}(k)$ computation steps in total. Lemma 2.7 guarantees that the ratio obtained this way is at most a factor of $\frac{4}{3}$ worse than the optimal ratio. We refer to this algorithm as *approximate IDS*. Note that the regret bound in Eq. (2.4) and all other bounds introduced in the next section scale directly with the information ratio. Regret bounds for the exact IDS policy based on these results therefore immediately translate to the approximate version.

2.2 GENERAL REGRET BOUNDS

We start by restating the regret bound Eq. (2.2) formally and in slightly generalized form.

Theorem 2.1 (IDS regret). *Let $G \subset [n]$ be a random subset of rounds such that the membership $t \in G$ is predictable on the filtration $\mathcal{F}_t$. Then*

$$\mathbb{E}\left[\sum_{t\in G} \hat{\Delta}_t(a_t)\right] \leq \sqrt{\mathbb{E}\left[\sum_{t\in G} \Psi_t(\mu_t)\right] \mathbb{E}\left[\sum_{t\in G} I_t(a_t)\right]}$$

Proof. The claim follows similar to Eq. (2.2) with the Cauchy-Schwarz inequality, the definition of the information ratio and two applications of the tower rule:

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in G} \hat{\Delta}_t(a_t)\right] &= \mathbb{E}\left[\sum_{t\in G} \hat{\Delta}_t(\mu_t)\right] = \mathbb{E}\left[\sum_{t\in G} \sqrt{\Psi_t I_t(\mu_t)}\right] \\ &\leq \sqrt{\mathbb{E}\left[\sum_{t\in G} \Psi_t(\mu_t)\right] \mathbb{E}\left[\sum_{t\in G} I_t(a_t)\right]}\,. \qquad \square\end{aligned}$$

Let us write down a few immediate consequences of Theorem 2.1 for later use. Note that these bounds hold for any policy $\pi_n = (\mu_t)_{t=1}^n$, whereas IDS is defined as the policy that myopically optimizes the upper bound. Recall that the total information gain is $\gamma_n = \sum_{t=1}^n I_t(a_t)$.

Corollary 2.1. *Assume that $\Psi_t(\mu_t) \leq \alpha_t$ holds for an $\mathcal{F}_t$-predictable sequence $(\alpha_t)_{t=1}^n$ and denote $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha_t$. Then*

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

Proof. Immediate from Theorem 2.1. $\square$

The previous result already reflects the fact that the information ratio is a time-dependent quantity. This plays a crucial role in deriving instance-dependent regret bounds. Denote by $\delta_t = \min_{a\in\mathcal{A}} \hat{\Delta}_t(a)$ the smallest estimated gap. It is reasonable to require that $\hat{\Delta}_t(a)$ is a consistent estimator of the true gaps, which in particular implies that $\delta_t \to 0$. If $2\delta_t \leq \min_{a\neq\hat{a}_t} \hat{\Delta}_t(a)$ a direct calculation with the help of Lemma 2.4 yields

$$\min_{\mu\in\mathscr{P}(\mathcal{A})} \Psi_t(\mu_t) \leq \min_{b\neq\hat{a}_t} \min_{p\in[0,1]} \frac{\hat{\Delta}_t((1-p)e_{\hat{a}_t} + pe_b)^2}{pI_t(b)} \leq 4\delta_t \min_{b\in\mathcal{A}} \frac{\hat{\Delta}_t(b) - \delta_t}{I_t(b)}\,.$$

We can therefore expect that the ratio can be bounded relative to the smallest estimated gap. We make use of this in the next two corollaries.

Corollary 2.2. *Let $B_t = \mathbb{1}\big(\hat{\Delta}_t(a) \geq \Delta(a) : \forall a \in \mathcal{A}\big)$ be the indicator of rounds where all gap estimates are conservative. Denote the smallest estimated gap by $\delta_t = \min_{a \in \mathcal{A}} \hat{\Delta}_t(a)$ and assume that $B_t \Psi_t(\mu_t) \leq 4\delta_t \zeta_t$ holds almost surely for a non-decreasing, $\mathcal{F}_t$-predictable sequence $(\zeta_t)_{t=1}^n$. Then*

$$\mathfrak{R}_n \leq 4\,\mathbb{E}[\zeta_n \gamma_n] + \sum_{t=1}^n \mathbb{E}[(1 - B_t)\Delta(a_t)]\,.$$

Proof. Define re-scaled gap estimates $\tilde{\Delta}_t(a) = B_t \Delta_t(a)$ and information gain $\tilde{I}_t(a) = \zeta_t I_t(a)$. The corresponding information ratio satisfies $\tilde{\Psi}_t \leq 4\delta_t$ by assumption. Using Theorem 2.1 we find

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^n \tilde{\Delta}_t(\mu_t)\right] &\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^n 4 B_t \delta_t\right] \mathbb{E}\left[\sum_{t=1}^n \zeta_t I_t(a_t)\right]} \\ &\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^n 4\tilde{\Delta}(a_t)\right] \mathbb{E}[\zeta_n \gamma_n]}\,,\end{aligned}$$

where the second inequality follows from $\delta_t \leq \tilde{\Delta}_t(a_t)$ and the assumption that ζ_t is non-decreasing. Squaring both sides and solving for the regret yields the claim. □

The next corollary strengthens the previous result by a factor 4 if the estimation errors can be controlled in stronger sense.

Corollary 2.3. *Assume that $\Psi_t \leq 4\delta_t \zeta_t$ holds almost surely for a non-decreasing, $\mathcal{F}_t$-predictable sequence $(\zeta_t)_{t=1}^n$. Then*

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

Proof. Using that $4xw \leq (x + w)^2$ for $x, w \in \mathbb{R}$, we find

$$\mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(a_t) - \delta_t\right] \leq \frac{1}{4}\mathbb{E}\left[\sum_{t=1}^n \delta_t\right]^{-1} \mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(a_t)\right]^2.$$

Applying Theorem 2.1 and the assumption $\Psi_t \leq 4\delta_t \zeta_t$, we get

$$\mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(a_t) - \delta_t\right] \leq \frac{1}{4}\mathbb{E}\left[\sum_{t=1}^n \delta_t\right]^{-1} \mathbb{E}\left[\sum_{t=1}^n 4\delta_t \zeta_n\right] \mathbb{E}[\gamma_n] = \zeta_n \mathbb{E}[\gamma_n]\,.$$

The claim follows. □

2.2.1 *Generalized Information Ratio*

The regret bound of Corollary 2.1 is at most $\mathcal{O}(\sqrt{n\mathbb{E}[\gamma_n]})$, provided that the estimation errors of the gaps can be controlled. For the information gain functions that we use, the total information gain typically contributes only logarithmically. In partial monitoring, however, the feedback can be such that the regret of the learner is at least $\Omega(n^{2/3})$ [16]. This suggests that one should use a different exponent in the information ratio and replace the Cauchy-Schwarz inequality with Hölder's inequality. To this end, Lattimore & György [102] introduced the *generalized information ratio*,

$$\Psi_{\lambda,t}(\mu) \triangleq \frac{\hat{\Delta}_t(\mu)^\lambda}{I_t(\mu)}\,. \tag{2.7}$$

It is straightforward to adapt the previous result to get the a regret rate of order $n^{1-1/\lambda}$ if the learner minimizes $\Psi_{\lambda,t}(\mu)$. That means, however, that in models where the achievable rate depends on the instance, we need to manually change the algorithm. Perhaps surprisingly, the following lemma shows that the IDS distribution obtained as a minimizer of $\Psi_{2,t}(\mu)$ remains close to a minimizer of $\Psi_{\lambda,t}(\mu)$ for $\lambda \geq 2$.

Lemma 2.8 (Lattimore & György [102, Lemma 21])**.** *Let μ_t^{IDS} be the IDS distribution $\mu_t^{\mathrm{IDS}} = \arg\min_{\mu\in\mathscr{P}(\mathcal{A})} \Psi_{2,t}(\mu)$. Then for all $\lambda \geq 2$,*

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

The next lemma generalizes the regret bound using the generalized information ratio. A similar bound on the Bayesian regret is given by Lattimore & György [102, Theorem 4].

Theorem 2.2. *Assume that $\Psi_{\lambda,t}(\mu_t) \leq \alpha_t$ holds almost surely for a $\mathcal{F}_t$-predictable sequence $(\alpha_t)_{t=1}^n$, and let $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha_t$. Then*

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

Proof. The claim follows along the lines of Theorem 2.1 and Corollary 2.2 and by using Hölder's inequality instead of Cauchy-Schwarz. In particular,

$$\mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^n (\Psi_{\lambda,t} I_t(\mu_t))^{1/\lambda}\right]$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \mathbb{E}\left[\sum_{t=1}^{n} \Psi_t^{\frac{1}{\lambda-1}}\right]^{1-\frac{1}{\lambda}} \mathbb{E}\left[\sum_{t=1}^{n} I_t(a_t)\right]^{\frac{1}{\lambda}} \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\sum_{t=1}^{n} \alpha_t^{\frac{1}{\lambda-1}}\right]^{1-\frac{1}{\lambda}} \mathbb{E}[\gamma_n]^{\frac{1}{\lambda}} \\
&\overset{(iii)}{\leq} \mathbb{E}\left[\Big(\sum_{t=1}^{n} \alpha_t\Big)^{\frac{1}{\lambda-1}} n^{\frac{\lambda-2}{\lambda-1}}\right]^{1-\frac{1}{\lambda}} \mathbb{E}[\gamma_n]^{\frac{1}{\lambda}} \\
&\overset{(iv)}{\leq} n^{\frac{\lambda}{\lambda-1}} \mathbb{E}[\bar{\alpha}_n]^{\frac{1}{\lambda}} \mathbb{E}[\gamma_n]^{\frac{1}{\lambda}} .
\end{aligned}
$$

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 $\bar{\alpha}_n$. □

2.2.2 *High-Probability Bounds*

Our definition of the regret $\mathfrak{R}_n$ includes the expectation over the randomness of the policy and the environment, and therefore provides no control on the tails of regret distribution. A remedy is to define the *pseudo regret*,

$$
\mathfrak{PR}_n \triangleq \sum_{t=1}^{n} \Delta(a_t) . \tag{2.8}
$$

In the frequentist, non-asymptotic analysis of algorithms for the linear bandit setting, high-probability bounds on the pseudo regret seem to be prevalent [3, 5, 7, 46]. One reason is that in practice the confidence level provides a convenient tuning parameter. It was also noted in [103, §9.2] that optimizing *just* the expected regret can lead to high variance of the regret distribution. Of course, high-probability bounds can be integrated to a bound in expectation.

In our case, statements on the expected regret are slightly simpler while preserving the main ideas. For completeness we also provide bounds on the pseudo regret but restrict ourselves to one basic result. The first observation is that for deterministic policies, the pseudo regret is bounded without extra work. Assume that the sampling distribution μ_t at time t is a Dirac on some action a_t. Using the Cauchy-Schwarz inequality, we find

$$
\sum_{t=1}^{n} \hat{\Delta}_t(a_t) = \sum_{t=1}^{n} \sqrt{\Psi_t(a_t) I_t(a_t)} \leq \sqrt{\sum_{t=1}^{n} \Psi_t(a_t) \sum_{t=1}^{n} I_t(a_t)} . \tag{2.9}
$$

Let us assume for simplicity that the gap estimates are chosen as upper bound to the true gaps, $\Delta(a_t) \leq \hat{\Delta}_t(a_t)$, for all $t \in [n]$ with probability $1-\delta$. Then the pseudo regret of any deterministic strategy with bounded information ratio $\Psi_t(a_t) \leq \alpha_0$ satisfies with probability $1-\delta$,

$$\mathfrak{PR}_n \leq \sqrt{\alpha_0 \gamma_n n}\,.$$

To obtain a meaningful bound, we still needs to control the total information gain $\gamma_n = \sum_{t=1}^n I_t(a_t)$, which is a random quantity. For the choices of the information gain considered in this thesis, we prove almost-sure worst-case bounds on γ_n that immediately lead to concrete bounds. In fact, Eq. (2.9) is essentially the same argument used in the analysis of UCB [46].

A similar high-probability regret bound for randomized policies such as IDS further needs to account for the randomness from sampling the actions from the distribution μ_t. The next results provides a bound with the same scaling. This assures that the randomness of the policy does not significance impact the tails of regret distribution. It is also important to note that the bound only appears in the analysis and does not affect the algorithm design.

Theorem 2.3 (IDS regret – high-probability bound)**.** *Let $\delta \in [0,1]$. Assume that $\Delta(a_t) \leq \hat{\Delta}_t(a_t)$ holds with probability at least $1-\frac{\delta}{2}$. Further assume that $\hat{\Delta}(a_t) \leq B$, $I_t(a_t) \leq J$ and $\Psi_t(\mu_t) \leq \alpha_0$ holds almost surely for all $t \in [n]$ and $J \geq 1$. Then, with probability at least $1-\delta$,*

$$\mathfrak{PR}_n \leq 2.5\sqrt{\alpha_0 n\big(\gamma_n + \mathcal{O}\big(J \log \tfrac{J}{\delta}\big)\big)} + \mathcal{O}\big(B \log \tfrac{n}{\delta}\big)$$

Proof. We start be expanding the estimated regret and using Cauchy-Schwarz.

$$\sum_{t=1}^n \hat{\Delta}_t(a_t) = \sum_{t=1}^n \hat{\Delta}_t(\mu_t) + \sum_{t=1}^n \big(\hat{\Delta}_t(a_t) - \hat{\Delta}_t(\mu_t)\big)\,.$$

The second sum is a martingale difference sequence. Denote the sum over the conditional variances by $V_n = \sum_{t=1}^n \mathbb{E}\Big[\big(\hat{\Delta}_t(a_t) - \hat{\Delta}_t(\mu_t)\big)^2\Big]$. Freedman's

inequality combined with a peeling argument (Lemma A.1, Appendix A) shows

$$\begin{aligned}\sum_{t=1}^{n}\big(\hat{\Delta}_t(a_t)-\hat{\Delta}_t(\mu_t)\big) &\leq \max\left\{4B\log\frac{2n+2}{\delta}, 2\sqrt{V_n\log\frac{2n+2}{\delta}}\right\}\\ &\overset{(i)}{\leq} \max\left\{4B\log\frac{2n+2}{\delta}, 2\sqrt{\sum_{t=1}^{n}B\hat{\Delta}_t(\mu_t)\log\frac{2n+2}{\delta}}\right\}\\ &\overset{(ii)}{\leq} 4B\log\frac{2n+2}{\delta}+\frac{1}{4}\sum_{t=1}^{n}\hat{\Delta}_t(\mu_t)\end{aligned}$$

The last step (ii) uses $2\sqrt{xw}\leq x+w$ for $x,w\geq 0$. Step (i) follows with the Bhatia-Davis inequality (Lemma A.4, Appendix A) and $\hat{\Delta}_t(a_t)\in[0,B]$,

$$\mathbb{E}\Big[\big(\hat{\Delta}_t(a_t)-\hat{\Delta}_t(\mu_t)\big)^2\Big]\leq\big(B-\hat{\Delta}_t(\mu_t)\big)\hat{\Delta}_t(\mu_t)\leq B\hat{\Delta}_t(\mu_t)\,.$$

We continue to bound the estimated regret,

$$\sum_{t=1}^{n}\hat{\Delta}_t(a_t)\leq\frac{5}{4}\sum_{t=1}^{n}\hat{\Delta}_t(\mu_t)\leq\frac{5}{4}\sqrt{\sum_{t=1}^{n}\Psi_t(\mu_t)\sum_{t=1}^{n}I_t(\mu_t)}\,.$$

It remains to bound another martingale difference sequence defined by the information gain. Using that $I_t(a_t)\in[0,J]$, Lemma A.3 in Appendix A leads to the following bound:

$$\sum_{t=1}^{n}I_t(\mu_t)\leq 2\sum_{t=1}^{n}I_t(a_t)+\mathcal{O}\Big(J\log\big(\tfrac{J}{\delta}\big)\Big)\,.$$

The claim follows with a union bound over all events such that the inequalities hold simultaneously. □

2.3 BAYESIAN INFORMATION-DIRECTED SAMPLING

Information-directed sampling was first introduced by Russo & Van Roy [135] and analyzed for Bayesian regret, which is formally defined below. We refer to this algorithm as *Bayesian IDS* to distinguish it from other variants that we introduce later in this thesis. The Bayesian setting uses a prior distribution over the parameters, which allows to define information theoretic concepts such of entropy and mutual information. The regret analysis is based on established tools from information theory and requires only few assumptions.

Assume that the parameter space $\mathcal{M}$ is associated with a sigma algebra and the learner is equipped with a prior distribution $\mathscr{F} \in \mathscr{P}(\mathcal{M})$. The *Bayesian regret* is defined in expectation over the prior,

$$\mathfrak{BR}_n(\pi, \mathscr{F}) \triangleq \mathbb{E}_{\theta\sim\mathscr{F}}[\mathfrak{R}_n(\pi, \theta)] = \mathbb{E}\left[\sum_{t=1}^{n} f_\theta(a^*) - f_\theta(a_t)\right]. \tag{2.10}$$

For the rest of the section, we assume that $k = |\mathcal{A}|$ is finite. Note that in the Bayesian interpretation, the optimal action $a^* = a^*(\theta)$ is random, since it depends on the realization of $\theta \sim \mathscr{F}$. Conditioned on the history, we define the probability that an action $a \in \mathcal{A}$ is optimal under the posterior as

$$q_t(a) \triangleq \mathbb{P}_t(a = a^*).$$

The Shannon entropy of a^* under the posterior distribution is

$$\mathbb{H}_t(a^*) \triangleq -\sum_{a\in\mathcal{A}} q_t(a) \log q_t(a).$$

The conditional entropy of a^* is defined using the conditional probabilities $q_t(a|a_t, y_t) \triangleq \mathbb{P}_t(a = a^*|a_t, y_t)$ as follows,

$$\mathbb{H}_t(a^*|a_t, y_t) \triangleq \mathbb{E}_t\left[-\sum_{a\in\mathcal{A}} q_t(a|a_t, y_t) \log q_t(a|a_t, y_t)\right],$$

where the conditional expectation is over the random outcome of a_t and y_t. The *mutual information* between the observation y_t and the optimal action a^* is defined as the entropy reduction

$$\mathbb{I}_t(a^*, y_t|a_t = a) \triangleq \mathbb{H}_t(a^*) - \mathbb{H}_t(a^*|y_t, a_t = a).$$

The Bayesian learner has the advantage of knowing the posterior distribution of a^*, which is used to define gap estimates and information gain,

$$\hat{\Delta}_t(a) = \mathbb{E}_t[\Delta(a^*) - \Delta(a)] \quad \text{and} \quad I_t^{\mathrm{MI}}(a) \triangleq \mathbb{I}_t(a^*; y_t|a_t = a). \tag{2.11}$$

Bayesian IDS is the policy that optimizes the information ratio in Eq. (2.1) defined with the quantities in Eq. (2.11). The analogue result to Theorem 2.1 is presented below. Note that Bayesian IDS myopically optimizes the bound.

Theorem 2.4 (Bayesian regret [135, Prop. 1]). *Let $\mathcal{A}$ be finite and $\pi_n = (\mu_t)_{t=1}^n$ be any policy. Then the Bayesian regret satisfies*

$$\mathfrak{BR}_n(\pi, \mathscr{F}) \le \sqrt{\bar{\alpha}_n \mathbb{H}(a^*) n},$$

where $\bar{\alpha}_n \triangleq \frac{1}{n}\sum_{t=1}^n \mathbb{E}[\Psi_t(\mu_t)]$ is the average expected information ratio.

Worst-case bounds on the Bayesian information ratio have been derived by Russo & Van Roy [135] for a variety of settings. Without making further assumptions other than bandit feedback $y_t = f(a_t) + \epsilon_t$ with independent noise, the information ratio defined by Eq. (2.11) satisfies almost surely

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) \leq \Psi_t(\mu_t^{\mathrm{TS}}) \leq \frac{|\mathcal{A}|}{2}\,, \tag{2.12}$$

where $\mu_t^{\mathrm{TS}}(a) \triangleq q_t(a)$ is the Thompson sampling policy π_n^{TS}, see [137, Prop. 3]. The entropy satisfies $\mathbb{H}(a^*) \leq \log(|\mathcal{A}|)$. Combining Theorem 2.4 and Eq. (2.12), we find that Bayesian regret of IDS satisfies

$$\mathfrak{BR}_n(\pi_n^{\mathrm{IDS}}, \mathscr{F}) \leq \sqrt{\tfrac{1}{2}|\mathcal{A}| \log(|\mathcal{A}|) n}\,.$$

Note that the argument implies the same bound for Thompson sampling. The bound is order optimal for general prior distributions up to the logarithmic factor [30]. The proof of the result is strikingly simple. It only relies on telescoping the information gain and bounding the information ratio for the Thompson sampling distribution using Pinsker's inequality. Tighter bounds on the information ratio are known for various settings, see [137] and there are also examples where the bound of IDS is significantly better than the bound for Thompson sampling [135].

2.4 CONTRIBUTIONS AND RELATED WORK

The foundations of information-directed sampling (IDS) are by Russo & Van Roy [135], who introduced the framework in the Bayesian bandit setting. The idea emerged from an elegant information-theoretic analysis of Thompson sampling by the same authors [137]. The central argument using the Cauchy-Schwarz inequality to decompose the cumulative regret appears already in the analysis of UCB by Dani, Hayes & Kakade [46].

Unless otherwise stated, the results presented in this chapter are based on work by the author and collaborators:

- Kirschner, J. & Krause, A. *Information Directed Sampling and Bandits with Heteroscedastic Noise* in *Proc. International Conference on Learning Theory (COLT)* (July 2018)

- Kirschner, J., Lattimore, T. & Krause, A. *Information Directed Sampling for Linear Partial Monitoring* in *Proc. International Conference on Learning Theory (COLT)* (July 2020)

- Kirschner, J., Lattimore, T., Vernade, C. & Szepesvári, C. *Asymptotically Optimal Information-Directed Sampling* in *Proc. International Conference on Learning Theory (COLT)* (Aug. 2021)

The generalized information ratio and regret bound is by Lattimore & György [102], introduced here for the sake of cleaner proofs. An earlier version of the regret bound in Theorem 2.2 for the special case $\lambda = 3$ is by the author and collaborators, [90, Theorem 3]. Lemma 2.7 is a novel result.

Beyond the application in bandits, the information ratio has been proven useful in other applications. Most notably, an extension of Sion's minimax theorem shows that the adversarial worst-case regret equals the minimax Bayesian optimal regret over all prior distributions. By deriving Bayesian algorithms with prior-independent bounds on information ratio, this argument leads to non-constructive regret bounds on the adversarial worst-case regret, see Bubeck & Eldan [29], Lattimore [101], and Lattimore & Szepesvári [105]. The information-ratio is also related to the stability term in follow-the-perturbed-leader (FTRL) and mirror descent, see Lattimore & György [102].

Lattimore & Szepesvári [105] analyze the information ratio with general divergences. An information gain based on the Tsallis-entropy is studied by Kalkanlı & Özgür [82]. An information-theoretic analysis of Thompson sampling on large action sets with linear reward functions is by Dong & Roy [52], and the result was recently extended to non-Gaussian distributions by Hamidi & Bayati [71].

3

HETEROSCEDASTIC LINEAR BANDITS

We now introduce the stochastic linear bandit setting. We start by reviewing confidence bounds for linear least-squares estimation. The caveat is that bandit algorithms collect data adaptively, hence results from i.i.d. estimation do not apply directly. Instead, we use a *online confidence set* that makes no assumption on how the sequence of actions is generated. Variants of this result will be used in most chapters that follow. Using the concentration result, we construct the gap estimates in a way such that the estimation error is controlled.

The frequentist information gain is defined by the log-determinant potential, which can be understood as measuring the volume reduction of the confidence ellipsoid. This leads us to a first *frequentist version* of information-directed sampling for stochastic linear bandits. Using the results from Chapter 2, we then derive bounds on the regret that match the best known bounds for UCB in this setting.

Beyond the standard assumptions, we address the case of *heteroscedastic noise*, where the variance of the observation noise depends on the chosen action. In the more general noise model, we show that IDS outperforms UCB and Thompson sampling on some instances by an arbitrarily large factor. This illustrates a limitation of optimistic approaches and Thompson sampling, which do use a measure of *informativeness* for their action choice.

SETTING In the linear bandit setting, we identify the actions with d-dimensional features $\mathcal{A} \subset \mathbb{R}^d$. The reward function is defined by a parameter $\theta \in \mathcal{M} \subset \mathbb{R}^d$ such that $f_\theta(a) = \langle a, \theta \rangle$. We make the following boundedness assumptions: The true parameter is bounded with $\|\theta\| \leq B$, and the actions are bounded with $\|a\| \leq L$ for all $a \in \mathcal{A}$. When the learner chooses an action $a_t \in \mathcal{A}$ at time t, the feedback is $y_t = \langle a_t, \theta \rangle + \epsilon_t$. We model heteroscedasticity by assuming that the noise ϵ_t is conditionally $\rho(a_t)^2$-sub-Gaussian for a fixed noise function $\rho : \mathcal{A} \to \mathbb{R}_{\geq 0}$ as defined in Eq. (1.2). Note that we use the frequentist bandit framework, where the reward function is fixed in advance and regret is defined as in Eq. (1.3).

3.1 ONLINE LINEAR LEAST-SQUARES

Most bandit algorithms rely on estimating the reward function in one way or another. In the linear setting with sub-Gaussian observation noise, a least-squares estimator for the unknown parameter is justified. Given observations $(a_s, y_s)_{s=1}^{t-1}$ from rounds $1, \dots, t-1$ and a regularizer $\lambda > 0$, the *regularized linear least-squares estimator* is

$$\hat{\theta}_t^{ls} \triangleq \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \left(\langle a_s, \theta \rangle - y_s \right)^2 + \lambda \|\theta\|^2 .$$

An analytic closed form is easily computed as

$$\hat{\theta}_t^{ls} = V_t^{-1} \sum_{s=1}^{t-1} a_s y_s , \quad \text{where} \quad V_t = \sum_{s=1}^{t-1} a_s a_s^\top + \lambda \mathbf{1}_d .$$

We use the convention that $V_1 = \lambda \mathbf{1}_d$. There is an abundant amount of work that studies least-squares estimation. In the bandit setting, care is required since the data is generated adaptively by the learner. In particular, the action a_t chosen at time t depends on the history $(a_s, y_s)_{s=1}^{t-1}$, and we cannot rely on results that require independent observations. The next lemma provides a *self-normalized concentration inequality* for the regularized least squares estimator that holds for adaptive data.

Lemma 3.1 (Abbasi-Yadkori, Pál & Szepesvári [3, Theorem 2]). *Let $(a_t)_{t=1}^{\infty}$ be a $\mathcal{F}_t$-adapted sequence in $\mathcal{A}$ with corresponding observations $y_t = \langle a_t, \theta \rangle + \epsilon_t$, where $\|\theta\| \le B$ and ϵ_t is conditionally ρ-sub-Gaussian, i. e.*

$$\forall \eta \in \mathbb{R}, \quad \mathbb{E}_t[\exp(\eta \epsilon_t) | a_t] \le \exp(\eta^2 \rho^2 / 2) .$$

Let $\mathcal{E}_{t,\delta}^{ls} \triangleq \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t^{ls}\|_{V_t}^2 \le \beta_{t,\delta}^{ls}\}$ be the confidence ellipsoid with confidence coefficient $\beta_{t,\delta}^{ls} = \left(\rho \sqrt{2 \log \frac{1}{\delta} + \log \left(\frac{\det(V_t)}{\det(V_1)} \right)} + \sqrt{\lambda} B \right)^2$. Then

$$\mathbb{P}\left[\forall t \ge 1, \theta \in \mathcal{E}_{t,\delta}^{ls} \right] \ge 1 - \delta .$$

Note, the confidence coefficient $\beta_{t,\delta}^{ls}$ is $\mathcal{F}_t$-predictable and can be directly used in the algorithm. For the analysis, it is often useful to derive upper bounds with an explicit dependence on problem parameters such as the dimension. The following bound follows from a simple calculation that relates the determinant and the trace [104, Lemma 19.4]:

$$\log\left(\frac{\det V_n}{\det V_1} \right) \le d \log\left(1 + \frac{n L^2}{d \lambda} \right) . \tag{3.1}$$

3.1.1 *Heteroscedastic Least-Squares*

When the observation noise is heteroscedastic, intuitively we expect that nosier observations carry less information and therefore should receive less weight in the estimation. The generalized Gauss-Markov theorem [8] suggests that the observation should receive a weight inversely proportional to the noise variance. The *weighted regularized least-squared estimator* is,

$$\hat{\theta}_t^{wls} \triangleq \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \frac{1}{\rho(a_s)^2} \big(\langle a_s, \theta \rangle - y_s\big)^2 + \lambda \|\theta\|^2 \,. \tag{3.2}$$

An analytic closed form is again readily computed as

$$\hat{\theta}_t^{wls} = W_t^{-1} \sum_{s=1}^{t-1} \frac{1}{\rho(a_s)^2} a_s y_s \,, \quad \text{where} \quad W_t = \sum_{s=1}^{t-1} \frac{1}{\rho(a_s)^2} a_s a_s^\top \,.$$

The next result extends Lemma 3.1 to the weighted least-squares estimator.

Lemma 3.2. *Let $(a_t)_{t=1}^\infty$ be a $\mathcal{F}_t$-adapted sequence in $\mathcal{A}$ with corresponding observations $y_t = \langle a_t, \theta \rangle + \epsilon_t$, where $\|\theta\| \le B$ and ϵ_t is conditionally $\rho(a_t)^2$-sub-Gaussian, i.e.*

$$\forall \eta \in \mathbb{R}, \quad \mathbb{E}_t[\exp(\eta \epsilon_t) | a_t] \le \exp(\eta^2 \rho(a_t)^2 / 2) \,.$$

Let $\mathcal{E}_{t,\delta}^{wls} \triangleq \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t^{wls}\|_{W_t}^2 \le \beta_{t,\delta}^{wls}\}$ be the confidence ellipsoid with weighted precision matrix W_t and $\beta_{t,\delta}^{wls} = \left(\sqrt{2 \log \frac{1}{\delta} + \log\big(\frac{\det(W_t)}{\det(W_1)}\big)} + \sqrt{\lambda} B\right)^2$. Then

$$\mathbb{P}\Big[\forall t \ge 1, \theta \in \mathcal{E}_{t,\delta}^{wls}\Big] \ge 1 - \delta \,.$$

Proof. Let $\tilde{a}_t = a_t / \rho(a_t)$ and $\tilde{y}_t = y_t / \rho(a_t)$. Note that $\tilde{y}_t = \langle \tilde{a}_t, \theta \rangle + \tilde{\epsilon}_t$, where $\tilde{\epsilon}_t = \epsilon_t / \rho(a_t)$ is 1-sub-Gaussian noise. The result follows as direct consequence of Lemma 3.1 applied to the sequence $(\tilde{a}_t, \tilde{y}_t)_{t=1}^\infty$. □

3.2 IDS FOR HETEROSCEDASTIC LINEAR BANDITS

For the remainder of this chapter, we settle on the weighted least-squares estimator defined in Eq. (3.2) and the confidence set of Lemma 3.2. To define information-directed sampling, we need an estimate $\hat{\Delta}_t(a)$ of each gap $\Delta(a) = \max_{b \in \mathcal{A}} \langle b - a, \theta \rangle$ and an information gain function $I_t(a)$. These

Algorithm 2: IDS for Heteroscedastic Linear Bandits

Input: Action set $\mathcal{A}$, regularizer $\lambda > 0$, norm bound B, noise function $\rho : \mathcal{A} \to \mathbb{R}_{\geq 0}$

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

$W_t \leftarrow \sum_{s=1}^{t-1} \rho(a_s)^{-2} a_s a_s^\top + \lambda \mathbf{1}_d$ `// least-squares estimation`

$\hat{\theta}_t^{wls} \leftarrow W_t^{-1} \sum_{s=1}^{t-1} \rho(a_s)^{-2} a_s y_s$

$\beta_t^{1/2} \leftarrow \sqrt{2\log(t^2) + \log\left(\frac{\det(W_t)}{\det(W_1)}\right)} + \sqrt{\lambda} B$

`// gap estimates`

$\hat{\Delta}_t(a) \leftarrow \max_{b \in \mathcal{A}} \langle b, \hat{\theta}_t^{wls} \rangle + \beta_t^{1/2} \|b\|_{W_t^{-1}} - \left(\langle a, \hat{\theta}_t^{wls} \rangle - \beta_t^{1/2} \|a\|_{W_t^{-1}} \right)$

$I_t(a) \leftarrow \frac{1}{2} \log\left(1 + \rho(a)^{-2} \|a\|^2_{W_t^{-1}}\right)$ `// information gain`

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

$a_t \sim \mu_t$

Choose a_t, observe $y_t \sim \langle a_t, \theta \rangle + \epsilon_t$

are introduced in the next two sections. The resulting version of IDS is summarized in Algorithm 2. We discuss implementation and computational complexity in Section 3.2.3, and present a bound on the worst-case regret that matches the best known bounds for UCB on homoscedastic noise in Section 3.2.4. In Section 3.2.5 we argue that on some instances with heteroscedastic noise, upper confidence bound algorithms and Thompson sampling perform arbitrarily worse than IDS. Variants of the algorithm and the analysis are collected in Section 3.3, including a gap-dependent logarithmic regret bound.

3.2.1 *Gap Estimates*

Anticipating the use of the IDS regret bound in Corollary 2.1, we need to make sure that the expected sum of estimation errors $\mathbb{E}\left[\sum_{t=1}^n \Delta(a_t) - \hat{\Delta}_t(a_t)\right]$ is bounded. This motivates a conservative choice of the gap estimate, which we define as

$$\hat{\Delta}_{t,\delta}(a) \triangleq \max_{b \in \mathcal{A}} \langle b, \hat{\theta}_t^{wls} \rangle + \beta_{t,\delta}^{1/2} \|b\|_{W_t^{-1}} - \left(\langle a, \hat{\theta}_t^{wls} \rangle - \beta_{t,\delta}^{1/2} \|a\|_{W_t^{-1}} \right). \quad (3.3)$$

Note, $\hat{\Delta}_{t,\delta}(a)$ is chosen as high-probability upper bound to the true gap $\Delta(a)$. Specifically, if the parameter estimate $\hat{\theta}_t^{wls}$ is well concentrated such that $\theta \in \mathcal{E}_{t,\delta}^{wls}$, we have

$$\begin{aligned}\Delta(a) &= \max_{b\in\mathcal{A}} \langle b-a, \hat{\theta}_t^{wls}\rangle + \langle b-a, \theta - \hat{\theta}_t^{wls}\rangle \\ &\le \max_{b\in\mathcal{A}} \langle b-a, \hat{\theta}_t^{wls}\rangle + \|b-a\|_{W_t^{-1}} \|\theta - \hat{\theta}_t^{wls}\|_{W_t} \\ &\le \max_{b\in\mathcal{A}} \langle b-a, \hat{\theta}_t^{wls}\rangle + \big(\|b\| + \|a\|_{W_t^{-1}}\big)\beta_{t,\delta}^{1/2} = \hat{\Delta}_t(a)\,. \end{aligned} \tag{3.4}$$

The first inequality is by Cauchy-Schwarz, and the second inequality uses that $\theta \in \mathcal{E}_{t,\delta}^{wls}$ and the triangle inequality. For the bound on the expected regret, we choose $\delta = 1/t^2$ and define $\hat{\Delta}_t(a) \triangleq \hat{\Delta}_{t,1/t^2}(a)$. Assume that the gaps are bounded, $\max_{a\in\mathcal{A}} \Delta(a) \le \Delta_{\max}$. By our boundedness assumption, we always have $\Delta_{\max} \le 2LB$. Using Eq. (3.4), we find

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^n \Delta(a_t) - \hat{\Delta}_t(a_t)\right] &\le \Delta_{\max} \sum_{t=1}^n \mathbb{P}[\Delta(a_t) > \hat{\Delta}_t(a_t)] \\ &\le \Delta_{\max} \sum_{t=1}^n \frac{1}{t^2} \le \mathcal{O}(\Delta_{\max})\,. \end{aligned} \tag{3.5}$$

3.2.2 *Information Gain*

Theorem 2.1 bounds the cumulative regret in terms of the information ratio and the total information gain. For the IDS policy $\pi_t = (\mu_t^{\mathrm{IDS}})$, we get

$$\mathbb{E}\left[\sum_{t=1}^n \hat{\Delta}_t(\mu_t^{\mathrm{IDS}})\right] \le \sqrt{\sum_{t=1}^n \min_{\mu\in\mathscr{P}(\mathcal{A})} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \sum_{t=1}^n I_t(a_t)}\,.$$

To obtain a meaningful bound, we need to choose the information gain such that the information ratio is not too large, and at the same time, the total information gain grows only slowly. The following choice is inspired from the UCB analysis [3, 46] and satisfies both criteria:

$$I_t^{\mathrm{DET}}(a) \triangleq \frac{1}{2}\log\Big(1 + \rho(a)^{-2}\|a\|_{W_t^{-1}}^2\Big)\,. \tag{3.6}$$

Note that the information gain scales naturally with the sub-Gaussian noise variance in a way that actions with noisier observations are less informative. With the help of the matrix determinant lemma (Lemma D.1),

we can rewrite the information gain as $I_t^{\mathrm{DET}}(a) = \frac{1}{2}\log(\det(W_{t+1})) - \frac{1}{2}\log(\det(W_t))$. Therefore, the total information gain is

$$\gamma_n = \sum_{t=1}^{n} I_t^{\mathrm{DET}}(a_t) = \frac{1}{2}\log\left(\frac{\det(W_{n+1})}{\det(W_1)}\right). \tag{3.7}$$

Note that the log-determinant ratio also appears in the confidence set and we can use Eq. (3.1) to derive an upper bound on γ_n that does not depend on the sequence of actions.

The information gain in Eq. (3.6) also has a Bayesian interpretation. For a Gaussian prior $\mathcal{N}(0, \lambda \mathbf{1}_d)$ on the true parameter θ and a Gaussian likelihood $y_t \sim \mathcal{N}(\langle a_t, \theta\rangle, \rho(a_t)^2)$, the posterior distribution of θ at time t is $\mathcal{N}(\hat{\theta}^{wls}, W_t)$. The mutual information between the parameter and the outcome is exactly

$$\begin{aligned}\mathbb{I}_t(\theta; y_t | a_t = a) &= \frac{1}{2}\log\left(\frac{\det\left(W_t + \rho(a)^{-2} a a^\top\right)}{\det(W_t)}\right) \\ &= \frac{1}{2}\log\left(1 + \rho(a)^{-2}\|a\|^2_{W_t^{-1}}\right) = I_t^{\mathrm{DET}}(a)\,.\end{aligned}$$

The same mutual information was already considered as a variant in the Bayesian IDS framework by Russo & Van Roy [135, Section 9.2].

3.2.3 *Computational Complexity*

Algorithm 2 is efficient for finite actions sets of size k. In each round, all quantities related to the least-squares estimator are calculated incrementally (Lemma D.2) using basic linear algebra operations in $\mathcal{O}(d^2)$. The gap estimates and the information gain are computed for each action using $\mathcal{O}(d^2 k)$ operations. Note that the maximum that appears in the definition of the gap estimate is attained by the UCB action and can be computed at the beginning of each round. Last, we need to compute the IDS distribution and sample from it. As discussed in Section 2.1.1, in general, this can be done in $\mathcal{O}(k^2)$ steps. Using the approximate version of IDS, the computation complexity is reduced to $\mathcal{O}(k)$ while preserving the guarantees up to a constant factor. In this case the overall complexity over n rounds is $\mathcal{O}(d^2 k n)$, which matches the complexity of UCB.

The same reasoning also applies when $\mathcal{A}$ is a polytope with k extreme points. For general compact action sets, the theory continues to hold but computation is more difficult. The proposed version of IDS requires to

compute the UCB action, which is known to be intractable for the elliptical confidence set without further approximations [46]. In addition, we need to minimize the information ratio. It is an open problem to determine if this step can be solved efficiently for a more general class of action sets, for example, using the fact that the information ratio is a convex function of the distribution (Lemma 2.3).

3.2.4 *Worst-Case Regret*

If we combine Corollary 2.1 and use Eq. (3.5) to bound the estimation error, we arrive at the following inequality:

$$\mathfrak{R}_n(\pi_n^{\mathrm{IDS}}, \theta) \leq \sqrt{\mathbb{E}\left[\sum_{t=1}^{n} \Psi_t(\mu_t^{\mathrm{IDS}})\right] \mathbb{E}[\gamma_n]} + \mathcal{O}(\Delta_{\max}) \tag{3.8}$$

To bound the information ratio, define $\rho_{\min} = \min_{a \in \mathcal{A}} \rho(a)$ and recall that $\|a\| \leq L$. In particular, $\|a\|_{W_t^{-1}}^2 \leq \|a\|_{W_0^{-1}}^2 \leq \frac{L^2}{\lambda}$. Using that $\log(1+x) \geq \frac{x}{2w}$ for all $w \geq 1$ and $x \in [0, w]$, we find that

$$\begin{aligned} I_t^{\mathrm{DET}}(a) &= \frac{1}{2} \log\Big(1 + \rho(a)^{-2} \|a\|_{W_t^{-1}}^2\Big) \\ &\geq \frac{1}{4} \min\Big(\lambda L^{-2}, \rho(a)^{-2}\Big) \|a\|_{W_t^{-1}}^2 . \end{aligned} \tag{3.9}$$

In the following, we abbreviate $\beta_t = \beta_{t,1/t^2}$ and we define the UCB action $a_t^{\mathrm{UCB}} \triangleq \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_t^{wls} \rangle + \beta_t^{1/2} \|a\|_{W^{-1}}$. The gap estimate of the UCB action is

$$\begin{aligned} \hat{\Delta}_t(a_t^{\mathrm{UCB}}) &= \max_{b \in \mathcal{A}} \langle b, \hat{\theta}_t^{wls} \rangle + \beta_t^{1/2} \|b\|_{W_t^{-1}} - \big(\langle a_t^{\mathrm{UCB}}, \hat{\theta}_t^{wls} \rangle - \beta_t^{1/2} \|a_t^{\mathrm{UCB}}\|_{W_t^{-1}}\big) \\ &= 2\beta_t^{1/2} \|a_t^{\mathrm{UCB}}\|_{W_t^{-1}} . \end{aligned}$$

With the last two displays combined, we arrive at the following bound on the information ratio:

$$\begin{aligned} \Psi_t(\mu_t^{\mathrm{IDS}}) = \min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) &\leq \frac{\hat{\Delta}_t(a_t^{\mathrm{UCB}})^2}{I_t^{\mathrm{DET}}(a_t^{\mathrm{UCB}})} \\ &\leq 16 \beta_t \max\Big(L^2/\lambda, \rho(a_t^{\mathrm{UCB}})^2\Big) . \end{aligned} \tag{3.10}$$

We complete the bound on the regret using Eq. (3.8):

$$\begin{aligned}\mathfrak{R}_n(\pi_n^{\mathrm{IDS}}, \theta) &\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^{n} \Psi_t(\mu_t^{\mathrm{IDS}})\right] \mathbb{E}[\gamma_n]} + \mathcal{O}(\Delta_{\max}) \\ &\leq 4\sqrt{\mathbb{E}\left[\sum_{t=1}^{n} \beta_t \max\left(L^2/\lambda, \rho(a_t^{\mathrm{UCB}})^2\right)\right] \mathbb{E}[\gamma_n]} + \mathcal{O}(\Delta_{\max}) \\ &\leq 4\max\left(L/\sqrt{\lambda}, \rho_{\max}\right) \sqrt{n\mathbb{E}[\beta_n]\mathbb{E}[\gamma_n]} + \mathcal{O}(\Delta_{\max})\,.\end{aligned}$$

For the last inequality, we used that β_t is a non-decreasing sequence and introduced the worst-case noise $\rho_{\max} = \max_{a \in \mathcal{A}} \rho(a)$. Note that the confidence coefficient satisfies $\beta_n = \sqrt{2\gamma_{n-1} + 4\log(n)} + \sqrt{\lambda}B$. Using the argument presented in Eq. (3.7) applied to W_n, the total information gain is at most $\gamma_n \leq \mathcal{O}(d \log(nL^2/(d\lambda\rho_{\min}^2)))$. The inverse scaling with the minimum noise $\rho_{\min} = \min_{a \in \mathcal{A}} \rho(a)$ comes from rescaling the actions inversely proportional to the noise variance. This is not totally unexpected, because for $\rho(a) \to 0$ we get $I_t(a) \to \infty$. On the other hand, simply thresholding the noise function away from zero avoids this dependency. The result is summarized in the following theorem.

Theorem 3.1 (Worst-Case Regret)**.** *Assume that* $\|\theta\| \leq B$, $\max_{a \in \mathcal{A}} \|a\| \leq L$ *and* $\max_{a \in \mathcal{A}} \rho(a) \leq \rho_{\max}$. *Then the regret of IDS as in Algorithm 2 satisfies*

$$\mathfrak{R}_n \leq 4\max\left(L/\sqrt{\lambda}, \rho_{\max}\right) \sqrt{n\mathbb{E}\left[\beta_{n,1/n^2}\right]\mathbb{E}[\gamma_n]} + \mathcal{O}(\Delta_{\max})\,.$$

In particular, $\mathfrak{R}_n \leq \mathcal{O}\big(\rho_{\max} LBd\sqrt{n}\log(n)\big)$.

If we keep only the horizon, the sub-Gaussian variance, and the dimension, the bound is $\mathfrak{R}_n \leq \mathcal{O}(d\rho_{\max}\sqrt{n}\log(n))$. For action sets that are exponentially large in the dimension, this is the best one can hope for up to the logarithmic factor [46]. On the other hand, when $k = |\mathcal{A}|$ is small our bound matches the bound of UCB, but a simple elimination algorithm [103, §22] has regret at most $\mathcal{O}(\sqrt{n\log(k)d})$. The extra $\sqrt{d}$ comes from the concentration bound in Lemma 3.1, which cannot be improved without using more specific properties of the action history [103, Exercise 20.1].

3.2.5 *A Limitation of Optimism and Thompson Sampling*

So far, we have proved a regret bound that depends on the worst-case noise value $\rho_{\max}$, reassuring in particular that the proposed algorithm is

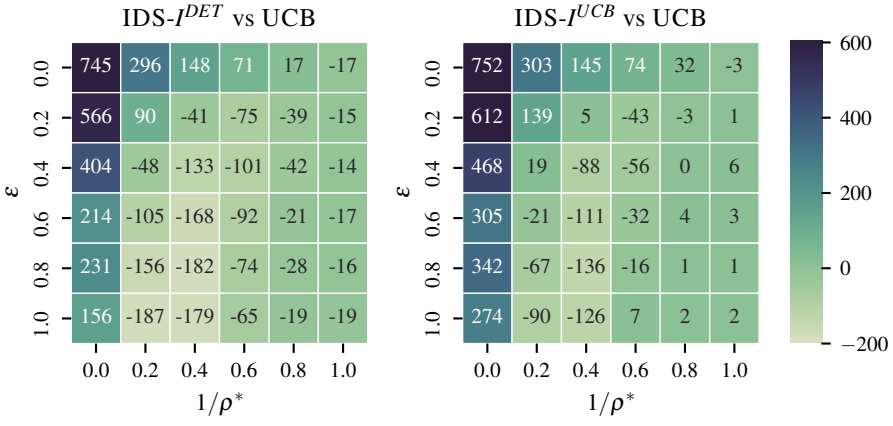


FIGURE 3.1: A numerical simulation of Example 7.5. The plots show the regret difference between IDS and UCB after 1000 steps, averaged over 100 runs. Darker colors / positive values correspond to instances where IDS outperforms UCB. The left plot shows IDS defined with I^{DET} defined in Eq. (3.6). The right plot shows IDS defined with I^{UCB} defined in Eq. (3.13) and discussed in Section 3.3.3. Note that as $\epsilon \to 0$, the gap of the informative actions gets smaller, whereas for $\rho \to \infty$, the best action becomes less informative.

a sensible approach in the standard linear bandit setting. We now argue that for some instances with heteroscedastic noise, IDS exploits the noise information in a way that UCB and Thompson sampling do not.

To formalize the optimistic approach for linear bandits, we assume that at time t, the learner has access to a confidence set $\mathcal{E}_t \subset \mathbb{R}^d$ that contains the true parameter with high probability. The confidence set is computed using all available data at time t and thus may include the noise information of the observation history. The upper confidence bound approach is to choose the action with the largest plausible reward, $a_t^{\mathrm{UCB}} = \arg\max_{a \in \mathcal{A}} \max_{\theta \in \mathcal{E}_t} \langle a, \theta \rangle$.

Example 3.1. Let $\mathcal{B}_\eta = \{a \in \mathbb{R}^d : \|a\| = \eta\}$ be the d-dimensional unit sphere with radius $\eta > 0$. As an illustrative example, we choose an action set consisting of two concentric spheres, $\mathcal{A} = \mathcal{B}_1 \cup \mathcal{B}_{1-\epsilon}$ for $\epsilon \in (0, 1)$. Note that for any parameter $\theta \neq 0$, an action $a \in \mathcal{B}_{1-\epsilon}$ is never optimal. This means that an optimistic approach only chooses actions on the outer sphere. When the noise is homoscedastic, one can argue that the actions with larger

norm provide a better signal-to-noise ratio. With heteroscedastic noise, however, this is no longer the case. Specifically, we can define

$$\rho(a) = \begin{cases} \rho^* & \text{if } a \in \mathcal{B}_1 \\ 1 & \text{if } a \in \mathcal{B}_\eta \,. \end{cases}$$

For $1 \ll \rho^*$, actions in $\mathcal{B}_{1-\epsilon}$ are much more informative and reduce the overall regret through the improved estimation. In an ideal world, we would now compute instance-dependent and minimax lower bounds and relate them to the regret of IDS. However, the version of IDS is not asymptotically optimal and a minimax calculation is rather involved and would require us to make artificial choices on the model class and the noise function. Instead, we restrict ourselves to two limiting considerations.

First, we let $\epsilon \to 0$. In this case, the regret of actions in $\mathcal{B}_{1-\epsilon}$ is only marginally larger than the regret of actions in $\mathcal{B}_1$. Among actions with (almost) the same regret, IDS prefers the action with larger information gain. When $1 \ll \rho^*$, the information gain of an action in $\mathcal{B}_{1-\epsilon}$ is larger than for actions in $\mathcal{B}_1$. Hence, IDS prefers $a \in \mathcal{B}_{1-\epsilon}$ and consequently the regret scales with $\mathcal{O}(d\sqrt{n}\log(n))$, whereas for UCB it is $\mathcal{O}(\rho^* d\sqrt{n}\log(n))$.

A second variant is to fix $\epsilon \in (0,1)$ and let $\rho^* \to \infty$. In this limit, the actions on $\mathcal{B}_1$ provide no information at all. This instance is a special case of a *globally observable* partial monitoring game, which we will discuss in detail in Chapter 6. For this case, one can show that Algorithm 2 has regret at most $\mathcal{O}((n\log(n))^{2/3} d^{1/3})$. The result essentially follows as a special case of Theorem 6.1. On the other hand, the regret of UCB is linear, since it obtains no information from the action it chooses. A numerical simulation of this example is in Fig. 3.1.

We note that the failure mode of Thompson sampling is very similar. Thompson sampling chooses $a_t = \arg\max_{a \in \mathcal{A}} \langle a, \tilde{\theta}_t \rangle$, where $\tilde{\theta}_t$ is a sample from a Bayesian posterior model of the parameter. That means that, as for UCB, Thompson sampling never chooses actions on $\mathcal{B}_\eta$, even if they are more informative.

We emphasize that both UCB and Thompson sampling use noise information for *estimation*, for example in the construction of confidence sets or the posterior distribution. The failure happens when trading regret and information: Optimistic algorithms and Thompson sampling only choose actions that appear plausible optimal, and neglect provably suboptimal actions even if they are very informative. It is the same limitation that causes these approaches to be asymptotically suboptimal in the linear bandit model

(even with homoscedastic noise) and ill-suited for the more general partial monitoring model, which we will see in Chapters 5 and 6.

3.3 VARIANTS

We present several variants of the regret bound presented in Theorem 3.1.

3.3.1 *High-Probability Bound*

Algorithm 2 satisfies a high-probability bound on the pseudo regret $\mathfrak{PR}_n$ defined in Eq. (2.8), if we replace the gap estimates with $\hat{\Delta}_{t,\delta}$, where $\delta \in [0,1]$ is the confidence level chosen by the user. The next results shows that this version of IDS satisfies a high-probability bound on the regret.

Theorem 3.2. *With the same assumptions as in Theorem 3.1, the pseudo regret of Algorithm 2 with gap estimates $\hat{\Delta}_{t,\delta}(a)$ satisfies with probability at least $1-\delta$,*

$$\mathfrak{PR}_n \leq \mathcal{O}\Big(d\rho_{\max} LB\sqrt{n}\log\left(\tfrac{n}{\delta}\right)\Big).$$

Proof. Along the lines of Eq. (3.10), it follows that

$$\Psi_t(\mu_t^{\mathrm{IDS}}) \leq \frac{\hat{\Delta}_t(a_t^{\mathrm{UCB}})^2}{I_t(a_t^{\mathrm{UCB}})} \leq 16\beta_{t,\delta}\max\Big(L^2/\lambda, \rho(a_t^{\mathrm{UCB}})^2\Big).$$

The claim follows from Theorem 2.3. □

3.3.2 *Gap-Dependent Bound*

Instance-dependent bounds are the topic of Chapter 5, but already here we can derive a gap-dependent logarithmic bound for finite action sets. We define the minimum gap $\Delta_{\min} = \arg\min_{a\neq a^*}\Delta(a)$, and require that the optimal action is unique. To obtain a gap-dependent bound, we need to implement the following immediate improvement to the algorithm. When the optimal action is uniquely identified with high probability, we deterministically choose the empirically best action $\hat{a}_t = \arg\min_{a\in\mathcal{A}}\hat{\Delta}_t(a)$. A sufficient condition to check is

$$M_t = \mathbb{1}\Big(\max_{b\neq\hat{a}_t}\langle b, \hat{\theta}^{wls}\rangle + \beta_{t,1/t^2}^{1/2}\|b\|_{W_t^{-1}} \geq \langle \hat{a}_t, \hat{\theta}^{wls}\rangle - \beta_{t,1/t^2}^{1/2}\|\hat{a}_t\|_{W_t^{-1}}\Big),$$

When $M_t = 0$ and assuming the parameter estimate is sufficiently concentrated such that $\theta \in \mathcal{E}_t$, then it holds that $\hat{a}_t = a^*$. We can preserve the algorithmic template of IDS by defining the gap estimate as follows:

$$\hat{\Delta}_t^+(a) = \begin{cases} \hat{\Delta}_t(a) M_t & \text{if } a = \hat{a}_t\,, \\ \hat{\Delta}_t(a) & \text{else.} \end{cases} \tag{3.11}$$

We refer to rounds with $M_t = 0$ as *exploitation rounds* and note that with the gap estimate $\hat{\Delta}_t^+(a)$, in such rounds IDS chooses $\hat{a}_t$ deterministically and $\Psi_t(\mu_t^{\mathrm{IDS}}) = 0$. Rounds with $M_t = 1$ are called *exploration rounds*. With the gap estimate Eq. (3.11), IDS satisfies a logarithmic, gap-dependent bound that is summarized in the next theorem. It is also easy to see that the result of Theorem 3.1 continues to hold.

Theorem 3.3 (Gap-Dependent Regret)**.** *For finite action sets with unique optimal action and minimum gap* $\Delta_{\min}$*, IDS defined with gap estimate Eq.* (3.11) *satsifies*

$$\mathfrak{R}_n \leq \mathcal{O}\big(\rho_{\max} \Delta_{\min}^{-1} d^2 \log(n)^2\big)$$

Proof. Corollary 2.2 combined with Eq. (3.5) implies that

$$\mathfrak{R}_n \leq 4\zeta \mathbb{E}\big[\beta_{n,1/n^2} \gamma_n\big] + \mathcal{O}(\Delta_{\max})\,,$$

provided that we show $\Psi_t(\mu_t) \leq 4\zeta \delta_t \beta_t$ in rounds t where $\Delta(a) \leq \hat{\Delta}_t(a)$ for all $a \in \mathcal{A}$. Hence, the result follows if we show that

$$\Psi_t \leq 64 \delta_t \beta_t \max(\lambda L^{-2}, \rho_{\max}) \Delta_{\min}^{-1}\,. \tag{3.12}$$

Note that by reusing the worst-case bound on the information ratio from Eq. (3.10), the result is immediate if $4\delta_t \geq \Delta_{\min}$. The bound also follows trivially if $M_t = 0$. On the contrary case, we get $4\delta_t \leq \Delta_{\min} \leq \hat{\Delta}_t(b)$ for all $b \neq a^*$ and the fact that we assume $\Delta(a) \leq \hat{\Delta}_t(a)$ for all $a \in \mathcal{A}$. In particular, it must be that $\hat{a}_t = a^*$. We can now upper bound the information ratio as follows:

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_t(\mu) \leq \min_{b \neq \hat{a}_t} \min_{p \in [0,1]} \frac{\hat{\Delta}_t((1-p)e_{\hat{a}_t} + p e_b)^2}{p I_t(b)} \leq \min_{b \neq \hat{a}_t} \frac{4\delta_t \hat{\Delta}_t(b)}{I_t(b)}\,.$$

The first inequality restricts the set of sampling distributions and drops the information gain of $\hat{a}_t$. The second inequality is satisfied with $p = \frac{\delta_t}{\hat{\Delta}_t(b) - \delta_t}$. We define the set of plausible maximizers,

$$\mathcal{M}_t = \big\{ a \in \mathcal{A} : \langle a, \hat{\theta}^{wls} \rangle + \beta_{t,1/t^2}^{1/2} \|a\|_{W_t^{-1}} \geq \max_{b \neq \hat{a}_t} \langle b, \hat{\theta}^{wls} \rangle - \beta_{t,1/t^2}^{1/2} \|b\|_{W_t^{-1}} \big\}\,.$$

Note that $M_t = 1$ implies $|\mathcal{M}_t| > 1$. Consider the action with the largest uncertainty among the plausible maximizers, $c_t = \arg\max_{c \in \mathcal{M}_t} \|c\|_{W_t}^{-1}$. Note that by our assumptions,

$$\Delta_{\min} \leq \hat{\Delta}_t(c_t) \leq 2\beta_t^{1/2} \|c_t\|_{W_t^{-1}} .$$

Combining the two displays and Eq. (3.9), we find

$$\Psi_t \leq 64 \delta_t \beta_t \max(\lambda L^{-2}, \rho_{\max}) \Delta_{\min}^{-1} .$$

This completes the proof. □

3.3.3 *Directed Information Gain*

The information gain in Eq. (3.6) is motivated from the worst-case analysis. Also, the Bayesian interpretation as mutual information $\mathbb{I}_t(\theta; y_t | a_t = a)$ suggests that it is a conservative choice since it incentivizes IDS to learn the parameter uniformly well. This intuition is confirmed in experiments, where the version of IDS shown in Algorithm 2 is often over-explorative. An improvement is to choose an information gain that reflects more closely what we actually care about: Identifying the optimal action. Of course the learner does not know the optimal action, but can use a surrogate. Certainly, the learner has to reduce the uncertainty about the value of the UCB action $a_t^{\mathrm{UCB}} = \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_t^{wls} \rangle + \beta_t^{1/2} \|a\|_{W^{-1}}$. This leads to the following information gain:

$$I_t^{\mathrm{UCB}}(a) \triangleq \frac{1}{2} \log\Big(\|a_t^{\mathrm{UCB}}\|^2_{V_t^{-1}} \Big) - \frac{1}{2} \log\Big(\|a_t^{\mathrm{UCB}}\|^2_{(W_t + \rho(a)^{-2} a a^\top)^{-1}} \Big) \qquad (3.13)$$

Relative to the information gain $I_t(a)$ defined in Eq. (3.6), $I_t^{\mathrm{UCB}}(a)$ satisfies the following properties. For any $a \in \mathcal{A}$. the new information gain is never larger, $I_t^{\mathrm{UCB}}(a) \leq I_t(a)$ with equality at the UCB action, $I_t^{\mathrm{UCB}}(a_t^{\mathrm{UCB}}) = I_t(a_t^{\mathrm{UCB}})$. Hence, the proof that we provided for Theorem 3.1 continues to hold true if we replace the information gain in Algorithm 2 with $I_t^{\mathrm{UCB}}(a)$. With homoscedastic noise, this algorithm is naturally biased towards UCB, and in experiments the two algorithms are hardly distinguishable. On the other hand, with heteroscedastic noise, IDS chooses actions that reduce the uncertainty about the outcome of the UCB action, even if the UCB action itself is not informative, for example in the instance we illustrated in Section 3.2.5.

3.3.4 *Deterministic Information-Directed Sampling*

An interesting observation is that the upper bound on the information ratio in Eq. (3.10) does not make use of randomization. This implies that in the bandit setting, it suffices to optimize the information ratio over Dirac distributions:

$$a_t^{\mathrm{DIDS}} = \operatorname*{arg\,min}_{a \in \mathcal{A}} \frac{\hat{\Delta}_t(a)^2}{I_t(a)} \,. \tag{3.14}$$

We refer to this algorithm as *deterministic* IDS. However, we remark already that randomization is a crucial ingredient for the asymptotic analysis and the bounds for partial monitoring in Chapters 5 and 6.

In the homoscedastic case, deterministic IDS defined in Eq. (3.14) is closely related to the UCB algorithm. For simplicity, assume that the regularization parameter λ is chosen large enough such that $\|a\|_{V_1^{-1}} \leq 1$ for all $a \in \mathcal{A}$. The information gain satisfies

$$\frac{1}{2}\|a\|^2_{V_t^{-1}} \leq 2I_t^{\mathrm{DET}}(a) = \log(1 + \|a\|^2_{V_t^{-1}}) \leq \|a\|^2_{V_t^{-1}} \triangleq \tilde{I}_t(a) \,,$$

which we also used in the proof. This implies that IDS defined with $\tilde{I}_t(a)$ satisfies the same bounds on the regret, up to constant factors. In the next lemma we show that deterministic IDS defined with the information gain $\tilde{I}_t$ is equivalent to the UCB algorithm. For a related result, see [170, Lemma 2.1].

Lemma 3.3. *The UCB action* $a_t^{\mathrm{UCB}} = \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_t \rangle + \beta_t^{1/2} \|a\|_{V_t^{-1}}$ *minimizes the deterministic information ratio,*

$$\Psi_t(a_t^{\mathrm{UCB}}) = \min_{a \in \mathcal{A}} \frac{\hat{\Delta}_t(a)^2}{\tilde{I}_t(a)} \,.$$

Proof. A directed calculation confirms that the information ratio of the UCB action is $\Psi_t(a^{\mathrm{UCB}}) = \frac{\hat{\Delta}_t(a_t^{\mathrm{UCB}})^2}{\tilde{I}_t(a_t^{\mathrm{UCB}})} = 4\beta_t$. Moreover, any $a \in \mathcal{A}$ satisfies

$$\begin{aligned}
\frac{\hat{\Delta}_t(a)^2}{\tilde{I}_t(a)} &= \frac{\left(\max_{b \in \mathcal{A}} \langle b, \hat{\theta}_t \rangle + \beta_t^{1/2} \|b\|_{V_t^{-1}} - \left(\langle a, \hat{\theta}_t \rangle - \beta_t^{1/2} \|a\|_{V_t^{-1}}\right)\right)^2}{\|a\|^2_{V_t^{-1}}} \\
&\geq 4\beta_t \,.
\end{aligned}$$

□

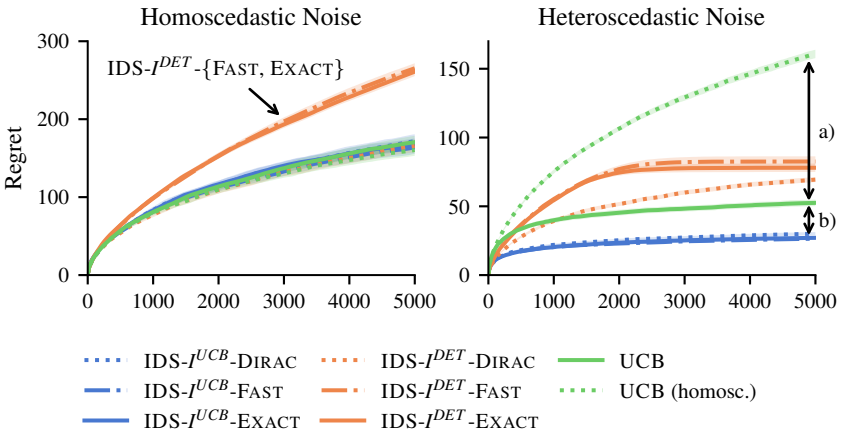


FIGURE 3.2: Performance on a randomly generated and fixed action set. With homoscedastic noise, shown on the left, all methods perform similarly. Only IDS with I^{DET} performs slighly worse unless the sampling distribution is restricted to a Dirac. The FAST-suffix indicates approximate IDS sampling (Section 2.1.1), which closely follows the exact version. Deterministic IDS (Eq. (3.14)) is specified with the DIRAC-suffix. With heteroscedastic noise, performance is more varied. Marker a) shows the improvement obtained with UCB when using a weighted least-squares estimator opposed to the homoscedastic baseline that uses a uniform noise upper bound. This improvement is from better *estimation*, but the noise information is not directly used to determine the action choice. Marker b) shows the improvement gained from better *exploration* on top of estimating the parameter with weighted least-squares. This demonstrates the importance of using the noise both for estimation and exploration

3.4 NUMERICAL RESULTS

We empirically demonstrate that using noise information can be beneficial on synthetic example. We choose a randomly generated instance with $k = 30$ actions in $\mathbb{R}^3$. The result is show in Fig. 3.2. For a systematic benchmark with Example 7.5, see Fig. 3.1.

3.5 CONTRIBUTIONS AND RELATED WORK

The frequentist version of information-directed sampling for heteroscedastic linear bandits and the worst-case bounds on the regret are published in:

- Kirschner, J. & Krause, A. *Information Directed Sampling and Bandits with Heteroscedastic Noise* in *Proc. International Conference on Learning Theory (COLT)* (July 2018)

The gap-dependent bounds presented in this chapter are novel, using the techniques by Kirschner, Lattimore, Vernade & Szepesvári [91].

Heteroscedastic bandits have been studied before, but mostly in the multi-armed bandit model. The lower bounds on the asymptotic regret apply directly to heteroscedastic noise distributions, see, for example, the work by Burnetas & Katehakis [31]. Cowan, Honda & Katehakis [44] study multi-armed bandits with unknown variance and derive an asymptotically optimal algorithm. Related results on active learning with heteroscedastic noise and a linear feedback model are by Chaudhuri, Jain & Natarajan [39].

Heteroscedasticity arises naturally in many applications and has been identified as a key challenge in applications of Bayesian optimization [45]. Note that the algorithm presented in this chapter can be kernelized. For details, we refer to the paper [87]. Kernelized methods are detailed in the more general partial monitoring formulation in Section 8.2.

4

A CONNECTION TO PRIMAL-DUAL METHODS

Previous work on information-directed sampling has focused primarily on worst-case analysis, and these results motivated the definition of the information ratio [135, 137]. This chapter's focus is on *asymptotic properties* of information-directed sampling, which, so far, are less well understood. The main result is a fundamental connection between the IDS distribution and a primal-dual approach to solve the asymptotic regret lower bound. The connection is a surprising result because IDS was not explicitly design for the asymptotic regime. Primal-dual methods recently attracted attention in the bandit literature for asymptotic and non-asymptotic exploration in structured bandits [48, 50, 160]. As a by-product, we provide efficient solvers for the asymptotic regret lower bound, which are required in several previous works [72, 79, 104].

SETTING We work in the *structured bandit* setting with a finite action set $\mathcal{A}$, and finite model class $\mathcal{M} \subset \mathbb{R}^{\mathcal{A}}$ consisting of functions mapping actions to reward. When the learner chooses an action $a_t \in \mathcal{A}$, the observation y_t is a sample from a specified distribution $\vartheta_{\theta,a_t} \in \mathscr{P}(\mathbb{R})$ that depends on the action and the instance $\theta \in \mathcal{M}$. For example, in the linear bandit setting with uni-variate Gaussian noise, $\vartheta_{\theta,a} = \mathcal{N}(\langle a, \theta \rangle, 1)$, but the results here do not rely on a specific distributional form.

The optimal action for $\theta \in \mathcal{M}$ is $a^*(\theta) = \arg\max_{a \in \mathcal{A}} f_\theta(a)$. The set of *alternative parameters* is $\mathcal{C}^*(\theta) = \{\nu \in \mathcal{M} : a^*(\theta) \neq a^*(\nu)\}$. A policy π is called *consistent* if $\mathfrak{R}_n(\pi, \theta) \leq o(n^p)$ for all $p > 0$ and all $\theta \in \mathcal{M}$.

Theorem 4.1 (Combes, Magureanu & Proutiere [43])**.** *The asymptotic regret of any consistent policy π on the instance $\theta \in \mathcal{M}$ is at least*

$$\liminf_{n \to \infty} \frac{\mathfrak{R}_n(\pi, \theta)}{\log(n)} \geq \mathfrak{c}^*(\theta)\,,$$

where $\mathfrak{c}^*(\theta)$ is specified by the following optimization problem:

$$\begin{aligned} \mathfrak{c}^*(\theta) = \inf_{\alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0}} & \sum_{a \in \mathcal{A}} \alpha(a)(f_\theta(a^*) - f_\theta(a)) \\ \text{s.t.} \quad & \min_{\nu \in \mathcal{C}^*(\theta)} \sum_{a \in \mathcal{A}} \alpha(a) D_{KL}(\vartheta_{a,\theta} \| \vartheta_{a,\nu}) \geq 1 \end{aligned} \tag{4.1}$$

Intuitively, the cost of the optimization corresponds to the regret when playing according to the allocation α, while the constraints require the policy to gather enough statistical evidence to infer the best action. The restriction to consistent polices is necessary, because otherwise we can define a policy to always play $a^* \in \mathcal{A}$, which results in zero regret when a^* is indeed optimal, but linear regret otherwise.

Our next goal is to derive sequential strategies to solve Eq. (4.1). We present the algorithmic results in the *oracle setting*, where the exact cost and constraint vectors are known. This way we can focus on the main ideas, while avoiding complications that arise from the statistical estimation errors. In Chapter 5, we implement the results for stochastic linear bandits.

Let us fix a true instance $\theta \in \mathcal{M}$ and write $a^* = a^*(\theta)$ and $\mathcal{C}^* = \mathcal{C}^*(\theta)$. Mainly for simplicity, we assume that a^* is unique. As before, the gaps are $\Delta(a) = f_\theta(a^*) - f_\theta(a)$. We define *constraint vectors* $h_\nu \in \mathbb{R}^{\mathcal{A}}_{\geq 0}$ as $h_\nu(a) = D_{\mathrm{KL}}(\vartheta_{\theta,a} \| \vartheta_{\nu,a})$ for all $\nu \in \mathcal{M}$. With this notation, the lower bound (4.1) is written as a *linear covering program* with $|\mathcal{C}^*|$ constraints,

$$\mathfrak{c}^* = \inf_{\alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0}} \sum_{a \in \mathcal{A}} \alpha(a)\Delta(a) \quad \text{s.t.} \quad \forall \nu \in \mathcal{C}^*, \ \sum_{a \in \mathcal{A}} \alpha(a) h_\nu(a) \geq 1 \,. \tag{4.2}$$

Note that there is no cost for allocating on the optimal action a^* since the corresponding gap is zero. Since the constraint vectors are also non-negative, we can always choose a solution with $\alpha(a^*) = \infty$ (defined in the appropriate limit). Following the terminology of Jun & Zhang [79], a constraint h_ν is *docile* if $h_\nu(a^*) > 0$. Docile constraints are satisfied by playing a^* alone which does not increase the regret.

In the following, we assume that $0 < \mathfrak{c}^* < \infty$, which requires the observation distributions $\vartheta_{\theta,a}$ to be such that the program is feasible. Further, the solution is non-trivial in the sense that the learner has to choose at least one sub-optimal action. For simplicity, we assume that $h_\nu(a) \leq 1$ for all $\nu \in \mathcal{C}^*$ and $a \in \mathcal{A}$, which can always be achieved by rescaling the constraints. We also require that $|\mathcal{C}^*| \geq 2$, as the approach is trivial for $|\mathcal{C}^*| = 1$. As a side remark, the case $\mathfrak{c}^* = 0$ is particularly challenging in the stochastic bandit setting because an asymptotically optimal learner needs to identify a^* within $o(\log(n))$ rounds, and then only play the optimal action [79].

A policy $\pi_n = (\mu_t)_{t=1}^n$ defines a cumulative allocation $\alpha_n = \sum_{t=1}^n \mu_t \in \mathbb{R}^{\mathcal{A}}$. We say an allocation is *asymptotically optimal and consistent* at rate β_n if

$$\lim_{n\to\infty} \frac{\Delta(\alpha_n)}{\beta_n} \leq \mathfrak{c}^*, \quad \text{and} \quad \forall \nu \in \mathcal{C}^*, \ \lim_{n\to\infty} \frac{h_\nu(\alpha_n)}{\beta_n} \geq 1 \,. \tag{4.3}$$

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

Algorithm 3: Primal-Dual Solver for Eq. (4.2)

Input: Action set $\mathcal{A}$, model class $\mathcal{M}$, instance θ, horizon n, rate β_n
Assume: No docile constraints, i. e. $h_\nu(a^*) = 0$ for all $\nu \in \mathcal{C}^*$.

$\alpha_0 \leftarrow 0 \in \mathbb{R}^{\mathcal{A}}$
$\eta \leftarrow \sqrt{2(\beta_n + 1)\log(|\mathcal{C}^*|)}$
$\Delta(a) \leftarrow \max_{b \in \mathcal{A}} f_\theta(b) - f_\theta(a), \quad \forall a \in \mathcal{A}$
$h_\nu(a) \leftarrow D_{\mathrm{KL}}(\vartheta_{\theta,a} \| \vartheta_{\nu,a}), \quad \forall a \in \mathcal{A}$ and $\nu \in \mathcal{M}$
for $t = 1, 2, 3, \ldots, n$ **do**
  if $\min_{\nu \in \mathcal{C}^*} h_\nu(\alpha_{t-1}) > \beta_n$ **then**
    $\mu_t \leftarrow e_{a^*}$
  else
    $q_t(\nu) \leftarrow \exp\big(-\eta h_\nu(\alpha_{t-1})\big), \forall \nu \in \mathcal{C}^*$
    $I_t(a) \leftarrow \sum_{\nu \in \mathcal{C}^*} q_t(\nu) h_\nu(a), \forall a \in \mathcal{A}$
    $c_t \leftarrow \arg\min_{c \neq a^*} \frac{\Delta(c)}{I_t(c)}$
    $\mu_t \leftarrow e_{c_t}$
  $\alpha_t \leftarrow \sum_{s=1}^{t} \mu_t$
return α_n

4.1 ONLINE CONVEX OPTIMIZATION

In this section, we review an approach due to Garg & Koenemann [63] and Arora, Hazan & Kale [11], which solves covering LPs – such as the oracle lower bound – using *online convex optimization* (OCO). A similar idea has recently inspired bandit algorithms for best arm identification [48] and regret minimization [50]. The approach sets up a fictitious two-player game that converges to the saddle point of the Lagrangian,

$$\max_{\lambda \geq \mathbb{R}^{\mathcal{C}^*}_{\geq 0}} \min_{\alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0}} \left\{ \mathcal{L}(\alpha, \lambda) \triangleq \Delta(\alpha) - \sum_{\nu \in \mathcal{C}^*} \lambda_\nu (h_\nu(\alpha) - 1) \right\}.$$

Strong duality holds and we can interchange the maximum and minimum. Note that the dual variables are on an unbounded space, but it turns out that we can normalize them. The Karush–Kuhn–Tucker conditions are

$$\Delta(x) - \sum_{\nu \in \mathcal{C}^*} \lambda_\nu h_\nu(x) = 0 \qquad \text{(stationarity)}$$

$$\forall \nu \in \mathcal{C}^*, \quad \lambda_\nu (h_\nu(\alpha) - 1) = 0 \qquad \text{(complementary slackness)}$$

Combining both and using that $\mathfrak{c}^* = \Delta(\alpha^*)$ we find $\mathfrak{c}^* = \sum_{\nu \in \mathcal{C}^*} \lambda_\nu$. This implies that $\mathfrak{c}^*$ normalizes the dual variables, and we define $q_\nu \triangleq \lambda_\nu / \mathfrak{c}^*$. The *normalized Lagrangian* is

$$\bar{\mathcal{L}}(\alpha, q) = \Delta(\alpha) - \mathfrak{c}^* \sum_{\nu \in \mathcal{C}^*} q_\nu (h_\nu(\alpha) - 1)\,, \tag{4.4}$$

where $q \in \mathscr{P}(\mathcal{C}^*)$ is now a distribution over the constraints. Recall that the allocation $\alpha_n = \sum_{t=1}^n \mu_t$ is chosen sequentially. In each iteration of the game, first the *dual player* (or q-learner) chooses a distribution $q_t \in \mathscr{P}(\mathcal{C}^*)$ over the constraints. Then, the response of the *primal player* is a distribution $\mu_t \in \mathscr{P}(\mathcal{A})$ over actions. The linear loss of the q-learner is defined by the response μ_t of the primal player,

$$l_t(q) \triangleq \sum_{\nu \in \mathcal{C}^*} q_t(\nu) h_\nu(\mu_t)\,. \tag{4.5}$$

Given the choice q_t of the q-learner, we define the combined constraint vector $I_t \triangleq \sum_{\nu \in \mathcal{C}^*} q_t(\nu) h_\nu \in \mathbb{R}^{\mathcal{A}}_{\geq 0}$, which satisfies $I_t(\mu_t) = l_t(q_t)$. The primal response that defines the policy $\pi_n = (\mu_t)_{t=1}^n$ is

$$\mu_t = \begin{cases} e_{a^*} & \text{if } \min\limits_{\nu \in \mathcal{C}^*} h_\nu(\alpha_{t-1}) > \beta_n\,, \\ e_{c_t} & \text{else, where } c_t = \arg\min_{a \in \mathcal{A} \setminus a^*} \frac{\Delta(a)}{I_t(a)}\,. \end{cases} \tag{4.6}$$

The approach is summarized in Algorithm 3. As written, the approach only works in the case *without* docile constraints, which can easily be fixed by allocating a^* explicitly. Note that the action c_t defined in Eq. (4.6) satisfies

$$\frac{\Delta(c_t)}{I_t(c_t)} = \min_{c \neq a^*} \frac{\Delta(c)}{I_t(c)} \overset{(i)}{=} \min\Big\{ \Delta(\alpha) : \alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0} \quad \text{s.t.} \quad I_t(\alpha) \geq 1 \Big\} \overset{(ii)}{\leq} \mathfrak{c}^*\,. \tag{4.7}$$

The equality (i) uses that there are no docile constraints, and that the optimization program with one constraint I_t is optimally solved by the allocation with $\alpha(a) = \mathbb{1}(a = c_t) I_t(c_t)^{-1}$. The inequality (ii) follows with the optimal allocation α^* as defined by Eq. (4.2). Using Eq. (4.7), we bound the regret,

$$\mathfrak{R}_n(\pi_n, \theta) = \Delta(\alpha_n) = \sum_{t=1}^n \Delta(\mu_t) \leq \sum_{t=1}^n \mathfrak{c}^* I_t(\mu_t) = \mathfrak{c}^* \sum_{t=1}^n l_t(q_t)\,.$$

So far, we have related the regret to the cumulative loss of the q-learner. We introduce the *q-learner regret*,

$$\mathfrak{Q}_n \triangleq \sum_{t=1}^n l_t(q_t) - \min_{\nu \in \mathcal{C}^*} \sum_{t=1}^n l_t(\nu) = \sum_{t=1}^n l_t(q_t) - \min_{\nu \in \mathcal{C}^*} h_\nu(\alpha_t)\,.$$

Note that the algorithm is defined to explicitly control $\min_{\nu\in\mathcal{C}^*} h_\nu(\alpha_{t-1})$. The difference is bounded by the assumption $h_\nu(a) \leq 1$, hence

$$\min_{\nu\in\mathcal{C}^*} \sum_{t=1}^{n} l_t(\nu) = \min_{\nu\in\mathcal{M}} h_\nu(\alpha_n) \leq \beta_n + 1\,.$$

This allows us to write the regret $\mathfrak{R}_n$ in terms of $\mathfrak{Q}_n$,

$$\mathfrak{R}_n(\pi,\theta) \leq \mathfrak{c}^*(\beta_n + 1 + \mathfrak{Q}_n)\,. \tag{4.8}$$

The literature on online convex optimization offers plenty of algorithms to turn the last display into a meaningful bound [124]. For concreteness, we choose the exponential weights learner [110, 168],

$$q_t(\nu) \propto \exp\left(-\eta \sum_{s=1}^{t-1} l_s(\nu)\right),$$

with a suitably chosen learning rate $\eta > 0$. A regret bound for this choice of q-learner is given in the next lemma.

Lemma 4.1. *With learning rate* $\eta = \sqrt{2\log(|\mathcal{C}^*|)(\beta_n+1)}$ *and assuming that the best loss in hindsight satisfies* $\min_{\nu\in\mathcal{C}^*}\sum_{t=1}^{n} l_t(\nu) \leq \beta_n + 1$*, the regret of the exponential weights learner is at most,*

$$\mathfrak{Q}_n \leq 2\sqrt{2\log(|\mathcal{C}^*|)(\beta_n+1)}\,.$$

Proof. The regret bound is a standard result, c. f. [124]. The proof is short enough to show it here. Denote the cumulative loss by $L_t(\nu) = \sum_{s=1}^{t-1} l_t(\nu)$. Exponential weights is equivalent to *follow the regularized leader* (FTRL),

$$q_t = \operatorname*{arg\,min}_{q\in\mathscr{P}(\mathcal{C}^*)} L_t(q) + \frac{1}{\eta}\psi(q)\,,$$

with the entropy function $\psi(q) = \sum_{\nu\in\mathcal{C}^*} q(\nu)\log(q(\nu))$ as a regularizer [146]. For learning rate $\eta > 0$, we define

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

The next inequality follows from telescoping Orabona [124, Lemma 7.1],

$$\mathfrak{Q}_n \leq -\frac{1}{\eta}\min_{q\in\mathscr{P}(\mathcal{C}^*)} \psi_\eta(q) + \sum_{t=1}^{n}\left([L_{t+1}+\psi_\eta](q_t) - [L_{t+1}+\psi_\eta](q_{t+1})\right).$$

For the first term, we immediately get $-\frac{1}{\eta}\min_q \psi(q) \leq \frac{\log(|\mathcal{C}^*|)}{\eta}$. The sum is often referred to as the stability term. The increments are bounded as follows,

$$[L_{t+1}+\psi_\eta](q_t) - [L_{t+1}+\psi_\eta](q_{t+1}) \leq [L_{t+1}+\psi_\eta](q_t) - [L_{t+1}+\psi_\eta](\tilde{q})\,,$$

where $\tilde{q}_t = \arg\min_{q\in\mathscr{P}(\mathcal{C}^*)} L_{t+1}+\psi_\eta \propto \exp(-\eta L_{t+1})$. After some algebraic manipulations, we arrive at

$$\begin{aligned}
&[L_{t+1}+\psi_\eta](q_t) - [L_{t+1}+\psi_\eta](\tilde{q}) \\
&= l_t(q_t) + \frac{1}{\eta}\log\left(\sum_{\nu\in\mathcal{C}^*} q_t(\nu)\exp\Big(-\eta(l_t(\nu))\Big)\right) \\
&\leq l_t(q_t) + \frac{1}{\eta}\left(\sum_{\nu\in\mathcal{C}^*} q_t(\nu)\exp\Big(-\eta(l_t(\nu))\Big) - 1\right) \\
&\leq \frac{\eta}{2}\sum_{\nu\in\mathcal{C}^*} q_t(\nu)(l_t(\nu))^2 \leq \frac{\eta}{2} l_t(q_t)\,.
\end{aligned}$$

The first inequality uses $\log(v) \leq v-1$ for all $v \geq 0$, and the second inequality uses $\exp(-v) \leq 1 - v + \frac{v^2}{2}$ for all $v \geq 0$. The last inequality uses again boundedness, $l_t(\nu) \leq 1$. All that remains is to solve for the regret:

$$\mathfrak{Q}_n \leq \frac{1}{\eta}\log(|\mathcal{C}^*|) + \frac{\eta}{2}\sum_{t=1}^n l_t(q_t) = \frac{1}{\eta}\log(|\mathcal{C}^*|) + \frac{\eta}{2}\left(\mathfrak{Q}_n + \min_{\nu\in\mathcal{C}^*} L_n(\nu)\right)$$

Provided that the best loss in hindsight satisfies $\min_{\nu\in\mathcal{C}^*} L_n(\nu) \leq \beta_n + 1$, a feasible choice of the learning rate is $\eta = \sqrt{2\log(|\mathcal{C}^*|)(\beta_n+1)}$. Hence

$$\mathfrak{Q}_n \leq 2\sqrt{2\log(|\mathcal{C}^*|)(\beta_n+1)}\,. \qquad \square$$

The result allows us to complete the bound on the policy regret Eq. (4.8). Further note that the approach is asymptotically consistent according to Eq. (4.3) if $\beta_n/n \to 0$. We summarize the result in the following theorem.

Theorem 4.2. *On any instance without docile constraints, Algorithm 3 returns an allocation that is asymptotically optimal and consistent. Furthermore, the regret of the corresponding policy π_n satisfies*

$$\mathfrak{R}_n(\pi_n,\theta) \leq \mathfrak{c}^*(\beta_n+1) + 2\mathfrak{c}^*\sqrt{2\log(|\mathcal{C}^*|)(\beta_n+1)}\,.$$

An immediate extension to the result is to replace the q-learner with AdaHedge [47, 54], which avoids the need to know the horizon while preserving the same scaling of the q-learner regret.

Algorithm 4: Information-Directed Sampling with Oracle Access

Input: Action set $\mathcal{A}$, model class $\mathcal{M}$, instance θ, horizon n, rate β_n, estimation error sequence $(\delta_t)_{t=1}^n$

$\alpha_0 \leftarrow 0 \in \mathbb{R}^{\mathcal{A}}$
$\eta \leftarrow \sqrt{2(\beta_n + 1)\log(|\mathcal{C}^*|)}$
for $t = 1, 2, 3, \ldots, n$ **do**
  if $\min_{\nu \in \mathcal{M}} h_\nu(\alpha_{t-1}) > \beta_n$ **then**
    $\mu_t \leftarrow e_{a^*}$
  else
    $\hat{\Delta}_t(a) \leftarrow \Delta(a) + \delta_t$, $\forall a \in \mathcal{A}$
    $q_t(\nu) \leftarrow \exp\big(-\eta h_\nu(\alpha_{t-1})\big)$, $\forall \nu \in \mathcal{C}^*$
    $I_t(a) \leftarrow \sum_{\nu \in \mathcal{C}^*} q_t(\nu) h_\nu(a)$, $\forall a \in \mathcal{A}$
    $\mu_t \leftarrow \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \dfrac{\hat{\Delta}_t(\mu_t)^2}{I_t(\mu_t)}$
  $\alpha_t \leftarrow \sum_{s=1}^t \mu_t$
return α_n

4.2 INFORMATION-DIRECTED SAMPLING AS PRIMAL-DUAL METHOD

We are now in the position to establish a connection between IDS and the primal-dual approach presented in the previous section. Let $(\delta_t)_{t=1}^n$ be a positive sequence with $\delta_n \to 0$ and $\sum_{t=1}^n \delta_t \to \infty$ as $n \to \infty$. We refer to δ_t as the *estimation error*. Assuming still that the reward function is known, we define *gap estimates* as $\hat{\Delta}_t(x) \triangleq \Delta(x) + \delta_t$. The choice anticipates the definition for the gap estimate that we will use in the next chapter. More importantly, $\hat{\Delta}(a) \geq \delta_t > 0$ ensures that IDS is defined in a meaningful way and does not degenerate to the greedy algorithm that just plays a^*. Using the same notation as in the previous section, the combined constraints $I_t = \sum_{\nu \in \mathcal{C}^*} q_t(\nu) h_\nu \in \mathbb{R}^{\mathcal{A}}_{\geq 0}$ define the *information gain*, where $q_t \in \mathscr{P}(\mathcal{C}^*)$ is computed by the q-learner. The information gain and the gap estimate define an *oracle version* of information-directed sampling,

$$\mu_t^{\mathrm{IDS}} = \operatorname*{arg\,min}_{\mu \in \mathscr{P}(\mathcal{A})} \left\{ \Psi_t(\mu) = \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu)} \right\}.$$

We follow this strategy as long as $\min_{\nu \in \mathcal{C}^*} h_\nu(\alpha_t) < \beta_n$. Once all constraints are satisfied, we choose the optimal action a^*. The approach is summarized in Algorithm 4.

A direct connection between IDS and the primal-dual method is obfuscated by the fact that the information ratio uses squared gaps, whereas we argued previously that one should play $\arg\min_{a\neq a^*} \Delta(a)/I_t(a)$. The link becomes clear when we compute the information ratio in the limit $\delta_t \to 0$ and $I_t(a^*) \to 0$. Both assumptions are reasonable in the statistical setting. Clearly, a consistent gap estimate of a^* has to approach zero. The information gain of a^* has to satisfiy $I_t(a^*) \leq o(1)$, if we expect a^* to be played $\Omega(n)$ times on a horizon n and require that the total information gain $\gamma_n = \sum_{t=1}^n I_t(a_t)$ grows at most sublinearly on the horizon. Specifically, we assume that

$$2\delta_t \leq \min_{a\neq a^*} \hat{\Delta}_t(a)\,, \tag{4.9}$$

or, equivalently, $\delta_t \leq \min_{a\neq a^*} \Delta(a) \triangleq \Delta_{\min}$. Requiring that the estimation error is smaller than the minimum gap is a natural assumption, since the learner has to identify a^* and allocate $\Omega(n)$ plays to the best action, whereas for large n, sub-optimal actions should receive at most $\mathcal{O}(\log(n))$ plays in the optimal allocation. Using Eq. (4.9) and assuming there are no docile constraints, we bound the information ratio with Lemma 2.5,

$$\min_{\mu\in\mathscr{P}(\mathcal{A})} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu_t)} \leq \min_{a\neq a^*} \frac{4\delta_t\hat{\Delta}_t(a)}{I_t(a)} = \min_{a\neq a^*} \frac{4\delta_t\Delta(a)}{I_t(a)} \leq 4\delta_t\mathfrak{c}^*\,. \tag{4.10}$$

The last inequality is by Eq. (4.7) and Eq. (2.6). Without docile constraints, it holds that $I_t(a^*) = 0$, and the IDS distribution is $\mu_t^{\mathrm{IDS}} = (1-p_t)e_{a^*} + p_t e_{c_t}$ where $c_t = \arg\min_{a\neq a^*} \Delta(a)/I_t(a)$ and the trade-off probability is $p_t = \delta_t/\Delta(c_t)$. Notably, the estimation error does not bias the cost/constraint ratio that determines c_t.

An argument along the lines of the analysis in the previous section provides a bound on the regret similar to Theorem 4.2. A more direct argument uses the generic IDS regret bound in Corollary 2.3, which states that when $\Psi_t(\mu_t) \leq 4\delta_t\zeta$, then

$$\mathfrak{R}_n(\pi_n,\theta) \leq \zeta\,\mathbb{E}[\gamma_n] + \sum_{t=1}^n \mathbb{E}\big[\Delta(a_t) - (\hat{\Delta}_t(a_t) - \delta_t)\big] = \zeta\,\mathbb{E}[\gamma_n]\,.$$

The second equality uses the definition of the oracle gap estimates. To keep the otherwise synthetic analysis simple, we assume that the ora-

cle algorithm is initialized with a estimation error sequence that satisfies $\delta_t \leq \min_{a \neq a^*} \Delta(a)$. The regret bound then reads

$$\mathfrak{R}_n(\pi_n, \theta) \leq \mathfrak{c}^* \sum_{t=1}^{n} I_t(\mu_t) \leq \mathfrak{c}^*(\beta_n + 1) + 2\mathfrak{c}^* \sqrt{2 \log(|\mathcal{C}^*|)(\beta_n + 1)}\,.$$

The inequality uses Lemma 4.1 and the fact that the best q-loss in hindsight is explicitly controlled by the design of the algorithm. Arguing that the allocation is consistent according to Eq. (4.3) is more delicate. In particular, when $\delta_t = 0$ or δ_t is approaching zero too fast, then IDS just plays a^* and the constraints are never satisfied. We noted before that the probability of sampling a sub-optimal action is $p_t = \delta_t / \Delta(c_t)$. Hence requiring that $\sum_{t=1}^{\infty} \delta_t = \infty$ while $\beta_n / n \to 0$ guarantees that the constraints are satisfied eventually. The result is summarized in the next theorem.

Theorem 4.3. *On an instance without docile constraints, the policy π_n defined by Algorithm 4 with estimation error sequence $(\delta_t)_{t=1}^n$ such that $2\delta_t \leq \min_{a \neq a^*} \hat{\Delta}_t(a)$ satisfies*

$$\mathfrak{R}_n(\pi_n, \theta) \leq \mathfrak{c}^*(\beta_n + 1) + 2\mathfrak{c}^* \sqrt{2 \log(|\mathcal{C}^*|)(\beta_n + 1)}\,.$$

Moreover, when $\lim_{n\to\infty} \sum_{t=1}^{n} \delta_t = \infty$ and $\beta_n / n \to 0$, the allocation defined by the policy is asymptotically consistent.

4.2.1 *Docile Constraints*

We can also analyze the case *with* docile constraints, where $h_\nu(a^*) > 0$ for one or multiple $\nu \in \mathcal{C}^*$. Denote by $\tilde{\alpha}^*(a) = \alpha^*(a)\mathbb{1}(a \neq a^*)$ the optimal allocation on suboptimal actions. For a trade-off parameter $\eta \in [0, 1]$, we define $\tilde{\mu}_t(\eta) \triangleq (1 - \eta)e_{a^*} + \eta \tilde{\alpha}^* \|\tilde{\alpha}^*\|_1^{-1}$ and $\eta_t^* = \delta_t \|\tilde{\alpha}^*\|_1 / \mathfrak{c}^*$ for which $\hat{\Delta}_t(\tilde{\mu}_t(\eta_t^*)) = 2\delta_t$. The distribution $\tilde{\mu}(\eta)$ explicitly randomizes between a^* and the normalized allocation over sub-optimal actions prescribed by the lower bound. Further, let $h_{\min} = \min\{h_\nu(a^*) : \nu \in \mathcal{C}^*, h_\nu(a^*) > 0\}$ be the smallest docile constraint coefficient. Then, using again Eq. (2.6),

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu_t)} \leq \frac{4\delta_t \hat{\Delta}_t(\tilde{\mu}(\eta_t^*))}{I_t(\tilde{\mu}(\eta_t^*))} \leq 4\delta_t \mathfrak{c}^* \max\left(1, \frac{\delta_t (\mathfrak{c}^* h_{\min})^{-1}}{1 - \frac{\delta_t \|\tilde{\alpha}^*\|_1}{\mathfrak{c}^*}}\right). \tag{4.11}$$

The second inequality follows from noting that $\Delta(\tilde{\mu}_t(\eta)) = \eta c^* \|\tilde{\alpha}^*\|_1^{-1}$, and

$$I_t(\tilde{\mu}_t(\eta)) \geq (1 - \eta) h_{\min} + \frac{\eta}{\|\tilde{\alpha}^*\|_1} \geq \min\left((1 - \eta) h_{\min}, \frac{\eta}{\|\tilde{\alpha}^*\|_1}\right).$$

Replacing Eq. (4.9) with a marginally stronger condition $3\delta_t \leq \min_{a \neq a^*} \hat{\Delta}_t(a)$, we get $2\delta_t \|\tilde{\alpha}^*\|_1 \leq \mathfrak{c}^*$, and Eq. (4.11) simplifies to

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \frac{\hat{\Delta}_t(\mu)^2}{I_t(\mu_t)} \leq \frac{4\delta_t \hat{\Delta}_t(\tilde{\mu}(\eta_t^*))}{I_t(\tilde{\mu}(\eta_t^*))} \leq 4\delta_t \mathfrak{c}^* \max\Big(1,\, 2\delta_t (\mathfrak{c}^* h_{\min})^{-1}\Big)\,. \tag{4.12}$$

Theorem 4.4. *For an instance with docile constraints, the regret of Algorithm 4 with estimation error sequence $(\delta_t)_{t=1}^n$ such that $3\delta_t \leq \min_{a \neq a^*} \hat{\Delta}_t(a)$ satisfies*

$$\mathfrak{R}_n(\pi_n, f) \leq \sum_{t=1}^{n} \mathbb{1}(2\delta_t > h_{\min}\mathfrak{c}^*) 2\delta_t + \mathfrak{c}^*(\beta_n + 1) + 2\mathfrak{c}^* \sqrt{2\log(|\mathcal{C}^*|)(\beta_n + 1)}$$

Proof. By Lemma 2.6, we always have $\hat{\Delta}_t(\mu_t^{\mathrm{IDS}}) \leq 2\delta_t$. The regret bound follows by we separately treating time steps where $2\delta_t(\mathfrak{c}^* h_{\min})^{-1} > 1$, and using Corollary 2.3 combined with Eq. (4.12) otherwise. □

Comparing Theorems 4.3 and 4.4 reveals a discontinuity as $h_{\min} \to 0$. Note that $h_{\min} > 0$ requires to play a^* while the constraints are not yet satisfies, which means the learner is still uncertain about the identity of a^*. On the other hand, the asymptotically optimal allocation is computed in the limit where the best action is known, effectively eliminating all docile constraints at no cost. Without docile constraints the situation is different, since the cost of identifying a^* solely depends on the structure of the sub-optimal actions.

4.2.2 *Worst-Case Regret*

Interestingly, we can also obtain a bound on the regret that does not depend on the instance. By Lemma 2.6,

$$\mathfrak{R}_n(\pi, f) \leq \sum_{t=1}^{n} \hat{\Delta}_t(\mu_t) \leq 2 \sum_{t=1}^{n} \delta_t\,.$$

The bound previews how IDS maintains control on the worst-case regret in the roll-in phase where $\delta_t \leq \min_{a \neq a^*} \Delta(a)$ is not satisfied. In the statistical setting with bandit information, it is reasonable to expect that $\delta_t \leq C\sqrt{\log(t)/t}$. The bound then implies that $\mathfrak{R}_n \leq C\sqrt{n \log(n)}$.

4.2.3 *IDS as Best-Response*

In hindsight, the primal-dual connection also manifests in Lemma 2.4, which characterizes the support of the IDS distribution. After rearranging, the lemma shows that any $a \in \operatorname{supp}(\mu_t^{\mathrm{IDS}})$ is a minimizer of the function

$$g_t(a) = \hat{\Delta}_t(a) - \frac{\Psi_t(\mu_t)}{2\hat{\Delta}_t(\mu_t)} I_t(a) \overset{n\to\infty}{\approx} \hat{\Delta}_t(a) - \mathfrak{c}^* I_t(a)\,.$$

The limiting statement holds provided that $\delta_t \to 0$ and $I_t(a^*) \to 0$, which implies $\Psi_t(\mu_t) \approx 4\mathfrak{c}^*\delta_t$ and $\hat{\Delta}_t(\mu_t) \approx 2\delta_t$. Therefore, the IDS distribution can be understood as best-response in the primal-dual game defined by the normalized Lagrangian Eq. (4.4), where the dual variables are chosen by the q-learner. Lemma 2.4 also implies that the best response is not necessarily unique. The information ratio imposes a particular trade-off that for $\delta_t = \min_{a\in\mathcal{A}} \hat{\Delta}_t(a) \to 0$ leads to randomization between the greedy action $\hat{a}_t = \arg\min_{a\in\mathcal{A}} \hat{\Delta}_t(a)$ and an informative action.

4.3 CONTRIBUTIONS AND RELATED WORK

The results in this chapter are based on Appendix D in the following work:

- Kirschner, J., Lattimore, T., Vernade, C. & Szepesvári, C. *Asymptotically Optimal Information-Directed Sampling* in *Proc. International Conference on Learning Theory (COLT)* (Aug. 2021)

The exposition here is more detailed and extends the previous work by explicitly allowing docile actions. The primal-dual approach to solve covering LPs is based on work by Garg & Koenemann [63] and was previously used for regret minimization in the bandit setting by Degenne, Shao & Koolen [50]. The formulation presented here differs from this work in that it avoids a re-parametrization of the allocation.

While our analysis provides a promising plan to derive asymptotically optimal algorithms for the structured bandit setting, there are many delicate technical challenges in the stochastic estimation setting that require further ideas. We will work out the details for the linear bandit setting in the next chapter, but a more generic analysis beyond the Gaussian linear models is left for future work.

5

ASYMPTOTIC OPTIMALITY

Unlike in the multi-armed bandit setting where Thompson sampling and UCB are instance-optimal, designing asymptotically optimal algorithms for linear bandits is much more challenging. It is known that algorithms based on optimism or Thompson sampling are *not* asymptotically optimal in the linear setting [104]. This fact is illustrated in Example 5.1 below. Asymptotic optimality of Bayesian IDS was already suggested by Russo & Van Roy [135, Section 7.3] based on numerical studies on multi-armed bandits, but so far no instance-dependent analysis is known. The version of IDS that we introduced in Chapter 3 satisfies a gap-dependent bound but is quite far from instance-optimal.

We now follow the plan outlined in the previous chapter: Using the connection between IDS and the primal-dual formulation of the lower bound, we design an information gain that leads to an *asymptotically optimal version of IDS* for linear bandits. Without much additional effort, we show that the same IDS algorithm is close to minimax optimal and satisfies a finite-time gap-dependent bound with benign lower order terms. Surprisingly, the information gain that we derive from the primal-dual setup is related to the mutual information used in the Bayesian IDS algorithm in the large data limit. This hints towards a deep connection between the information theoretic analysis of Bayesian IDS and the optimal frequentist regret.

SETTING Recall that in the linear bandit setting, actions are represented by d-dimensional features $\mathcal{A} \subset \mathbb{R}^d$, and the reward function $f_\theta(a) = \langle a, \theta \rangle$ is linearly parameterized by $\theta \in \mathcal{M} \subset \mathbb{R}^d$. When the learner chooses an action $a_t \in \mathcal{A}$ in round t, the observation is $y_t = \langle a_t, \theta \rangle + \epsilon_t$, where ϵ_t is ρ-sub-Gaussian observation noise. In line with all previous work focusing on the asymptotic setting, we assume that the action set is finite with $k \triangleq |\mathcal{A}|$ and the optimal action $a^*(\theta) = \arg\max_{a \in \mathcal{A}} \langle a, \theta \rangle$ is unique. Eliminating these assumptions is a delicate and possibly nontrivial challenge left for future work. For technical reasons, we assume that the parameter set $\mathcal{M} \subset \mathbb{R}^d$ is a polytope and we require $\operatorname{diam}(\mathcal{A}) \leq 1$ and $\operatorname{diam}(\mathcal{M}) \leq 1$. Recall the definition of the sub-optimality gap $\Delta(a) = \langle a^* - a, \theta \rangle$ and the smallest gap, $\Delta_{\min} = \min_{a \neq a^*} \Delta(a)$. For actions $a, b \in \mathcal{A}$, we denote by

$\mathcal{H}_a^b = \{\nu \in \mathcal{M} : \langle a - b, \nu \rangle \geq 0\}$ the half-space of parameters where the reward of a is at least the reward of b. The set of *alternative parameters*

$$\mathcal{C}^*(\theta) = \{\nu \in \mathcal{M} : a^*(\nu) \neq a^*(\theta)\} = \cup_{a \neq a^*(\theta)} \mathcal{H}_a^{a^*(\theta)}$$

contains all parameters where the optimal action is different from $a^*(\theta)$. We omit the dependence on the instance θ when there is no ambiguity.

ASYMPTOTIC LOWER BOUND We already introduced the asymptotic lower bound for structured bandits in Theorem 4.1. Here we specialize the result to the linear setting with Gaussian noise. For an allocation $\alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0}$ over actions we define the covariance matrix $V(\alpha) = \sum_{a \in \mathcal{A}} \alpha(a) a a^\top$. Let $\mathfrak{c}^*$ be the solution to the following convex program,

$$\mathfrak{c}^*(\theta) \triangleq \inf_{\alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0}} \sum_{a \in \mathcal{A}} \alpha(a) \langle a^* - a, \theta \rangle \quad \text{s.t.} \quad \min_{\nu \in \mathcal{C}^*} \tfrac{1}{2} \|\nu - \theta\|^2_{V(\alpha)} \geq 1 \,. \tag{5.1}$$

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

Theorem 5.1 (Asymptotic Lower Bound)**.** *Any consistent algorithm π for the linear bandit setting with Gaussian noise has regret $\mathfrak{R}_n(\pi, \theta)$ at least*

$$\liminf_{n \to \infty} \frac{\mathfrak{R}_n(\theta, \pi)}{\log(n)} \geq \mathfrak{c}^*(\theta) \,.$$

The result as stated here is by Combes, Magureanu & Proutiere [43]. It follows from a more general result by Graves & Lai [68].

Example 5.1 (End of Optimism)**.** This example of a 2-dimensional linear bandit was used by Lattimore & Szepesvári [104] to show that algorithms based on optimism and Thompson sampling are not asymptotically optimal in the linear setting. The paper is titled "The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits". Since then, the instance is known as the *'end of optimism'* example, although it was already

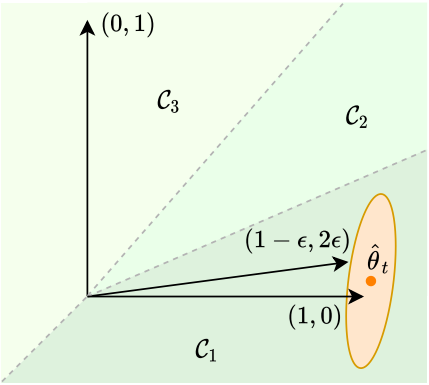


FIGURE 5.1: The 'end of optimism' example.

introduced earlier by Soare, Lazaric & Munos [150, Appendix A]. There are three arms $a_1 = (1,0)$, $a_2 = (1-\epsilon, 2\epsilon)$ and $a_3 = (0,1)$ with a tuning variable $\epsilon > 0$. The true parameter is $\theta = (1,0)$ which makes action a_1 optimal. The situation is illustrated in Fig. 5.1. The colored regions $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ depict *cells*, defined as the subset of parameters in $\mathbb{R}^2$ for which a_1, a_2 or a_3 is optimal respectively. Let $\hat{\theta}_t$ be the least-square estimate after t rounds and $V_t = \sum_{i=1}^{3} \alpha_t(i) a_i a_i^\top$ the covariance matrix, where after t rounds action $i \in \{1,2,3\}$ has been played $\alpha_t(i)$ times in total. When the confidence ellipsoid $\mathcal{E}_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_t}^2 \leq c \log(n)\}$ is contained in the cell $\mathcal{C}_1$, the learner has identified the best action with high probability. Algorithms based on optimism and Thompson sampling quickly rule out the suboptimal arm a_3 and just play either a_1 or a_2. The twist is that the third arm is still informative for determining a^*, and in fact an asymptotically optimal algorithm plays only on $\{a_1, a_3\}$. To see why, note that any no-regret learner plays a^* a lot, therefore the parameter is well-estimated along the direction $(1,0)$. It remains to shrink the confidence ellipsoid approximately along the direction $(0,1)$. Choosing arm a_2 means the learner updates the covariance with $V_{t+1} \leftarrow V_t + a_2 a_2^\top$, which implies that the V_t-norm of $(0,1)$ increases about ϵ^2 while the instantaneous regret suffered is ϵ. On the other hand, when the learner chooses a_3, the regret is 1 while the V_t-norm also increases by 1. Hence, an optimistic algorithm has asymptotic regret that scales with $\mathfrak{R}_n \approx \log(n)/\epsilon$, but the regret of an optimal algorithm is only $\mathfrak{R}_n \approx 1 \cdot \log(n)$.

Algorithm 5: Asymptotically Optimal IDS

Input: Finite action set $\mathcal{A}$

$s \leftarrow 1$

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

  $V_s \leftarrow \sum_{i=1}^{s-1} a_i a_i^\top + \mathbf{1}_d$

  $\hat{\theta}_s \leftarrow V_s^{-1} \sum_{i=1}^{s-1} a_i y_i$ `// least-squares estimate`

  $\hat{a}_s \leftarrow \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_s \rangle$ `// empirically best action`

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

  $a_s^{\mathrm{UCB}} \leftarrow \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_s \rangle + \beta_{s,1/s^2}^{1/2} \|a\|_{V_s^{-1}}$ `// UCB action`

  `// gap estimates`

  $\hat{\Delta}_s(a) \leftarrow \left(\max_{b \in \mathcal{A}} \langle b, \hat{\theta}_s \rangle + \beta_{s,1/s^2}^{1/2} \|b\|_{V_s^{-1}} \right) - \langle b, \hat{\theta}_s \rangle$

  $\hat{v}_s(c) \leftarrow \arg\min_{\nu \in \mathcal{H}_c^{\hat{a}_s}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ `// see Eq. (5.9)`

  $m_s \leftarrow \min_{c \neq \hat{a}_s} \frac{1}{2} \|\hat{v}_s(c) - \hat{\theta}_s\|_{V_s}^2$

  $\eta_s \leftarrow \min_{l \leq s} m_l^{-1/2} \log(k)$

  $q_s(c) \leftarrow \exp(-\eta_s \|\hat{v}_s(c) - \hat{\theta}_s\|_{V_s}^2)$

  `// information gain`[†]

  $I_s(a) \leftarrow \frac{1}{2} \sum_{c \neq \hat{a}_s} q_s(c) \left(|\langle \hat{v}_s(c) - \hat{\theta}_s, a \rangle| + \beta_{s,1/s^2}^{1/2} \|a\|_{V_s^{-1}} \right)^2$

  if $m_s \geq \frac{1}{2} \beta_{s,1/(t \log(t))}$ **then**

    Choose $\hat{a}_s$ `// exploitation (disregard data)`

  else

    $\mu_s \leftarrow \arg\min_{\mu \in \mathscr{P}(\mathcal{A})} \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)}$ `// IDS distribution`

    Sample $a_s \sim \mu_s$, observe $y_s = \langle a_s, \theta^* \rangle + \epsilon_s$

    $s \leftarrow s + 1$ `// exploration step counter`

[†] We normalize the q-weights in the analysis, but this is not required for the algorithm.

5.1 ASYMPTOTICALLY OPTIMAL INFORMATION-DIRECTED SAMPLING

The learner interacts the environment is on rounds $t = 1, \ldots, n$ where the horizon n is unknown. We distinguish between *exploration* and *exploitation* rounds. In exploitation rounds, a^* is identified with high probability and the algorithm plays the action it estimates to be optimal. In exploration rounds, we sample from the IDS distribution. Exploration steps are indexed by $t_1, \ldots, t_{s_n}$, where s_n is the total number of exploration rounds.

We refer to s and t as *local* and *global time* respectively, and to s_n as the *effective horizon*. To avoid double indexing, the mapping $s \mapsto t_s$ is implicit. The convention is that an s-index refers to the local time quantities, whereas a t-index refers to global time quantities. For example, the action chosen at time t_s is a_s and the observed reward is y_s. Similarly, an action a_s at local time s has a global time correspondence $a_t = a_{t_s}$. In exploration rounds, the algorithm is defined to sample the action a_s from the IDS distribution μ_s^{IDS},

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

The exploitation conditions, and the gap estimates $\hat{\Delta}_s(a)$ and information gain $I_s(a)$, defined for each $s \geq 1$, are introduced in the following. The complete procedure is summarized in Algorithm 5.

5.1.1 *Least-Squares and Exploitation Rounds*

All estimated quantities are defined using data collected in exploration rounds, whereas observation data from exploitation rounds is discarded. Let $\hat{\theta}_s \triangleq V_s^{-1} \sum_{i=1}^{s-1} a_i y_i$ be the regularized least squares estimator with covariance matrix $V_s \triangleq \sum_{i=1}^{s-1} a_i a_i^\top + \mathbf{1}_d$, computed with data $\{(a_i, y_i)\}_{i=1}^{s-1}$. The *empirically best action* is $\hat{a}_s \triangleq \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_s \rangle$. We assume that the learner has a *concentration coefficient* $\beta_{s,\delta}$ that satisfies

$$\mathbb{P}[\exists\, s \geq 1 \text{ with } \|\hat{\theta}_s - \theta\|_{V_s}^2 \geq \beta_{s,\delta}] \leq \delta\,. \tag{5.2}$$

For concreteness, we use the choice provided in Lemma 3.1, which is

$$\beta_{s,\delta}^{1/2} \triangleq \sqrt{2 \log \delta^{-1} + \log\det(V_s)} + 1\,. \tag{5.3}$$

The reader might be worried about the log determinant term, which is known to create an asymptotically suboptimal dependence on the dimension, and can be improved with a different choice of the confidence coefficient [104, 160]. On the other hand, $\beta_{s,\delta} \leq 2\log\delta^{-1} + \mathcal{O}(d\log(s))$ by Eq. (3.1), and we circumvent this shortcoming by limiting the amount of data the algorithm collects to $s_n = \mathcal{O}\big(\operatorname{poly}(\log(n))\big)$. In this case, the log determinant only contributes to lower order terms, $\log\det(V_{s_n}) \leq \mathcal{O}(d\log(s_n)) \leq \mathcal{O}(d\log\log(n))$. We also exploit this property for other steps in the analysis, but it is unclear whether it is essential or not.

For all $c \neq \hat{a}_s$, let $\hat{\nu}_s(c) = \arg\min_{\nu \in \mathcal{H}_c^{\hat{a}_s}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ be the closest parameter to $\hat{\theta}_s$ in V_s-norm for which c is better than $\hat{a}_s$. This is a strongly convex

objective over the convex set $\mathcal{H}_c^{\hat{a}_s}$, hence $\hat{v}_s(c)$ can be computed efficiently. In practice, we can drop the constraints on the parameter set (i.e. set $\mathcal{M} = \mathbb{R}^d$), in which case $\hat{v}_s(c)$ can be computed in closed form, see Eq. (5.9) below. Exploitation rounds t (with corresponding local time $s = s_t$) are defined by the *exploitation condition*,

$$m_s \triangleq \tfrac{1}{2} \min_{c \neq \hat{a}_s} \|\hat{v}_s(c) - \hat{\theta}_s\|_{V_s}^2 \geq \tfrac{1}{2} \beta_{s_t, 1/(t\log(t))} \triangleq \tfrac{1}{2} \beta_t^{\mathrm{glob}} , \tag{E}$$

which guarantees that with probability $(t\log(t))^{-1}$ there exists no plausible alternative parameter $\nu \neq \hat{\theta}_s$, such that an action $a \neq \hat{a}_s$ is optimal for ν.

5.1.2 *Gap Estimates*

Define $\beta_s \triangleq \beta_{s,1/s^2}$. At local time s, the *gap estimate* is

$$\hat{\Delta}_s(a) \triangleq \max_{b \in \mathcal{A}} \langle b - a, \hat{\theta}_s \rangle + \beta_s^{1/2} \|b\|_{V_s^{-1}} . \tag{5.4}$$

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

$$\begin{aligned} \Delta(a) &= \max_{b \in \mathcal{A}} \langle b - a, \theta \rangle = \max_{b \in \mathcal{A}} \langle b - a, \theta - \hat{\theta}_s \rangle + \langle b - a, \hat{\theta}_s \rangle \\ &\overset{(i)}{\leq} \max_{b \in \mathcal{A}} \|b - a\|_{V_s^{-1}} \|\theta - \hat{\theta}_s\|_{V_s} + \langle b - a, \hat{\theta}_s \rangle \\ &\overset{(ii)}{\leq} \max_{b \in \mathcal{A}} \langle b, \hat{\theta}_s \rangle + \beta_s^{1/2} \|b\|_{V_s^{-1}} - (\langle a, \hat{\theta}_s \rangle - \beta_s^{1/2} \|a\|_{V_s^{-1}}) \overset{(iii)}{\leq} 2\hat{\Delta}_s(a) . \end{aligned} \tag{5.5}$$

Inequality (i) follows from the Cauchy-Schwarz inequality, (ii) uses the definition of the confidence scores and the triangle inequality, and (iii) uses $\hat{\Delta}_s(a) \geq \beta_s^{1/2} \|a\|_{V_s^{-1}}$. The gap estimate of the empirically best action $\hat{a}_s$ is $\delta_s \triangleq \hat{\Delta}_s(\hat{a}_s)$. Equivalently, the gap estimate can be written as $\hat{\Delta}_s(a) = \langle \hat{a}_s - a, \hat{\theta}_s \rangle + \delta_s$, and therefore we refer to δ_s as the *estimation error*. The UCB action is $a_s^{\mathrm{UCB}} \triangleq \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_s \rangle + \beta_s^{1/2} \|a\|_{V_s^{-1}}$.

The choice of the confidence coefficient in exploration and exploitation rounds is justified in the following lemma. It shows that the regret accumulated in rounds where the estimate is inaccurate is negligible.

Lemma 5.1. *Define the indicator* $B_s = \mathbb{1}\left(\beta_s \geq \|\hat{\theta}_s - \theta\|^2_{V_s}\right)$ *for rounds* s *where the confidence bounds at level* β_s *are valid. Let* $\Delta_{\max} = \max_{a \in \mathcal{A}} \Delta(a)$*. Then*

$$\mathfrak{R}_n \leq \mathbb{E}\left[\sum_{s=1}^{s_n} \Delta_s(a_s) B_s\right] + \mathcal{O}\left(\Delta_{\max} \log\log(n)\right).$$

Proof. Abbreviate $\chi_s \triangleq \|\hat{\theta}_s - \theta\|^2_{V_s}$. Naturally, the regret decomposes into exploration and exploitation rounds.

$$\begin{aligned}
\mathfrak{R}_n &= \mathbb{E}\left[\sum_{s=1}^{n} \Delta(a_t)\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\sum_{s=1}^{s_n} \Delta(a_s) B_s + \Delta_{\max}\Big(\sum_{s=1}^{s_n} \mathbb{1}(\beta_s < \chi_s) + \sum_{t=1}^{n} \mathbb{1}\left(\beta_t^{\text{glob}} < \chi_{s_t}\right)\Big)\right] \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\sum_{s=1}^{s_n} \Delta(a_s) B_s + \Delta_{\max} \sum_{s=1}^{s_n} \frac{1}{s^2} + \Delta_{\max} \sum_{t=1}^{n} \frac{1}{t \log t}\right] \\
&\overset{(iii)}{\leq} \mathbb{E}\left[\sum_{s=1}^{s_n} \Delta(a_s) B_s\right] + \mathcal{O}(\Delta_{\max} \log\log(n)).
\end{aligned}$$

For (i) we used the boundedness assumption on the gaps. For (ii), note that by Eq. (5.2), $\mathbb{P}[\beta_s < \chi_s] \leq \frac{1}{s^2}$ and $\mathbb{P}[\beta_t^{\text{glob}} < \chi_{s_t}] < \frac{1}{t \log t}$. Lastly, (iii) bounds the sums. □

5.1.3 *Information Gain*

Recall that $\hat{v}_s(c) = \arg\min_{\nu \in \mathcal{H}_c^{\hat{a}_s}} \|\nu - \hat{\theta}_s\|^2_{V_s}$ is the closest alternative to $\hat{\theta}_s$ in V_s-norm for which $\hat{a}_s$ is not optimal. The *asymptotic information gain* is

$$I_s^{\text{A}}(a) \triangleq \frac{1}{2} \sum_{c \neq \hat{a}_s} q_s(c) \left(|\langle \hat{v}_s(c) - \hat{\theta}_s, a\rangle| + \beta_s^{1/2} \|a\|_{V_s^{-1}}\right)^2, \tag{5.6}$$

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

$$q_s(c) \propto \begin{cases} 0 & \text{if } c = \hat{a}_s, \\ \exp\left(-\frac{\eta_s}{2} \|\hat{v}_s(c) - \hat{\theta}_s\|^2_{V_s}\right) & \text{otherwise}. \end{cases} \tag{5.7}$$

As we explained in Chapter 4, the q-weights are interpreted as dual variables that are chosen by an exponential weights learner. The learning rate is

set to $\eta_s \triangleq \min_{l \le s} m_l^{-1/2} \log(k)$, where $m_s \triangleq \frac{1}{2} \min_{c \neq \hat{a}_s} \|\hat{v}_s(c) - \hat{\theta}_s\|^2_{V_s}$. The definition of the learning rate is to ensure that η_s is monotonically decreasing, which is used explicitly in the proof. The weights q_s can be understood as soft-min approximation of the minimum constraint value,

$$m_s \le \frac{1}{2} \sum_{c \neq \hat{a}_s} q_s(c) \|\hat{v}_s(c) - \hat{\theta}_s\|^2_{V_s} \le m_s + \frac{\log(k)}{\eta_s} . \tag{5.8}$$

The statement follows from Lemma 5.2 below. We remark that the $\log(k)$ factor in the learning rate is chosen to cancel the $\log(k)$ in the soft-min bound, which makes our worst-case regret bound independent of k. Other trade-offs that affect the lower order terms in the analysis are possible. Since η_s appears in multiple bounds, there is no uniquely optimal choice.

Note that up to the standard boundedness assumptions, the proposed algorithm is essentially hyper-parameter free. Nonetheless, the confidence parameter $\beta_{s,1/\delta}$ and the learning rate η_s used in the definition of I_s provide some tuning knobs to improve performance in practice. Empirically, the exact value of η_s has only a minor impact on the performance in our experiments, as shown in Figs. 5.3 and 5.4 and discussed in Section 5.3.

Lemma 5.2 (Softmin approximation)**.** *$A_1, \dots A_k \ge 0$ be a sequence of positive numbers and $A_{\min} = \min_{i \in [k]} A_i$. Let $q_i \propto \exp(-\eta A_i)$ be exponential mixing weights with $\eta > 0$. Then*

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

Furthermore, the mixing weights q_i are bounded as follows,

$$\frac{1}{k} \exp\big(-\eta(A_i - A_{\min})\big) \le q_i \le \exp\big(-\eta(A_i - A_{\min})\big) .$$

Proof. Let $\psi^*_\eta(A) = \frac{1}{\eta} \log\Big(\sum_{i \in [k]} \exp(\eta A_i)\Big)\Big)$ be the Fenchel conjugate of the normalized entropy function. A small calculation confirms that $q = \nabla_A \psi^*_\eta(-A)$. By convexity of ψ^*_η,

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

The last inequality follows from

$$\begin{aligned} \psi^*_\eta(-A) &= \eta^{-1} \log\Big(\sum_i \exp(-\eta A_i)\Big) \\ &\ge \eta^{-1} \log\Big(\exp(-\eta \min_i A_i)\Big) = -A_{\min} . \end{aligned}$$

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

$$\exp(-\eta A_{\min}) \leq \sum_i \exp(-\eta A_i) \leq k \exp(-\eta A_{\min}) . \qquad \square$$

5.1.4 *Computational Aspects*

There are three kinds of operations in the algorithm. First, using elementary matrix operations, we can update V_s^{-1}, $\det(V_s)$ and $\hat{\theta}_s$ incrementally (Lemma D.2), and note that the s-index terms only need to be updated after exploration rounds. It can be checked that $\mathcal{O}(kd^2 s_n)$ operations are needed over all n rounds to compute this part. Second, the optimization problem that defines the alternative parameters $\hat{v}_s(c)$ is a quadratic program with d variables and linear constraints $\langle \hat{v}_s(c), c - \hat{a}_s \rangle \geq 0$ and $\hat{v}_s(c) \in \mathcal{M}$. Such optimization problems can be solved very efficiently in practice and in $O(ld^3)$ time in the worst case for model sets $\mathcal{M}$ with l constraints. Note, the analysis suggests that we can tolerate an additive numerical error on the information gain of order $\mathcal{O}((s \log(s))^{-1})$. In practice, we can drop the constraints on $\mathcal{M}$, in which case $\hat{v}_s(c)$ has an analytical closed form,

$$\hat{v}_s(c) = \hat{\theta}_s - \frac{\langle \hat{\theta}_s, \hat{a}_s - c \rangle}{\|\hat{a}_s - c\|^2_{V_s^{-1}}} V_s^{-1} (\hat{a}_s - c) . \tag{5.9}$$

Third, computation of the IDS distribution is done in $\mathcal{O}(k^2)$ steps or approximated in $\mathcal{O}(k)$ as discussed in Section 2.1.1. A closer inspection of the asymptotic analysis reveals that the bound is attained on a distribution that randomizes between the greedy action $\hat{a}_s$ and some other (informative) action, hence the approximate IDS distribution is sufficient to achieve asymptotic optimality.

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

5.1.5 *Regret Bounds*

The regret bounds for Algorithm 5 come in three flavours:

- In Theorem 5.2, we show a rate-optimal worst-case regret bound of $\mathfrak{R}_n \leq \mathcal{O}(d\sqrt{n}\log(n))$. The proof bounds the information ratio by a constant, and the information gain using the elliptic potential lemma.

- In Theorem 5.3, we derive a gap-dependent logarithmic regret bound of $\mathfrak{R}_n \le \mathcal{O}\big(d^3 \Delta_{\min}^{-1} \log(n)^2\big)$. The proof uses an instance-dependent bound on the information ratio, that is attained by a distribution closely related to Thompson sampling.
- In Theorem 5.4, we show that the proposed algorithm is asymptotically optimal, that is $\mathfrak{R}_n \le \mathfrak{c}^* \log(n) + o(\log(n))$. The result relies on improving the bounds on the information ratio and the information gain in the asymptotic regime.

Theorem 5.2 (Worst-case regret)**.** *The regret of Algorithm 5 is bounded by*

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

We remark that our bound matches the bound of UCB, but is worse by a factor $\sqrt{d}$ than basic elimination algorithms [103, §23], that achieve $\mathcal{O}(\sqrt{d \log(k) n})$ when the number of actions is small. Before the proof, we show a worst-case bound on the total information $\gamma_n = \sum_{s=1}^{s_n} I_s^{\mathrm{A}}(a_s)$.

Lemma 5.3 (Total information)**.** *For any sequence of actions $a_1, \dots, a_{s_n}$, the total information gain $\gamma_n = \sum_{s=1}^{s_n} I_s^A(a_s)$ is bounded as follows,*

$$\gamma_n \le 2\Big(\beta_n^{\mathrm{glob}} + \big(\beta_n^{\mathrm{glob}}\big)^{1/2} + \beta_{s_n}\Big) d \log(s_n) \le \mathcal{O}\big((d \log(n))^2\big)\,.$$

Proof. Note that

$$\begin{aligned}
\sum_{s=1}^{s_n} I_s^{\mathrm{A}}(a_s) &= \tfrac{1}{2} \sum_{s=1}^{s_n} \Bigg(\sum_{c \neq \hat{a}_s} q_s(c) |\langle \hat{v}_s(c) - \hat{\theta}_s, a_s \rangle| + \beta_s^{1/2} \|a_s\|_{V_s^{-1}} \Bigg)^2 \\
&\le \sum_{s=1}^{s_n} \sum_{c \neq \hat{a}_s} q_s(c) \langle \hat{v}_s(c) - \hat{\theta}_s, a_s \rangle^2 + \beta_s \|a_s\|_{V_s^{-1}}^2 \\
&\overset{(i)}{\le} \sum_{s=1}^{s_n} \sum_{c \neq \hat{a}_s} q_s(c) \big(\|\hat{v}_s(c) - \hat{\theta}_s\|_{V_s}^2 + \beta_s \big) \|a_s\|_{V_s^{-1}}^2 \\
&\overset{(ii)}{\le} \sum_{s=1}^{s_n} \left(\min_{c \neq \hat{a}_s} \|\hat{v}_s(c) - \hat{\theta}_s\|_{V_s}^2 + \frac{2 \log(k)}{\eta_s} + \beta_s \right) \|a_s\|_{V_s^{-1}}^2 \\
&\overset{(iii)}{\le} \Big(\beta_n^{\mathrm{glob}} + (\beta_n^{\mathrm{glob}})^{1/2} + \beta_{s_n} \Big) \sum_{s=1}^{s_n} \|a_s\|_{V_s^{-1}}^2 \\
&\overset{(iv)}{\le} \Big(\beta_n^{\mathrm{glob}} + (\beta_n^{\mathrm{glob}})^{1/2} + \beta_{s_n} \Big) 2d \log\det(V_{s_n})\,.
\end{aligned}$$

Step (i) uses the Cauchy-Schwarz inequality, and (ii) the soft-min bound for the q-weights in Eq. (5.8). For (iii), $m_s = \frac{1}{2}\min_{c\neq\hat{a}_s}\|\hat{v}_s(c) - \hat{\theta}_s\|^2_{V_s} \leq \frac{1}{2}\beta_n^{\text{glob}}$ holds by definition in all exploration rounds and the choice $\eta_s = \min_{l\leq s} m_l^{-1/2}\log(k)$ and lastly, (iv) bounds the elliptic potential as stated below in Lemma 5.4. Considering that $\log\det(V_{s_n}) \leq \mathcal{O}(d\log(s_n))$ and $\beta_{s_n,1/\delta} = 2\log\frac{1}{\delta} + \mathcal{O}(d\log(s_n))$ completes the proof. □

Lemma 5.4 (Elliptical potential lemma)**.** *Assume that* $\|a_s\|^2_{V_s^{-1}} \leq 1$ *and* $\|a_s\| \leq 1$*. Then*

$$\sum_{s=1}^{s_n}\|a_s\|^2_{V_s^{-1}} \leq 2\log\det(V_{s_n}) \leq 2d\log\Big(1 + \frac{s_n}{d}\Big)\,.$$

A proof can be found in [3, Lemma 11]. Note that by $\operatorname{diam}(\mathcal{A}) \leq 1$ and the choice $V_0 = \mathbf{1}_d$, the assumptions of the lemma are always satisfied for our setting.

Proof of Theorem 5.2. Provided an almost-sure bound $\Psi_s(\mu_s^{\text{IDS}}) \leq \alpha_0$ on the information ratio, by Corollary 2.1 the regret satisfies

$$\mathfrak{R}_n \leq \sqrt{\alpha_0\mathbb{E}[\gamma_n]n} + \sum_{t=1}^{n}\mathbb{E}\big[\Delta(a_t) - \hat{\Delta}_t(a_t)\big]\,,$$

Lemma 5.1 shows that with our choice of confidence level, the error term is at most $\mathcal{O}(\Delta_{\max}\log\log(n))$ where $\Delta_{\max} = \max_{a\in\mathcal{A}}\Delta(a)$ is the maximum gap. Further, the total information gain satisfies $\gamma_n = \sum_{s=1}^{s_n} I_s^{\text{A}}(a_s) \leq \mathcal{O}(d^2\log(n)^2)$ by Lemma 5.3 above. We complete the proof with a bound on the information ratio. Since μ_s^{IDS} is chosen by IDS to minimize Ψ_s,

$$\Psi_s(\mu_s^{\text{IDS}}) = \min_{\mu\in\mathscr{P}(\mathcal{A})}\Psi_s(\mu) \leq \frac{\hat{\Delta}_s(a_s^{\text{UCB}})^2}{I_s^{\text{A}}(a_s^{\text{UCB}})} \leq 2\,. \tag{5.10}$$

The last inequality follows from the fact that $\hat{\Delta}_s(a_s^{\text{UCB}}) = \beta_s^{1/2}\|a_s^{\text{UCB}}\|_{V_s^{-1}}$ and bounding

$$\begin{aligned} I_s^{\text{A}}(a_s^{\text{UCB}}) &= \frac{1}{2}\sum_{c\neq\hat{a}_s} q_s(c)\big(|\langle\hat{v}_s(c) - \hat{\theta}_s, a_s^{\text{UCB}}\rangle| + \beta_s^{1/2}\|a_s^{\text{UCB}}\|_{V_s^{-1}}\big)^2 \\ &\geq \tfrac{1}{2}\beta_s\|a_s^{\text{UCB}}\|^2_{V_s^{-1}}\,, \end{aligned}$$

where we used that q_s is defined as a distribution supported on $\mathcal{A}\setminus\hat{a}_s$. □

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

Theorem 5.3 (Gap-dependent regret)**.** *The regret of Algorithm 5 is bounded by*

$$\mathfrak{R}_n \leq \mathcal{O}\Big(\Delta_{\min}^{-1} d^3 \log(n)^2\Big)\,.$$

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

Lemma 5.5. *Let s be a local time with $\|\hat{\theta}_s - \theta\|_{V_s}^2 \leq \beta_s$. Then the optimal information ratio is bounded as follows,*

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

Proof. Let $\lambda \geq 2$ be a constant to be chosen later. If $2\lambda\delta_s \geq \Delta_{\min}$, then $\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_s(\mu) \leq \frac{4\lambda\delta_s}{\Delta_{\min}}$ by (5.10). Hence we may assume $2\lambda\delta_s \leq \Delta_{\min}$ in the following. By Eq. (5.5), for all s with $\|\hat{\theta}_s - \theta\|_{V_s}^2 \leq \beta_s$ and $a \neq a^*$, it holds that $\Delta_{\min} \leq 2\hat{\Delta}_s(a)$, so in particular $\hat{a}_s = a^*$. Define $\tilde{\mu}_s = \frac{1}{2}e_{\hat{a}_s} + \frac{1}{2}q_s$ to be the uniform mixture[1] of q_s and a Dirac at $\hat{a}_s$. Let $\bar{\Delta}_s(a) = \langle \hat{\theta}_s, \hat{a}_s - a \rangle$ and note that $\bar{\Delta}(\tilde{\mu}_s) \geq (\lambda - 1)\delta_s \geq \delta_s$ by the assumption $\lambda \geq 2$. Therefore, by Lemma 2.5 and Eq. (2.6),

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_s(\mu) \leq \min_{p \in [0,1]} \frac{(1-p)\delta_s + p\hat{\Delta}_s(\tilde{\mu}_s)}{pI_s(\tilde{\mu}_s)} \leq \frac{4\delta\bar{\Delta}_s(\tilde{\mu}_s)}{I_s(\tilde{\mu}_s)}\,. \tag{5.11}$$

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

$$\begin{aligned} I_s^{\mathrm{A}}(\tilde{\mu}_s) &\geq \frac{1}{2}\sum_{a \in \mathcal{A}} \tilde{\mu}_s(a) \sum_{c \neq \hat{a}_s} q_s(c) \langle \hat{v}_s(c) - \hat{\theta}_s, a \rangle^2 \\ &= \frac{1}{2}\sum_{c \neq \hat{a}_s} q_s(c) \min_{\nu \in \mathcal{H}_c^{\hat{a}_s}} \|\nu - \hat{\theta}_s\|_{V(\tilde{\mu}_s)}^2\,. \end{aligned} \tag{5.12}$$

1 With a Laplace approximation, the weights $q_s(c)$ correspond to the posterior probability of an action c being preferred over $\hat{a}_s$ by the Bayesian model with Gaussian prior and likelihood (Section 5.1.7). As such, the distribution $\tilde{\mu}_s$ resembles the *top-two Thompson sampling* approach proposed by Russo [133].

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

$$\begin{aligned}\langle \hat{\theta}_s, \hat{a}_s - a \rangle &= \min_{\nu : \langle \nu, a - \hat{a}_s \rangle \geq 0} \|\nu - \hat{\theta}_s\|_{V(\tilde{\mu}_s)} \|\hat{a}_s - a\|_{V(\tilde{\mu}_s)^{-1}} \\ &\leq \min_{\nu \in \mathcal{H}_a^{\hat{a}_s}} \|\nu - \hat{\theta}_s\|_{V(\tilde{\mu}_s)} \|\hat{a}_s - a\|_{V(\tilde{\mu}_s)^{-1}} .\end{aligned}$$

The inequality follows from $\mathcal{H}_a^{\hat{a}_s} \subset \{v : \langle v, a - \hat{a}_s \rangle \geq 0\}$. Combining the last display with the definition of $\tilde{\mu}_s$, the fact that $\hat{a}_s = a^*$ and Cauchy-Schwarz,

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

For (i) we used that $\sum_{a \neq \hat{a}_s} q_s(a) \|a\|^2_{V(\tilde{\mu})^{-1}} \leq 2 \sum_{a \neq \hat{a}_s} q_s(a) \|a\|^2_{V(q_s)^{-1}} = 2d$ and $\|\hat{a}_s\|^2_{V(\tilde{\mu}_s)^{-1}} \leq 2$, and (ii) follows from Eq. (5.12). Next, for $a \neq \hat{a}_s$,

$$\bar{\Delta}_s(a) = \hat{\Delta}_s(a) - \delta_s \geq \frac{1}{2}\Delta_{\min} - \delta_s \geq \frac{1}{2}\Big(1 - \frac{1}{\lambda}\Big)\Delta_{\min} .$$

By the definition of $\tilde{\mu}_s$ we have $\bar{\Delta}_s(\tilde{\mu}_s) \geq \frac{1}{4}(1 - 1/\lambda)\Delta_{\min}$ and with Eq. (5.11),

$$\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_s(\mu) \leq \frac{4\delta_s \bar{\Delta}_s(\tilde{\mu}_s)}{I_s(\tilde{\mu}_s)} = \frac{4\delta_s \bar{\Delta}_s(\tilde{\mu}_s)^2}{\bar{\Delta}_s(\tilde{\mu}_s) I_s(\tilde{\mu}_s)} \leq \frac{32\delta_s(1+d)}{\Delta_{\min}\big(1 - \frac{1}{\lambda}\big)} .$$

The claim follows with $\lambda = 8(1+d) + 1$. □

Proof of Theorem 5.3. Denote $\chi_s \triangleq \|\hat{\theta}_s - \theta\|^2_{V_s}$. Using the bound on the information ratio from Lemma 5.5 and Theorem 2.1, we get

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(a_s) \mathbb{1}(\chi_s \leq \beta_s)\right]^2 \leq \frac{4(8d+9)}{\Delta_{\min}} \mathbb{E}\left[\sum_{t=1}^{n} \delta_s\right] \mathbb{E}[\gamma_n] .$$

Using that $\delta_s \leq \hat{\Delta}_s(a_s)$ and solving for the regret yields

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \hat{\Delta}_s(a_s) \mathbb{1}(\chi_s \leq \beta_s)\right] \leq \frac{4(8d+9)}{\Delta_{\min}} \mathbb{E}[\gamma_n] .$$

The estimation error towards the actual regret is bounded by Lemma 5.1, and the claim follows with $\Delta(a_s)\mathbb{1}(\chi_s \leq \beta_s) \leq 2\hat{\Delta}_s(a_s)\mathbb{1}(\chi_s \leq \beta_s)$. □

Our next result shows that Algorithm 5 is asymptotically optimal. The proof requires that the optimal action is not zero, which is used in Lemma 5.8 to bound the effective horizon s_n. The lemma states that the logarithm of the expected horizon is a lower order term, $\mathbb{E}[\log(s_n)] \leq \mathcal{O}(\log\log(n)\|a^*\|^{-1})$, which, for example, is used to bound the log-determinant term in the confidence coefficient Eq. (5.2). We point out that for $\|a^*\| = 0$, the geometry of the lower bound is fundamentally different, because the optimal action provides no information. Whether the assumption is necessary or an artifact of the analysis remains to be determined. Alternatively, we can also replace the gap estimates with thresholded gaps $\hat{\Delta}_s^+(a) = \langle \hat{a}_s - a, \hat{\theta}_s \rangle + \delta_s^+$, where $\delta_s^+ = \max(\delta_s, s^{-1/2})$. We think that with this definition of the gap estimate and minor changes in the proofs, the statement of the next theorem holds without restrictions and Theorems 5.2 and 5.3 remain valid. Since it is unclear if the assumptions is required and to keep the exposition simple, we present analysis for the gap estimates without thresholding.

Theorem 5.4 (Asymptotic regret)**.** *Algorithm 5 is asymptotically optimal,*

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

where $\mathfrak{c}^*$ is the solution to the lower bound Eq. (5.1) *and we assume that $\|a^*\| > 0$.*

The proof exploits the primal-dual interpretation from Chapter 4 and is given in Section 5.2. In particular, we show that the information ratio satisfies $\Psi_s(\mu_s) \leq 4\delta_s\big(\mathfrak{c}^* + \mathcal{O}(\chi_s^{1/2} m_s^{-1/2} + \delta_s)\big)$ asymptotically, where $\chi_s = \|\hat{\theta}_s - \theta\|_{V_s}^2$ is the self-normalized estimation error. Further, we relate the definition of the information gain to the regret of an exponential weights learner, which leads to an improved bound on the total information gain $\gamma_n = \sum_{s=1}^{s_n} I_s^{\mathrm{A}}(a_s) \leq \log(n) + o(\log(n))$. Up to a few technicalities, the regret bound then follows from Corollary 2.3.

A noteworthy feature of the proof is that it avoids all but one concentration inequality (used in the definition of the gap estimate), which makes many steps of the analysis significantly simpler. We also remark that the lower order terms on the regret can be obtained explicitly from the analysis. Similar to the concurrent work by Tirinzoni *et al.* [160], the lower order terms contain additive polynomial functions of instance-dependent quantities such as the inverse minimum gap. When included as a multiplicative constant in front of the $\log(n)$ term, these terms create an exponential dependency on some instance-dependent parameters. However, it should be stressed that the gap-dependent bound in Theorem 5.3 has much milder

dependencies in the lower order terms, suggesting that the transition to the asymptotic regime is in fact well controlled.

5.1.6 *Optimistic Information Gain*

Our definition of the information gain ensures that asymptotically,

$$I_s^{\mathrm{A}}(a) \approx \frac{1}{2} \sum_{c \neq \hat{a}_s} q_s(c) \langle \hat{v}_s(c) - \hat{\theta}_s, a \rangle^2 .$$

In finite time, however, the mean estimates can be inaccurate. The definition of the information gain in Eq. (5.6) can be interpreted as an optimistic version of the asymptotic term. This term is an essential ingredient in the proof of Theorem 5.2, which in turn is used in the logarithmic bounds. Note that the optimistic term is closely related to the information gain that we used in Chapter 3. Since this choice is mainly motivated from a worst-case perspective, empirically it can lead to over-exploration in the finite-time regime.

A closer inspection of the worst-case regret proof, in particular Eq. (5.10), reveals that the optimistic term is only needed for the UCB action. This motivates the following definition:

$$I_s^{\text{A-UCB}}(a) = \frac{1}{2} \sum_{c \neq \hat{a}_s} q_s(c) \left(|\langle \hat{v}_s(c) - \hat{\theta}_s, a \rangle| + \mathbb{1}\left(a = a_s^{\mathrm{UCB}}\right) \beta_s^{1/2} \|a\|_{V_s^{-1}} \right)^2 . \tag{5.13}$$

The optimistic term for the UCB actions is similar to the variant that we already proposed in Section 3.3.3. With a few additional steps in the proof of Lemma 5.10 and Theorem 5.4, the resulting algorithm is shown to satisfy the same regret bounds as presented in Theorems 5.2 to 5.4. Since the proofs are very similar, we omit the details. We compare both information gain functions numerically in Section 5.3.

Another idea is to tighten the definition of the alternative parameters. The *cell* of $a \in \mathcal{A}$ is the set $\mathcal{C}_a = \{\nu \in \mathcal{M} : \max_{c \in \mathcal{A}} \langle \nu, c - a \rangle = 0\}$ that contains all parameters ν with $a^*(\nu) = a$. This motivates the definition

$$\tilde{v}_s(a) = \operatorname*{arg\,min}_{\nu \in \mathcal{C}_a} \|\nu - \hat{\theta}_s\|_{V_s}^2 .$$

For $c \neq \hat{a}_s$ we let $\tilde{q}(c) \propto \exp(-\eta \|\tilde{v}_s(c) - \hat{\theta}_s\|_{V_s}^2)$ and define

$$I_s^{\text{A-CELL}}(a) \triangleq \frac{1}{2} \sum_{c \neq \hat{a}_s} \tilde{q}_s(c) \left(|\langle \tilde{v}_s(c) - \hat{\theta}_s, a \rangle| + \beta_s^{1/2} \|a\|_{V_s^{-1}} \right)^2 . \tag{5.14}$$

Note, all bounds that we obtain hold true for IDS defined with $I_s^{\text{A-CELL}}$ as well. The only change in the proof is to replace $\mathcal{H}_a^{\hat{a}_s}$ with $\mathcal{C}_a$. This is possible because $\mathcal{C}^* = \cup_{a \neq a^*} \mathcal{C}_a = \cup_{a \neq a^*} \mathcal{H}_a^{a^*}$, hence the change is simply a different decomposition of the set of alternative parameters $\mathcal{C}^*$ into convex regions. One might expect faster convergence from the fact that $\tilde{q}_s$ is more concentrated, but empirically we find little difference compared to I_s. The numerical results are in Fig. 5.7 and discussed below in Section 5.3. On the other hand, for unconstrained parameter sets $\mathcal{M}$, we can compute $\hat{v}_s(c)$ in closed form (Eq. 5.9), whereas $\tilde{v}_s(c)$ can only be computed by solving a positive definite quadratic program with k linear constraints for each action $c \neq \hat{a}_s$. Interestingly, however, the information gain (5.14) is related to the Bayesian mutual information $\mathbb{I}_s(y_s; a^* | a_s = a)$ as we explain next.

5.1.7 *Approximations of the Mutual Information*

We introduced the Bayesian IDS algorithm in Section 2.3. The information gain function that was primarily analyzed in the Bayesian framework by Russo & Van Roy [135] is the mutual information

$$I_t^{\text{MI}}(a) = \mathbb{I}_t(y_t; a^* | a_t = a) = \mathbb{H}_t(a^*) - \mathbb{H}_t(a^* | y_t, a_t = a)\,.$$

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

As a remedy, Russo & Van Roy [135] proposed the following *variance-based* information gain

$$I_t^{\text{VAR}}(a) \triangleq \mathbb{E}_t\Big[\big(\mathbb{E}_t[\langle a, \theta \rangle | a^*] - \mathbb{E}_t[\langle a, \theta \rangle]\big)^2\Big] = \mathbb{E}_t\Big[\langle \bar{v}_t(a^*) - \hat{\theta}_t, a \rangle^2\Big]\,. \quad (5.15)$$

The last step uses that $\mathbb{E}_t[\theta] = \hat{\theta}_t$ and we defined $\bar{v}_t(a) = \mathbb{E}_t[\theta | a^* = a]$. Russo & Van Roy further showed that the variance-based information gain is a lower bound to the mutual information, $I_t^{\text{MI}}(a) \geq 2 I_t^{\text{VAR}}(a)$, while, at the same time, the information ratio is still bounded in the Bayesian setting with linear reward [135, Proposotion 7]. Importantly, Eq. (5.15) can be approximated for a moderate number of actions using samples from the posterior distribution.

The variance-based information gain is related to $I_t^{\text{A-CELL}}$ in Eq. (5.14). We compute the posterior probability $\bar{q}_t(c) \triangleq \mathbb{P}_t[a^* = c]$ with a Laplace approximation of the integral over the cell $\mathcal{C}_c = \{\theta \in \mathcal{M} : a^*(\theta) = c\}$,

$$\begin{aligned}\bar{q}_t(c) &= \frac{1}{\sqrt{(2\pi)^d \det(V_t)}} \int_{\mathcal{C}_c} \exp\Big(-\tfrac{1}{2}\|\nu - \hat{\theta}_t\|_{V_t}^2\Big) d\nu \\ &\approx Q_c^{-1} \exp\Big(-\tfrac{1}{2}\|\tilde{\nu}_t(c) - \hat{\theta}_t\|_{V_t}^2\Big)\,,\end{aligned}$$

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

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

which resembles the definition of the cell-based information gain in Eq. (5.14).

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

$$\bar{q}_t^\tau(c|a) \triangleq \bar{q}_t(c) \exp\Big(-\tfrac{\tau}{2} \langle \tilde{\nu}_t(c) - \hat{\theta}_t, a \rangle^2\Big)\,,$$

and $q_t^\tau(\bar{a}_t^*|a) \triangleq 1 - \sum_{c \neq \bar{a}_t^*} q_t^\tau(c|a)$. Using the approximate posterior probabilities, the entropy reduction up to first order is

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

Using that $-x \log x \geq x$ for $x \ll 1$, the last expression can be lower bounded to arrive at a form similar to the cell-based information gain Eq. (5.14).

Lastly, we numerically compare the different information gain functions on Example 5.1, where the UCB strategy is asymptotically suboptimal. The result is in Fig. 5.2 for the following information gain functions:

- $I_t^{\text{A-CELL}}$, defined in Eq. (5.14).

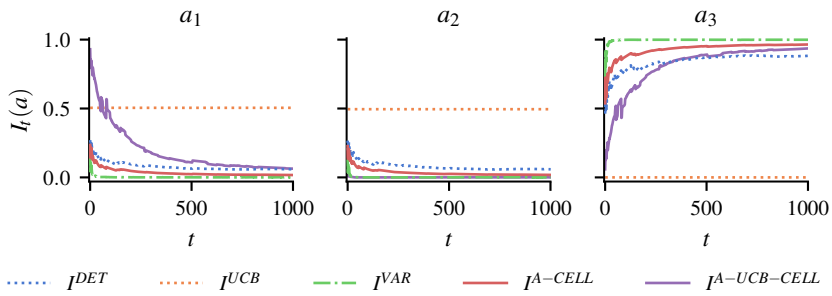


FIGURE 5.2: Comparison of information gain functions on the 'end of optimism' example with $\epsilon = 0.01$. The information gain functions are evaluated on the same trajectory of IDS with $I_t^{\text{A-UCB}}$, and normalized such that $\sum_{a \in A} I_t(a) = 1$. On this instance, a_1 is optimal, a_2 is ϵ-suboptimal, and a_3 is 1-suboptimal, but asymptotically more informative than action a_2. The learner immediately identifies action a_3 as suboptimal, and the UCB action is one of the first two actions. Hence I_t^{UCB}, which measures entropy reduction of the UCB action, is large for actions a_1 and a_2. All other information gain functions assign a vanishing score to a_1, because a_1 is played a $\Omega(t)$ times and consequently the direction is well estimated. Also visible is that the lower-order terms of the I_t^{A} and $I_t^{\text{A-UCB}}$ are increasingly dominated by the asymptotic term. I_t^{VAR} is approximated using 10^4 samples from the posterior distribution, and converges much faster than the information gain functions based on the q-learner, which uses a more conservative learning rate. Not shown is that the approximation with posterior samples is unstable on a larger horizon without increasing the number of samples accordingly. Finally, I_t^{DET} shows a similar behaviour as the asymptotic information gain functions on this example, but it is easy to construct action sets where they differ.

- $I_t^{\text{A-UCB-CELL}}$, as in Eq. (5.13) with cell-based alternatives.
- I_t^{VAR}, defined in Eq. (5.15).
- I_t^{DET}, defined in Eq. (3.6).
- I_t^{UCB}, defined in Eq. (3.13).

The asymptotic information gain based on half-spaces in Eqs. (5.6) and (5.13) is not shown since it was empirically identical with the cell-based variant.

While our reasoning here is rather informal, we think that it warrants a more formal investigation in the future. Such results could be fruitful in two directions. First, interpreting the mutual information as an approximation of a dual loss could lead to an instance-dependent analysis for the

Bayesian IDS algorithm, either on the frequentist or Bayesian regret. Second, the Bayesian information gain might serve as a starting point to design more effective information gain functions in the frequentist framework, for example adapted to other likelihood functions and regularizers.

5.2 ASYMPTOTIC REGRET: PROOFS

The proof of Theorem 5.4 relies on improved bounds on the information ratio and the total information gain in the asymptotic regime. These are provided after the main proof in Sections 5.2.1 and 5.2.2.

Proof of Theorem 5.4. We denote $\chi_s \triangleq \|\hat{\theta}_s - \theta\|_{V_s}^2$ and $B_s \triangleq \mathbb{1}(\chi_s \leq \beta_s)$. With Lemma 5.1 we get

$$\mathfrak{R}_n \leq \mathbb{E}\left[\sum_{s=1}^{s_n} \Delta(a_s) B_s\right] + \mathcal{O}\big(\log\log(n)\big)$$

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

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

In particular, we can write

$$\mathbb{E}\left[\sum_{s=1}^{s_n} \Delta(a_s) B_s\right] = \mathbb{E}\left[\sum_{s \in S_1} \Delta(a_s)\right] + \mathbb{E}\left[\sum_{s \in S_2} \Delta(a_s)\right] + \mathbb{E}\left[\sum_{s \in S_3} \Delta(a_s)\right].$$

We address the three terms in order. We will see that the last term (S_3) contributes $\mathfrak{c}_n^* \log(n) + o(\log(n))$, whereas all other terms are lower order.

SUM OVER S_1: By Theorem 2.1, we have

$$\mathbb{E}\left[\sum_{s \in S_1} \hat{\Delta}(a_s)\right]^2 \leq \mathbb{E}\left[\sum_{s \in S_1} \Psi_s(\mu_s^{\mathrm{IDS}})\right] \mathbb{E}\left[\sum_{s \in S_1} I_s^{\mathrm{A}}(a_s)\right].$$

To bound the information-ratio, the definition of S_1 implies the conditions of Lemma 5.5, which combined with $\delta_s \leq \hat{\Delta}_s(a_s)$ yields

$$\sum_{s \in S_1} \mathbb{E}\Big[\Psi_s(\mu_s^{\mathrm{IDS}})\Big] \leq \mathcal{O}\Big(\tfrac{d}{\Delta_{\min}}\Big) \sum_{s \in S_1} \mathbb{E}\big[\hat{\Delta}_s(a_s)\big] .$$

The total information gain on S_1 is bounded using the same steps as in the proof of Lemma 5.3,

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

where (i) follows because $m_s < \beta_s \lambda^{-1}$ for $s \in S_1$ and (ii) from the elliptic potential (Lemma 5.4) and using that $\log(k)\eta_s^{-1} \leq (\beta_n^{\mathrm{glob}})^{1/2}$. Combining and rearranging the last three displays and using $\Delta(a_s)B_s \leq 2\hat{\Delta}_s(a_s)B_s$ with $B_s = 1$ for $s \in S_1$ yields

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

SUM OVER S_2: First note that $\chi_s \leq \beta_s < m_s$ implies $\hat{a}_s = a^{\mathrm{UCB}} = a^*$. For any $a \in \mathcal{A}$,

$$\begin{aligned}
\beta_s^{-1/2}\delta_s - \|a\|_{V_s^{-1}} &\overset{(i)}{=} \|a^*\|_{V_s^{-1}} - \|a\|_{V_s^{-1}} \overset{(ii)}{\leq} \|a^* - a\|_{V_s^{-1}} \\
&\overset{(iii)}{\leq} \frac{1}{(2m_s)^{1/2} - \chi_s^{1/2}} \overset{(iv)}{\leq} \frac{2}{m_s^{1/2}} \overset{(v)}{<} \frac{\delta_s}{2\beta_s^{1/2}} ,
\end{aligned} \tag{5.16}$$

where (i) follows because $\hat{a}_s = a_s^{\mathrm{UCB}} = a^*$, implying that $\delta_s = \beta_s^{1/2}\|a^*\|_{V_s^{-1}}$. (ii) follows from the triangle inequality, (iii) from Lemma 5.6 and (iv) because $\chi_s \leq m_s/4$. Finally, (v) holds since $\delta_s^2/16 > \beta_s/m_s$. With $a = a_s$ and rearranging yields $\delta_s \leq 2\beta_s^{1/2}\|a_s\|_{V_s^{-1}}$ and hence

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

where (i) uses $\hat{\Delta}_s(\mu_s) \leq 2\delta_s$ (Lemma 2.6). From here, we can apply Cauchy-Schwarz in a similiar manner as in Theorem 2.1, to get

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

For the last inequality, we used that $4\beta_s^{1/2}m_s^{-1/2} \leq \delta_s \leq \hat{\Delta}_s(a_s)$, the elliptic potential (Lemma 5.4) and $m_s \leq \beta_{s_n,n\log(n)} \leq \mathcal{O}(\log(n) + d\log(s_n))$. Hence, combining the last two displays and $\Delta_s(a_s) \leq 2\hat{\Delta}_s(a_s)$, we get

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

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

$$\mathbb{E}\left[\sum_{s\in S_3} \Delta(a_s)\right] \leq \mathbb{E}\left[\sum_{s\in S_3} \bar{\Delta}_s(a_s)\right] + \mathbb{E}\left[\sum_{s\in S_3} \chi_s^{1/2}\|a^* - a_s\|_{V_s^{-1}}\right]. \tag{5.17}$$

For the second sum, note that by Lemma 5.7 the information gain of $a_s \neq a^*$ is lower bounded by a constant, $I_s^{\mathrm{A}}(a_s) = \Omega\big(\Delta_{\min}^2 d^{-1}\big)$. As in Eq. (5.16), Lemma 5.6 implies

$$\chi_s^{1/2}\|a^* - a_s\|_{V_s^{-1}} \leq 2\chi_s^{1/2}m_s^{-1/2}\mathbb{1}(a_s \neq a^*) \leq \mathcal{O}\big(\lambda^{1/2}d\Delta_{\min}^{-2}I_s^{\mathrm{A}}(a_s)\big)\,.$$

Summing the last display inside the expectation and using Lemma 5.3 yields

$$\mathbb{E}\left[\sum_{s\in S_3} \chi_s^{1/2}\|a^* - a_s\|_{V_s^{-1}}\right] \leq \mathcal{O}\big(\lambda^{1/2}d\log(n)\mathbb{E}[\log(s_n)]\big)\,.$$

For the first sum in Eq. (5.17), we use $4xz \leq (x+z)^2$ and the fact that $\hat{\Delta}_s(\mu_s) = \bar{\Delta}_s(\mu_s) + \delta_s$, to get

$$\mathbb{E}\left[\sum_{s\in S_3} \bar{\Delta}_s(\mu_s)\right] \leq \frac{1}{4}\mathbb{E}\left[\sum_{s\in S_3} \delta_s\right]^{-1}\mathbb{E}\left[\sum_{s\in S_3} \hat{\Delta}_s(\mu_s)\right]^2$$
$$\leq \frac{1}{4}\mathbb{E}\left[\sum_{s\in S_3} \delta_s\right]^{-1}\mathbb{E}\left[\sum_{s\in S_3} \Psi_s(\mu_s)\right]\mathbb{E}\left[\sum_{s\in S_3} I_s(a_s)\right] \tag{5.18}$$

The second inequality is by Theorem 2.1. The *main steps* follow. Lemma 5.10 in Section 5.2.1 bounds the information ratio,

$$\Psi_s(\mu_s) \le 4\delta_s(\mathfrak{c}^* + \mathcal{O}(\delta_s + \chi_s^{1/2} m_s^{-1/2}) \le 4\delta_s(\mathfrak{c}^* + \mathcal{O}(\lambda^{1/2}))\,,$$

where the last inequality uses $\delta_s/4 \le \beta_s^{1/2} m_s^{-1/2} \le \lambda^{1/2}$. In particular,

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

It remains to bound the information gain on S_3. We denote $l_s(q_s) = \sum_{c\neq a^*} q_s(c)\langle \hat{v}_s(c) - \hat{\theta}_s, a_s\rangle^2$. Note that since $\hat{a}_s = a^*$ on S_3, $l_s(q_s) = I_s^{\mathrm{A}}(a_s)$. Further, let $J_s = \mathbb{1}\Big(24^2\eta_s\chi_s\|a_s\|^2_{V_s^{-1}} \le 1; \chi_s\|a_s\|^2_{V_s^{-1}} \le 1\Big)$. It is easy to verify that for small enough λ, $J_s = 1$ for all $s \in S_3$. Hence, by Lemma 5.12 in Section 5.2.2 and using that $m_s \le \log(n) + \log\log(n) + \mathcal{O}(d\log(s_n))$,

$$\begin{aligned}\mathbb{E}\left[\sum_{s\in S_3} I_s^{\mathrm{A}}(a_s)\right] = \mathbb{E}\left[\sum_{s\in S_3} l_s(q_s)\right] &\le \mathbb{E}\left[\sum_{s=1}^{s_n} J_s l_s(q_s)\right] \\ &\le \log(n) + \mathcal{O}\big(\log(n)^{1/2}\mathbb{E}\Big[\log(s_n)^2\Big]\big)\,.\end{aligned}$$

With the bounds on the information ratio and information gain, we find

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

Hence we conclude

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

Finally, with Lemma 5.8, we get that $\mathbb{E}[\log(s_n)^b] \le \mathcal{O}(\log\log(n))$. Therefore, with $\lambda = \log(n)^{-2/3}$ all terms except for $\mathfrak{c}^*\log(n)$ are of lower order and the claim follows. □

The following technical calculation relates the minimum constraint value m_s to the norm $\|a^* - a\|_{V_s^{-1}}$ for any $a \in \mathcal{A}$.

Lemma 5.6. *Let $m_s = \frac{1}{2}\min_{c\neq\hat{a}_s}\|\hat{v}(c) - \hat{\theta}_s\|^2_{V_s}$. Assume that $\|\hat{\theta}_s - \theta\|^2_{V_s} < 2m_s$ and $\max_{a\in\mathcal{A}}\Delta(a) \le 1$. Then $\hat{a}_s = a^*$ and further, for all $a \in \mathcal{A}$,*

$$\big((2m_s)^{1/2} - \|\hat{\theta}_s - \theta\|_{V_s}\big)\|a^* - a\|_{V_s^{-1}} \le 1\,.$$

Proof. Since $m_s = \frac{1}{2}\min_{a\neq\hat{a}_s}\min_{\nu\in\mathcal{C}_a}\|\nu-\hat{\theta}_s\|_{V_s}^2$, the assumption that $\|\hat{\theta}_s-\theta\|_{V_s}^2 < 2m_s$ implies that $\hat{a}_s = a^*(\hat{\theta}_s) = a^*(\theta)$. Further, for any $a\in\mathcal{A}$,

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

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

It is useful to know that when IDS chooses actions other than a^*, the information gain is large as quantified by the next lemma.

Lemma 5.7 (Constant information gain)**.** *Assume that $\hat{a}_s = a^*$ and $2\delta_s \leq \hat{\Delta}_s(a)$ for all $a\neq\hat{a}_s$. If $c_s\neq a^*$ is contained in the support of the IDS distribution $\operatorname{supp}(\mu_s)$, then the information gain of c_s is at least a constant,*

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

Proof. Note that by $c_s\in\operatorname{supp}(\mu_s)$ and Lemma 2.4,

$$I_s^{\mathrm{A}}(c_s) = \big(2\hat{\Delta}_s(c_s)-\hat{\Delta}_s(\mu_s)\big)\frac{\hat{\Delta}_s(\mu_s)}{\Psi_s(\mu_s^{\mathrm{IDS}})} \geq \frac{2\delta_s\big(\hat{\Delta}_s(c_s)-\delta_s\big)}{\Psi_s(\mu_s^{\mathrm{IDS}})} \geq \frac{\hat{\Delta}_s(c_s)\delta_s}{\Psi_s(\mu_s^{\mathrm{IDS}})}\,.$$

We first used that $\delta_s\leq\hat{\Delta}_s(\mu)\leq 2\delta_s$ (Lemma 2.6) and then the assumption that $2\delta_s\leq\hat{\Delta}_s(c_s)$. Further, $2\hat{\Delta}_s(c_s)\geq\Delta_{\min}$, and by Lemma 5.5,

$$\Psi_s(\mu_s^{\mathrm{IDS}}) \leq \frac{4\delta_s(8d+9)}{\Delta_{\min}}\,.$$

Combining the inequalities yields the result. □

The next lemma bounds the effective horizon. It is the only result that makes use of the assumption $\|a^*\|>0$.

Lemma 5.8. *Assume that $\|a^*\|>0$. Then the number of exploration steps s_n in Algorithm 5 is bounded in expectation,*

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

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

Proof. By Theorem 5.3,

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

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

$$\delta_s = \max_{b \in \mathcal{A}} \langle b - a^*, \hat{\theta}_s \rangle + \beta_s^{1/2} \|b\|_{V_s^{-1}} \ge \beta_s^{1/2} \|a^*\|_{V_s^{-1}} \ge \|a^*\| s^{-1/2}\,.$$

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

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

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

$$\begin{aligned}
\mathbb{E}[\log(s_n)^x] = 2^x \mathbb{E}\Big[\log(s_n^{1/2})^x\Big] &\le 2^x \mathbb{E}\Big[\log\big(\max(s_n^{1/2}, \exp(x-1))\big)^x\Big] \\
&\le 2^x \log(\mathbb{E}\Big[s_n^{1/2}\Big] + \exp(x-1))^x \\
&\le \mathcal{O}(\log(\log(n))^x)\,. \qquad \square
\end{aligned}$$

5.2.1 *Asymptotic Bound on the Information Ratio*

We define $\alpha^* \in (\mathbb{R}_{\ge 0} \cup \{\infty\})^k$ as the allocation that attains the lower bound in Eq. (5.1), obtained in the appropriate limit. Note that the optimal allocation satisfies $\alpha(a^*) = \infty$. Denote by $\tilde{\alpha}^*(a) = \alpha^* \mathbb{1}(a \neq a^*)$ the optimal allocation on the sub-optimal actions, which is always finite. The next lemma quantifies the increase in regret if we truncate the optimal allocation on $\alpha(a^*)$ to a finite value. Recall that $\mathcal{C}^* = \{\nu \in \mathcal{M} : a^*(\nu) \neq a^*(\theta)\}$ is the set of alternative parameters defined for the true paramater $\theta \in \mathcal{M}$.

Lemma 5.9 (Truncated optimal allocation)**.** *Let $\alpha_\lambda^*(a) = \tilde{\alpha}^* + \lambda \mathbb{1}(a = a^*)$ be the optimal allocation, truncated on a^* such that $\alpha_\lambda^*(a^*) = \lambda$. There exists a constant $C(\theta, \mathcal{A})$ that depends only on the instance and the action set, and such that for all $\nu \in \mathcal{C}^*$,*

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

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

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

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

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

We expect the same holds approximately for ν with $\langle \nu - \theta^*, a^* \rangle^2 \leq 2\lambda^{-1}$. To make this formal, we require that $\mathcal{M}$ is a polytope. Lemma 5.11 with an appropriate shift of the parameter space and $\lambda_{\max}(V(\tilde{\alpha}^*)) \leq \|\tilde{\alpha}^*\|_1$ imply

$$\begin{aligned} \min_{\nu^* \in \mathcal{H}_a^*} \|\nu - \nu^*\|^2_{V(\tilde{\alpha}^*)} &\leq C(\theta, \mathcal{A})\|\tilde{\alpha}^*\|_1 \langle \nu - \theta, a^* \rangle^2 \\ &\leq 2\lambda^{-1} C(\theta, \mathcal{A})\|\tilde{\alpha}^*\|_1 \leq 1 , \end{aligned} \tag{5.19}$$

where the last two inequalities use the case assumptions. Let us fix ν^* to be the minimizer of the left-hand side. By the reverse triangle inequality,

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

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

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

For (i) we use the choice ν^* and Eq. (5.19). For (ii) we minimize over $\langle \nu - \theta, a^* \rangle$. This completes the proof. □

The previous result equips us to derive a bound on the information ratio that relates to the asymptotic regret $\mathfrak{c}^*$.

Lemma 5.10 (Asymptotic information ratio)**.** *Define* $\chi_s \triangleq \|\hat{\theta}_s - \theta\|^2_{V_s}$ *and* $m_s = \frac{1}{2}\min_{c \neq a^*} \|\hat{v}_s(c) - \hat{\theta}_s\|^2_{V_s}$. *Assume that* $4\chi_s \leq m_s$ *and* $\chi_s \leq \beta_s$. *Then,*

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

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

Proof. Not surprisingly, the proof strategy is to bound the information ratio with a distribution informed from the lower bound in Eq. (5.1). First note that the assumption $m_s \geq 4\chi_s$ implies $\hat{a}_s = a^*$ by Lemma 5.6. We introduce a shorthand $\bar{\Delta}_s(a) = \langle \hat{\theta}_s, \hat{a}_s - a\rangle$ for the estimated mean gap and let $\tilde{\mu} \in \mathscr{P}(\mathcal{A})$ be a distribution with $2\delta_s \leq \hat{\Delta}(\tilde{\mu}) = \delta_s + \bar{\Delta}_s(\tilde{\mu})$. Then, by Lemma 2.5,

$$\begin{aligned}\min_{\mu \in \mathscr{P}(\mathcal{A})} \Psi_s(\mu) &\leq \min_{0 \leq p \leq 1} \frac{\big((1-p)\hat{\Delta}_s(a^*) + p\hat{\Delta}(\tilde{\mu})\big)^2}{pI_s(\tilde{\mu})} \\ &= \frac{4\delta_s(\hat{\Delta}_s(\tilde{\mu}) - \delta_s)}{I_s(\tilde{\mu})} = \frac{4\delta_s\bar{\Delta}_s(\tilde{\mu})}{I_s(\tilde{\mu})}\,. \end{aligned} \tag{5.20}$$

Note that the last ratio is invariant in constant rescaling $\tilde{\mu}$, so we easily get rid of normalization factors. Recall that $\tilde{\alpha}^*$ is the optimal allocation over suboptimal actions, as defined at the beginning of the section. We let $\alpha^*_\lambda(a) = \tilde{\alpha}^*(a) + \lambda \mathbb{1}(a = a^*)$ be the truncated optimal allocation and $\tilde{\mu}_\lambda = \alpha^*_\lambda / (\|\tilde{\alpha}^*\|_1 + \lambda)$ be the corresponding normalized distribution. Using Cauchy-Schwarz, we get

$$\begin{aligned}\Delta(\tilde{\mu}_\lambda) - \bar{\Delta}_s(\tilde{\mu}_\lambda) &\leq \|\hat{\theta}_s - \theta^*\|_{V_s} \max_{a \neq a^*} \|a^* - a\|_{V_s^{-1}} \\ &\leq \frac{\chi_s^{1/2}}{(2m_s)^{1/2} - \chi_s^{1/2}} \leq \frac{2\chi_s^{1/2}}{m_s^{1/2}}\,.\end{aligned}$$

The second inequality uses Lemma 5.6 and the definition $\chi_s = \|\hat{\theta}_s - \theta\|^2_{V_s}$. The last inequality simplifies the expression with the assumption $4\chi_s \leq m_s$. Note that $\Delta(\tilde{\mu}_\lambda) = \frac{\mathfrak{c}^*}{\|\tilde{\alpha}^*\|_1 + \lambda}$. Hence, to satisfy $\delta_s \leq \bar{\Delta}_s(\tilde{\mu}_\lambda)$, it is sufficient to satisfy the following constraint on λ,

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

At equality, we get

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

Note that as $\delta_s \to 0$ and $m_s \to \infty$, we get an unbounded allocation as expected. Next, we compute the approximation errors. Using Lemma 5.6,

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

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

$$\begin{aligned}I_s(\alpha_\lambda^*) &= \tfrac{1}{2} \sum_{a \in \mathcal{A}} \alpha_\lambda^*(a) \sum_{c \neq a^*} q_s(c) \Big(|\langle \hat{v}_s(c) - \hat{\theta}_s, a\rangle| + \beta_s^{1/2} \|a\|_{V_s^{-1}} \Big)^2 \\ &\geq \tfrac{1}{2} \sum_{a \in \mathcal{A}} \alpha_\lambda^*(a) \sum_{c \neq a^*} q_s(c) \langle \hat{v}_s(c) - \theta, a\rangle^2 \\ &= \tfrac{1}{2} \sum_{c \neq a^*} q_s(c) \|\hat{v}_s(c) - \theta\|_{V(\alpha_\lambda^*)}^2 \\ &\geq 1 - 2C(\mathcal{A}, \theta)\|\tilde{\alpha}^*\|_1 \lambda^{-1} .\end{aligned}$$

The last step is by Lemma 5.9. Finally, the proof is completed by plugging α_λ^* in Eq. (5.20) and using $\frac{\mathfrak{c}^* + A}{1-B} = \mathfrak{c}^* + \frac{A + \mathfrak{c}^* B}{1-B}$:

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

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

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

The error bound in Lemma 5.9 makes use of the following technical lemma on convex polytopes.

Lemma 5.11 (Convex Polytopes)**.** *Let $\mathcal{K}$ be a convex polytope. For unit vector $\eta \in \mathbb{R}^d$, let $\mathcal{K}_0 = \{v \in \mathcal{K} : \langle v, \eta\rangle = 0\}$ be the intersection of $\mathcal{K}$ with a $(d-1)$-dimensional hyperplane, which is assumed to be non-empty. Then there exists a constant $C > 0$ such that for all $w \in \mathcal{K}$,*

$$\min_{v_0 \in \mathcal{K}_0} \|v_0 - w\|_2 \leq C\langle w, \eta\rangle .$$

Proof. Let $\mathcal{P} = \{v \in \mathcal{K} : \langle v, \eta \rangle \geq 0\}$, which is also a convex polytope. We first show there exists a $C > 0$ such that for all $w \in \mathcal{P}$,

$$\min_{v_0 \in \mathcal{K}_0} \|v_0 - w\|_2 \leq C \langle w, \eta \rangle \,. \tag{5.21}$$

The result follows from a symmetric argument for $\{v \in \mathcal{K} : \langle v, \eta \rangle \leq 0\}$. To establish Eq. (5.21), let $V \subset \mathbb{R}^d$ be the vertices of $\mathcal{P}$, which is a finite set. Define $h : \mathcal{P} \setminus \mathcal{K}_0 \to \mathbb{R}$ by

$$h(w) = \max_{v \in \mathcal{K}_0} \frac{\langle \eta, w - v \rangle}{\|w - v\|} \,.$$

Clearly, $1/C \triangleq \min_{v \in V : \langle v, \eta \rangle > 0} h(v) > 0$. Hence, the mapping $\varphi : V \to \mathcal{K}_0$ such that $\varphi(v) = v$ for $v \in \mathcal{K}_0$ and $\varphi(v) = \arg\max_{x \in \mathcal{K}_0} \frac{\langle \eta, v - x \rangle}{\|v - x\|}$ satisfies $\|v - \varphi(v)\|_2 \leq C \langle \eta, v - \varphi(v) \rangle$. Given any $w \in \mathcal{P}$, let α be a probability distribution on V such that $w = \sum_{v \in V} \alpha(v) v$ and let $v_0 = \sum_{v \in V} \alpha(v) \varphi(v) \in \mathcal{K}_0$. Then,

$$\begin{aligned}
\|w - v_0\|_2 &= \left\| \sum_{v \in V} \alpha(v) v - \sum_{v \in V} \alpha(v) \varphi(v) \right\|_2 \\
&\leq \sum_{v \in V} \alpha(v) \|v - \varphi(v)\|_2 \\
&\leq C \sum_{v \in V} \alpha(v) \langle \eta, v - \varphi(v) \rangle \\
&= C \langle \eta, w \rangle \,.
\end{aligned}$$

□

5.2.2 *Asymptotic Bound on the Information Gain*

The next lemma uses the interpretation of the information gain as the loss of the exponential weights learner to tighten the bound on the total information gain in the asymptotic regime. In the statistical setting, the connection is only approximately true. For instance, on time steps s where $a^* \neq a^*(\hat{\theta}_s)$, the q_s-weights are not even defined on the same support as the q-weights used in the primal-dual setup. This is the reason for the fairly technical assumptions in the lemma, that are also chosen with their use in the proof of Theorem 5.4 in mind.

Lemma 5.12. *Let $q_s^*(c) \propto \exp(-\eta_s \|\hat{v}_s(c) - \hat{\theta}_s\|_{V_s}^2)$ be mixing weights defined on $\mathcal{A} \setminus a^*$ (i.e. also when $\hat{a}_s \neq a^*$), with $\hat{v}_s(c) = \arg\min_{v \in \mathcal{H}_c^{a^*}} \|v - \hat{\theta}_s\|_{V_s}^2$ for*

all $c \neq a^*$. *Define the loss* $l_s(q_s) = \sum_{c \neq a^*} q_s^*(c) \langle \hat{v}_s(c) - \theta_s \rangle^2$ *and the indicator* $J_s = \mathbb{1}\Big(24^2 \eta_s \chi_s \|a_s\|^2_{V_s^{-1}} \leq 1; \chi_s \|a_s\|^2_{V_s^{-1}} \leq 1\Big)$, *where* $\chi_s \triangleq = \|\hat{\theta}_s - \theta\|^2_{V_s}$. *Then*

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

The statement is a regret bound for the exponential weights learner that defines the q_s^*-weights, which excludes the loss of the learner on steps where $J_s = 0$. The difference to standard online learning bounds is that we are interested in the baseline loss $L_s(a) = \frac{1}{2}\|\hat{v}_s(a) - \hat{\theta}_s\|^2_{V_s}$, which does not exactly equal the sum of instantaneous loss $\sum_{s=1}^{s_n} l_s(a)$.

Proof of Lemma 5.12. We make use of the formulation of the exponential weights algorithm in the mirror descent framework, in particular the *follow the regularized leader* (FTRL) algorithm [146]. To this end, let $\psi(q) = \sum_{c \neq a^*} q(c) \log(q(c))$ be the entropy function defined for $q \in \mathscr{P}(\mathcal{A} \setminus a^*)$. For learning rate $\eta > 0$, we define

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

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

$$q_s^* = \operatorname*{arg\,min}_{q \in \mathscr{P}(\mathcal{A} \setminus a^*)} L_s(q) + \psi_s(q)\,.$$

Denote $\mathfrak{Q}_n = \sum_{s=1}^{s_n} J_s l_s(q_s^*) - \min_{c \neq a^*} \|\hat{v}_{s_n}(c) - \hat{\theta}_{s_n}\|^2_{V_{s_n}}$. The following inequality is easily verified by telescoping [124, Lemma 7.1],

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

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

$$\begin{aligned} &[L_s + l_s + \psi_s](q_s^*) - [L_{s+1} + \psi_{s+1}](q_{s+1}^*) \\ &\leq [L_{s+1} + \psi_s](q_s^*) - [L_{s+1} + \psi_s](\tilde{q}_{s+1}) + [L_s + l_s - L_{s+1}](q_s^*)\,. \end{aligned} \tag{5.22}$$

Note that L_{s+1} exhibits an intricate dependence on the outcome y_s, whereas all other quantities appearing in the last display are $\mathcal{F}_s$-predictable. Using

that $\tilde{q}_{s+1}$ is a minimizer of $L_{s+1} + \psi_s$ and the definition of the Bregman divergence $D_\psi(p\|q) = \psi(p) - \psi(q) - \langle \nabla\psi(q), p - q\rangle$, we find

$$\begin{aligned}[L_{s+1} + \psi_s](q_s^*) - [L_{s+1} + \psi_s](\tilde{q}_{s+1}) &= \frac{1}{\eta_s} D_{\psi_s}(q_s^*, \tilde{q}_{s+1}) \\ &= \frac{1}{\eta_s} \sum_{c \neq a^*} q_s^*(c) \log \frac{q_s^*(c)}{\tilde{q}_{s+1}(c)} \,.\end{aligned}$$

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

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

A technical calculation which directly bounds the moments of the subgaussian noise under the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_s]$ with the condition $J_s = 1$, is summarized in Lemma B.3 in Appendix B. This yields

$$\begin{aligned}&\sum_{s=1}^{s_n} J_s \mathbb{E}_s[[L_{s+1} + \psi_s](q_s^*) - [L_{s+1} + \psi_s](\tilde{q}_{s+1})] \\ &\leq \sum_{s=1}^{s_n} \frac{J_s}{\eta_s} \sum_{c \neq a^*} q_s^*(c) \mathbb{E}_s\left[\sum_{i=2}^{\infty} \frac{(-\eta_s (L_{s+1}(c) - L_s(c)))^i}{i!} \right] \\ &\leq \sum_{s=1}^{s_n} \sum_{c \neq a^*} q_s^*(c) \mathcal{O}\Big(\eta_s \big(\chi_s \|a_s\|^2_{V_s^{-1}} + \|\hat{v}_s(c) - \hat{\theta}_s\|^2_{V_s} \|a_s\|^2_{V_s^{-1}} \big) \Big) \\ &\leq \mathcal{O}\Big(\log(n)^{1/2} \log(s_n)^2 \Big)\end{aligned}$$

The last step makes use of Lemma 5.2, $\eta_s m_s \leq \beta^{1/2}_{s_n, 1/(n \log n)} \leq \mathcal{O}(\log(n)^{1/2} + \log(s_n)^{1/2})$ and Lemma 5.2. Going back to Eq. (5.22), still for the case where

$J_s = 1$, it remains to bound the shift term $S_s(a) \triangleq L_s(a) + l_s(a) - L_{s+1}(a)$. We have

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

Here, (i) follows from the Lemma B.1 in Appendix B, Cauchy-Schwarz and taking the expectation; (ii) is Jensen's inequality and (iii) is the softmin inequality (Lemma 5.2). Hence, using that $m_s \leq \beta_{s_n, n\log(n)} \leq \mathcal{O}(\log(n) + \log(s_n))$ and the elliptic potential lemma (Lemma 5.4), we find

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

Lastly, we address Eq. (5.22) for the case $J_s = 0$, which then reads

$$[L_s + \psi_s](q_s) - [L_{s+1} + \psi_{s+1}](q_{s+1}) \leq L_s(q_{s+1}) - L_{s+1}(q_{s+1})\,. \quad (5.23)$$

We can reuse Lemma B.2 to find,

$$\begin{aligned}\mathbb{E}_s[L_s(q_{s+1}) - L_{s+1}(q_{s+1})] &\leq \mathcal{O}\big(\chi_s\|a_s\|^2_{V_s^{-1}} + |\langle\hat{v}_s - \hat{\theta}_s, a_s\rangle| + \langle\hat{v}_s - \hat{\theta}_s, a_s\rangle^2\big)\\ &\leq \mathcal{O}\big(\chi_s\|a_s\|^2_{V_s^{-1}} + 1)\big)\,.\end{aligned}$$

Using that when $J_s = 0$ we have $1 \leq \chi_s\|a_s\|^2_{V_s}$, or $1 \leq 24^2\eta_s\chi_s\|a_s\|^2_{V_s^{-1}}$, with Lemma 5.4 we sum up these terms to

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

The claim follows. □

5.3 NUMERICAL RESULTS

We compare IDS with UCB [3] and SOLID [160], the latter being the only other method that is both asymptotically and worst-case optimal. Refer to [160] for further experiments where SOLID is compared to OAM [72] and Thompson sampling.

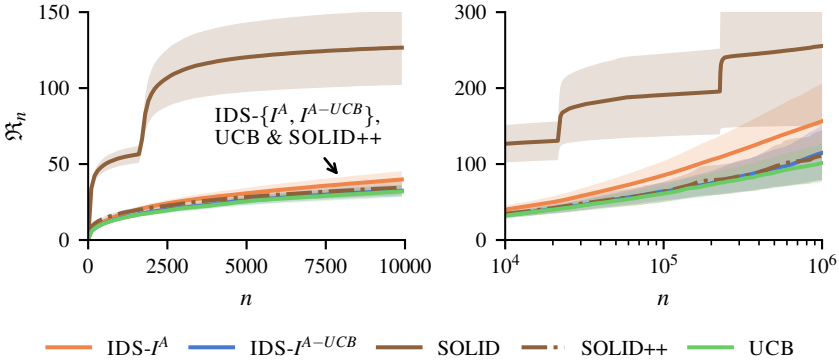


FIGURE 5.3: Averaged performance on randomly generated action sets in $\mathbb{R}^2$ with $k = 6$ actions. Note that SOLID++ and IDS-$I_t^{\text{A-UCB}}$ are competitive with UCB.

To enable a fair comparison, we use the same confidence coefficient β_t defined in Eq. (5.2) for all algorithms. We also run the same experiments with the confidence coefficient derived by [160], but we found no significant difference in the results. For SOLID, we use the default hyper-parameters suggested by Tirinzoni *et al.* [160, Appendix K]. As recommended by the authors, we further implemented a heuristic variant, SOLID++, which is optimized for better performance in finite time and does not reset the sampling distribution at the beginning of each phase.

IDS is implemented as in Algorithm 5 but with unconstrained parameter set ($\mathcal{M} = \mathbb{R}^d$), which allows us to compute the parameter $\hat{\nu}_s(x)$ in closed form. We also use the approximate IDS sampling strategy (Section 2.1.1), which we show performs similarly compared to exact sampling. Except for in the experiment where we empirically optimize η_s and β_s, all frequentist variants of IDS satisfy the theoretical guarantees presented in this chapter.

We set the variance of the noise to $\rho^2 = 0.1$ and scale β_s accordingly, which is chosen so that the asymptotic regime is observed after fewer rounds relative to $\rho^2 = 1$. All results are averaged over 100 runs and we display 95% confidence regions.

RANDOM ACTION SETS In the first experiment, we randomly sample 6 actions drawn uniformly from the unit sphere at the beginning of each run. The results are shown in Fig. 5.3. All policies except for SOLID have comparable averaged performances, but the latter is not optimized for

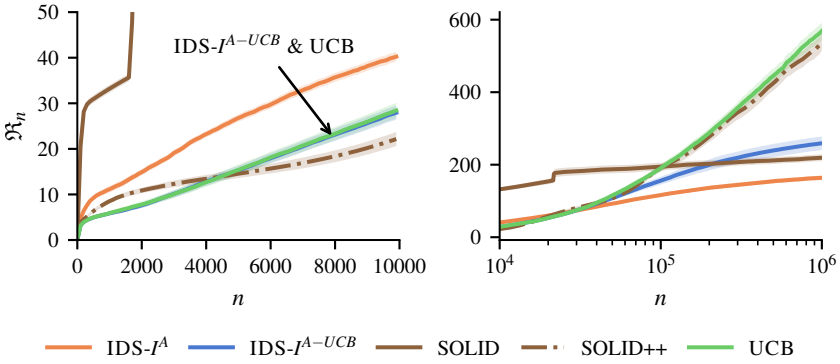


FIGURE 5.4: The 'end of optimism' example with $\epsilon = 0.01$. Note that SOLID, IDS-I_t^{A} and IDS-$I_t^{\text{A-UCB}}$ clearly outperform UCB in the asymptotic regime. The heuristically optimized SOLID++ does enter the asymptotic regime in our simulations.

worst-case regret. IDS-I^{UCB} performs similar to UCB, followed by IDS with information gain I^{A}.

THE END OF OPTIMISM? Example 5.1 is an instance where optimistic algorithms such as UCB behave sub-optimally. The action set has three actions $x_1 = [1.0, 0.0]$, $x_2 = [1 - \epsilon, 4\epsilon]$, $x_3 = [0.0, 1.0]$ for a tuning variable $\epsilon \in (0, 1)$ and $\theta = [1.0, 0.0]$. The lower bound constant is $\mathfrak{c}_n^* = 6.4$, independent of ϵ and computed for noise variance $\rho^2 = 0.1$, whereas for UCB, the regret is at least $\Omega(1/\epsilon)$ as $\epsilon \to 0$. In our experiment, we choose $\epsilon = 0.01$ which suffices to highlight the difference in the asymptotic regret on the horizon $n = 10^6$.

Results in this setting are shown in Fig. 5.4. As expected, UCB's asymptotics show a suboptimal log-slope, but it is surprisingly followed by SOLID++. Despite our attempts, we are presently not able to provide a good explanation for this result and it might require a more involved analysis of the SOLID++ heuristic. However, both versions of IDS and the theoretical SOLID reach the optimal asymptotic around $t = 10^5$ (10^4 for SOLID), and significantly outperform UCB on that problem. An interesting observation is that IDS-$I^{\text{A-UCB}}$ performs better in finite time, whereas IDS-I^{A} reaches the asymptotic regime earlier. This is in line with the empirical behavior of the information gain functions, shown in Fig. 5.2.

TUNING η_s AND β_s As presented, Algorithm 5 is hyperparameter free. In practice, it is possible to improve performance significantly by empirically

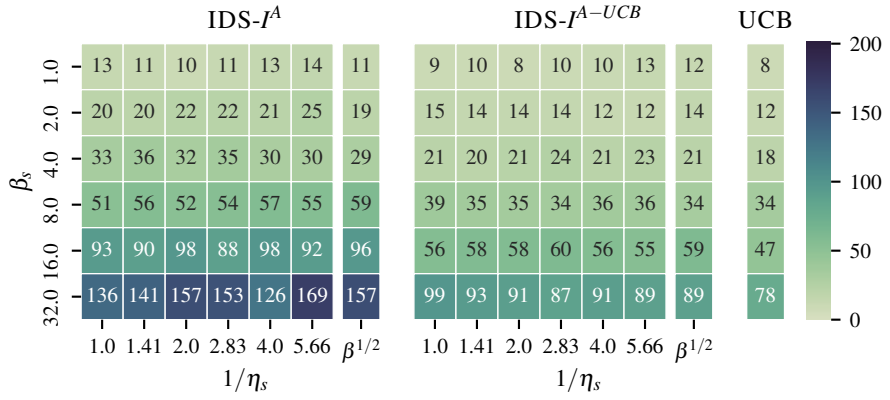


FIGURE 5.5: The matrix shows the regret on randomly generated action sets after $n = 10^4$ steps for different values of β_s and η_s. The first observation is that the regret can be *significantly* reduced by choosing a smaller value of β_s. On the other hand, tuning the q-learning rate η_s affects performance marginally. Tuning *only* β_s and setting $\eta_s = 1/\sqrt{\beta_s}$ as suggested by the theory leads to near optimal results.

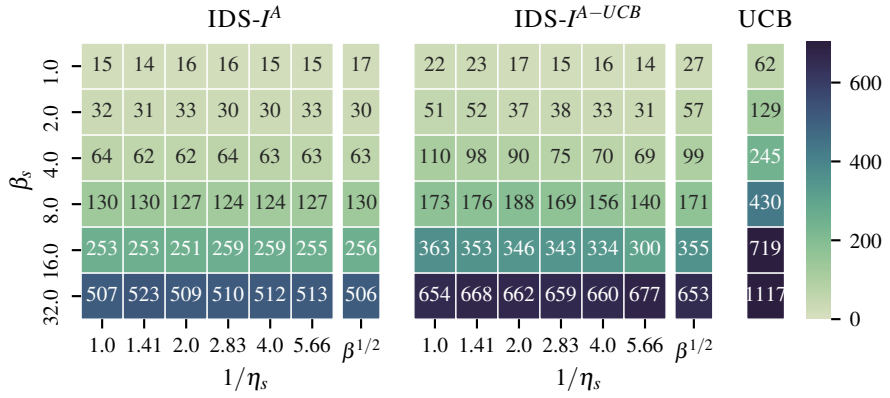


FIGURE 5.6: The matrix shows the regret on the 'end of optimism' example after $n = 10^6$ steps for different values of β_s and η_s. The observations are similar as for Fig. 5.5. Note that IDS is consistently better than UCB for any value of β_s.

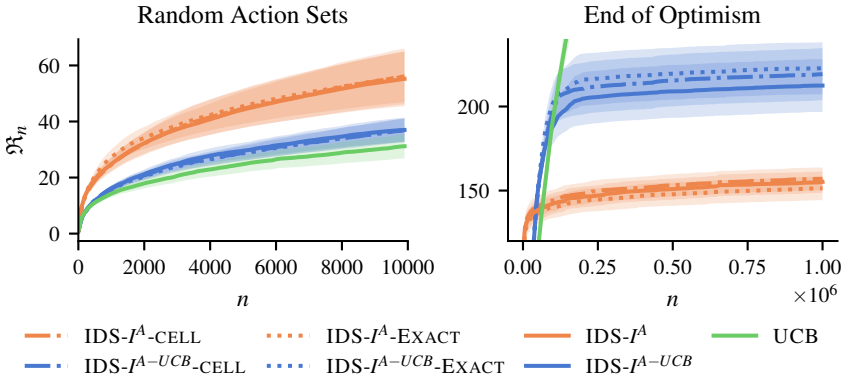


FIGURE 5.7: Comparison of information gain functions defined on cells and halfspaces respectively, as well as exact and approximate sampling from the IDS distribution. Note that all variants with the same information gain achieve similar performance within the standard error. In the right plot, the y-axis is scaled to make the difference visible.

optimizing the confidence parameter β_s and the q-learning rate η_s. We show how both parameters affect the performance on randomly generated action sets and the 'end of optimism' example in Figs. 5.5 and 5.6. The main finding is that the learning rate η_s only marginally affects performance. On the other hand, the theoretical value of β_s appears to be conservative, and performance can be improved by at least one order of magnitude with a much smaller value for β_s than validated by the theory. It is commonly known that a smaller confidence parameter improves performance of the UCB algorithm, but so far little theory is known to explain this effect [c. f. 70].

CELL-BASED INFORMATION GAIN On the same examples, we compare the cell-based information gain in Eq. (5.14) with the information gain defined on halfspaces in Eq. (5.6). We further study if the approximate sampling strategy described in Section 2.1.1 impacts the performance compared to sampling the IDS distribution exactly. The results in Fig. 5.7 show that on our (arguably small) examples, there is no statistically significant difference between the different variants.

COMPARISON WITH BAYESIAN METHODS In our last empirical benchmark, we include Bayesian methods, specifically Thompson sampling (TS) and an approximation of Bayesian IDS. Our implementation of Bayesian

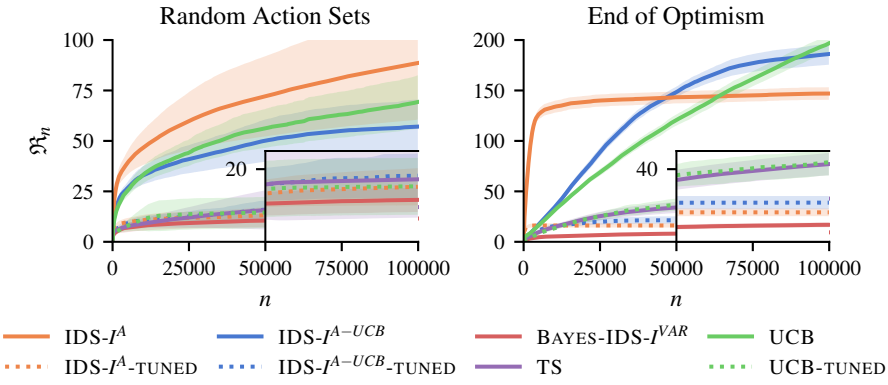


FIGURE 5.8: Comparison with Bayesian methods. Our results show that Bayesian IDS outperforms frequentist methods, even when setting $\beta_s = 1$.

Algorithm	$d = 2, k = 6$	$d = 5, k = 50$
BAYES-IDS-I^{VAR}	561.7±58.8	2560.0±78.4
BAYES-IDS-I^{VAR}-FAST	544.4±69.7	1771.9±40.5
IDS-I^{A-UCB}	50.5±22.6	798.5±233.5
IDS-I^{A-UCB}-FAST	45.7±18.8	106.8±28.6
IDS-I^{DET}	52.7±11.6	888.5±75.6
IDS-I^{DET}-FAST	42.0±9.7	76.1±19.6
UCB	26.9±7.7	23.9±5.7
TS	21.6±5.9	22.2±6.9

TABLE 5.1: Runtime comparison on random action sets with horizon $n = 10^5$. The table shows mean and standard-deviation of the runtime in seconds on 50 runs, computed on a single core at 2.30GHz. The FAST-suffix indicates the approximate IDS sampling as in Section 2.1.1.

IDS uses the variance-based information gain defined in Eq. (5.15), and we approximate the Bayesian gap estimates and information gain using 10^4 posterior samples per round as suggested in [135, Algorithm 6]. The performance plots are in Fig. 5.8. Thompson sampling significantly outperforms UCB and the frequentist IDS variants, unless we set $\beta_s = 1$, which, as noted before, improves performance of the frequentist methods. The approximation of Bayesian IDS is the most effective on our benchmark, outperforming the best frequentist method on the 'end of optimism' example roughly by a factor two. Lastly, we show runtime of all methods on a horizon $n = 10^6$ in Table 5.1. Note that despite the approximation, Bayesian IDS is computationally much more demanding, whereas the frequentist IDS is only about a factor of 5 slower than Thompson sampling on instances in $\mathbb{R}^5$ with $k = 50$ actions.

5.4 CONTRIBUTIONS AND RELATED WORK

The asymptotic analysis of information-directed sampling is by the following authors:

- Kirschner, J., Lattimore, T., Vernade, C. & Szepesvári, C. *Asymptotically Optimal Information-Directed Sampling* in *Proc. International Conference on Learning Theory (COLT)* (Aug. 2021)

Besides the overall invigorating collaboration, the author gladly acknowledges that the proof of Lemma 5.11 was contributed by Tor Lattimore. The numerical experiments in Section 5.3 are joint work with Claire Vernade.

As we mentioned in the introduction, a sequence of work establishes asymptotically optimal algorithms for linear bandits [43, 72, 79, 104]. Arguably, these methods are not very practical and also not worst-case optimal without further modifications. The first work that explores primal-dual methods for regret minimization in structured bandits is by Degenne, Shao & Koolen [50]. It is not know if this approach is worst-case optimal. In our notation, their algorithm corresponds to choosing the action with the best information-regret trade-off $c_s = \arg\min_{c \in \mathcal{A}} \hat{\Delta}_s(c) / I_s(c)$. IDS instead asymptotically randomizes between a^* and c_s, which allows it to maintain the worst-case regret bound.

The IDS algorithm presented here is among the first algorithms known to be both asymptotically optimal and (nearly) worst-case optimal. The only other approach to achieve this is by Tirinzoni *et al.* [160], which appeared online concurrently with our preprint. Interestingly, their method,

named SOLID, is also a primal-dual approach, which on a high level shares similarity with our analysis. On the other hand, there are many important differences. The SOLID approach is based on a different formulation of the Lagrangian that keeps the minimum over $\mathcal{C}^*$ in Eq. (5.1). That means the dual variable is one-dimensional, but also the Lagrangian is no longer smooth in the primal variable. SOLID is defined by alternating optimistic sub-gradient steps on the allocation and the dual variable.

6

LINEAR STOCHASTIC PARTIAL MONITORING

Partial monitoring is a framework for sequential decision-making in which the learner does not directly observe the reward [139]. Instead, the learner observes a feedback from a distribution that is correlated with the reward in a way specified by the model. Decoupling the reward from the observation emphasizes the exploration-exploitation trade-off. It also adds a great amount of flexibility, and a majority of stateless online decision-making problems can be viewed as a special case of partial monitoring.

Our focus is on the stochastic version of the problem with a linear reward and observation structure, which is sometimes referred to as *combinatorial partial monitoring* [40, 109]. Linear partial monitoring strictly generalizes linear bandits. It also captures full-information and semi-bandit feedback, and more exotic models such as dueling bandits. We will discuss more examples in Section 6.1, after we have introduced the model formally.

The main contribution in this chapter is a version of IDS for linear partial monitoring. We show that IDS achieves the optimal worst-case regret rate in terms of the horizon in all games with finitely many actions. The result is complemented with a classification theorem showing that, up to logarithmic factors, the minimax regret of linear finite-action games is either 0, $\tilde{\Theta}(n^{1/2})$, $\tilde{\Theta}(n^{2/3})$ or $\Omega(n)$. All upper bounds are achieved with the same IDS algorithm and without tuning hyper-parameters.

SETTING Let $\mathcal{A}$ be a compact action set. To add flexibility, we use an explicit action-feature mapping $\phi_a : \mathcal{A} \to \mathbb{R}^d, a \mapsto \phi_a$. As before, a fixed parameter $\theta \in \mathbb{R}^d$ defines the reward function $f_\theta(a) = \langle \phi_a, \theta \rangle$ and gaps $\Delta(a) = \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \theta \rangle$. We assume that the model set $\mathcal{M}$ contains the Euclidean ball $\mathcal{M} \supset \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$ and is bounded with $\max_{\theta \in \mathcal{M}} \|\theta\| \leq B$ for some $B \geq 1$. We refer to such parameter sets as *directionally unconstrained*. The inclusion of the unit ball is explicitly used in the construction of the lower-bounds, whereas the upper bounds only require the norm bound.

For each action $a \in \mathcal{A}$, the observation is specified by a linear *feedback map* $M_a : \mathbb{R}^d \to \mathbb{R}^m$, written as a matrix in $\mathbb{R}^{m \times d}$. If the learner chooses an action $a_t \in \mathcal{A}$ in round t, the observation is an m-dimensional vector

$y_t = M_{a_t}\theta + \epsilon_t$. The noise vector $\epsilon_t \in \mathbb{R}^m$ is assumed to be conditionally independent and ρ-sub-Gaussian, that is

$$\forall\, u \in \mathbb{R}^d\,,\ \|u\| = 1\,,\ \forall\, \eta > 0\,, \quad \mathbb{E}_t[\exp(\eta u^\top \epsilon_t)|a_t] \leq \exp(\eta^2\rho^2/2)\,.$$

Throughout, we make the following boundedness assumptions. The feedback maps satisfy $\|M_a\|_2 \leq L$ where $\|\cdot\|_2$ is the operator norm, and the action set satisfies $\operatorname{diam}(\mathcal{A}) \leq 1$. Combined with the assumption that $\|\theta\|_2 \leq B$, this implies $\Delta(a) \leq B$ for all $a \in \mathcal{A}$. The functions $a \mapsto \phi_a$ and $a \mapsto M_a$ are continuous to ensure existence of the IDS distribution.

6.1 EXAMPLES

The framework of linear partial monitoring captures many applications and models for sequential decision-making that have been studied independently in the literature. We provide some examples below.

Example 6.1 (Linear Bandits)**.** The linear bandits model and the notation we used in previous chapters is recovered by setting $\phi_a = a = M_a^\top$. Heteroscedastic bandits (Chapter 3) with $\mathcal{A} \subset \mathbb{R}^d$ and noise function $\rho : \mathcal{A} \to \mathbb{R}_{\geq 0}$ are modeled with $\phi_a = a$ and $M_a = \phi_a/\rho(a)$.

Example 6.2 (Full Information Feedback)**.** The observation operators can be defined to yield *more* information than in the bandit case, up to revealing the parameter in each round, e. g. $M_a = \mathbf{1}_d$. Unlike in online learning where the data is adversarial, full information feedback in the stochastic setting reduces to an estimation problem. In particular, nothing prevents the learner from playing the estimated greedy action in each round. This is exactly what IDS does, since the information gain is the same for all actions.

Example 6.3 (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* [34, 115]. When choosing an action, the learner observes the reward of all adjacent actions in the graph defined on the action set. 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,

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

We ignore the technicality that the observation dimension now in general depends on the action. 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) and full information (fully connected graph). We point out that [111] studied a version of Bayesian IDS in this setting.

Example 6.4 (Dueling Bandits)**.** In dueling bandits, the learner chooses pairs of actions and receives feedback on which of the two actions has higher reward [156, 175]. We emphasize that the dueling bandit model has many intricate variants and a vast literature on its own [22]. By far not every dueling bandit is easily modeled as a partial monitoring game. With additional assumptions, Gajane & Urvoy [62] 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, which we detail on in Section 8.2.4.

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 reward difference $y_t = \langle \phi_{a_t} - \phi_{b_t}, \theta \rangle + \epsilon_t$. The learner collects (unobserved) reward for both actions, corresponding to 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 6.5 (Localized Dueling Bandits)**.** We propose a variant of the dueling bandit that combines the graph structure from Example 6.3 with dueling feedback. Concretely, we allow utility-based dueling feedback as in Example 6.4 *only* for actions that are connected in the graph. 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 6.6 (Combinatorial Partial Monitoring)**.** In the *combinatorial* bandit problem the typical requirement is that the offline problem $\arg\max_{a \in \mathcal{A}} \langle \phi_a, \theta \rangle$ can be solved efficiently for all $\theta \in \mathcal{M}$. The action set can be exponentially large, and learning algorithms for this scenario are designed to only use the solver for the offline problem. The IDS algorithm we introduce in

this chapter is *not* oracle efficient, although the theory still applies. The combinatorial setting is the motivation for the linear partial monitoring setting in the work by Lin *et al.* [109] and Chaudhuri & Tewari [40], and both previously proposed methods are oracle efficient.

The combinatorial version of the multi-armed bandit setting [36] makes more specific assumptions on the feedback sturucture. Let $\mathcal{I}$ be an index set with associated features ϕ_a for $a \in \mathcal{A}$. The action set is a set of subsets $\mathcal{A} \subset 2^{\mathcal{I}}$. 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 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 B actions at once, i. e. $\mathcal{A} = \{a \subset \mathcal{I} : |a| = B\}$. The example also displays the exponential blow-up of the action set. Finding an version of IDS that exploits the combinatorial structure for efficient computation is an interesting question for future work.

Example 6.7 (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 recently proposed by Fiez *et al.* [56] in the context of best arm identification. A toy example that fits into this category is that of *"apple tasting"* [35]. 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. The reader who prefers *whisky tastings* is encouraged to read the paper by Fiez *et al.* [56], which also contains additional examples.

6.2 LOCAL AND GLOBAL OBSERVABILITY

How fast the learner can determine a near optimal action in a partial monitoring game depends on the geometric structure of actions and feedback. Some more terminology is useful. An action $a \in \mathcal{A}$ is called *Pareto optimal* if ϕ_a is an extreme point of the convex hull of the features, $\operatorname{conv}(\phi_a : a \in \mathcal{A})$.

A linear partial monitoring game is called *finite* if it has finitely many Pareto optimal actions. The set of actions that are optimal for $\theta \in \mathcal{M}$ is

$$\mathcal{A}^*(\theta) = \{a \in \mathcal{A} : \langle \phi_a, \theta \rangle = \max_{b \in \mathcal{A}} \langle \phi_b, \theta \rangle\},$$

which is defined on sets $\mathcal{E} \subseteq \mathbb{R}^d$ by $\mathcal{A}^*(\mathcal{E}) = \cup_{\theta \in \mathcal{E}} \mathcal{A}^*(\theta)$. A game is called *globally observable* if

$$\forall a, b \in \mathcal{A}, \quad \phi_a - \phi_b \in \operatorname{span}(\operatorname{im}(M_c^\top) : c \in \mathcal{A}). \tag{6.1}$$

The condition for global observability in Eq. (6.1) ensures that for any two actions $a, b \in \mathcal{A}$ and parameter $\theta \in \mathcal{M}$, the reward difference $\langle \phi_a - \phi_b, \theta \rangle$ can be estimated from data collected by the learner using appropriate actions. A game is called *locally observable* if for every convex set $\mathcal{E} \subset \mathbb{R}^d$,

$$a, b \in \mathcal{A}^*(\mathcal{E}), \quad \phi_a - \phi_b \in \operatorname{span}(\operatorname{im}(M_c^\top) : c \in \mathcal{A}^*(\mathcal{E})). \tag{6.2}$$

Local observability is a stronger assumption than global observability, and leads to improved regret. The definition implies that the learner can estimate any reward difference among actions that appear plausible optimal for confidence set $\mathcal{E} \subset \mathcal{M}$, by playing only on the same set of actions.

In Section 6.4.1 we give several equivalent definitions of local and global observability, which are used in the construction of the lower bounds and are sometimes easier to check. The reader familiar with partial monitoring is assured that our definitions coincide with the classical notion based on the neighborhood graph, which we formally show in the same section.

6.2.1 *Alignment Constants*

The regret upper bounds rely on quantifying the notions of global and local observability. In particular, the conditions in Eqs. (6.1) and (6.2) can be satisfied while the signal-to-noise ratio of the observations from the feedback maps M_a is arbitrarily small. The constants that appear in the analysis depend on the degree to which the learner can efficiently gain information, which roughly depends on how well the observation operator M_{a_t} is aligned with a direction $\phi_a - \phi_b$ in which we aim to improve the accuracy of our estimation.

For a convex set $\mathcal{E} \subset \mathcal{M}$, we define the *extended plausible maximizer set*,

$$\mathcal{A}^+(\mathcal{E}) \triangleq \{a \in \mathcal{A} : \phi_a \in \operatorname{conv}(\phi_b : b \in \mathcal{A}^*(\mathcal{E})). \tag{6.3}$$

One easily checks that any $a \in \mathcal{A}^+(\mathcal{E})$ satisfies $\Delta(a) \le \max_{b \in \mathcal{A}^*(\mathcal{E})} \Delta(b)$, hence $\mathcal{A}^+(\mathcal{E})$ contains actions for which the regret is not larger than for any action in $\mathcal{A}^*(\mathcal{E})$. The *worst-case alignment* for $\mathcal{E}$ is

$$\alpha(\mathcal{E}) \triangleq \max_{\nu \in \mathbb{R}^d} \max_{a,b \in \mathcal{A}^*(\mathcal{E})} \min_{c \in \mathcal{A}^+(\mathcal{E})} \frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\|M_c \nu\|^2} \,. \tag{6.4}$$

The fact that the direction of ν is unconstrained in the maximization reflects the assumption that the affine hull of $\mathcal{M}$ spans $\mathbb{R}^d$. Note that the fraction is scale-invariant in ν. One should think of $\nu \in \mathbb{R}^d$ as a small perturbation of the parameter θ that flips the sign of the loss differences $\langle \phi_a - \phi_b, \theta + \nu \rangle$, whereas the denominator captures how efficiently the learner can detect such a change with a statistical test.

The definition of $\alpha(\mathcal{E})$ relates to global and local observability in a precise way. In particular, a game is globally observable if and only if $\alpha(\mathcal{M}) < \infty$. A game is locally observable, if and only if for all convex $\mathcal{E} \subset \mathcal{M}$, we get $\alpha(\mathcal{E}) < \infty$. The equivalence follows from Lemma 6.1 below and Lemma 6.6 in Section 6.4.1. On locally observable games with infinite action sets, note that $\sup_{\mathcal{E} \subset \mathcal{M}} \alpha(\mathcal{E}) = \infty$ is possible, which can slow down learning arbitrarily without further assumptions. This is one of the reasons the classification result in Section 6.4 holds only for finite action sets.

The next lemma provides an upper bound on the alignment constant that depends directly on the feedback maps.

Lemma 6.1. *Let $\mathcal{E} \subset \mathcal{M}$ be convex and locally observable in the sense that*

$$\forall\, a, b \in \mathcal{A}^*(\mathcal{E}) \,, \quad \phi_a - \phi_b \in \operatorname{span}(\operatorname{im}(M_c^\top) : c \in \mathcal{A}^+(\mathcal{E})) \,.$$

Let $l = |\mathcal{A}^+(\mathcal{E})|$ be the number of actions in the extended plausible maximizer set of $\mathcal{E}$. Denote by $M = (M_c^\top : c \in \mathcal{A}^+(\mathcal{E}))^\top \in \mathbb{R}^{ml \times d}$ the matrix that stacks the feedback maps and by $r = \operatorname{rank}(M) \le d$ the rank of M. Further, let $\mathcal{B} \subset \mathcal{A}^+(\mathcal{E})$ be any subset of $|\mathcal{B}| \le r$ actions such that the matrix stack $B = (M_c^\top : c \in \mathcal{A}^+(\mathcal{E}))^\top \in \mathbb{R}^{mr \times d}$ has $\operatorname{rank}(B) = r \le d$. Then

$$\alpha(\mathcal{E}) \le \max_{a,b \in \mathcal{A}^*(\mathcal{E})} \min_{\substack{w \in \mathbb{R}^{ml} \\ M^\top w = \phi_a - \phi_b}} \left(\sum_{c \in \mathcal{A}^+(\mathcal{E})} \|w_c\| \right)^2 \le r \lambda_{\min}(B^\top B)^{-1} \,.$$

Proof. Let $a, b \in \mathcal{A}^*(\mathcal{E})$ with $a \ne b$. By assumption, there exists a $w \in \mathbb{R}^{lm}$ such that $\phi_a - \phi_b = M^\top w$ with $w \ne 0$. Then,

$$\langle \phi_a - \phi_b, \nu \rangle^2 = \langle M^\top w, \nu \rangle^2 = \langle w, M\nu \rangle^2 = \left(\sum_{c \in \mathcal{A}^+(\mathcal{E})} \langle w_c, M_c \nu \rangle \right)^2 ,$$

where we denote by $w_c \in \mathbb{R}^m$ the weights corresponding to M_c. An application of the Cauchy-Schwarz inequality shows:

$$\frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\max_{c \in \mathcal{A}^+(\mathcal{E})} \|M_c \nu\|^2} \leq \frac{\left(\sum_{c \in \mathcal{A}^+(\mathcal{E})} \|w_c\| \|M_c \nu\|\right)^2}{\max_{c \in \mathcal{A}^+} \|M_c \nu\|^2} \leq \left(\sum_{c \in \mathcal{A}^+(\mathcal{E})} \|w_c\|\right)^2 .$$

The first claim in the lemma follows by optimizing over the estimation vectors w. We compute a specific solution with the matrix B specified in the lemma statement. Let $\tilde{w} = (BB^\top)^\dagger B(\phi_a - \phi_b)$ where $(BB^\top)^\dagger$ denotes the pseudo inverse of $(BB^\top)$. Therefore, using the properties of the pseudo inverse and $\|\phi_a - \phi_b\| \leq 1$,

$$\left(\sum_{c \in \mathcal{B}} \|\tilde{w}_c\|\right)^2 \leq r\|\tilde{w}\|^2 \leq r\lambda_{\max}\left(B^\top (BB^\top)^\dagger (BB^\top)^\dagger B\right) = r\lambda_{\min}(B^\top B)^{-1} .$$

□

6.2.2 *Examples*

Example 6.8 (Locally Observable Games)**.** From the examples mentioned in Section 6.1, the following models are *locally observable*.

- With *bandit* feedback, the reward difference of any two actions $a, b \in \mathcal{A}$ can be estimated by playing both actions, which means bandits are locally observable. Specifically $w_a \phi_a + w_b \phi_b = \phi_a - \phi_b$ is satisfied for $w_a = -w_b = 1$, hence with Lemma 6.1, we get $\alpha(\mathcal{E}) \leq (\|w_a\| + \|w_b\|)^2 = 4$ for any $\mathcal{E} \subset \mathcal{M}$. The heteroscedastic bandit feedback we discussed in Chapter 3 can be modeled by letting $\phi_a = a$ and $M_a^\top = \rho(a)^{-1}\phi_a$.
- Semi-bandit and full-information feedback contain *more* information than the bandit observation, hence are locally observable.
- The dueling bandit model as defined in Example 6.4 is locally observable. To compute the alignment constant, assume that $\phi_{a,b}, \phi_{a',b'} \in \mathcal{A}^*(\mathcal{E})$ for some $\mathcal{E} \subset \mathcal{M}$. But then also $\phi_{c,c} \in \mathcal{A}^*(\mathcal{E})$ for all $c \in \{a, a', b, b'\}$ since two different actions are never uniquely optimal. It follows that $(a, a'), (b, b') \in \mathcal{A}^+(\mathcal{E})$. Estimating the reward difference between (a, b) and (a', b') with observations from $\mathcal{A}^+(\mathcal{E})$ is done as follows:

$$\phi_{a,b} - \phi_{a',b'} = \phi_a + \phi_b - (\phi_{a'} + \phi_{b'}) = w_1 M_{a,a'} + w_2 M_{b,b'} ,$$

which is satisfied for $w_1 = w_2 = 1$. Therefore, $\alpha(\mathcal{E}) \leq 4$ by Lemma 6.1. We note that the upper bounds directly depend on the alignment constants and are tight for the linear bandit model. In this sense, the utility-based dueling bandit model is not more difficult than the bandit setting.

Example 6.9 (Globally Observable Games)**.** The following games are globally observable, but not necessarily locally observable.

- For a (simplified) version of the localized dueling bandit (Example 6.5), let $d > 3$ and consider the action set $\mathcal{A} = [d]$ with features $\phi = e_a$ and feedback maps defined for $a < d$ as

$$M_a = e_a - e_{a+1}\,,$$

and $M_d = 0$. The learner only observes the reward difference of two consecutive actions in the natural ordering. Clearly, for any $a < b \in \mathcal{A}$, we can write $\phi_a - \phi_b = \sum_{c=a}^{b-1} \phi_c$, hence the game is globally observable with $\alpha(\mathcal{M}) \leq d^2$. On the other hand define the following convex set:

$$\mathcal{E} = [1,2] \times \underbrace{[-1,0] \times \cdots \times [-1,0]}_{d-2 \text{ times}} \times [1,2]$$

When $\theta \in \mathcal{E}$, playing either the first or the last action is optimal, i. e. $\mathcal{A}^*(\mathcal{E}) = \{1, d\}$. However, estimating the reward difference $\phi_1 - \phi_d$ is only possible by playing the other actions, which are provably suboptimal for $\theta \in \mathcal{E}$. This means the game is not locally observable.

- Transductive games (Example 6.7) are not locally observable in general, since the learner suffers a constant regret for information.

6.3 IDS FOR LINEAR PARTIAL MONITORING

In this section, we introduce a version of IDS for linear partial monitoring. The first step is to generalize the previous definitions of the gap estimate and the information gain function to the new feedback model. In round t, the least-squares estimator with regularizer $\lambda > 0$ is

$$\hat{\theta}_t = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \|M_{a_s}\theta - y_s\|^2 + \lambda\|\theta\|^2 = V_t^{-1} \sum_{s=1}^{t-1} M_{a_s}^\top y_s\,, \qquad (6.5)$$

where the covariance matrix is $V_t = \sum_{s=1}^{t-1} M_{a_s}^\top M_{a_s} + \lambda \mathbf{1}_d$. For the analysis we require that $\lambda \geq L^2$, which implies that $\|M_a V_t^{-1} M_a^\top\|_2^2 \leq \frac{1}{\lambda}\|M_a M_a^\top\|_2^2 \leq 1$.

Algorithm 6: IDS for Linear Partial Monitoring

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

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

$V_t \leftarrow \sum_{s=1}^{t-1} M_{a_s} M_{a_s}^\top + \lambda \mathbf{1}_d$ `// least-squares estimation`

$\hat{\theta}_t \leftarrow V_t^{-1} \sum_{s=1}^{t-1} M_{a_s}^\top y_s$

$\beta_t^{1/2} \leftarrow \rho \sqrt{\log \det(V_t) + 2 \log(t^2)} + \sqrt{\lambda} B$

`// gap estimates`

$\hat{\Delta}_t(a) \leftarrow \min\Big\{ \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \hat{\theta}_t \rangle + \beta_t^{1/2} \|\phi_b - \phi_a\|_{V_t^{-1}}, B \Big\}$

$I_t^{\mathrm{DET}}(a) \leftarrow \frac{1}{2} \log\Big(\mathbf{1}_m + M_a V_t^{-1} M_a^\top \Big)$ `// information gain`

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

$a_t \sim \mu_t$

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

This assumption can be lifted with a minor modification of the proof while preserving the same scaling. The corresponding confidence set is

$$\mathcal{E}_{t,\delta} = \{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t}^2 \le \beta_{t,\delta} \} .$$

A direct extension of Lemma 3.1 shows that with the confidence coefficient $\beta_{t,\delta}^{1/2} = \rho \sqrt{\log\left(\frac{\det(V_t)}{\det(\lambda \mathbf{1}_d)} \right) + 2 \log \frac{1}{\delta}} + \sqrt{\lambda} B$ and provided that $\|\theta\| \le B$,

$$\mathbb{P}[\forall t \ge 1, \theta \in \mathcal{E}_{t,\delta}] \ge 1 - \delta .$$

In the algorithm we use $\beta_t \triangleq \beta_{t,1/t^2}$ and $\mathcal{E}_t \triangleq \mathcal{E}_{t,1/t^2}$. The gap estimate is

$$\hat{\Delta}_t(a) = \min \Big\{ \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \hat{\theta}_t \rangle + \beta_{t,1/t^2}^{1/2} \|\phi_b - \phi_a\|_{V_t^{-1}},\, B \Big\} , \tag{6.6}$$

which is explicitly truncated using that $\Delta(a) \le B$. The estimate is chosen so that with high probability $\Delta(a) \le \hat{\Delta}_t(a)$ for all $a \in \mathcal{A}$ and all rounds t. For the information gain we use

$$I_t^{\mathrm{DET}}(a) = \frac{1}{2} \log \det\Big(\mathbf{1}_m + M_a V_t^{-1} M_a^\top \Big) . \tag{6.7}$$

The definition generalizes the information gain function we used in Section 3.2.2. The Bayesian analogue is the mutual information $\mathbb{I}_t(\theta; y_t | a_t = a)$

when using a Gaussian prior and likelihood function. As before, the total information gain is computed with the help of the matrix determinant lemma (Lemma D.1) and telescoping,

$$\gamma_n = \sum_{t=1}^{n} I_t^{\mathrm{DET}}(a_t) = \frac{1}{2}\big(\log\det(V_{n+1}) - \log\det(\lambda \mathbf{1}_d)\big)\,.$$

Note that $\beta_n^{1/2} = \rho\sqrt{\gamma_{n-1} + \log(\frac{1}{\delta})} + \sqrt{\lambda}B$. Along the same lines of (3.1), it follows that

$$\gamma_n \le d\log\Big(1 + \tfrac{nmL^2}{d\lambda}\Big)\,. \tag{6.8}$$

The complete IDS algorithm for linear partial monitoring is summarized in Algorithm 6. Since the maximum in the definition of $\hat{\Delta}_t(a)$ depends on a, in general we need $\mathcal{O}(d^2|\mathcal{A}|^2)$ operations per round to compute all gaps on a finite action set. In Section 6.3.2, we show that a different definition can be used, which reduces the computation complexity to $\mathcal{O}(d^2|\mathcal{A}|)$ while preserving the theoretical guarantees.

6.3.1 *Regret Bounds*

The first result is a regret bound for globally observable games. Assuming only global observability means that the learner has to play clearly suboptimal actions to obtain information in general. As an instructive example, considers the case where a single action $\tilde{a} \in \mathcal{A}$ provides full-information feedback, $M_{\tilde{a}} = \mathbf{1}_d$, whereas all other actions, including the optimal action, provide no information (e.g. $M_a = 0$). A short calculation and Lemma 2.5 confirm that IDS samples the informative action $\tilde{a}$ with probability

$$\tilde{p}_t = \min\left(\frac{\delta_t}{\hat{\Delta}_t(\tilde{a}) - \delta_t}, 1\right),$$

where $\delta_t = \min_{a\in\mathcal{A}} \hat{\Delta}_t(a)$ is the smallest gap estimate. Let $T_{\tilde{a}}(t)$ be the number of times the informative action has been chosen up to step t. Ignoring log factors and constants, a reasonable scaling for the smallest gap is $\delta_t = T_{\tilde{a}}(t)^{-1/2}$. Since the cost for $\tilde{a}$ is at least constant, we can approximately fix $p_t = \delta_t$, and compute the update step $\mathbb{E}_t[T_{\tilde{a}}(t+1)] = T_{\tilde{a}}(t) + T_{\tilde{a}}(t)^{-1/2}$. If we initialize with $T_{\tilde{a}}(1) = 1$ and simulate the dynamics (or solve the limiting differential equation), we find that $T_{\tilde{a}}(n) \approx n^{2/3}$. Hence, in the example the regret of IDS is $\mathfrak{R}_n \approx n^{2/3}$.

The example captures the most difficult scenario, since in any globally observable game, the learner can estimate all gaps and the cost per step is bounded. The theorem below follows this intuition, and confirms that the regret of IDS in all globally observable games is at most $\mathfrak{R}_n \leq \tilde{\mathcal{O}}(n^{2/3})$.

Theorem 6.1. *On any globally observable game the regret of IDS (Algorithm 6) satisfies,*

$$\mathfrak{R}_n \leq n^{2/3}(32\alpha(\mathcal{M})B\mathbb{E}[\beta_n]\mathbb{E}[\gamma_n])^{1/3} + \mathcal{O}(B)\,.$$

With the bound on β_n and γ_n from Eq. (6.8) and $\lambda \geq L^2$, the regret bound translates to $\mathfrak{R}_n \leq \mathcal{O}((\alpha(\mathcal{M})LB)^{1/3}(dn\log(n)^{2/3}))$. Bounds on the alignment constant generally depend on the feedback maps as in Lemma 6.1. The lower bounds in Section 6.4 show that the dependence on n cannot be improved for globally observable finite games that are not locally observable, up to logarithmic factors. In infinite games, the situation is more complicated. Curvature of the action set can lead to fast rates, even if the local observability condition is not satisfied, as we show in work with collaborators [90, §2.4].

Proof of Theorem 6.1. We start by establishing an inequality which shows the existence of an action with large information gain compared to the gap estimate of the greedy action. Let $\hat{a} = \arg\max_{a\in\mathcal{A}}\langle\phi_a, \hat{\theta}_t\rangle$. Clearly,

$$\begin{aligned}\min_{a\in\mathcal{A}}\hat{\Delta}_t(a) &\leq \max_{b\in\mathcal{A}}\langle\phi_b - \phi_{\hat{a}}, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi_b - \phi_{\hat{a}}\|_{V_t^{-1}} \\ &\leq \beta_t^{1/2}\max_{a,b\in\mathcal{A}}\|\phi_a - \phi_b\|_{V_t^{-1}}\,.\end{aligned}$$

We continue with the squared norm and the shorthand $\alpha = \alpha(\mathcal{M})$,

$$\begin{aligned}\max_{a,b\in\mathcal{A}}\|\phi_a - \phi_b\|^2_{V_t^{-1}} &= \max_{\nu\in\mathbb{R}^d}\max_{a,b\in\mathcal{A}}\frac{\langle\phi_a - \phi_b, V_t^{-1/2}\nu\rangle^2}{\|\nu\|} \\ &\overset{(i)}{\leq} \alpha\max_{\nu\in\mathbb{R}^d}\max_{c\in\mathcal{A}}\frac{\|M_c^\top V_t^{-1/2}\nu\|^2}{\|\nu\|^2} \\ &= \alpha\max_{c\in\mathcal{A}}\lambda_{\max}(M_c^\top V_t^{-1}M_c) \\ &\overset{(ii)}{\leq} 2\alpha\max_{c\in\mathcal{A}}\log\det(\mathbf{1}_m + M_c^\top V_t^{-1}M_c) \\ &= 4\alpha\max_{c\in\mathcal{A}} I_t^{\mathrm{DET}}(c)\,.\end{aligned}$$

For (i) we used the definition of the alignment constant in Eq. (6.4). For (ii), we used the inequality $x \le 2\log(1+x)$ for $x \in [0,1]$ and that the eigenvalues of $M_c^\top V_t^{-1} M_c$ are bounded in $[0,1]$ by the assumption that $\lambda \ge L^2$ and $\|M_c\| \le L$. Combining the last two displays shows that

$$\min_{a\in\mathcal{A}} \hat{\Delta}_t(a)^2 \le 4\alpha\beta_t \max_{c\in\mathcal{A}} I_t^{\mathrm{DET}}(c)\,. \tag{6.9}$$

It remains to bound the information ratio.

$$\Psi_t(\mu_t^{\mathrm{IDS}}) \overset{(i)}{\le} \min_{c\in\mathcal{A}} \frac{4\delta_t \hat{\Delta}_t(c)}{I_t^{\mathrm{DET}}(c)} \overset{(ii)}{\le} \min_{c\in\mathcal{A}} \frac{4\delta_t B}{I_t^{\mathrm{DET}}(c)} \overset{(iii)}{\le} \frac{16\alpha\beta_t B}{\hat{\Delta}_t(\hat{a}_t)} \overset{(iv)}{\le} \frac{32\alpha\beta_t B}{\hat{\Delta}_t(\mu_t^{\mathrm{IDS}})} \tag{6.10}$$

Step (i) follows from (2.6) and for (ii) we used that $\hat{\Delta}_t(c) \le B$. Then, (iii) follows from Eq. (6.9), and the last step (iv) uses Lemma 2.6. Rearranging shows that the generalized information ratio (Eq. (2.7)) is bounded,

$$\Psi_{3,t}(\mu_t^{\mathrm{IDS}}) = \frac{\hat{\Delta}_t(\mu_t^{\mathrm{IDS}})^3}{I_t^{\mathrm{DET}}(\mu_t^{\mathrm{IDS}})} \le 32\alpha\beta_t B\,.$$

The result follows from Theorem 2.2 and noting that the confidence level is chosen such that $\sum_{t=1}^n \mathbb{E}\big[\Delta(a_t) - \hat{\Delta}_t(a_t)\big] \le \mathcal{O}(B)$. □

Local observability is a stronger assumption that greatly eases learning and allows for faster regret rates. We say a game is *uniformly local observable* if $\alpha_0 = \sup_{\mathcal{E}\subset\mathcal{M} \text{ convex}} \alpha(\mathcal{E}) < \infty$. Finite locally observable games are always uniformly local observable. The next theorem shows that the regret of IDS on uniformly local observable games is at most $\mathcal{O}(\sqrt{\alpha_0 \mathbb{E}[\beta_n]\mathbb{E}[\gamma_n] n})$. More precisely, α_0 can be replaced with the average alignment $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha(\mathcal{E}_t) \le \alpha_0$, defined on the sequence of confidence sets realized by the algorithm. This shows how IDS adapts towards the current instance of the partial monitoring game, and can sometimes lead to faster rates even on games that are not globally observable.

From the proof of the theorem it is evident that randomization is not necessary to bound information ratio in locally observable games. Deterministic IDS, which optimizes the ratio over a deterministic action choice $a_t^{\mathrm{DIDS}} = \arg\min_{a\in\mathcal{A}} \Psi_t(a)$, achieves the same upper bound with our analysis. Randomization is however essential for globally observable games.

Theorem 6.2. *On any locally observable game, IDS satisfies*

$$\mathfrak{R}_n \le 4\sqrt{\mathbb{E}[\bar{\alpha}_n\beta_n]\mathbb{E}[\gamma_n] n} + \mathcal{O}(B)\,,$$

where $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha(\mathcal{E}_t)$ *is the averaged realized local alignment constant.*

Further bounding γ_n and β_n using Eq. (6.8) and $\lambda \geq L^2$ yields regret $\mathfrak{R}_n \leq \mathcal{O}\big(\sqrt{n\mathbb{E}[\alpha_n]L}\, d \log(n) + B\big)$. In particular, for bandits and dueling bandits we have $\alpha(\mathcal{E}_t) \leq 4$ and the regret is $\mathfrak{R}_n \leq \mathcal{O}(\sqrt{nd}\log(n))$. On linear bandits, the bound is the same as for UCB and matches the lower bound on large action sets up to logarithmic factors [103, §24].

Proof of Theorem 6.2. Local observability allows to restrict the action choice to plausible maximizers $\mathcal{A}^*(\mathcal{E}_t)$. Fix any $c \in \mathcal{A}^*(\mathcal{E}_t)$ and let $\nu_c \in \mathcal{E}_t$ be a paramater for which c is optimal. Then

$$\begin{aligned}
\hat{\Delta}_t(c) &\overset{(i)}{\leq} \max_{b\in\mathcal{A}} \max_{\nu\in\mathcal{E}_t} \langle \phi_b - \phi_c, \nu\rangle \\
&\overset{(ii)}{=} \max_{b\in\mathcal{A}^*(\mathcal{E}_t)} \max_{\nu\in\mathcal{E}_t} \langle \phi_b - \phi_c, \nu - \nu_c\rangle + \langle \phi_b - \phi_c, \nu_c\rangle \\
&\overset{(iii)}{\leq} \max_{b\in\mathcal{A}^*(\mathcal{E}_t)} \max_{\nu\in\mathcal{E}_t} \langle \phi_b - \phi_c, \nu - \nu_c\rangle \\
&\overset{(iv)}{\leq} 2\beta_t^{1/2} \max_{b\in\mathcal{A}^*(\mathcal{E}_t)} \|\phi_b - \phi_c\|_{V_t^{-1}} .
\end{aligned}$$

For (i) we used the definition of the gap estimate and dropped the truncation. For (ii) observe that the first maximum is attained on $\mathcal{A}^*(\mathcal{E}_t)$. Step (iii) is by the choice of ν_c, and (iv) uses Cauchy-Schwarz and the definition of $\mathcal{E}_t$. Hence, for all $c \in \mathcal{A}^*(\mathcal{E}_t)$,

$$\hat{\Delta}_t(c) \leq 2\beta_t^{1/2} \max_{a,b\in\mathcal{A}^*(\mathcal{E}_t)} \|\phi_a - \phi_b\|_{V_t^{-1}} .$$

The same inequality holds for $c \in \mathcal{A}^+(\mathcal{E}_t)$ by the definition of the extended plausible maximizer set in Eq. (6.3), and because the function $\phi \mapsto \max_{\nu\in\mathcal{E}_t, b\in\mathcal{A}} \langle \phi_b - \phi, \nu\rangle$ is convex. Similar to Eq. (6.9) and using the definition of the local alignment constant, we get

$$\hat{\Delta}_t(c)^2 \leq 16\beta_t \alpha(\mathcal{E}_t) \max_{\tilde{c}\in\mathcal{A}^+(\mathcal{E}_t)} I_t^{\mathrm{DET}}(\tilde{c}) .$$

We bound the information ratio by optimizing over Dirac distributions supported on $\mathcal{A}^+(\mathcal{E}_t)$,

$$\Psi_t(\mu_t^{\mathrm{IDS}}) \leq \min_{c\in\mathcal{A}^+(\mathcal{E}_t)} \frac{\hat{\Delta}_t(c)^2}{I_t^{\mathrm{DET}}(c)} \leq 16\beta_t \alpha(\mathcal{E}_t) .$$

The result follows from Corollary 2.1 and bounding the estimation error. □

6.3.2 *A Faster Gap Estimate*

The maximum that appears in the definition of the gap estimate $\hat{\Delta}_t(a)$ in Eq. (6.6) cannot be computed independently of a. This can be improved by introducing the empirical maximizer $\hat{a}_t = \arg\max_{a\in\mathcal{A}}\langle \phi_a, \hat{\theta}_t\rangle$:

$$\begin{aligned}\hat{\Delta}_t(a) &= \max_{b\in\mathcal{A}}\langle \phi_b - \phi_a, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi_b - \phi_a\|_{V_t^{-1}} \\ &\le \max_{b\in\mathcal{A}}\langle \phi_b - \phi_a, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi_b - \phi_{\hat{a}_t}\|_{V_t^{-1}} + \beta_t^{1/2}\|\phi_{\hat{a}_t} - \phi_a\|_{V_t^{-1}} \\ &\le 2\Big(\max_{b\in\mathcal{A}}\langle \phi_b - \phi_a, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi_b - \phi_{\hat{a}_t}\|_{V_t^{-1}}\Big)\,.\end{aligned}$$

The upper bound warrants an alternative definition of the gap estimate,

$$\tilde{\Delta}_t(a) \triangleq \max_{b\in\mathcal{A}}\langle \phi_b - \phi_a, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi_b - \phi_{\hat{a}_t}\|_{V_t^{-1}}\,. \tag{6.11}$$

In particular, $u_t \triangleq \max_{b\in\mathcal{A}}\langle \phi_b, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi_b - \phi_{\hat{a}_t}\|_{V_t^{-1}}$ is independent of a and we can compute the gap estimate via $\tilde{\Delta}_t(a) = u_t - \langle \hat{\theta}_t, \phi_a\rangle$.

The argument above is easily repeated to show $\tilde{\Delta}_t(a) \le 2\hat{\Delta}_t(a)$. Hence, Algorithm 6 defined with the gap estimate $\tilde{\Delta}_t(a)$ is immediately seen to satisfy the same regret bounds up to a constant multiplicative factor as in Theorems 6.1 and 6.2. The constants can be improved by more carefully reproducing the steps in the proof. The asymptotic analysis in Chapter 5 uses a similarly relaxed estimator and the analysis suggests that nothing is lost in the limit of a large horizon. Combined with the approximate version of IDS (Section 2.1.1) and incremental updates for the least-square estimate, the overall complexity is thus reduced to $\mathcal{O}(|\mathcal{A}|d^2 n)$ over n rounds, compared to $\mathcal{O}(|\mathcal{A}|^2 d^2 n)$ required for the direct implementation.

6.3.3 *Directed Information Gain*

The definition of information gain function in Eq. (6.7) is a conservative choice that is primarily motivated by the worst-case analysis. It captures the increase of the log-determinant potential and does not depend on the current estimate of the parameter or an estimate of the best action.

For a fixed $w \in \mathbb{R}^d$ we define the *directed information gain*,

$$I_t(a; w) \triangleq \frac{1}{2}\Big(\log\big(\|w\|^2_{V_t^{-1}}\big) - \log\big(\|w\|^2_{(V_t|M_a)^{-1}}\big)\Big)\,. \tag{6.12}$$

The definition corresponds to the mutual information $\mathbb{I}(\langle w, \theta\rangle; y_t | a_t = a)$ which measures the Gaussian entropy reduction of θ projected onto the

subspace spanned by w. The next lemma shows that the new information gain is strictly smaller, $I_t(a;w) \leq I_t(a)$, which can be interpreted as an *information processing inequality*. In particular, the total information gain can be bounded as before.

Lemma 6.2 (Information processing)**.** *For any $w \in \mathbb{R}^d$ and $a \in \mathcal{A}$, it holds that $I_t(a;w) \leq I_t(a)$.*

Proof. The proof uses basic linear algebra, specifically the Sherman-Morrison formula (Lemma D.2) and the matrix determinant lemma (Lemma D.1).

$$
\begin{aligned}
&2I_t(a;w) \\
&= \log\left(\frac{\|w\|^2_{V_t^{-1}}}{\|w\|^2_{(V_t|M_a)^{-1}}}\right) \\
&= -\log\left(1 - \frac{w^\top V_t^{-1} M_a (\mathbf{1}_m + M_a^\top V_t^{-1} M_a)^{-1} M_a^\top V_t^{-1} w}{\|w\|^2_{V_t^{-1}}}\right) \\
&\leq \max_{v \in \mathbb{R}^d : \|v\|_2 = 1} -\log\left(v^\top v - v^\top V_t^{-1/2} M_a (\mathbf{1}_m + M_a^\top V_t^{-1} M_a)^{-1} M_a^\top V_t^{-1/2} v\right) \\
&= \max_{v \in \mathbb{R}^d : \|v\|_2 = 1} -\log\left(v^\top \left(\mathbf{1}_d - V_t^{-1/2} M_a (\mathbf{1}_m + M_a^\top V_t^{-1} M_a)^{-1} M_a^\top V_t^{-1/2}\right) v\right) \\
&= \log\left(\lambda_{\max}\left(\left(\mathbf{1}_d - V_t^{-1/2} M_a (\mathbf{1}_m + M_a^\top V_t^{-1} M_a)^{-1} M_a^\top V_t^{-1/2}\right)^{-1}\right)\right). \\
&\leq \log\left(\det\left(\mathbf{1}_d - V_t^{-1/2} M_a (\mathbf{1}_m + M_a^\top V_t^{-1} M_a)^{-1} M_a^\top V_t^{-1/2}\right)^{-1}\right) \\
&= \log\det\left(\mathbf{1}_m + M_a^\top V_t^{-1} M_a\right) = 2I_t^{\mathrm{DET}}(a)\,.
\end{aligned}
$$

The second inequality follows because all eigenvalues of the matrix inside the determinant are not smaller than 1, and then using the generalized matrix determinant lemma to rewrite the expression. □

A direction $w \in \mathbb{R}^d$ that reliably measures the progress toward idenfiying the true maximizer is the most uncertain direction in the set of plausible maximizers,

$$
w_t = \operatorname*{arg\,max}_{\substack{w = \phi_a - \phi_b \\ a,b \in \mathcal{A}^*(\mathcal{E}_t)}} \|w\|^2_{V_t^{-1}}\,. \tag{6.13}
$$

Intuitively, as long as $\|w_t\|_{V_t^{-1}} > 0$, there is ambiguity in the plausible maximizer set, that if not resolved, can lead to linear regret of the learner.

From a lower bound perspective, one could argue that w_t is the direction in which the adversary can increase the regret the most with a perturbation of the parameter $\theta + \eta w_t$ for $\eta \in \mathbb{R}$. This motivates the information gain $I_t(a; w_t)$. Note that the definition relies on computing the set of plausible maximizers, which is a more expensive operation compared to computing the undirected information gain $I_t(a)$.

We know already that the information gain $I_t(a; w_t)$ is smaller than $I_t(a)$, but it is not yet clear if the information ratio is still bounded. The next lemma provides an affirmative answer.

Lemma 6.3. *Let $\mathcal{E} \subset \mathcal{M}$ be a convex subset and let $w = \phi_a - \phi_b$ for some $a, b \in \mathcal{A}^*(\mathcal{E})$. Then*

$$\|w\|^2_{V_t^{-1}} \leq 4\alpha(\mathcal{E}_t) \max_{c \in \mathcal{A}^+(\mathcal{E})} I_t(c; w)\,.$$

Proof. First, note that $\frac{1}{2}(\mathbf{1}_d + M_c^\top V_t^{-1} M_c) \preceq \mathbf{1}_d$ by our assumption that $\|M_c\|_2 \leq L$ and $\lambda \geq L^2$. Hence

$$\|M_c^\top V_t^{-1} w\|^2 \leq 2 w^\top V_t^{-1} M_c (\mathbf{1}_m + M_c^\top V_t^{-1} M_c)^{-1} M_c^\top V_t^{-1} w\,.$$

By definition of the alignment constant (Eq. (6.4)),

$$\alpha(\mathcal{E}) \geq \min_{c \in \mathcal{A}^+(\mathcal{E})} \frac{\langle \phi_a - \phi_b, V_t^{-1} w\rangle^2}{\|M_c^\top V_t^{-1} w\|^2} = \min_{c \in \mathcal{A}^+(\mathcal{E})} \frac{\|w\|^4_{V_t^{-1}}}{\|M_c^\top V_t^{-1} w\|^2}\,.$$

Hence, combining the last two displays and rearranging,

$$\begin{aligned}
\|w\|^2_{V_t^{-1}} &\leq \alpha(\mathcal{E}) \max_{c \in \mathcal{A}^+(\mathcal{E})} \frac{\|M_c^\top V_t^{-1} w\|^2}{\|w\|^2_{V_t^{-1}}} \\
&\leq 2\alpha(\mathcal{E}) \max_{c \in \mathcal{A}^+(\mathcal{E})} \frac{w^\top V_t^{-1} M_c (\mathbf{1}_m + M_c^\top V_t^{-1} M_c)^{-1} M_c^\top V_t^{-1} w}{\|w\|^2_{V_t^{-1}}}\,.
\end{aligned}$$

Since $x \leq -\log(1-x)$ for all $x \in [0, 1]$, for the fraction we get

$$\begin{aligned}
&\frac{w^\top V_t^{-1} M_c (\mathbf{1}_m + M_c^\top V_t^{-1} M_c)^{-1} M_c^\top V_t^{-1} w}{\|w\|^2_{V_t^{-1}}} \\
&\leq \log\left(1 - \frac{w^\top V_t^{-1} M_c (\mathbf{1}_m + M_c^\top V_t^{-1} M_c)^{-1} M_c^\top V_t^{-1} w}{\|w\|^2_{V_t^{-1}}}\right) \\
&= \log\left(\|w\|^2_{V_t^{-1}}\right) - \log\left(\|w\|^2_{(V_t | M_c)^{-1}}\right) = 2 I_t(a; w)\,.
\end{aligned}$$

This completes the proof. □

The inequality provided by the lemma is the essential step in the regret analysis that bounds the information ratio. The lemma implies that Algorithm 6 defined with information gain $I_t^{\mathrm{PM}}(a) \triangleq I_t(a, w_t)$ achieves the same regret bounds as in Theorems 6.1 and 6.2.

Corollary 6.1. *Define IDS with the directed information gain $I_t^{PM}(xa; w_t)$ as in Eqs.* (6.12) *and* (6.13)*. Then, on globally observable games, the regret satisfies*

$$\mathfrak{R}_n \leq n^{2/3}(32\alpha(\mathcal{M})B\mathbb{E}[\beta_n]\mathbb{E}[\gamma_n])^{1/3} + \mathcal{O}(B)\,.$$

On locally observable games, the regret satisfies

$$\mathfrak{R}_n \leq 4\sqrt{\mathbb{E}[\bar{\alpha}_n\beta_n]\mathbb{E}[\gamma_n]n} + \mathcal{O}(B)\,,$$

where $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^{n}\alpha(\mathcal{E}_t)$ is the averaged realized local alignment constant.

6.4 CLASSIFICATION OF FINITE LINEAR GAMES

The upper bounds for IDS show that for globally observable games the regret is $\tilde{O}(n^{2/3})$, while for locally observable games it is $\tilde{O}(n^{1/2})$. Of course, if there is only one Pareto optimal action (or duplicates, i. e. actions a, b with $\phi_a = \phi_b$), then the regret vanishes for any algorithm that just plays this action.

Assumption 6.1. *For the remainder of this section we assume that $\mathcal{A}$ is finite.*

The classification theorem complements the upper bounds by showing that this is the best the learner can hope for in linear games with finitely many actions. The *minimax regret* on a model set $\mathcal{M}$ is defined by taking the infimum of the worst-case regret over all possible policies $\pi = (\pi_t)_{t=1}^n$,

$$\mathfrak{R}_n^* = \inf_{\pi}\sup_{\theta\in\mathcal{M}}\mathfrak{R}_n(\pi, \theta)\,.$$

Theorem 6.3 (Classification)**.** *The minimax regret of any finite linear partial monitoring game with $\mathcal{M} = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$ satisfies*

$$\mathfrak{R}_n^* = \begin{cases} 0 & \text{if there is only one Pareto optimal action or duplicates,} \\ \tilde{\Theta}(n^{1/2}) & \text{for locally observable games,} \\ \tilde{\Theta}(n^{2/3}) & \text{for globally observable games,} \\ \Omega(n) & \text{otherwise}\,. \end{cases}$$

Proof. Finite partial monitoring games can be completely classified by considering a graph structure known as the *neighborhood graph* [106]. The classification theorem follows by proving that for games that are not globally observable, the regret is linear in the worst case. For globally observable games that are not locally observable the regret is $\Omega(n^{2/3})$ and that for locally observable games with more than one Pareto optimal action it is $\Omega(n^{1/2})$. The lower bounds are supplied in Appendix C. The proof is completed by combining upper and lower bounds, and carefully checking that all cases have been covered. □

6.4.1 *Neighborhood Graph*

In the context of the lower bounds, it is useful to know that our definitions of local and global observability coincide with the classical notions derived from the *neighborhood graph*, that are now standard in finite partial monitoring [16, 106]. The definition of global observability is further equivalent with the existence of a *global observer set* [109], defined a set of actions $\mathcal{B} \subset \mathcal{A}$ such that $\operatorname{span}(\operatorname{im}(M_b^\top) : b \in \mathcal{B}) = \operatorname{span}(\phi_a - \phi_b : a, b \in \mathcal{A})$.

Let $\Phi = \{\phi_a : a \in \mathcal{A}\}$ be the set of features. Recall that an action $a \in \mathcal{A}$ is Pareto optimal if ϕ_a is an extreme point of $\operatorname{conv}(\Phi)$. An action a is called *degenerate* if the feature ϕ_a is on the boundary of $\operatorname{conv}(\Phi)$, but not an extreme point. Degenerate actions can be optimal, but never uniquely so. Actions with features in the interior of $\operatorname{conv}(\Phi)$ are called *dominated* and are never optimal.

The situation is illustrated in Figure 6.1. Given an action $a \in \mathcal{A}$, the *cell* of a is the set of parameters for which action a is optimal:

$$\mathcal{C}_a = \{\theta \in \mathcal{M} : a = a^*(\theta)\}\,.$$

Since $\mathcal{A}$ is finite, $\operatorname{conv}(\Phi)$ is a polytope and $\mathcal{C}_a$ is either the singleton $\{0\}$ or a polyhedral cone intersected with $\mathcal{M}$. An action a is Pareto optimal if $\dim(\mathcal{C}_a) = d$, which can be seen by observing that $\mathcal{C}_a$ is the normal cone of a with respect to the convex body $\operatorname{conv}(\Phi)$. Pareto optimal actions a and b are called neighbours if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d - 1$, where the dimension of a polytope is defined as the dimension of the smallest affine space containing it. The neighbourhood relation defines a connected graph on the set of Pareto optimal actions. For neighboring Pareto optimal actions a and b let $\mathcal{N}_{ab} = \{c \in \mathcal{A} : \mathcal{C}_a \cap \mathcal{C}_b \subseteq \mathcal{C}_c\}$. Note that, besides a and b, $\mathcal{N}_{ab}$ contains only degenerate actions c with $\dim(\mathcal{C}_c) = d - 1$.

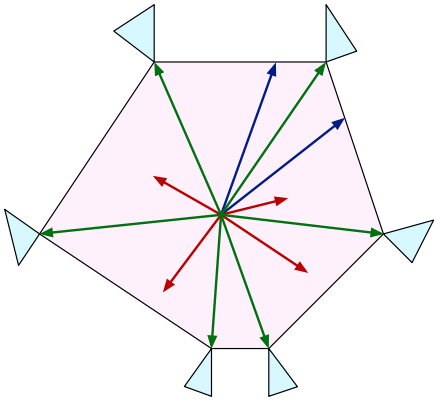


FIGURE 6.1: Green vectors are Pareto optimal, blue ones are degenerated and red are dominated. The light blue cones associated with each Pareto optimal action indicate the direction of θ for which that action is optimal.

Lemma 6.4. *The following conditions equivalently characterize globally observable games with directionally unconstrained parameter sets.:*

i) For all actions $a, b \in \mathcal{A}$, $\phi_a - \phi_b \in \operatorname{span}(\operatorname{im}(M_c^\top) : c \in \mathcal{A})$.

ii) For all Pareto optimal actions $a, b \in \mathcal{P}$, $\phi_a - \phi_b \in \operatorname{span}(\operatorname{im}(M_c^\top) : c \in \mathcal{A})$.

iii) There exists a global observer set.

Proof. For the implication (ii $\Rightarrow$ i), note that Pareto optimal actions are the extreme points of $\operatorname{conv}(\Phi)$, therefore any ϕ_a for $a \in \mathcal{A}$ can be written as a convex combination of Pareto optimal actions. (i $\Rightarrow$ iii) follows by taking $\mathcal{A}$ as global observer set. (iii $\Rightarrow$ ii) immediately follows from the definition of a global observer set. □

The next lemma clarifies the relation of neighboring actions and local observability. Note that the result does not make any assumptions about the model set, which will be used later.

Lemma 6.5. *Let* $\mathcal{A}$ *be finite and* $\mathcal{E} \subset \mathbb{R}^d$ *be any convex set. Then*

i) All Pareto optimal actions in $\mathcal{A}^*(\mathcal{E})$ *are connected on the neighborhood graph.*

ii) For two Pareto optimal actions $a, b \in \mathcal{A}^*(\mathcal{E})$ *it holds that* $\mathcal{N}_{ab} \subset \mathcal{A}^*(\mathcal{E})$.

iii) For any $a \in \mathcal{A}^*(\mathcal{E})$, ϕ_a *can be written as convex combination of Pareto optimal actions in* $\mathcal{A}^*(\mathcal{E})$.

Proof. For points $x, y \in \mathbb{R}^d$, we denote $[x, y] = \{tx + (1-t)y : t \in [0,1]\}$.

i) The proof is intuitively simple. Take any Pareto optimal actions $a, b \in \mathcal{A}^*(\mathcal{E})$ and let $\theta_a \in \mathcal{C}_a$ and $\theta_b \in \mathcal{C}_b$. Then take the chord $[\theta_a, \theta_y] \subset \mathcal{C}$ and consider the path $(a_i)_{i=1}^n$ defined by the cells that intersect $[\theta_a, \theta_b] \cup \mathcal{C}_{a_i} \neq \emptyset$. There is a technicality that this chord may pass through intersections of cells that have dimension $d-2$. A perturbation and dimension argument fixes the proof. For a formal proof, see [90, Lemma 28] or the similar result in [105, Lemma 23].

ii) Let $a, b \in \mathcal{A}^*(\mathcal{E})$ be Pareto optimal actions. Pick any $\theta \in \mathcal{C}_a \cap \mathcal{C}_b \cap \mathcal{C}$. If $c \in \mathcal{N}_{ab}$, we have $\mathcal{C}_a \cap \mathcal{C}_b \subset \mathcal{C}_c$, hence $\theta \in \mathcal{C}_c$ and c is optimal for θ. Therefore $c \in \mathcal{A}^*(\mathcal{E})$.

iii) Let F be the lowest dimensional face of $\mathrm{conv}(\Phi)$ containing $a \in \mathcal{A}^*(\mathcal{E})$. Assume that a is in the interior of F (otherwise it would be an extreme point and so Pareto optimal). Then let θ be a parameter such that a is optimal. $H = \{v : \langle \phi_a - v, \theta \rangle = 0\}$ is a supporting hyperplane of $\mathrm{conv}(F)$. Hence F is a subset of H. Note that $H \cap \{\phi_a : a \in \mathcal{A}^*(\theta)\}$ contains features from actions that are optimal for θ. Therefore all extreme points of F are in $\mathcal{A}^*(\mathcal{E})$ and since ϕ_a is in the convex hull of the extreme points of F the result follows. □

The next lemma shows that observability can be characterized in terms of the neighborhood relation.

Lemma 6.6. *The following conditions equivalently characterize locally observable games:*

i) For all convex $\mathcal{E} \subset \mathcal{M}$ and $a, b \in \mathcal{A}^*(\mathcal{E})$,

$$\phi_a - \phi_b \in \mathrm{span}(\mathrm{im}(M_c^\top) : c \in \mathcal{A}^+(\mathcal{E}))\,.$$

ii) For any convex $\mathcal{E} \subset \mathcal{M}$ and $a, b \in \mathcal{A}^*(\mathcal{E})$,

$$\phi_a - \phi_b \in \mathrm{span}(\mathrm{im}(M_c^\top) : c \in \mathcal{A}^*(\mathcal{E}))\,.$$

iii) For any two neighboring Pareto optimal actions $a, b \in \mathcal{P}$,

$$\phi_a - \phi_b \in \mathrm{span}(\mathrm{im}(M_c^\top) : c \in \mathcal{N}_{ab})\,.$$

Proof. For neighouring Pareto optimal actions a, b, consider the set of parameters $\mathcal{E}_{ab} \triangleq \mathrm{conv}(\mathrm{relint}(\mathcal{C}_a \cup \mathcal{C}_b)) \subset \mathcal{M}$. The proof is based on the

observation that $\mathcal{N}_{ab} = \mathcal{A}^*(\mathcal{E}_{ab}) = \mathcal{A}^+(\mathcal{E}_{ab})$. For general $\mathcal{E}$, we can write $\mathcal{A}^*(\mathcal{E}) = \mathcal{E}^*(\cup_{a,b\in\mathcal{A}^*(\mathcal{E}),a,b\text{neighbours}}\mathcal{E}_{ab})$ and use a chaining argument.

"i) $\Rightarrow$ ii)". Immediate since $\mathcal{A}^*(\mathcal{E}) \subset \mathcal{A}^+(\mathcal{E})$.

"ii) $\Rightarrow$ iii)". Let $a, b \in \mathcal{A}$ be neighboring Pareto optimal actions. Pick $\theta \in \operatorname{relint}(\mathcal{C}_a \cap \mathcal{C}_b)$. Then $\mathcal{N}_{ab} = \mathcal{A}^*(\theta)$ by Lemma 6.5 and therefore $\phi_a - \phi_b \in \operatorname{span}\{M_a : x \in \mathcal{A}^*(\mathcal{E})\}$ by i).

"iii) $\Rightarrow$ i)". Let $a, b \in \mathcal{A}^*(\mathcal{E})$. First note that by Lemma 6.5, iii), $\phi_a - \phi_b$ can be written as linear combination of Pareto optimal actions in $\mathcal{A}^*(\mathcal{E})$. Therefore we can assume that a, b are Pareto optimal. By Lemma 6.5, i), there exists a sequence $(a_i)_{i=1}^l$ of Pareto optimal actions with $a_1 = a$ and $a_l = b$, such that a_i, a_{i+1} are neighbors and $\{a_i : i = 1, \dots, l\} \subset \mathcal{A}^*(\mathcal{E})$. By assumption, $\phi_{a_i} - \phi_{a_{i-1}} \in \operatorname{span}\{M_c : c \in \mathcal{N}_{a_i a_{i-1}}\}$. Since $\phi_a - \phi_b = \sum_{i=1}^{l-1} \phi_{a_i} - \phi_{a_{i+1}}$ the claim follows by noting that $\mathcal{N}_{a_i a_{i-1}} \subset \mathcal{A}^+(\mathcal{E})$. □

6.5 REMARKS ON ASYMPTOTIC OPTIMALITY

The information gain functions in this chapter are motivated from the worst-case perspective and are conservative in practice. Assuming a gap separation condition, it is possible to prove a logarithmic gap-dependent bound on the regret, using the technique introduced in Section 3.3.2. However, as in the linear case, these bounds are far from instance-optimal in general.

The asymptotic lower bound for linear partial monitoring with finitely many actions follows from existing results. As in the linear bandit setting, the asymptotic regret is defined by the convex optimization problem similar to Eq. (5.1). The only difference that for allocations $\alpha \in \mathbb{R}^{\mathcal{A}}_{\geq 0}$, the covariance $V(\alpha) = \sum_{a\in\mathcal{A}} \alpha(a) M_a^\top M_a$ is defined using the feedback maps:

$$\mathfrak{c}^*(\theta) \triangleq \inf_{\alpha\in\mathbb{R}^{\mathcal{A}}_{\geq 0}} \sum_{a\in\mathcal{A}} \alpha(a)\langle \phi_{a^*} - \phi_a, \theta\rangle \quad \text{s.t.} \quad \min_{\nu\in\mathcal{C}^*(\theta)} \tfrac{1}{2}\|\nu - \theta\|^2_{V(\alpha)} \geq 1\,.$$

The set of alternative parameters is $\mathcal{C}^*(\theta) = \cup_{a\neq a^*(\theta)} \mathcal{C}_a$, where the cell $\mathcal{C}_a$ is defined in Eq. (7.1) and depends on $\mathcal{M}$. The asymptotic lower bound can be similarly stated as in Theorem 5.1: For any consistent policy π_n, the asymptotic regret is at least $\liminf_{n\to\infty} \mathfrak{R}(\pi_n, \theta)/\log(n) \geq \mathfrak{c}^*$.

An educated guess to extend the information gain function in Eq. (5.6) is

$$I_s^{\mathrm{A}}(a) \triangleq \frac{1}{2} \sum_{c\neq \hat{a}_s} q_s(c) \Big(\|M_a(\hat{v}_s(c) - \hat{\theta}_s)\| + \beta_s^{1/2} \|M_a\|_{V_s^{-1}} \Big)^2 , \tag{6.14}$$

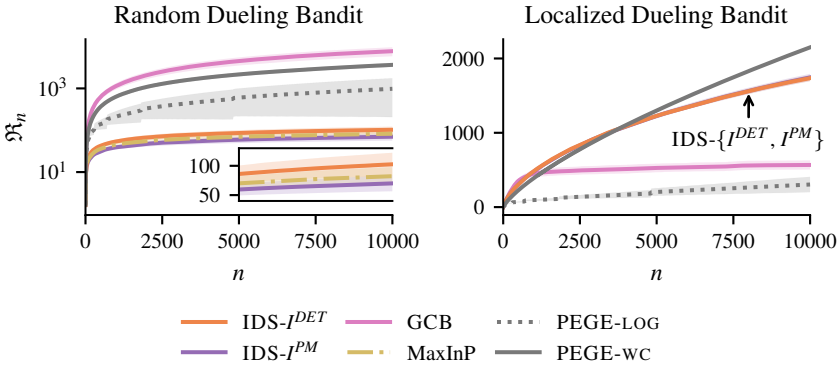


FIGURE 6.2: Simulation of randomly generated dueling bandit instances with $d = 2, k = 6$ (left) and the localized dueling bandit setting (Example 6.9) with $d = 3$ (right). Note that the y-axis on the left has logarithmic scale. In the dueling bandit setting IDS with I_t^{PM} and I_t^{DET} is competitive with MAXINP. The GCB and PEGE approach are designed for the globally observable setting, and are much more conservative. In the localized dueling bandit setting GCB and PEGE tuned for logarithmic regret out perform IDS. This is not totally unexpected, since both approaches uniformly explore with a global observer set that, in this example, is close to the optimal allocation. In general, however, these methods are not competitive with IDS unless the exploration distribution is optimized. Results are averaged over 100 runs and the confidence region shows the standard error.

where we define the alternative parameters in the same way as before,

$$\hat{\nu}_s(c) = \operatorname*{arg\,min}_{\theta:\langle \phi_a - \phi_{\hat{a}_s}, \theta\rangle \geq 0} \|\hat{\theta}_s - \nu\|_{V_s}^2,$$

and $\|M_a\|_{V_s^{-1}} \triangleq \|M_a V_s^{-1} M_a^\top\|_2$ for the operator norm. The mixing weights $q(c)$ are defined as in Eq. (5.7). The notion of the local time s introduced in Section 5.1 can be used to ensure that the optimistic term $\beta_s^{1/2}\|M_a\|_{V_s^{-1}}$ is asymptotically vanishing. The gap estimate can be defined as in Eq. (6.11). We leave it as a challenge for the future to work out potentially non-trivial details on asymptotic optimality in the general partial monitoring setting.

6.6 NUMERICAL RESULTS

We present numerical results in the utility-based dueling bandit setting, defined in Example 6.4. This game is locally observable. For a globally observable game that is not locally observable, we simulate the localized

dueling bandit setting described in Example 6.9. As baselines we compare to GCB by Lin *et al.* [109], and PEGE by Chaudhuri & Tewari [40]. For the latter, the authors propose different variants tuned for worst-case regret (PEGE-WC) and for logarithmic regret (PEGE-LOG). Both GCB and PEGE require a *global observer set*, defined as a subset of actions such that $\dim(\operatorname{im}(M_c^\top) : c \in \mathcal{A}) = \dim(\phi_a - \phi_b : a, b \in \mathcal{A})$. Finding efficient observer sets and exploration distributions is itself a difficult question. As suggested in the previous work, we randomly order the action set and incrementally add actions to the global observer set until the dimensionality condition is satisfied. Recent progress on designing more efficient exploration distribution in the (semi-)bandit setting is by Wagenmaker, Katz-Samuels & Jamieson [169], but adaptation of these results to the partial monitoring setting is still outstanding. In the dueling bandit setting, we additionally compare to the *maximum informative pair* (MAXINP) strategy by Saha & Gopalan [140]. The approach plays the actions corresponding to the most uncertain pair defined in Eq. (6.13). It is similar to IDS with the directed information gain $I_t^{\mathrm{PM}}(a; w_t)$ (Eq. (6.12)). However, the strategy fails in the globally observable case, since it only explores plausibly optimal actions. The results are shown in Fig. 6.2.

6.7 CONTRIBUTIONS AND RELATED WORK

The results present in this chapter are based on the following publication:

- Kirschner, J., Lattimore, T. & Krause, A. *Information Directed Sampling for Linear Partial Monitoring* in *Proc. International Conference on Learning Theory (COLT)* (July 2020)

Contributions entirely by Tor Lattimore are the lower bounds (Appendix C), which are included for completeness, Fig. 6.1 and regret bounds on convex action sets (not included, for details see the conference paper). For related work in the partial monitoring setting, see Section 1.2.3.

7

CONSTRAINED PARAMETER SETS

In the previous chapter, we assumed that the parameter set contains the unit ball in $\mathbb{R}^d$, which we refer to as the *directionally unconstrained* setting. Of course, the regret upper bounds are still valid for smaller parameter sets. However, additional constraints on $\mathcal{M}$ mean that the adversarial construction in the lower bound is more restricted, which in some cases implies faster rates. This also suggests that there is room to improve the upper bounds. We start with two examples.

Example 7.1. Define $\mathcal{M} = [0,1]^d$ and let the action set $\mathcal{A} = \{a, b\}$ contain two actions with features $\phi_a = -\phi_b = (1, \dots, 1) \in \mathbb{R}^d$. Given the constraints on $\mathcal{M}$, action b is never uniquely optimal. A good algorithm should therefore not attempt to estimate the gap $\langle \phi_a - \phi_b, \theta \rangle$, and always play action a, independent of the feedback maps. On the other hand, with a directionally unconstrained parameter set, action b is clearly optimal for $\theta = \phi_b$. Hence, without feedback the game is hopeless, and the learner suffers linear regret in the worst-case.

Example 7.2. The situation is even more delicate if the affine hull of $\mathcal{M}$ has dimension smaller than d. Let $\mathcal{M} = \{\lambda \cdot e_1 : \lambda \in \mathbb{R}\}$ contain just the first coordinate axis. For actions $\mathcal{A} = \{a, b\}$ with features $\phi_a = -\phi_b = e_1 \in \mathbb{R}^d$, either a or b is optimal depending on the sign of $\langle e_1, \theta \rangle$. Consider feedback maps $M_a = -M_b = (1, \dots, 1)$. Clearly, the feedback suffices to determine the optimal action, whereas in the directionally unconstrained case, the game is not globally observable according to Eq. (6.1).

Another prominent example that requires constraints is the linear formulation of finite partial monitoring, which we discuss in more detail in Section 7.3.

SETTING The notation is the same as in the previous chapter. We assume that $\mathcal{M} \subset \mathbb{R}^d$ is convex, non-empty and compact. For the algorithm, we require that linear optimization over $\mathcal{M}$ is feasible and that the Euclidean projection onto $\mathcal{M}$ can be computed efficiently. To keep the notation simple, we also assume $\|M_a\|_2 \le 1$, $\operatorname{diam}(\{\phi_a : a \in \mathcal{A}\}) \le 1$ and $\max_{\theta \in \mathcal{M}} \|\theta - \theta_0\| \le 1$ for some prior estimate $\theta_0 \in \mathcal{M}$.

7.1 LOCAL AND GLOBAL OBSERVABILITY

To classify Examples 7.1 and 7.2 and correctly, we need to refine the definitions of local and global observability to take the geometry of $\mathcal{M}$ into account. The linear span of parameter differences is denoted by

$$\mathcal{V} = \operatorname{span}(\{\omega - \nu : \omega, \nu \in \mathcal{M}\}) .$$

From a lower bound perspective, this is the set of directions available to an adversary to perturb the parameter in way that changes the optimal action, but is hard to detect for the learner. The orthogonal projection onto $\mathcal{V}$ is denoted by $P_{\mathcal{V}} : \mathbb{R}^d \to \mathbb{R}^d$, defined such that $\operatorname{im}(P_{\mathcal{V}}) = \mathcal{V}$.

Recall that the cell of $a \in \mathcal{A}$ is defined as the set of parameters for which $a \in \mathcal{A}$ is optimal,

$$\mathcal{C}_a \triangleq \{\theta \in \mathcal{M} : a^*(\theta) = a\} . \tag{7.1}$$

An action is called Pareto optimal if $\dim(\mathcal{C}_a) = \dim(\mathcal{V})$. An action is called degenerate if $\dim(\mathcal{C}_a) = \dim(\mathcal{V}) - 1$, and dominated if $\mathcal{C}_a = \emptyset$. The set of all Pareto optimal actions is $\mathcal{P}$. The reader may check that the notions are equivalent with the definitions that we introduced in Section 6.2 in the directionally unconstrained setting.

The definition of global observability is extended to the constrained case by introducing the projection onto $\mathcal{V}$ and checking the inclusions there. A game with parameter constraints is called globally observable if

$$\forall\, a, b \in \mathcal{P} , \quad P_{\mathcal{V}}(\phi_a - \phi_b) \in \operatorname{span}(\operatorname{im}(P_{\mathcal{V}} M_c^\top) : c \in \mathcal{A}) . \tag{7.2}$$

The global alignment constant is

$$\alpha^{\mathcal{V}} \triangleq \max_{\nu \in \mathcal{V}} \max_{a,b \in \mathcal{P}} \min_{c \in \mathcal{A}} \frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\|M_c \nu\|^2} . \tag{7.3}$$

For the locally observable case, we first need a more precise version of the extended maximizer set. First, recall the set of plausible maximizers for $\mathcal{E} \subset \mathcal{M}$ is $\mathcal{A}^*(\mathcal{E}) = \{a \in \mathcal{A} : \max_{\nu \in \mathcal{E}} \max_{b \in \mathcal{A}} \langle \phi_b - \phi_a, \nu \rangle = 0\}$. The set of *plausible Pareto optimal actions* for $\mathcal{E} \subset \mathcal{M}$ is

$$\mathcal{P}^*(\mathcal{E}) = \{a \in \mathcal{P} : a \in \mathcal{A}^*(\mathcal{E})\} .$$

The *extended plausible Pareto* set is defined as

$$\mathcal{P}^+(\mathcal{E}) = \{a \in \mathcal{A} : P_{\mathcal{V}} \phi_a \in \operatorname{conv}(P_{\mathcal{V}} \phi_b : b \in \mathcal{P}^*(\mathcal{E}))\} .$$

We say $\mathcal{E} \subset \mathcal{M}$ is observable, if

$$a, b \in \mathcal{P}^*(\mathcal{E}), \quad P_{\mathcal{V}}(\phi_a - \phi_b) \in \operatorname{span}(\operatorname{im}(P_{\mathcal{V}} M_c^\top) : c \in \mathcal{P}^+(\mathcal{E})). \tag{7.4}$$

A game is called locally observable if all convex subsets $\mathcal{E} \subset \mathcal{M}$ are observable. The local alignment constant for $\mathcal{E}$ is

$$\alpha^{\mathcal{V}}(\mathcal{E}) \triangleq \max_{\nu \in \mathcal{V}} \max_{a,b \in \mathcal{P}^*(\mathcal{E})} \min_{c \in \mathcal{P}^+(\mathcal{E})} \frac{\langle \phi_a - \phi_b, \nu \rangle^2}{\|M_c \nu\|^2}. \tag{7.5}$$

As a caveat to the reader, our definitions should be understood as sufficient conditions for the upper bounds. We have not yet proven lower bounds or a classification theorem for general constrained parameter sets. Even though our definitions are motivated from the construction of the lower bounds, it is possible that the conditions can be weakened. However, for the special case of finite partial monitoring, we show in Section 7.3.2 that our definitions are equivalent to the established classification.

7.2 IDS WITH PARAMETER CONSTRAINTS

The key step to improve the upper bounds is to intersect the confidence ellipsoid $\mathcal{E}_t$ with the model set $\mathcal{M}$, and define the gap estimate accordingly,

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

Without further modifications, using this gap estimate in Algorithm 6 leads to improved regret rates, including in the examples given at the beginning of this section. However, the regret bound obtain this way still scales unfavorably with the dimension d, instead of quantities such as $\dim(\mathcal{V})$ or $\dim(\operatorname{span}(\operatorname{im}(M_c^\top) : c \in \mathcal{A}))$ that can be much smaller than d. To improve the dependency on the dimension, we define the subspace of $\mathcal{V}$ observed by the feedback maps,

$$\mathcal{W} = \operatorname{span}(\operatorname{im}(P_{\mathcal{V}} M_c^\top) : c \in \mathcal{A}). \tag{7.6}$$

The dimension of $r \triangleq \dim(\mathcal{W})$ satisfies

$$r \leq \min\{\dim(\mathcal{V}), \dim(\operatorname{im}(M_c^\top) : c \in \mathcal{A})\} \leq \min\{d, mk\}. \tag{7.7}$$

Naturally, the learner can parameterize the estimate directly in $\mathcal{W}$, since any direction $\nu^\perp \in \mathcal{W}^\perp$ is either not observed or the parameters in $\mathcal{M}$

are invariant in this direction. Let $W : \mathbb{R}^r \to \mathbb{R}^d$ be the linear embedding with $\mathrm{im}(W) = \mathcal{W}$ and $W^\top W = \mathbf{1}_r$. The least-squares estimate on $\mathcal{W}$ with regularizer $\lambda > 1$ and prior estimate $\theta_0 \in \mathcal{M}$ is

$$\hat{\vartheta}_t = \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^r} \sum_{s=1}^{t-1} \|M_{a_s} W \vartheta + (\mathbf{1}_d - WW^\top)\theta_0 - y_s\|^2 + \lambda \|\vartheta - W^\top \theta_0\|^2 . \tag{7.8}$$

Let $W_t = \sum_{s=1}^{t-1} (M_{a_t} W)^\top (M_{a_t} W) + \mathbf{1}_r$. The confidence set in $\mathbb{R}^d$ is defined by

$$\mathcal{E}_{t,\delta} = \{\theta \in \mathbb{R}^d : \|W^\top \theta - \hat{\vartheta}_t\|_{W_t}^2 \le \beta_{t,\delta}^{\mathcal{W}}\} ,$$

where $\beta_{t,\delta}^{\mathcal{W}} \triangleq (\sigma^2 \sqrt{2 \log \frac{1}{\delta} + \log \frac{\det(W_t)}{\det(\lambda \mathbf{1}_r)}} + \sqrt{\lambda})^2$. Note, the confidence set is unconstrained on $\mathcal{W}^\perp$ and satisfies $\mathbb{P}[\forall t \ge 1, \theta \in \mathcal{E}_{t,\delta}] \ge 1 - \delta$. As before, we set $\beta_t^{\mathcal{W}} = \beta_{t,1/t^2}^{\mathcal{W}}$ and $\mathcal{E}_t = \mathcal{E}_{t,1/t^2}$.

To define a gap estimate that can be computed in $\mathcal{O}(|\mathcal{A}|)$ steps, let $\hat{\theta}_t^{\mathcal{M}} \triangleq \arg\min_{\theta \in \mathcal{M}} \|W^\top \theta - \hat{\vartheta}_t\|_{W_t}^2$ be a parameter in the model set, that is closest to the mean estimate $\hat{\vartheta}_t$ on $\mathcal{W}$ in W_t-norm. The empirically best action is $\hat{a}_t \triangleq \arg\max_{a \in \mathcal{A}} \langle \phi_a, \hat{\theta}_t^{\mathcal{M}} \rangle$. The gap estimate is defined using the relaxation at $\hat{a}_t$,

$$\hat{\Delta}_t^{\mathcal{W}}(a) \triangleq \delta_t + \max_{\theta \in \mathcal{E}_t \cap \mathcal{M}} \langle \phi_{\hat{a}_t} - \phi_a, \theta \rangle, \text{ where } \delta_t \triangleq \max_{\theta \in \mathcal{E}_t \cap \mathcal{M}} \max_{b \in \mathcal{A}} \langle \phi_b - \phi_{\hat{a}_t}, \theta \rangle . \tag{7.9}$$

Note that the gap estimate can be computed by solving $2k$ linear programs over $\mathcal{M}$ with positive semi-definite quadratic constraints. The gap estimate satisfies $\Delta(a) \le \hat{\Delta}_t^{\mathcal{W}}(a)$ with probability $1 - 1/t^2$ in all rounds. Lastly, the information gain is

$$I_t^{\mathcal{W}}(a) = \frac{1}{2} \log \det \left(\mathbf{1}_m + (M_a W) W_t^{-1} (M_a W)^\top\right) , \tag{7.10}$$

and the total information gain is $\gamma_n^{\mathcal{W}} = \frac{1}{2} \log \det(W_{t+1})$. The main advantage of introducing the projection is that the bounds on $\beta_t^{\mathcal{W}}$ and $\gamma_n^{\mathcal{W}}$ scale with $r \le d$. Further, the learner directly works with r-dimensional quantities instead of the over-parameterized parameter in $\mathbb{R}^d$. The improvement is important in finite partial monitoring, where the dimension d is often exponential in r.

It remains to check that the information ratio is still bounded. The worst-case regret bound for the globally observable case is summarized in the next theorem.

Theorem 7.1. *On any game with constrained parameter set and bounded global alignment $\alpha^{\mathcal{V}}$ as defined in Eq. (7.3), the regret of IDS with gap estimates $\hat{\Delta}_t^{\mathcal{W}}$ (Eq. (7.9)) and information gain $I_t^{\mathcal{W}}$ (Eq. (7.10)) satisfies,*

$$\mathfrak{R}_n \leq n^{2/3}\Big(32\alpha^{\mathcal{V}}\mathbb{E}\Big[\beta_n^{\mathcal{W}}\Big]\mathbb{E}\Big[\gamma_n^{\mathcal{W}}\Big]\Big)^{1/3} + \mathcal{O}(1)\,.$$

Proof. The proof follows along the lines of Theorem 6.1, where we need to check that the projections do not remove important information and using the definition of the alignment constant in Eq. (7.3).

The only difference is in the way we upper bound the gap estimate. Note that for $\hat{a}_t = \arg\max_{a\in\mathcal{A}}\langle \phi_a, \hat{\theta}_t^{\mathcal{M}}\rangle$ it holds that

$$\begin{aligned}\delta_t = \hat{\Delta}_t^{\mathcal{W}}(\hat{a}_t) &= \max_{\theta\in\mathcal{E}_t\cap\mathcal{M}}\max_{b\in\mathcal{P}}\langle \phi_b - \phi_{\hat{a}_t}, \theta\rangle \\ &\leq \max_{\theta\in\mathcal{E}_t\cap\mathcal{M}}\max_{b\in\mathcal{P}}\langle \phi_b - \phi_{\hat{a}_t}, \theta - \hat{\theta}_t^{\mathcal{M}}\rangle\,.\end{aligned}$$

Without loss of generality, we choose $\hat{a}_t \in \mathcal{P}$. Using the definition of the global alignment constant $\alpha^{\mathcal{V}}$, we find

$$\begin{aligned}\hat{\Delta}_t^{\mathcal{W}}(\hat{a}_t)^2 &\leq \max_{\theta\in\mathcal{E}_t\cap\mathcal{M}}\max_{b\in\mathcal{P}}\langle \phi_b - \phi_{\hat{a}_t}, \theta - \hat{\theta}_t^{\mathcal{M}}\rangle^2 \\ &\leq \alpha^{\mathcal{V}}\max_{\theta\in\mathcal{E}_t\cap\mathcal{M}}\max_{c\in\mathcal{A}}\|M_c(\theta - \hat{\theta}_t^{\mathcal{M}})\|^2 \\ &= \alpha^{\mathcal{V}}\max_{\theta\in\mathcal{E}_t\cap\mathcal{M}}\max_{c\in\mathcal{A}}\|M_cWW^\top(\theta - \hat{\theta}_t^{\mathcal{M}})\|^2 = (\star)\,.\end{aligned}$$

The equality follows from $M_cWW^\top(\theta - \hat{\theta}_t^{\mathcal{M}}) = M_c(\theta - \hat{\theta}_t^{\mathcal{M}})$, which holds by the definition of W, and the fact that $(\theta - \hat{\theta}_t^{\mathcal{M}}) \in \mathcal{V}$. Continuing,

$$\begin{aligned}(\star) &\overset{(i)}{\leq} \alpha^{\mathcal{V}}\max_{\theta\in\mathcal{E}_t\cap\mathcal{M}}\max_{c\in\mathcal{A}}\max_{\tilde{v}\in\mathbb{R}^r}\frac{\|M_cWV_t^{-1/2}\tilde{v}\|}{\|\tilde{v}\|^2}\|V_t^{1/2}W^\top(\theta - \hat{\theta}_t^{\mathcal{M}})\|^2 \\ &\overset{(ii)}{\leq} 4\alpha^{\mathcal{V}}\beta_t^{\mathcal{W}}\max_{c\in\mathcal{A}}\lambda_{\max}\big(M_cWV_t^{-1}(M_cW)^\top\big) \\ &\leq 16\alpha^{\mathcal{V}}\beta_t^{\mathcal{W}}\max_{c\in\mathcal{A}} I_t(c)\,. \qquad (7.11)\end{aligned}$$

The upper bounds (i) and (ii) follow from basic linear algebra, and (ii) further uses that $\{\theta, \hat{\theta}_t^{\mathcal{W}}\} \subset \mathcal{E}_t$. The remaining steps for bounding the information ratio and the regret are the same as in Eq. (6.10). □

The locally observable case is summarized in the next theorem.

Theorem 7.2. *On any game with constrained parameter set the regret of IDS with gap estimates $\hat{\Delta}_t^{\mathcal{W}}$ (Eq. (7.9)) and information gain $I_t^{\mathcal{W}}$ (Eq. (7.10)) satisfies,*

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

where $\bar{\alpha}_n^{\mathcal{V}} = \frac{1}{n}\sum_{t=1}^n \alpha^{\mathcal{V}}(\mathcal{E}_t \cap \mathcal{M})$ is the average realized local alignment (Eq. (7.5)).

Proof. Again, the proof is very similar to the proof of Theorem 6.2. Let $c \in \mathcal{P}^+(\mathcal{E}_t \cap \mathcal{M})$ be any action in the extended plausible Pareto set. Repeating the steps leading to Eq. (7.11) for the local version, we find

$$\hat{\Delta}_t^{\mathcal{W}}(c) \leq 2 \max_{\nu,\omega \in \mathcal{E}_t \cap \mathcal{M}} \max_{a,b \in \mathcal{P}^*(\mathcal{E}_t \cap \mathcal{M})} \langle \phi_a - \phi_b, \nu - \omega \rangle\,.$$

And consequently,

$$\begin{aligned}
\hat{\Delta}_t(c)^2 &\leq 4 \max_{\nu,\omega \in \mathcal{E}_t \cap \mathcal{M}} \max_{c' \in \mathcal{P}^+(\mathcal{E}_t \cap \mathcal{M})} \|M_{c'}(\nu - \omega)\|^2 \\
&\leq 16\beta_t^{\mathcal{W}}\alpha^{\mathcal{V}}(\mathcal{E}_t \cap \mathcal{M}) \max_{c' \in \mathcal{P}^+(\mathcal{E}_t \cap \mathcal{M})} \lambda_{\max}\big((M_{c'}W)V_t(M_{c'}W)^\top\big) \\
&\leq 64\beta_t^{\mathcal{W}}\alpha^{\mathcal{V}}(\mathcal{E}_t \cap \mathcal{M}) \max_{c' \in \mathcal{P}^+(\mathcal{E}_t \cap \mathcal{M})} I_t(c')\,.
\end{aligned}$$

The bound on the information ratio is now immediate. Note, the increase by a factor 2 compared to Theorem 6.2 is from using the faster version of the gap estimate. □

7.3 FINITE STOCHASTIC PARTIAL MONITORING

In this section, we complete the picture by deriving bounds for IDS in the classical *finite partial monitoring* setting, using the results established for constrained parameter sets. As in the unconstrained case, the regret bounds for IDS match the established classification of finite partial monitoring. The result follows by showing that the classical definitions match the definitions we introduced for the constrained setting in Section 7.1. We further derive specialized bounds on the alignment constants.

Historically, 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. A finite partial monitoring game consists of *actions* $\mathcal{A} = [k]$, a finite 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 and feedback are defined by,

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

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

The learner has access to both R and Φ. In each round $t = 1, \dots, n$ 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.

Let us introduce 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 $S_a \in \{0, 1\}^{s \times d}$ 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 $\vartheta \in \mathscr{P}(\mathcal{X})$ 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 $y_t = e_{\sigma_t} = S_{a_t} e_{x_t} \in \mathbb{R}^m$. The best action is

$$a^* = \operatorname*{arg\,max}_{a \in \mathcal{A}} \left\{ \mathbb{E}_{x \sim \vartheta}[R(a, x)] = \langle \phi_a, \vartheta \rangle \right\},$$

which is chosen arbitrarily if it is not unique.

7.3.1 *Examples*

Let us first give a few examples typically encountered in the finite setting.

Example 7.3 (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. However, the subspace $\mathcal{W}$ in Eq. (7.6) has dimension $r \leq 2k$.

Example 7.4 (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:

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

Example 7.5 (Linear Counter Example)**.** 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, the rewards are such that the first action is always optimal, so in finite partial monitoring a good algorithm has zero regret. On the other hand, an algorithm for the linear setting has to account for the case when $\vartheta = (-\sqrt{2}, -\sqrt{2})$, and the second action is optimal. Consequently, any learner suffers linear regret on at least one of the two cases. The different assumptions on ϑ mean that this game is hopeless in the directionally unconstrained linear setting, while in finite partial monitoring it is trivial.

7.3.2 *Classification of Finite Partial Monitoring*

As in the linear case, the smallest possible regret that a learner can hope to achieve depends on the structure of the reward and signal functions. For example, in bandit games the optimal worst-case regret is $\Theta(n^{1/2})$, but for dynamic pricing it is $\Theta(n^{2/3})$ [16]. It is also easy to construct examples where the learner immediately knows the best action and suffers no regret;

as well as instances where the feedback is insufficient to determine the optimal action, and consequently linear worst-case regret is unavoidable.

Significant effort has been put into classifying all possible cases, and it is now understood that any finite game belongs to one of four categories: *trivial*, *easy*, *hard* and *hopeless*. The definitions of the game categories is based on the geometric structure of R and Φ. Our presentation follows the standard terminology [e.g., 18]. As before, the cell $\mathcal{C}_a$ of action $a \in [k]$ is defined as the set of outcome distributions for which a is optimal,

$$\mathcal{C}_a = \{\lambda \in \mathscr{P}(\mathcal{X}) : \forall b \in [k], (e_a - e_b)^\top R\lambda \geq 0\}\,.$$

An action with $\dim(\mathcal{C}_a) = d - 1$ is called *Pareto optimal* and *degenerate* otherwise if $\mathcal{C}_a$ is not empty. Degenerate actions can be optimal, but not uniquely so. The set of Pareto optimal actions is $\mathcal{P} \subset \mathcal{A}$. Actions a and b are called a *duplicate* if $\phi_a = \phi_b$. Note that duplicate actions may differ on the signal function. A game is called *non-degenerate* if it has no degenerate actions and no duplicates. Two pareto optimal actions a, b are called *neighbors* if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d - 2$. The *neighborhood* $\mathcal{N}_{ab}$ of two neighbouring actions a and b is the set of actions

$$\mathcal{N}_{ab} = \{c \in \mathcal{A} : \dim(\mathcal{C}_a \cap \mathcal{C}_b \cap \mathcal{C}_c) = d - 2\}\,.$$

Some illustrated examples are given by Lattimore & Szepesvari [103, §37]. Recall that the features in finite partial monitoring are defined as $\phi_a \triangleq R^\top e_a$. A finite partial monitoring game is called *globally observable* if

$$\forall a, b \in \mathcal{P}: \quad \phi_a - \phi_b \ \text{span}(\text{im}(S_c^\top) : c \in \mathcal{A})\,. \tag{7.12}$$

A finite partial monitoring game is called *locally observable* if

$$\forall \text{ neighboring } a, b \in \mathcal{P}, \quad \phi_a - \phi_b \ \text{span}(\text{im}(S_c^\top) : c \in \mathcal{N}_{ab})\,. \tag{7.13}$$

Using the notion of local and global observability, the game categories are defined as follows:

TRIVIAL: Games with only one Pareto optimal action or duplicates.

EASY: Locally observable games that are not trivial.

HARD: Globally observable games that are not easy or trivial.

HOPELESS: Games that are not globally observable or trivial.

The classification theorem establishes the regret rate for each category.

Theorem 7.3 (Classification of finite partial monitoring, [16, 105])**.** *For any finite partial monitoring game, the minimax regret $\mathfrak{R}_n^*$ satisfies*

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

The classification theorem for finite stochastic partial monitoring was proved separately by a number of authors, many of whom ultimately collaborated in writing a comprehensive journal paper [16], which left only logarithmic factors on the table. These were later resolved by Lattimore & Szepesvári [105].

As we have pointed out already, the classification result for the linear case in Section 6.4 and the finite case do not imply each other. Naturally, the upper bounds in the unconstrained linear case apply in the finite setting but are not necessarily tight, as evident by Example 7.5. Another difference is that the linear setting permits infinite observation (and action) spaces, which are not usually covered by existing results in the finite setting.

7.3.3 *Simplex Constraints*

In finite partial monitoring, the parameter that defines the reward $\langle \phi_a, \vartheta \rangle$ is a distribution $\vartheta \in \mathscr{P}(\mathcal{X})$ over outcomes. Therefore, the model set is the simplex $\mathcal{M} = \{\theta \in [0,1]^d, \|\theta\|_1 = 1\}$. The difference space $\mathcal{V} = \operatorname{span}(\{\nu - \omega : \nu, \omega \in \mathcal{M}\}$ is such that $\mathcal{V}^\perp = \operatorname{span}(1_d)$, where $1_d \in \mathbb{R}^d$ is the vector of all ones. Consequently, the projection map $P_{\mathcal{V}}$ satisfies $\ker(P_{\mathcal{V}}) = \operatorname{span}(1_d)$.

In the setting with general parameter constraints, the estimation conditions in Eqs. (7.2) and (7.4) are based on projected features and feedback maps. Coincidentally, the direction 1_d removed by the projection $P_{\mathcal{V}}$ is the only direction that is always observed with the feedback maps S_a. In other words, the zero-one structure of S_a implies that $1_d \in \operatorname{span}(\operatorname{im}(S_a^\top) : c \in \mathcal{A})$. It also explains why the projection does not appear in the classical definitions of local and global observability for finite games in Eqs. (7.12) and (7.13). The projection is however essential to correctly classify games with general parameter constraints such as Examples 7.1 and 7.2.

The next lemma clarifies the relation between global observability in finite games defined in Eq. (7.12), and the definition of the global alignment

constant $\alpha^{\mathcal{V}}$ in Eq. (7.3). Let $S = (S_c^\top)_{c\in\mathcal{A}}^\top \in \mathbb{R}^{mk\times d}$ be the matrix formed by vertically stacking the observation matrices. An *estimation vector* for a pair of pareto optimal actions a and b is a vector $w_{ab} \in \mathbb{R}^{mk}$ such that $S^\top w_{ab} = \phi_a - \phi_b$. By definition, for any $x \in \mathcal{X}$,

$$R(a,x) - R(b,x) = \langle \phi_a - \phi_b, e_x\rangle = \sum_{c\in\mathcal{A}} \langle w_{ab}^c, S_c e_x\rangle\,,$$

where $w_{ab}^c \in \mathbb{R}^m$ are the coordinates of w_{ab} associated with action c. Similarly to Lemma 6.1, we can bound the alignment constant by bounding the norm of the estimation vector.

Lemma 7.1. *A finite partial monitoring game is globally observable if and only if $\alpha^{\mathcal{V}} < \infty$. In this case further*

$$\alpha^{\mathcal{V}} \le \max_{a,b\in\mathcal{P}} \min_{S^\top w_{ab} = \phi_a - \phi_b} \left(\sum_{c\in\mathcal{A}} \|w_{ab}^c\|\right)^2 \le mdk^{d+2}\,,$$

where the minimum is over estimation vectors $w_{ab} \in \mathbb{R}^{mk}$ and $w_{ab}^c \in \mathbb{R}^m$ denotes the subset of coordinates corresponding to action $c \in \mathcal{A}$.

Proof of Lemma 7.1. The proof of the first inequality is almost identical to the unconstrained linear case Lemma 6.1. Note that by the definition of global observability, there exists a vector $w \in \mathbb{R}^{mk}$ such that $\phi_a - \phi_b = w_{ab}^\top S$. The inequality follows by optimizing over the estimation vectors,

$$\begin{aligned}
\alpha^{\mathcal{V}} &= \max_{a,b\in\mathcal{P}} \max_{\nu:\langle\nu,1\rangle=0} \min_{c\in\mathcal{A}} \frac{\langle \phi_a - \phi_b, \nu\rangle^2}{\|S_c\nu\|^2} \\
&= \max_{a,b\in\mathcal{P}} \max_{\nu:\langle\nu,1\rangle=0} \min_{c\in\mathcal{A}} \frac{\langle w_{ab}^\top S, \nu\rangle^2}{\|S_c\nu\|^2} \\
&= \max_{a,b\in\mathcal{P}} \max_{\nu:\langle\nu,1\rangle=0} \frac{\left(\sum_{c\in\mathcal{A}} \langle w_{ab}^c, S_c\nu\rangle\right)^2}{\max_{c\in\mathcal{A}} \|S_c\nu\|^2} \\
&\le \max_{a,b\in\mathcal{P}} \left(\sum_{c\in\mathcal{A}} \|w_{ab}^c\|\right)^2 .
\end{aligned}$$

To obtain the second inequality, note that by [103, Proposition 37.18], w_{ab}^c can be chosen so that $\|w_{ab}^c\|_\infty \le d^{1/2}k^{d/2}$. The result follows from the Cauchy-Schwarz inequality to bound $\|w_{ab}^c\| \le \sqrt{m}\|w_{ab}^c\|_\infty$.

The equivalence follows if we show that $a, b \in \mathcal{P}$ with $\phi_a - \phi_b \notin \operatorname{span}(\operatorname{im}(S_c^\top) : c \in \mathcal{A})$ implies $\alpha^{\mathcal{V}} = \infty$. Note that because a, b are Pareto

optimal, there exists distributions $\nu_a, \nu_b \in \mathscr{P}(\mathcal{X})$ such that $|\langle \phi_a - \phi_b, \nu_a - \nu_b \rangle| > 0$. We can choose ν as the projection of $\nu_a - \nu_b$ onto $\phi_a - \phi_b$, for which we get $S_c \nu = 0$ for all $c \in \mathcal{A}$ by assumption, hence $\alpha^{\mathcal{V}} = \infty$. □

In the next lemma, we bound $\alpha^{\mathcal{V}}(\mathcal{E})$ for locally observable games and make a connection to the definition of local observability. For a pair of Pareto optimal actions $a, b \in \mathcal{P}$, let $\mathcal{E}_{ab}$ be the convex hull of the relative interior of $\mathcal{C}_a \cup \mathcal{C}_b$.

Lemma 7.2. *A finite partial monitoring game is locally observable if and only if for all pairs of neighbors a and b, $\alpha^{\mathcal{V}}(\mathcal{E}_{ab}) < \infty$. In this case,*

$$\alpha^{\mathcal{V}}(\mathcal{E}_{ab}) \leq \min_{w_{ab}^\top S = \phi_a - \phi_b} \left(\sum_{c \in \mathcal{E}_{ab}} \|w_{ab}^c\| \right)^2 \leq md|\mathcal{N}_{ab}|^{d+2},$$

where $w_{ab} \in \mathbb{R}^{mk}$. Further, for convex $\mathcal{E} \subset \mathscr{P}(\mathcal{X})$ with non-empty interior and assuming that $\mathcal{P}^*(\mathcal{E})$ contains more than one non-duplicate Pareto optimal action, $\mathcal{P}^*(\mathcal{E}) = \cup_{a,b \in \mathcal{P}^*(\mathcal{E}), \text{neighbors}} \mathcal{E}_{ab}$ and

$$\alpha^{\mathcal{V}}(\mathcal{E}) \leq \left(\sum_{\substack{a,b \in \mathcal{P}^*(\mathcal{E}) \\ a,b \text{ neighbors}}} \alpha^{\mathcal{V}}(\mathcal{E}_{ab})^{1/2} \right)^2 \leq 4mdk^{d+2}.$$

For non-degenerate games, $\alpha^{\mathcal{V}}(\mathcal{E}) \leq 4k^2m^3$.

Proof. The proof of the equivalence and the bound on $\alpha^{\mathcal{V}}(\mathcal{E}_{ab})$ follows along the lines of Lemma 7.2. The decomposition of $\mathcal{P}^*(\mathcal{E})$ as a union over the neighborhoods of pareto optimal action $a, b \in \mathcal{P}^*(\mathcal{E})$ follows from Lemma 6.5. The bound on $\alpha^{\mathcal{V}}(\mathcal{E})$ is obtained as direct consequence of the decomposition. The bound on $\alpha^{\mathcal{V}}(\mathcal{E})$ for non-degenerate games follows from [103, Proposition 37.18], which shows that for these games w_{ab}^c can be chosen to be zero for $c \notin \{a, b\}$ and $\|w_{ab}^c\|_\infty \leq m$ otherwise. □

7.3.4 *IDS for Finite Partial Monitoring*

We are now in the position to apply IDS for the constrained setting to finite partial monitoring. The definitions for the gap estimate and the information gain are specialized from Eqs. (7.9) and (7.10), with the simplex as model set, $\mathcal{M} = \{\theta \in [0,1]^d, \|\theta\|_1 = 1\}$. The space generated by difference vectors in $\mathcal{M}$ is $\mathcal{V} = \{v \in \mathbb{R}^d, \langle v, 1_d \rangle = 0\}$, where $1_d \in \mathbb{R}^d$ is the vector of all

Algorithm 7: IDS for Finite Partial Monitoring

Input: Regularizer λ, prior estimate $\theta_0 \in \mathscr{P}(\mathcal{X})$, norm bound $\|\theta - \theta_0\| \le B$, embedding $W : \mathbb{R}^r \to \mathbb{R}^d$

for $t = 1, 2, 3, \dots$ **do**

`// Embedded regularized least-squares:`

$\hat{\vartheta}_t^{\mathcal{W}} \leftarrow \arg\min_{\vartheta \in \mathbb{R}^r} \sum_{s=1}^{t-1} \|S_{a_s} W\vartheta - y_s\|^2 + \lambda \|\vartheta - W^\top \theta_0\|^2$

$\hat{\theta}_t^{\mathcal{M}} \leftarrow \arg\min_{\theta \in \mathscr{P}(\mathcal{X})} \|W\hat{\theta}_t^{\mathcal{W}} - \hat{\vartheta}_t^{\mathcal{W}}\|_{W_t}^2$ `// projected estimate`

$\beta_t^{1/2} \leftarrow 2\sqrt{2\log(t) + \log\left(\frac{\det(W_t)}{\det(\lambda \mathbf{1}_r)}\right)} + \sqrt{\lambda} B$

$\mathcal{E}_t = \{\theta \in \mathbb{R}^d : \|W^\top \theta - \hat{\vartheta}_t^{\mathcal{W}}\|_{W_t}^2 \le \beta_t\}$ `// confidence set`

$\hat{a}_t \leftarrow \arg\max_{a \in \mathcal{A}} \langle \phi_a, \hat{\theta}_t^{\mathcal{M}} \rangle$ `// best action`

$\delta_t \triangleq \max_{\theta \in \mathcal{E}_t \cap \mathscr{P}(\mathcal{X})} \max_{b \in \mathcal{A}} \langle \phi_b - \phi_{\hat{a}_t}, \theta \rangle$ `// estimation error`

$\hat{\Delta}_t^{\mathcal{W}}(a) \leftarrow \delta_t + \max_{\theta \in \mathcal{E}_t \cap \mathscr{P}(\mathcal{X})} \langle \phi_{\hat{a}_t} - \phi_a, \theta \rangle$ `// gap estimates`

$I_t^{\mathcal{W}}(a) \leftarrow \frac{1}{2} \log\det\left(\mathbf{1}_m + (M_a W) W_t^{-1} (M_a W)^\top\right)$ `// info. gain`

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

Choose $a_t \sim \mu_t$, observe $e_{\sigma_t} \sim S_{a_t} \vartheta$

ones. Also recall the definition of $\mathcal{W}$ in Eq. (7.6) and $r = \dim(\mathcal{W})$. Refer to Algorithm 7 for the complete algorithm.

Before the regret bounds, we address two remaining technical differences in finite partial monitoring. The first is that the observation likelihood is defined by the outcomes. This means, the noise is added on the parameter, whereas in the linear formulation we assumed that the noise is added to the evaluation. Fortunately, the proof of the confidence set in Lemma 3.1 is flexible enough to accommodate this change. The new result is summarized in the next lemma.

Lemma 7.3. *Assume the outcome distribution $\vartheta \in \mathscr{P}(\mathcal{X})$ satisfies $\|\vartheta - \vartheta_0\| \le B$. Let $(x_t)_{t=1}^n$ be a i.i.d. sequence sampled from ϑ. For any sequence of actions $a_1, \dots a_n$ with corresponding observations $y_t = S_{a_t} e_{x_t}$, the least squares estimator $\hat{\vartheta}_t$ defined in (7.8) satisfies for all $\delta \in [0, 1]$,*

$$\mathbb{P}[\forall t \ge 1,\ \|\hat{\vartheta} - W^\top \vartheta\|_{W_t}^2 \le \beta_{t,1/\delta}] \ge 1 - \delta\,,$$

where $\beta_{t,1/\delta}^{1/2} = 2\sqrt{2\log\frac{1}{\delta} + \log\left(\frac{\det(W_t)}{\det(\lambda \mathbf{1}_r)}\right)} + \sqrt{\lambda} B.$

Although not quite an immediate result, the claim follows along the lines of [3, Theorem 2]. The proof is deferred to the end of the section.

The second difference is that we can use the special structure of the feedback maps S_a to derive more specific bounds on the information gain and the confidence coefficient.

Lemma 7.4. *The total information gain* $\gamma_n = \sum_{t=1}^n I_t(a_t)$ *for Eq.* (7.10) *with* $M_a = S_a$ *satisfies*

$$\gamma_n \le r \log\left(1 + \sum_{t=1}^{n} \operatorname{rank}(S_{a_t})\right) \le r \log\Big(1 + \frac{rn}{\lambda}\Big)\,.$$

Proof. Telescoping shows that $\gamma_n = \log\det W_{n+1} - \log\det(\lambda \mathbf{1}_r)$. Denote by $w_1, \dots w_r$ the eigenvalues of W_{n+1}. By the inequality of arithmetic and geometric means,

$$\det(W_{n+1}) = \prod_{i=1}^{r} w_i \le \left(\frac{1}{r}\operatorname{tr}(W_{n+1})\right)^r = \left(\lambda + \frac{1}{r}\sum_{t=1}^{n-1} \operatorname{tr}(W^\top S_{a_t}^\top S_{a_t} W)\right)^r$$

The result follows by noting that

$$\begin{aligned}\operatorname{tr}(W^\top S_{a_t}^\top S_{a_t} W) = \|S_{a_t} W\|_F^2 &\le \operatorname{rank}(S_{a_t})\|S_{a_t} W\|_2^2 \\ &\le \operatorname{rank}(S_{a_t})\|S_{a_t}\|_1 \|W\|_\infty \le \operatorname{rank}(S_{a_t}) r \le r^2\,.\end{aligned}$$

For the last step, we used that $\|S_{a_t}\|_1 \le 1$, $\|W\|_\infty \le \sqrt{r}\|W\|_2$, and lastly $W^\top W = \mathbf{1}_r$, hence $\|W\|_2 \le \sqrt{r}$. □

With the last two lemmas, the regret bounds for finite partial monitoring follow from Theorems 7.1 and 7.2.

Corollary 7.1. *On* globally observable *finite partial monitoring games, IDS (Algorithm 7) satisfies*

$$\mathfrak{R}_n \le \mathcal{O}\Big((mdk^{d+2})^{1/3}(rn\log(nr))^{2/3}\Big)\,.$$

On locally observable *finite partial monitoring games, the regret is at most*

$$\mathfrak{R}_n \le \mathcal{O}\Big(\sqrt{mdk^{d+2}n}\, r \log(nr)\Big)\,.$$

On non-degenerate, *locally observable finite partial monitoring games, the regret is at most*

$$\mathfrak{R}_n \le \mathcal{O}\Big(ks^{3/2}\sqrt{n}\, r \log(nr)\Big)\,.$$

On trivial games, $\min_{a\in\mathcal{A}} \hat{\Delta}_t(a) = 0$ *in all rounds, and IDS always plays the Pareto optimal action.*

The upper bounds match the classification result in Theorem 7.3. Notably, the algorithm does not make explicit use of the cell structure and adapts to the different game categories automatically. Lastly, we provide the proof of the confidence bound with parameter noise.

Proof of Lemma 8.3. The proof is almost identical to [3, Theorem 2], therefore we sketch only the key steps. For outcome x_t, define the parameter-noise vector $\xi_t = e_{x_t} - \vartheta$, where $\vartheta \in \mathscr{P}(\mathcal{X}) \subset \mathbb{R}^d$ is the outcome distribution. First, using the closed form solution of the least-squares estimator, we find

$$\|\hat{\vartheta}_t - W^\top \vartheta\|_{W_t} \le \|\textstyle\sum_{s=1}^{t-1} W^\top S_{a_s}^\top S_{a_s}\xi_s\|_{W_t^{-1}} + \sqrt{\eta}\|\vartheta - \vartheta_0\|\,.$$

Consider any $a \in \mathcal{A}$ and note that for a unit vector $u \in \mathbb{R}^m$,

$$|u^\top S_a \xi_t| \le \|u\|_\infty \|S_a\xi_t\|_1 \le \|S_a e_{x_t} - S_a\vartheta\|_1 \le 2\,.$$

Hence $S_a\xi_t$ is a 4-subgaussian random vector in $\mathbb{R}^m$. For any $\nu \in \mathbb{R}^r$ define

$$Q_t(\nu) = \exp\left(\frac{1}{2}\langle \nu, \sum_{s=1}^{t-1} W^\top S_{a_s}^\top S_{a_s}\xi_s\rangle - \frac{1}{2}\|\nu\|_{W_t}^2\right).$$

Note that Q_t is a super-martingale:

$$\mathbb{E}_t[Q_{t+1}] = Q_t \mathbb{E}_t\left[\exp\left(\frac{1}{2}\langle \nu, W^\top S_{a_t}^\top S_{a_t}\xi_t\rangle - \frac{1}{2}\|W\nu\|_{S_{a_t}^\top S_{a_t}}^2\right)\right] \le Q_t\,.$$

The last step follows from,

$$\begin{aligned}
\mathbb{E}_t\left[\exp\left(\frac{1}{2}\langle \nu, W^\top S_{a_t}^\top S_{a_t}\xi_t\rangle\right)\right] &= \mathbb{E}_t\left[\exp\left(\frac{1}{2}\langle S_{a_t}W\nu, S_{a_t}\xi_t\rangle\right)\right] \\
&= \mathbb{E}_t\left[\exp\left(\frac{1}{2}\|S_{a_t}W\nu\|\langle \frac{S_{a_t}W\nu}{\|S_{a_t}W\nu\|}, S_{a_t}\xi_t\rangle\right)\right] \\
&\le \exp\left(\frac{1}{2}\|S_{a_t}W\nu\|^2\right) = \exp\left(\|W\nu\|_{S_{a_t}^\top S_{a_t}}^2\right).
\end{aligned}$$

The inequality follows because $S_{a_t}\xi_t$ is a 4-subgaussian vector. Finally, let $h = \mathcal{N}(0, (\eta \mathbf{1}_r)^{-1})$. Then the following Gaussian integral can be computed in closed-form,

$$\bar{Q}_t = \int_{\mathbb{R}^d} Q_t(\nu) dh = \left(\frac{\det(\eta \mathbf{1}_r)}{\det(W_t)}\right)^{1/2} \exp\left(\frac{1}{2}\|\sum_{s=1}^{t-1} W^\top S_{a_s}^\top S_{a_s}\xi_s\|_{W_t^{-1}}^2\right).$$

The result follows using a maximal inequality on $\sup_{t\ge 1} \log \bar{Q}_t$. □

7.4 CONTRIBUTIONS AND RELATED WORK

The results on finite partial monitoring and constrained parameter sets were developed together with Tor Lattimore and Andreas Krause. For related work on partial monitoring, see Section 1.2.3.

8

EXTENSIONS AND APPLICATIONS

In the last chapter before the conclusion, we present two extensions of the IDS framework, demonstrating the generality of the results. First, in Section 8.1, we introduce a novel contextual formulation of linear partial monitoring that generalizes the well-known linear contextual bandit setting. When the context is sampled from a fixed distribution, we extend the definition of the information ratio to include the expectation over the randomness of the context. By optimizing the contextual information ratio, we obtain IDS sampling distributions that include the context distribution for exploration.

Second, in Section 8.2, we derive a kernelized version of IDS. This significantly increases the range of possible applications, for example, by making use of practically relevant smoothness priors. We also supply a variety of example applications that illustrate the way the framework can be used in practice. The examples include kernelized dueling bandits as well as a novel bias-robust algorithm for regret minimization.

8.1 CONTEXTUAL PARTIAL MONITORING

In the contextual bandit problem, the learner receives a context in each round *before* choosing an action. In applications, the context represents additional information available to the learner, such as, for example, temperature measurements, daytime, or the profile of a user visiting a website. These factors can change from round to round and are not controlled by the learner, but the reward depends on both the context and the chosen action.

Formally, let $\mathcal{Z}$ be a context set and $z_t \in \mathcal{Z}$ the context presented at time t. To avoid measure theoretic complications and ensure computability, we assume that the action and context sets are finite. The reward function $f : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$ is extended with the contextual argument. For each context $z \in \mathcal{Z}$, a subset of actions $\mathcal{A}^z \subset \mathcal{A}$ is available for playing. The learner competes with the best action $a^*(z) = \arg\max_{a \in \mathcal{A}^z} f(a, z)$ chosen

in hindsight for each the context $z \in \mathcal{Z}$ among the set $\mathcal{A}^z$. For a sequence of contexts $(z_t)_{t=1}^n$, we define the *contextual regret*,

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

A *contextual linear partial montioring game* is defined by three functions:

$$\begin{aligned} A &: \mathcal{Z} \to 2^{\mathcal{A}},\ z \mapsto \mathcal{A}^z && \text{(available actions)} \\ \phi &: \mathcal{A} \times \mathcal{Z} \to \mathbb{R}^d,\ (a, z) \mapsto \phi_a^z && \text{(reward features)} \\ M &: \mathcal{A} \times \mathcal{Z} \to \mathbb{R}^{m \times d},\ (a, z) \mapsto M_a^z && \text{(observation maps)} \end{aligned}$$

We further define the joint action space $\mathcal{A}^{\mathcal{Z}} \triangleq \times_{z \in \mathcal{Z}} \mathcal{A}^z$. The reward function is parameterized with a single $\theta \in \mathbb{R}^d$ shared among all contexts such that $f_\theta(a, z) = \langle \phi_a^z, \theta \rangle$. The feedback in round t for action a_t and context z_t is $y_t = M_a^z \theta + \epsilon_t$, where $\epsilon_t \in \mathbb{R}^d$ is conditionally independent ρ-sub-Gaussian noise. The linear contextual partial monitoring strictly generalizes the linear contextual bandit setting.

In the next two sections, we develop IDS policies for the contextual partial monitoring setting. The first variant directly extends the algorithm proposed in Section 6.3 and optimizes the sampling distribution of each context. For the second variant, we assume that the context is sampled from a fixed distribution. We will see that this allows for much weaker conditions where no-regret is possible.

For simplicity, we focus on the case with directionally unconstrained set $\mathcal{M} = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$ and assume boundedness of the observation and reward features $\|M_a^z\|_2 \leq 1$ and $\operatorname{diam}(\phi(a, z) : a \in \mathcal{A}, z \in \mathcal{Z}) \leq 1$. Note that estimation can be done in the same way as before, using the least-squares estimator $\hat{\theta}_t$ and confidence set $\mathcal{E}_t$ defined in Section 6.3.

8.1.1 *Conditional IDS*

In the first step, we directly extend the definitions of the gap estimate and the information gain to the contextual argument. For the gap estimate, we use the computationally faster version from Section 6.3.2. Let $\hat{a}_t(z) = \arg\max_{a \in \mathcal{A}} \langle \phi_a^z, \hat{\theta}_t \rangle$ be the empirically best action for context $z \in \mathcal{Z}$. The gap estimate is

$$\hat{\Delta}_t(a, z) = \min\left\{ \max_{b \in \mathcal{A}^z} \langle \phi_b^z - \phi_a^z, \hat{\theta}_t \rangle + \beta_t^{1/2} \|\phi_b^z - \phi_{\hat{a}_t(z)}^z\|_{V_t^{-1}}, 1 \right\}. \tag{8.1}$$

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

$$I_t^{\mathrm{DET}}(a, z) = \frac{1}{2} \log \det(\mathbf{1} + M_a^{z\top} V_t^{-1} M_a^z) \,. \tag{8.2}$$

Conditional IDS is the policy that optimizes the information ratio for the observed context,

$$\mu_t^{\mathrm{IDS}}(z_t) = \operatorname*{arg\,min}_{\mu \in \mathscr{P}(\mathcal{A}^{z_t})} \left\{ \Psi(\mu, z_t) = \frac{\hat{\Delta}_t(\mu, z_t)^2}{I_t(\mu, z_t)} \right\}. \tag{8.3}$$

The computational complexity required to find the minimizer of the information ratio is the same as in the non-contextual case, since the IDS distribution is computed only for the given context. For the analysis, we extend the notion of the alignment constant for any convex $\mathcal{E} \subset \mathcal{M}$ with the contextual argument,

$$\alpha^z(\mathcal{E}) = \max_{\nu \in \mathbb{R}^d} \max_{a,b \in \mathcal{A}_z^*(\mathcal{E})} \min_{c \in \mathcal{A}_z^+(\mathcal{E})} \frac{\langle \phi_a^z - \phi_b^z, \nu \rangle^2}{\|M_c^z \nu\|^2} \,, \tag{8.4}$$

where we define contextual extensions of the (extended) plausible maximizer as follows:

$$\mathcal{A}_*^z(\mathcal{E}) = \cup_{\theta \in \mathcal{E}} \{ a \in \mathcal{A}^z : \langle \phi_a^z, \theta \rangle = \max_{b \in \mathcal{A}^z} \langle \phi_b^z, \theta \rangle \} \,,$$
$$\mathcal{A}_+^z(\mathcal{E}) = \{ a \in \mathcal{A}^z : \phi_a^z \in \mathrm{conv}(\phi_b^z : b \in \mathcal{A}_*^z(\mathcal{E})) \} \,.$$

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

Corollary 8.1. *For any $\mathcal{F}_t$-predictable sequence $(z_t)_{t=1}^n$ in $\mathcal{Z}$, the regret of conditional IDS satisfies,*

$$\mathfrak{R}_n \leq n^{2/3} (32 \mathbb{E}[\bar{\alpha}_n \beta_n] \mathbb{E}[\gamma_n])^{1/3} + \mathcal{O}(1) \,,$$

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

Proof. The claim follows along the lines of Theorem 6.1, noting that the generalized information ratio (Eq. (2.7)) is bounded as follows:

$$\Psi_{3,t}(\mu_t^{\mathrm{IDS}}) = \frac{\hat{\Delta}_t(\mu_t^{\mathrm{IDS}})^3}{I_t^{\mathrm{DET}}(\mu_t^{\mathrm{IDS}})} \leq 32 \alpha^{z_t}(\mathcal{M}) \beta_t \,.$$

For the last step of the proof, we make use of Theorem 2.2 using a time-dependent bound on $\Psi_{3,t}$. □

Corollary 8.2. *For any $\mathcal{F}_t$-predictable sequence $(z_t)_{t=1}^n$ in $\mathcal{Z}$, the regret of conditional IDS satisfies*

$$\mathfrak{R}_n \leq 4\sqrt{\mathbb{E}[\bar{\alpha}_n \beta_n]\mathbb{E}[\gamma_n] n} + \mathcal{O}(1)\,,$$

where $\bar{\alpha}_n = \frac{1}{n}\sum_{t=1}^n \alpha^{z_t}(\mathcal{E}_t)$ 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 6.2. □

In particular, in the bandit setting $\alpha^{z_t}(\mathcal{E}_t) \leq 4$ holds independently of the context (c. f. Example 6.1). Therefore, Corollary 8.2 recovers the same bounds as UCB in the linear contextual bandit setting. Moreover, we immediately get a regret bound for the contextual extension of the dueling bandit setting in Example 6.4.

8.1.2 *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 $\chi \in \mathscr{P}(\mathcal{Z})$. If the distribution is unknown, it is natural to replace χ with an online estimate of the context distribution, c. f. [160]. It is instructive to think about some examples:

Example 8.1 (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 $a \in \mathcal{A}^z$; we call these non-informative contexts. For such non-informative contexts, the only sensible choice is the greedy action. 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 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 8.2 (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 [20, 72]. One can think of the context as part of the action space, where the sampling distribution is imposed by the environment.

The conditional version of IDS in Eq. (8.3) does not depend on the context distribution χ. It is easy to see that it behaves sub-optimally in both examples, and the bounds in Corollaries 8.1 and 8.2 become vacuous when a context occurs where the information gain is zero for all actions. To understand how the randomness of the context affects the regret bounds, we include the contextual distribution in the information ratio. By optimizing the marginals of the joint distribution over action and context, we obtain an IDS algorithm that leverages 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 joint space of marginal sampling distributions for each context. For context distribution $\chi \in \mathscr{P}(\mathcal{Z})$ and marginals $\xi \in \mathscr{P}_\times(\mathcal{A}^\mathcal{Z})$, we define the integrated features

$$\phi^\chi_\xi \triangleq \int_\mathcal{Z}\int_{\mathcal{A}^z} \phi^z_a d\xi(a,z)d\chi(z)\,,$$

and we let $\phi^\chi_a \triangleq \phi^z_{e_a}$, where e_a for $a \in \mathcal{A}^\mathcal{Z}$ is a deterministic choice in each context. The gap estimate for context distribution χ and marginals ξ is defined as

$$\hat{\Delta}^\chi_t(a,z) \triangleq \min\left\{\max_{b\in\mathcal{A}^\mathcal{Z}}\langle\phi^\chi_b - \phi^z_a, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi^\chi_b - \phi^\chi_{\hat{a}_t}\|_{V_t^{-1}}, 1\right\}. \tag{8.5}$$

Up to the truncation, Jensen's inequality shows that this gap estimate is never worse than the conditional gap defined in Eq. (8.1):

$$\begin{aligned}&\max_{b\in\mathcal{A}^\mathcal{Z}}\langle\phi^\chi_b - \phi^\chi_\xi, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi^\chi_b - \phi^\chi_{\hat{a}_t}\|_{V_t^{-1}}\\ \le &\int_\mathcal{Z}\int_\mathcal{A}\max_{b\in\mathcal{A}^\mathcal{Z}}\langle\phi^z_b - \phi^z_a, \hat{\theta}_t\rangle + \beta_t^{1/2}\|\phi^z_b - \phi^z_{\hat{a}_t}\|_{V_t^{-1}}d\xi(a,z)d\chi(z)\,.\end{aligned} \tag{8.6}$$

Note that the right-hand side of the last equation can be computed in $\mathcal{O}(|\mathcal{A}||\mathcal{Z}|)$ steps, compared to the left-hand side, which requires $\mathcal{O}(|\mathcal{A}|^{|\mathcal{Z}|})$ steps in general.

As information gain we use the same definition as in Eq. (8.2) with the convention that

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

Contextual IDS is defined to optimize the marginals $\xi \in \mathscr{P}^{\mathcal{Z}}$,

$$\xi_t^{\mathrm{IDS}} \triangleq \operatorname*{arg\,min}_{\xi \in \mathscr{P}_\times(\mathcal{A}^{\mathcal{Z}})} \left\{ \Psi_t^\chi(\xi) = \frac{\hat{\Delta}_t^\chi(\xi)^2}{I_t^\chi(\xi)} \right\}. \tag{8.7}$$

The action $a_t \sim \xi_t^{\mathrm{IDS}}(\cdot, z_t)$ is sampled from the marginal corresponding to the observed context z_t. In the joint minimization of the information ratio, the contextual distribution now χ contributes to exploration and a smaller information ratio.

Note that optimizing the marginals is computationally more demanding than just optimizing the sampling distribution conditioned on the context. Since the information ratio is a convex function of the distribution (Lemma 2.3), we can optimize the marginals using standard solvers. A particularly simple implementation is with the Frank-Wolfe algorithm [59, 76], as solving linear functions over $\mathscr{P}_\times(\mathcal{A}^{\mathcal{Z}})$ is immediate. The contextual IDS algorithm with Frank-Wolfe is summarized in Algorithm 8. A numerical demonstration of the contextual setting is in Fig. 8.1.

Before presenting the regret bounds, we extend the definition of the alignment constant. Let $\mathcal{E} \subset \mathbb{R}^d$ be a convex set. The definition of plausible maximizers is extended to the product space $\mathcal{A}_*^{\mathcal{Z}}(\mathcal{E}) \triangleq \times_{z \in \mathcal{Z}} \mathcal{A}_*^z(\mathcal{E})$ and $\mathcal{A}_+^{\mathcal{Z}}(\mathcal{E}) \triangleq \times_{z \in \mathcal{Z}} \mathcal{A}_+^z(\mathcal{E})$. The expected alignment for $\chi \in \mathscr{P}(\mathcal{Z})$ is

$$\alpha^\chi(\mathcal{E}) \triangleq \max_{\omega \in \mathbb{R}^d} \max_{a,b \in \mathcal{A}_*^{\mathcal{Z}}(\mathcal{E})} \min_{c \in \mathcal{A}_+^{\mathcal{Z}}(\mathcal{E})} \frac{\langle \phi_a^\chi - \phi_b^\chi, \omega \rangle^2}{\int_{\mathcal{Z}} \| M_{c(z)}^z \omega \|^2 d\chi(z)}. \tag{8.8}$$

The next lemma shows that this definition is strictly better than the conditional alignment constant $\alpha^z(\mathcal{E})$ for fixed $z \in \mathcal{Z}$. Moreover, the lemma shows that $\alpha^\chi(\mathcal{E})$ is bounded as long as for all $z \in \mathcal{Z}$ and $a, b \in \mathcal{A}^z$, there exists *some* other context $z' \in \mathcal{Z}$ that occurs with positive probability and an action $c \in \mathcal{A}^{z'}$ such that the gap $\langle \phi_a^z - \phi_b^z, \nu \rangle$ can be estimated from observations of c in context z'. This includes cases where $\alpha^z(\mathcal{E}) = \infty$, for example, when there is no information available in context z. The more general definition captures the intuition that the learner can wait for a realization of the context z' to collect the data for estimating the gap $\langle \phi_a^z - \phi_b^z, \nu \rangle$ more easily and at low cost.

Lemma 8.1. *Let $\alpha^\chi(\mathcal{E})$ be the expected alignment* (8.8) *and $\alpha^z(\mathcal{E})$ the conditional alignment* (8.4)*. Then, for any $\mathcal{E} \subset \mathbb{R}^d$ convex and $\chi \in \mathscr{P}(\mathcal{Z})$ it holds that*

$$\alpha^\chi(\mathcal{E}) \leq \int_{\mathcal{Z}} \alpha^z(\mathcal{E}) d\chi(z) \leq \max_{z \in \mathrm{supp}(\chi)} \alpha^z(\mathcal{E}),$$

Algorithm 8: Contextual IDS with Frank-Wolfe

Input: Action set $\mathcal{A}$, context set $\mathcal{Z}$, context distribution $\chi \in \mathscr{P}(\mathcal{Z})$, Frank-Wolfe steps l_{FW}.

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

`// gap estimates & information gain`

$\hat{\Delta}_t(a, c)$, $\forall a \in \mathcal{A}, z \in \mathcal{Z}$ `// Eq. (8.1) or (8.5)`

$I_t(a, z)$, $\forall a \in \mathcal{A}, z \in \mathcal{Z}$ `// Eq. (8.2)`

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

for $l = 2, \dots, l_{FW}$ **do**

$\bar{\Delta}^{(l)} \leftarrow \sum_{z \in \mathcal{Z}, a \in \mathcal{A}} \lambda(z) \xi_t^{(l-1)}(a, z) \Delta_t(a, z)$

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

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

$G^{(l)}(a, z) \leftarrow 2\lambda(z) \Delta_t(a, z) \bar{\Delta}^{(l)} \bar{I}^{(l)} - \lambda(z) I_t(a, z) (\bar{\Delta}^{(l)})^2$

`// Frank-Wolfe step`

for $z \in \mathcal{Z}$ **do**

$a_z^* \leftarrow \arg\max_{a \in \mathcal{A}} G^{(l)}(a, z)$

$\xi_t^{(l)}(a, z) \leftarrow (1 - \frac{1}{l}) \xi_t^{(l-1)}(a, z), \forall a \in \mathcal{A}$

$\xi_t^{(l)}(a_z^*, z) \leftarrow \frac{1}{l}$

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

`Sample action:` $a_t \sim \xi_t^{(l_{\mathrm{FW}})}(\cdot, z_t)$

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

and,

$$\alpha^\chi(\mathcal{E}) \leq \max_{\nu \in \mathbb{R}^d} \max_{z \in \mathcal{Z}, a, b \in \mathcal{A}_z^*(\mathcal{E})} \min_{z' \in \mathcal{Z}, c \in \mathcal{A}_{z'}^+(\mathcal{E})} \frac{\langle \phi_a^z - \phi_b^z, \nu \rangle^2}{\chi(z') \|M_c^{z'} \nu\|^2}.$$

Further, for any Dirac distribution e_z on $z \in \mathcal{Z}$, equality holds: $\alpha^{e_z}(\mathcal{E}) = \alpha^z(\mathcal{E})$.

Proof. The first claim follows from noting that the map $(p, q) \mapsto p^2/q$ is convex on $\mathbb{R} \times \mathbb{R}_{\geq 0}$ and Jensen's inequality applied to the probability measure χ. The second claim follows with $\langle \phi_a^\chi - \phi_b^\chi, \nu \rangle^2 \leq \max_{z \in \mathcal{Z}} \langle \phi_a^z - \phi_b^z, \nu \rangle^2$ and $\int_{\mathcal{Z}} \|M_{c(z)}^z \omega\|^2 d\chi(z) \geq \chi(z') \|M_{c(z')}^{z'} \omega\|^2$ for any $z' \in \mathcal{Z}$. The last claim is immediate from the definitions. □

Theorem 8.1. *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. (8.7)) satisfies,*

$$\mathfrak{R}_n \leq n^{2/3}(32\mathbb{E}[\bar{\alpha}_n\beta_n]\mathbb{E}[\gamma_n])^{1/3} + \mathcal{O}(1)\,,$$

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

Proof. The proof is very similar to the proof of Theorem 6.1. Note that for a fixed $a \in \mathcal{A}^{\mathcal{Z}}$, we can interpret the gap estimate $\hat{\Delta}_t^\chi(a)$ as a non-contextual gap on the extended action space $\mathcal{A}^{\mathcal{Z}}$ with features ϕ_a^χ. Recall that $\hat{a}_t \in \mathcal{A}^{\mathcal{Z}}$ is defined as $\hat{a}_t(z) = \arg\max_{a\in\mathcal{A}^z}\langle \phi_a^z, \theta_t\rangle$. It follows from the same steps in the aforementioned theorem,

$$\hat{\Delta}_t^\chi(\hat{a}_t) \leq \beta_t^{1/2} \max_{a,b\in\mathcal{A}_*^{\mathcal{Z}}(\mathcal{M})} \|\phi_a^\chi - \phi_b^\chi\|_{V_t^{-1}}\,.$$

We abbreviate $\alpha = \alpha^\chi(\mathcal{M})$ and continue the bound on the norm of the feature difference,

$$\begin{aligned}
\max_{a,b\in\mathcal{A}^{\mathcal{Z}}} \|\phi_a^\chi - \phi_b^\chi\|_{V_t^{-1}}^2 &= \max_{\omega\in\mathbb{R}^d} \max_{a,b\in\mathcal{A}_*^{\mathcal{Z}}(\mathcal{M})} \frac{\langle \phi_a^\chi - \phi_b^\chi, V_t^{-1/2}\omega\rangle^2}{\|\omega\|^2} \\
&\leq \alpha \max_{\omega\in\mathbb{R}^d} \max_{c\in\mathcal{A}_+^{\mathcal{Z}}(\mathcal{M})} \frac{\int_{\mathcal{Z}} \|M_{c(z)}^z V_t^{-1/2}\omega\|^2 d\chi(z)}{\|\omega\|^2} \\
&\leq \alpha \max_{c\in\mathcal{A}_+^{\mathcal{Z}}(\mathcal{M})} \int_{\mathcal{Z}} \lambda_{\max}(M_{c(z)}^z V_t^{-1} M_{c(z)}^{z\top}) d\chi(z) \\
&\leq 4\alpha \max_{c\in\mathcal{A}_+^{\mathcal{Z}}(\mathcal{M})} I_t^\chi(c)
\end{aligned}$$

Note also that in the directionally unconstrainted case, $\mathcal{A}_+^{\mathcal{Z}}(\mathcal{M}) = \mathcal{A}^{\mathcal{Z}}$. The information ratio is bounded by optimizing the trade-off between $\hat{a}_t$ and $c_t = \arg\max_{c\in\mathcal{A}^{\mathcal{Z}}} I_t^\chi(c)$. Similar to Eq. (2.6),

$$\Psi_t(\xi_t^{\mathrm{IDS}}) \leq \min_{c\in\mathcal{A}} \frac{4\delta_t\hat{\Delta}_t(c)}{I_t^{\mathrm{DET}}(c)} \overset{(i)}{\leq} \min_{c\in\mathcal{A}} \frac{4\delta_t}{I_t^{\mathrm{DET}}(c)} \overset{(ii)}{\leq} \frac{16\alpha\beta_t}{\hat{\Delta}_t(\hat{a}_t)} \overset{(iii)}{\leq} \frac{32\alpha\beta_t}{\hat{\Delta}_t(\xi_t^{\mathrm{IDS}})}\,.$$

For (i) we used that $\hat{\Delta}_t(c) \leq 1$ and (ii) follows from the previous two displays combined. Lastly, (iii) uses $\hat{\Delta}_t^\chi(\xi_t^{\mathrm{IDS}}) \leq 2\hat{\Delta}_t^\chi(\hat{a}_t)$, which follows from a similar argument as in the proof of lemma Lemma 2.6. Clearly, by

definition, the information ratio cannot be improved by shifting mass from the marginals ξ_t^{IDS} to $\hat{a}_t$ in each context,

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

Hence, the gradient of $\psi(p)$ cannot be negative at $p = 0$, which yields

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

The claimed inequality follows by rearranging. This completes the bound on the generalized information ratio (Eq. (2.7)):

$$\Psi_{3,t}^\chi(\xi_t^{\mathrm{IDS}}) = \frac{\hat{\Delta}_t^\chi(\xi_t^{\mathrm{IDS}})^3}{I_t^\chi(\xi_t^{\mathrm{IDS}})} \leq 32\alpha\beta_t\,.$$

The proof is concluded with Theorem 2.2 and bounding the estimation error. □

Theorem 8.2. *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. (8.7)) satisfies,*

$$\mathfrak{R}_n \leq 4\sqrt{\mathbb{E}[\bar{\alpha}_n\beta_n]\mathbb{E}[\gamma_n]n} + \mathcal{O}(1)\,,$$

where $\bar{\alpha}_n = \frac{1}{n}\alpha^\chi(\mathcal{E}_t)$ 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 refer to the proof of Theorem 6.2 for more details. Note that for any $c \in \mathcal{A}^+(\mathcal{E}_t)$,

$$\hat{\Delta}_t^\chi(\hat{a}_t) \leq 2\beta_t^{1/2} \max_{a,b\in\mathcal{A}_*^{\mathcal{Z}}(\mathcal{E}_t)} \|\phi_a^\chi - \phi_b^\chi\|_{V_t^{-1}}$$

Then, using a similar argument as in the proof of Theorem 6.1, it follows from the definition of $\alpha^\chi(\mathcal{E}_t)$ that

$$\max_{a,b\in\mathcal{A}_*^{\mathcal{Z}}(\mathcal{E}_t)} \|\phi_a^\chi - \phi_b^\chi\|_{V_t^{-1}}^2 \leq 4\alpha^\chi(\mathcal{E}_t) \max_{c\in\mathcal{A}_+^{\mathcal{Z}}(\mathcal{E}_t)} I_t^\chi(c)$$

Hence, deterministically choosing $c_t = \max_{c\in\mathcal{A}_+^{\mathcal{Z}}(\mathcal{E}_t)} I_t^\chi(c)$ in each context leads to a bounded information ratio,

$$\Psi_t^\chi(\xi_t^{\mathrm{IDS}}) \leq \frac{\hat{\Delta}_t^\chi(c_t)^2}{I_t^\chi(c_t)} \leq 16\beta_t\alpha^\chi(\mathcal{E}_t)\,.$$

The claim follows from Corollary 2.1 and bounding the estimation error. □

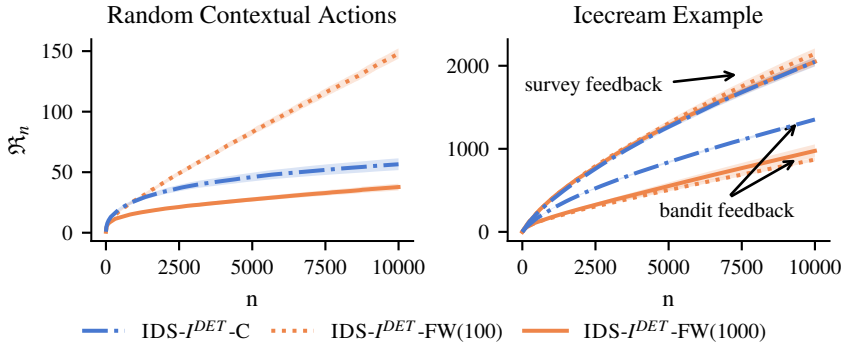


FIGURE 8.1: Numerical results in the contextual setting. The left plots shows randomly generated instances in $\mathbb{R}^4$ with 20 actions, 5 contexts and uniform context distribution. The right plot shows a simulation of the customer survey setting described in Section 8.1.4 using a ice cream rating data set. Results are averaged over 100 runs and the confidence region shows standard error. The FW-suffix indicates the contextual version where we optimize the marginals with 100 and 1000 Frank-Wolfe steps, and C-suffix is conditional IDS. Optimizing the marginal distribution shows a clear advantage on randomly generated instances and on the ice cream data set with bandit feedback. However, 100 Frank-Wolfe steps are not sufficient to find the IDS distribution on the random instances, and the learner suffers linear regret after $\sim$ 1000 steps.

8.1.3 *Illustrating Example: Improving Air-Quality and Traffic Flow*

A naturally varying context in many application captures environmental conditions such as temperature or humidity. As a concrete example, consider the problem of optimizing traffic flow in a city. On each day, we can test a set of design parameters $a \in \mathcal{A}$ that affect the traffic flow. The objective is based on air-quality measurements from sensors spread across locations $\mathcal{X} \subset \mathbb{R}^2$ in the city. Lastly, forecasted weather conditions $z \in \mathcal{Z}$ can be used to predict the air quality. We assume access to features $\phi^z_{a,x} \in \mathbb{R}^d$ that model the air quality $f_x(a, z) = \langle \phi^z_{a,x}, \theta \rangle$ at each measurement station $x \in \mathcal{X}$ for design $a \in \mathcal{A}$ and weather conditions $z \in \mathcal{Z}$.

A weighted sum of air-quality measurements $f(a, z) = \sum_{x \in \mathcal{X}} w_x f_x(a, z)$ serves as target objective. Cumulative regret is a sensible metric in this case, because experimentation on the design space directly affects the air-quality. Moreover, the linear model can be used to account for correlation among the measurements. For example, nearby measurement stations are expected to yield similar measurements. Lastly, weather conditions z are assumed to

follow a known distribution $\chi \in \mathscr{P}(\mathcal{Z})$, that, for example, can be estimated from historical data.

Note that the learner has access to the individual measurement stations $f_x(a_t, z_t)$ for all $x \in \mathcal{X}$, and not all of them are necessarily used in the weighted target. From the partial monitoring perspective, the setup is modeled using the following reward features and feedback maps:

$$\phi_a^\chi = \sum_{x \in \mathcal{X}} w_x \int_{\mathcal{Z}} \phi_a^z d\chi(z)\,,$$
$$M_a^z = [\phi_{a,x}^z]_{x \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}| \times d}\,.$$

8.1.4 *Illustrating Example: Best Ice Cream in Town*

Our next example is about optimizing customer surveys. For concreteness, assume we are helping a *Gelateria* to improve ice cream recipes based on feedback from their customers. There are l different flavours of ice cream and each flavour $z \in [l]$ has an associated design space $\mathcal{I}_z$. The simplest case is A/B testing, where each $\mathcal{I}_z$ is restricted to two options. The joint design space is $\mathcal{A}^{\text{design}} = \times_{z=1}^{l} \mathcal{I}_z$. An ice cream recipe for flavour z is associated with a feature vector $\phi_a^z \in \mathbb{R}^d$ that depends on $a \in \mathcal{A}^{\text{design}}$.

Customers choose ice cream according to a fixed and known distribution $\chi \in \mathscr{P}(\mathcal{I})$. Since the ice cream is for take-away, in general, the customers do not come back for feedback. However, offering ice cream for free, customers can be persuaded to provide detailed feedback on the recipe corresponding to the chosen flavour. To accommodate the two options, we extend the action space to $\tilde{\mathcal{A}} = \mathcal{A}^{\text{design}} \times \{\text{ask}, \text{sell}\}$. The cost of feedback is captured in the reward function, which is defined for action $\tilde{a}_t = (a_t, \text{action}_t)$ as

$$f(\tilde{a}_t, z_t) = \langle \phi_{a_t}^{z_t}, \theta \rangle \mathbb{1}(\text{action}_t = \text{sell})\,.$$

The features and feedback maps are defined correspondingly as

$$\phi_{a,\text{sell}}^\chi = \sum_{z=1}^{l} \chi(z) \phi_a^z\,, \qquad M_{a,\text{sell}}^z = 0$$
$$\phi_{a,\text{ask}}^\chi = 0\,, \qquad M_{a,\text{ask}}^z = \phi_a^z$$

Numerical results for the *survey feedback* setting are shown in Fig. 8.1 using a data set of ice cream ratings[1]. We simulate the setting as defined above for 5 flavors with 4 design options each, and compare to *bandit feedback*, where the learner observes the reward for each action at no additional cost.

1 https://www.kaggle.com/tysonpo/ice-cream-dataset

8.2 KERNELIZED PARTIAL MONITORING

Linear partial monitoring is a highly flexible framework, considering that there is no assumption on the action-feature maps ϕ_a other than compactness of the feature space. This only really becomes a restriction if $\mathcal{A}$ is not finite. Pushing this to an extreme, we 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 reward function with a vector in $\theta \in \mathcal{H}$, and choose a $\mathcal{H}$-valued action-feature map $a \mapsto \phi_a$. By the Riesz representation theorem, we can equivalently identify ϕ_a with a linear functional in the dual $\mathcal{H}^*$. Similarly, the feedback maps are modeled as linear operators $M_a : \mathcal{H} \to \mathbb{R}^d$ that observe the parameter on the subspace $\mathrm{im}(M_a^*)$, where M_a^* is the adjoint mapping. While this adds a great amount of flexibility, it poses two additional challenges.

First, the feature dimension can be large or infinite, which renders regret bounds with a dependence on the dimension vacuous. In previous work on kernelized bandits, this is addressed by replacing the dimension by an appropriate notation of an *effective dimension* [164] or directly bounding the log-determinate that appears in the confidence bounds and the total information gain [152, 163]. We will return to this point in Section 8.2.2.

Second, calculating the least-square estimate in the feature space in general requires $\mathcal{O}(d^2)$ memory and computation steps, which becomes prohibitive if d is large. Kernelized methods circumvent this limitation using a *representer theorem* [64, 85, 142], also known as to as *kernel trick*. Kernel methods are widely popular in machine learning [143], and several kernelized bandit algorithms have been analyzed [1, 40, 152, 164, 165]. More broadly, by interpreting kernel regression as a Gaussian process [83, 130], the field of Bayesian optimization is understood to solve a closely related problem [119, 152].

8.2.1 *Kernel Regression for Partial Monitoring Feedback*

The observations $y_t = M_{a_t}\theta + \epsilon_t$ define a least-square estimate with regularizer $\lambda > 0$ in the Hilbert space $\mathcal{H}$,

$$\hat{\theta}_t \triangleq \operatorname*{arg\,min}_{\theta\in\mathcal{H}} \sum_{s=1}^{t-1} \|M_{a_s}\theta - y_s\|^2 + \lambda\|\theta\|_{\mathcal{H}}^2$$

An immediate observation is that the regularized least-squares solution is always contained in a finite-dimensional subspace spanned by the data,

$$\hat{\theta}_t \in \operatorname{span}(\operatorname{im}(M_{a_s}^*) : s \in [t-1]) .$$

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

$$J_{a,b} : \mathcal{H} \to \mathbb{R}^{m+1}, \quad f \mapsto [\phi_a f, (M_b f)^\top]^\top . \tag{8.9}$$

The lemma below guarantees that the reward estimate $\hat{f}_t(a) \triangleq \langle \phi_a, \hat{\theta}_t \rangle$ can be computed efficiently from finite-dimensional quantities, provided that we can compute 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 required operators are summarized by a *kernel function*, that naturally defines a vector-valued *reproducing kernel Hilbert space* (RKHS) [10, 126]. For a modern introduction, see also [33, Definition 2.1].

Lemma 8.2 (Partial Monitoring Representer Theorem)**.** *Assume that the subspace* $\operatorname{span}(\operatorname{im}(J_{a,b}^*) : a, b \in \mathcal{A}) \subset \mathcal{H}$ *is a* $\mathbb{R}^{m+1}$*-valued RKHS over* $\mathcal{A}$ *with evaluation functionals* $J_{a,b}$ *defined in Eq.* (8.9) *and a known corresponding kernel*

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

For reward evaluations of the least-square solution, the following holds:

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

which is defined in terms of finite-dimensional expressions:

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

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

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

If $\mathcal{H}$ is a real-valued RKHS over k with kernel $k : \mathcal{A} \times \mathcal{A}$ and kernel features $k_a \in \mathcal{H}$, the lemma recovers standard kernel regression for bandit feedback with $\phi_a = M_a = k_a$. Below, we illustrate in several examples how the extra generality is useful in settings beyond the standard bandit model.

Proof of Lemma 8.2. We define the map

$$\Phi_t : \mathcal{H} \to \mathbb{R}^{(t-1)m}, \quad \theta \mapsto [(M_{a_1}\theta)^\top, \dots, (M_{a_{t-1}}\theta)^\top]^\top \tag{8.10}$$

as the stack of evaluation maps in the observation history. Ignoring computability issues, the least-square solution is $\hat{\theta}_t = V_t^{-1}\Phi_t^* \boldsymbol{y}_t$, where $V_t : \mathcal{H} \to \mathcal{H}, \theta \mapsto (\Phi_t^*\Phi_t + \lambda \mathbf{1}_{\mathcal{H}})\theta$ is an invertable linear map 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. □

In order to make use of the estimator in the IDS algorithm, we also need a kernelized statement of the confidence bounds, which we provide in the next lemma.

Lemma 8.3. *Let* $V_t = \sum_{s=1}^{s-1} M_{a_s}^* M_{a_s} + \lambda \mathbf{1}_{\mathcal{H}}$ *and* $\mathcal{E}_{t,\delta} = \{\theta \in \mathcal{H} : \|\theta\|_{V_t}^2 \le \beta_{t,\delta}\}$ *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* $\|\theta\|_{\mathcal{H}} \le B$, *then*

$$\mathbb{P}[\forall t \ge 1, \theta \in \mathcal{E}_t] \ge 1 - \delta\,.$$

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

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

where $\psi_t(a,b) \triangleq \frac{1}{\lambda}\Big(\psi(a,b) - (k_t(a) - k_t(b))^\top (K_t + \lambda\mathbf{1})^{-1}(k_t(a) - k_t(b))\Big)$ *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 Lemma 8.2.*

Proof. The confidence set is the same as [1, 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 (Lemma D.2),

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

where Φ_t is defined as in Eq. (8.10). □

8.2.2 *Kernelized Information-Directed Sampling*

Equipped with the representer theorem and the confidence bound from the previous section, we can define kernelized gap estimates and kernelized information gain functions for information-directed sampling. We use the relaxed gap estimate from Section 6.3.2, which is computationally simpler. Let $\hat{f}_t(a)$ as defined in Lemma 8.2, and $\beta_t \triangleq \beta_{t,1/t^2}$ and $\psi_t(a,b)$ as defined in Lemma 8.3. The kernelized gap estimate is

$$\hat{\Delta}_t(a) = \hat{f}_t(\hat{a}_t) - \hat{f}_t(a) + (\beta_t \psi_t(\hat{a}_t, a))^{1/2}\,, \tag{8.11}$$

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. (6.7) is

$$I_t^{\mathrm{DET}}(a) = \frac{1}{2} \log\det\Big(\mathbf{1}_m + \frac{1}{\lambda}\big(k_M(a,a) - L_t(a) K_t^{-1} L_t(a)^\top\big)\Big)\,, \tag{8.12}$$

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. (8.10). The total information gain is

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

The theoretical guarantees for IDS Theorems 6.1 and 6.2 stated in terms of the confidence coefficient β_n and the total information gain $\gamma_n = \sum_{t=1}^n I_t^{\mathrm{DET}}(a_t)$ are not affected by the change of representation. The log-determinant $\log\det(\mathbf{1} + \lambda^{-1} K_t)$ can often be bounded independent 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 [121, 152], summarized in [163]. A kernelized version of the directed information gain Eq. (6.12) can be derived similarly.

8.2.3 *Example: Kernelized Dueling Bandits*

We present a kernelized version of the linear dueling bandit variant introduced in Example 6.4. 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

$$y_t = g(a_t^1) - g(a_t^2) + \epsilon_t\,, \tag{8.13}$$

Algorithm 9: Approximate IDS for Dueling Feedback

Input: Ground set $\mathcal{I}$

for $t = 1, 2, 3, \dots$ **do**

  $\hat{a}_t \leftarrow \arg\max_{a \in \mathcal{I}} \hat{g}_t(a)$ `// Eq. (8.14)`

  $\hat{\Delta}_t^g(a) \leftarrow \delta_t + \hat{g}_t(\hat{a}_t) - \hat{g}_t(a)$ `// Eq. (8.15)`

  $I_t(a, c) \leftarrow \frac{1}{2} \log(1 + \psi_t(a, c))$ `// Eq. (8.16)`

  $c_t, p_t \leftarrow \arg\min_{c \in \mathcal{I}, p \in [0,1]} \frac{\left((1-p)\delta_t + p(\delta_t + \hat{\Delta}_t^g(\hat{a}_t, c))\right)^2}{p I_t(\hat{a}_t, c)}$

  $B_t \sim \text{Bernoulli}(p_t)$

  if $B_t == 1$, **then**

    $(a_t^1, a_t^2) \leftarrow (\hat{a}_t, c_t)$

    $y_t \leftarrow$ `DuelingFeedback`(a_t^1, a_t^2) `// Eq. (8.13)`

  else

    $(a_t^1, a_t^2) \leftarrow (\hat{a}_t, \hat{a}_t)$ `// No feedback`

and suffers instantaneous regret $f(a_t^1, a_t^2) = g(a_t^1) + g(a_t^2)$ for both actions. Provided that $\|g\|_\infty \leq 1$, the sub-Gaussian noise ϵ_t can be defined such that the feedback is binary $y_t \in \{-1, 1\}$ as in the standard dueling model, but this is not a requirement.

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 $\|g\|_{\mathcal{H}} \leq B$. 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

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

With the above, we can directly apply Algorithm 6 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. (8.11) and the

approximate IDS distribution. This can be improved, by directly estimating the utility function g and using the dueling structure, as we explain next.

KERNELIZED DUELING IDS Recall that $\boldsymbol{y}_t = [y_1, \dots, y_{t-1}]^\top$ is the vector that collects the observations and let $\{(a_s^1, a_s^2)\}_{s=1}^{t-1}$ be the action history. Note that for any $a \in \mathcal{I}$ we can recover estimates for the utility $g(a)$ via $\hat{g}_t(a) = \frac{1}{2}\hat{f}_t\big((a,a)\big)$. Using evaluation weights $k_t^g(a) \triangleq [k(a, a_s^1) - k(a, a_s^2)]_{s=1,\dots,t-1}$, we define the estimate

$$\hat{g}_t(a) = k_t^g(a)^\top (K_t + \lambda \mathbf{1}_{t-1})^{-1} \boldsymbol{y}_t \,. \tag{8.14}$$

The maximizer of the estimated utility function is $\hat{a}_t = \arg\max_{a \in \mathcal{I}} \hat{g}_t(a)$. Note that the pair $(\hat{a}_t, \hat{a}_t) \in \mathcal{A}$ is the greedy action for pair-wise evaluations. Define

$$\delta_t \triangleq \max_{b \in \mathcal{I}} \hat{g}_t(b) - \hat{g}_t(\hat{a}_t) + (\beta_t \psi_t^g(b, \hat{a}_t))^{1/2} \,,$$

where we use $\beta_t = \beta_{t,1/t^2}$ as in Lemma 8.3 and let $\psi_t^g(a,b) \triangleq \|k_a - k_b\|^2_{V_t^{-1}}$, which is computed with the kernel metric $\psi^g(a,b) \triangleq k(a,a) + k(b,b) - 2k(a,b)$ as follows:

$$\psi_t^g(a,b) = \frac{1}{\lambda}\big(\psi^g(a,b) - (k_t^g(a) - k_t^g(b))^\top (K_t + \lambda \mathbf{1}_{t-1})^{-1} (k_t^g(a) - k_t^g(b))\big) \,.$$

The gap estimate for the utility function is defined for any $a \in \mathcal{I}$,

$$\hat{\Delta}_t^g(a) \triangleq \delta_t + \hat{g}_t(\hat{a}_t) - \hat{g}_t(a) \,, \tag{8.15}$$

which is directly extended to pairs $\hat{\Delta}_t(a^1, a^2) \triangleq \hat{\Delta}_t^g(a^1) + \hat{\Delta}_t^g(a^2)$. Lastly, the information gain Eq. (8.12) for a pair of actions (a^1, a^2) simplifies to

$$I_t(a^1, a^2) = \frac{1}{2} \log\big(1 + \psi_t^g(a^1, a^2)\big) \,, \tag{8.16}$$

The key observation is that it suffices to optimize the information ratio over the set $\mathcal{B} = \{(\hat{a}_t, b) : b \in \mathcal{I}\}$. Note that $\mathcal{B}$ contains the greedy action choice $(\hat{a}_t, \hat{a}_t)$ for which $I_t(\hat{a}_t, \hat{a}_t) = 0$. The approximate IDS algorithm restricted to the subset $\mathcal{B}$ therefore randomizes between the greedy pair $(\hat{a}_t, \hat{a}_t)$ and some other action pair in $(\hat{a}_t, b) \in \mathcal{B}$, that provides information on the reward difference $g(\hat{a}_t) - g(b)$. Concretely, let

$$c_t, p_t \triangleq \operatorname*{arg\,min}_{c \in \mathcal{I}, p \in [0,1]} \frac{\big((1-p)2\delta_t + p(\delta_t + \hat{\Delta}_t^g(c))\big)^2}{p I_t(\hat{a}_t, c)} \,. \tag{8.17}$$

The trade-off probability $p_t(c)$ is obtained with Lemma 2.5 in closed-form,

$$p_t(c) = \min\left(\frac{2\delta_t}{\hat{\Delta}_t^g(c) - \delta_t}, 1\right).$$

The approximate dueling policy then samples $a_t = (\hat{a}_t, \hat{a}_t)$ with probability $(1 - p_t)$ and $a_t = (\hat{a}_t, c_t)$ with probability p_t. We denote the corresponding sampling distribution by μ_t^{duel}. To bound the information ratio for μ_t^{duel}, we let $\mathcal{B}_t^* = \{b \in \mathcal{I} : \hat{g}_t(\hat{a}_t) - g(b) \leq (\beta_t \psi_t(b, \hat{a}_t))^{1/2}\}$. Note that

$$\begin{aligned}
\delta_t &= \max_{b \in \mathcal{I}} \hat{g}_t(b) - \hat{g}_t(\hat{a}_t) + (\beta_t \psi_t^g(b, \hat{a}_t))^{1/2} \\
&= \max_{b \in \mathcal{B}_t^*} \hat{g}_t(b) - \hat{g}_t(\hat{a}_t) + (\beta_t \psi_t^g(b, \hat{a}_t))^{1/2} \\
&\leq \max_{b \in \mathcal{B}_t^*} (\beta_t \psi_t^g(b, \hat{a}_t))^{1/2}. \qquad (8.18)
\end{aligned}$$

The inequality follows by noting that $\hat{g}_t(b) \leq \hat{g}_t(\hat{a}_t)$ by definition of $\hat{a}_t$. The information-ratio is bounded by playing deterministically on $\mathcal{B}_t^*$,

$$\begin{aligned}
\Psi_t(\mu_t^{\text{duel}}) = \frac{\hat{\Delta}_t(\mu_t^{\text{duel}})^2}{I_t(\mu_t^{\text{duel}})} &\leq \min_{b \in \mathcal{B}_t^*} \frac{\hat{\Delta}_t(\hat{a}_t, b)^2}{I_t(\hat{a}_t, b)} = \min_{b \in \mathcal{B}_t^*} \frac{(\delta_t + \hat{\Delta}_t^g(b))^2}{I_t(\hat{a}_t, b)} \\
&\leq \frac{9\beta_t \psi_t(a_t, b)}{I_t(a_t, b)}.
\end{aligned}$$

The second inequality follows from Eq. (8.18) and noting that any $b \in \mathcal{B}_t^*$ satisfies $\hat{\Delta}_t^g(b) \leq (2\beta_t \psi_t(b, \hat{a}_t))^{1/2}$. Lastly, assuming that the regularizer is chosen large enough such that $\psi_t(a, b) \leq 1$ holds for all $a, b \in \mathcal{I}$, we get $I_t(\hat{a}_t, b) \geq 4\psi_t(\hat{a}_t, b)$. Therefore, $\Psi_t(\mu_t^{\text{duel}}) \leq 36\beta_t$. A regret bound for approximate dueling IDS follows with the established technique. The next corollary summarizes the result.

Corollary 8.3. *Assume that the feedback satisfies Eq.* (8.13) *for a function* $g \in \mathcal{H}(\mathcal{I})$ *with bounded norm* $\|g\|_{\mathcal{H}} \leq B$ *and* ρ*-sub-Gaussian observation noise. Then the regret of the dueling IDS (Algorithm 9) satisfies*

$$\mathfrak{R}_n \leq \mathcal{O}(\sqrt{n\beta_n\gamma_n}).$$

Using the same argument as in Section 3.3.2 and assuming that the optimal action is unique, it is also possible to show a gap-dependent regret $\mathcal{R} \leq \mathcal{O}(\Delta_{\min}^{-1}\beta_n\gamma_n)$, where $\Delta_{\min}$ is the minimum gap of the utility function g. This translates to $\mathcal{O}(\Delta_{\min}^{-1} d^2 \log(n)^2)$ regret for finite-dimensional parameter.

We remark that the worst-case analysis justifies the use of a deterministic action choice. The main difference is that the deterministic algorithm only chooses the greedy pair $(\hat{a}_t, \hat{a}_t)$ when $\delta_t = 0$, i. e. when the optimal action has been identified with probability $1 - t^{-2}$. On the other hand, the randomized IDS distribution samples the greedy pair as soon as $\delta_t \leq \hat{\Delta}_t^g(c_t)$.

The efficient version of IDS for dueling bandits is summarized in Algorithm 9. In conclusion, by directly using the dueling structure, we reduced the computation complexity per round from $\mathcal{O}(|\mathcal{I}|^2)$ to $\mathcal{O}(|\mathcal{I}|)$, which makes the approach applicable to much larger actions.

8.2.4 *Example: Bias-Robust Bayesian Optimization*

A perhaps unexpected application of the utility-based dueling bandit model is in robust regret minimization. We discuss a bandit setting, where the feedback is adversarially biased in way that is not directly observed by the learner. *Additive confounded bandit feedback* is defined as follows:

$$\check{y}_t = f(a_t) + \kappa_t + \epsilon_t \,, \tag{8.19}$$

where ϵ_t is conditionally independent ρ-sub-Gaussian noise and $\kappa_t \in \mathbb{R}$ is a time-dependent confounding term that is chosen adversarially and is hidden from the learner.

Applications of the confounded feedback model include robust optimization with feedback from physical sensing devices that are subject to calibration errors or time-dependent drift [80, 93, 151]. For example, feedback drift is a substantial challenge in some applications of Bayesian optimization, such as tuning design parameters of particle accelerators [92, 116, 141]. In the context of mobile health applications, a similar model was investigated in [69] to handle non-stationary user feedback. Another application area is adversarial attacks on bandit algorithms, which was recently studied with various assumptions on the confounding sequence [24, 25, 108, 114]. For further discussion of potential applications, see [89].

Clearly, sublinear regret in the feedback model Eq. (8.19) is not possible without further restrictions. Here, we assume that κ_t is chosen by the adversary at the beginning of round t, independent of the action choice of the learner. This is an extension of the *confounded linear bandit* setting, proposed and analyzed by Krishnamurthy, Wu & Syrgkanis [97] for linear f. More concretely, we consider the following two cases:

a) The bias κ_t is bounded, $|\kappa_t| \leq C_{max}$ and fixed at the beginning of round t, but can otherwise arbitrarily depend on $(a_s, y_s)_{s=1}^{t-1}$.

b) The difference between two consecutive bias terms is bounded, $|\kappa_t - \kappa_{t-1}| \leq D_{max}$ and κ_t is fixed at the beginning of round $t-1$, but can otherwise arbitrarily depend on $(a_s, y_s)_{s=1}^{t-2}$.

Note that the best action $a^* = \arg\max_{a \in \mathcal{A}} f(a)$ and the regret can be defined as before, since both quantities are invariant under additive shift of the reward function.

REDUCTION TO DUELING BANDITS The observation model (8.19) allows the adversary to absorb any additive shift of the reward, i.e. $\tilde{f}(a) = f(a) + C$ for $C \in \mathbb{R}$, hence rendering feedback for $f(a)$ and $\tilde{f}(a)$ indistinguishable. In general, the learner can only hope to recover the true function up to an additive constant.

To obtain informative feedback, the learner has to randomize the action choice, since otherwise the adversary can predict the action and choose the confounding term κ_t in a way that $\check{y}_t = \epsilon_t$. Below, we propose two randomized sampling schemes that reduce confounded bandit evaluation to the utility-based dueling feedback defined in Eq. (8.13). That way, we can directly apply the dueling IDS algorithm from Section 8.2.3.

TWO-POINT REDUCTION The first scheme uses two confounded observations to construct a single dueling evaluation. Given inputs $a_t^1, a_t^2 \in \mathcal{A}$ in round t, we obtain two confounded observations, where the order of evaluation is uniformly randomized. The two observations are

$$\check{y}_t^1 = f(a_t^1) + \kappa_{2t+i_t} + \epsilon_{2t+i_t}\,,$$
$$\check{y}_t^2 = f(a_t^2) + \kappa_{2t+1-i_t} + \epsilon_{2t+1-i_t}\,,$$

where $i_t \sim \text{Bernoulli}(0.5)$. We then define

$$y_t^{\text{duel-2}} = \check{y}_t^1 - \check{y}_t^2\,. \tag{8.20}$$

Assuming that κ_{2t} and κ_{2t+1} are fixed before either of a_t^1, a_t^2 is chosen by the learner and using that the observation noise ϵ_t is zero-mean, one easily confirms that $\mathbb{E}\left[y^{\text{duel-2}}\right] = f(a_t^1) - f(a_t^2)$. We further make use of the following properties of sub-Gaussian random variables. A bounded random variable X such that $X \in [-B, B]$ is B^2-sub-Gaussian. Two independent random variables X_1, X_2 that are B_1^2- and B_2^2-sub-Gaussian respectively, $X_1 + X_2$ is $(B_1^2 + B_2^2)$-sub-Gaussian. Hence if $|\kappa_{2t} - \kappa_{2t+1}| \leq D_{\max}$, it follows that the effective observation noise

$$y^{\text{duel-2}} - \mathbb{E}\left[y^{\text{duel-2}}\right] = \kappa_{2t+i_t} + \epsilon_{2t+i_t} - (\kappa_{2t+1-i_t} + \epsilon_{2t+1-i_t})$$

is $\sqrt{D_{\max}^2 + 2\rho^2}$-sub-Gaussian.

ONE-POINT REDUCTION It is also possible to construct the dueling bandit feedback from a *single* randomized evaluation. For given inputs $a_t^1, a_t^2 \in \mathcal{X}$ we choose one point uniformly at random and evaluate the confounded function (8.19) to obtain a the observation

$$\check{y}_t = f(a_t^{(1+i_t)}) + \kappa_t + \epsilon_t\,,$$

where $i_t \sim \text{Bernoulli}(0.5)$. The dueling bandit feedback is

$$y^{\text{duel-1}} = (-1)^{i_t} 2 y_t\,. \tag{8.21}$$

Again, we get an unbiased observation of the reward difference, $\mathbb{E}\left[y^{\text{duel-1}}\right] = f(a_t^1) - f(a_t^2)$. Further, if $|\kappa_t| \leq C_{\max}$, then $y^{\text{duel-1}} - \mathbb{E}\left[y^{\text{duel-1}}\right]$ is $2\sqrt{C_{\max}^2 + \rho^2}$-sub-Gaussian. Compared to the two-point reduction, here the sub-Gaussian variance ρ depends on the absolute value $|\kappa_t|$ of the confounding term instead of the difference $|\kappa_t - \kappa_{t+1}|$. On the other hand, the one-point sampling scheme only requires κ_t to be fixed before the choice of a_t, but may depend on *all* previous actions and observations.

REGRET GUARANTEES Using either reduction, we can directly apply the (kernelized) dueling IDS algorithm (Algorithm 9). Note that we require knowledge of the bounds $C_{\max}$ or $D_{\max}$ to scale the noise constant in the confidence coefficient appropriately. To make the dependence explicit, note that for ρ-sub-Gaussian noise, $\beta_n(\rho^2) = \mathcal{O}(\rho^2 \gamma_n)$. The result is summarized in the following corollary.

Corollary 8.4. *For biased observations according to Eq.* (8.13) *with ρ^2-sub-Gaussian observation noise and dueling feedback obtained via the one-point reduction in Eq.* (8.21)*, the regret of Algorithm 9 satisfies*

$$\mathfrak{R}_n \leq \mathcal{O}\big((C_{\max} + \rho)\sqrt{n\gamma_n}\big)\,,$$

assuming that $\max_{t\in[n]} \kappa_t \leq C_{\max}$ and the adversary is allowed to choose b_t depending on all previous actions and observations, $\{a_s, y_s\}_{s=1}^{t-1}$.

With the two-point reduction in Eq. (8.20)*, the regret of Algorithm 9 satisfies ,*

$$\mathfrak{R}_n \leq \mathcal{O}\big((D_{\max} + \rho)\sqrt{n\gamma_n}\big)\,,$$

assuming that $\max_{t\in[n]} |\kappa_{2t} - \kappa_{2t+1}| \leq D_{\max}$ and the adversary is allowed to choose κ_t depending on all but the last two actions and observations, $\{a_s, y_s\}_{s=1}^{t-2}$.

The bounds for the confounded bandit setting match the bounds in the *unconfounded* setting up to the larger noise constant, and further match the minimax lower bound in the linear bandit setting up to logarithmic factors.

A CONNECTION TO DOUBLY-ROBUST ESTIMATION The linear case $f(a) = \langle \phi_a, \theta \rangle$ with finite-dimensional parameter $\theta \in \mathbb{R}^d$ and features $\phi_a \in \mathbb{R}^d$ was previously investigated by Krishnamurthy, Wu & Syrgkanis [97]. To estimate the unknown parameter θ directly from confounded observations $\{\check{y}_s = \langle \phi_{a_s}, \theta \rangle + \kappa_s + \epsilon_s)\}_{s=1}^{t-1}$, they propose a doubly-robust estimator $\hat{\theta}_t^{\mathrm{dr}}$. For centered feature vectors $\bar{\phi}_t = \int_{\mathcal{A}} \phi_a d\mu_t(a)$ and regularizer $\lambda > 0$, they define

$$\Gamma_t \triangleq \sum_{s=1}^{t-1} (\phi_{a_s} - \bar{\phi}_s)(\phi_{a_s} - \bar{\phi}_s)^\top + \lambda \mathbf{1}_d \,,$$

$$\hat{\theta}_t^{\mathrm{dr}} \triangleq \Gamma_t^{-1} \sum_{s=1}^{t-1} (\phi_{a_s} - \bar{\phi}_s)\check{y}_s. \tag{8.22}$$

They further derived a self-normalized confidence bound:

$$\|\hat{\theta}_t^{\mathrm{dr}} - \theta\|_{\Gamma_t}^2 \leq \mathcal{O}(d \log(n) + \log(n/\delta) + \lambda) \,. \tag{8.23}$$

The scaling of this confidence set is close to optimal for adaptive data in general, even without confounding [103, Exercise 20.2]. Interestingly, when $\mu_t = \mathrm{Uniform}(\{a^1, a^2\})$ is chosen to randomize between two actions $a^1, a^2 \in \mathcal{A}$, then this estimator coincides with the least-squares estimator that we obtain for the dueling bandit feedback. This follows immediately by observing that $2(\phi_{a_t} - \bar{\phi}_t) = \phi_{a^1} - \phi_{a^2}$ for $a_t \in \{a^1, a^2\}$. Up to constants and logarithmic factors, both confidence sets result in the same regret bound $\mathfrak{R}_n \leq \tilde{\mathcal{O}}(d\sqrt{n})$. The main difference is that the reduction to dueling feedback avoids the reformulation with the doubly-robust estimator and benefits from improved constants, while the concentration bound in Eq. (8.23) is a more general result.

The BOSE algorithm by Krishnamurthy, Wu & Syrgkanis [97] chooses the sampling distribution μ_t^{BOSE} directly over plausible maximizers such that the variance $\|\phi_t - \bar{\phi}_t\|_{\Gamma_t^{-1}}^2$ of the doubly-robust estimator is well-behaved. The approach requires solving a convex-quadratic feasibility problem over the space of sampling distributions, which is computationally much more expensive than dueling IDS. Further, the μ_t^{BOSE} distribution is supported on $d+1$ points in general, which makes a direct kernelization of the approach difficult. Lastly, we note that Kim & Paik [84] analyzed Thompson sampling

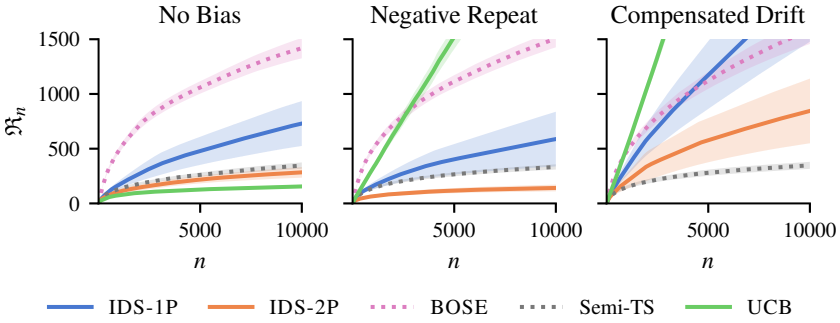


FIGURE 8.2: In the unconfounded setting (left), UCB works best, while SEMITS and IDS-2P are not much worse. When the adversary subtracts the observation from the last round as bias (middle), UCB has linear regret, whereas all other methods, that are designed for this setting, have sublinear regret. IDS-2P works best, followed by SEMI-TS and IDS-1P. In the setting with drift (right), the situation is similar, but now SEMI-TS works best. Note that we compensate the drift in the observations by subtracting the last observation. Otherwise, the bias term is unbounded, and all methods expect for IDS-2P suffer linear regret.

with a similar doubly-robust estimator, that coincides with our estimation scheme in the same way.

NUMERICAL EXPERIMENTS We evaluate the proposed method with the one-point reduction (IDS-1P) and the two-point reduction (IDS-2P) in two numerical experiments with confounded observations. To allow a fair comparison with the two-sample scheme, we account for the regret of both evaluations and scale the number of rounds appropriately.

LINEAR REWARD In the first experiment, we use a linear reward function $f(x) = \langle x, \theta \rangle$. In this setting, the BOSE method by Krishnamurthy, Wu & Syrgkanis [97] and the semi-parametric variant of Thompson sampling (SEMI-TS) by Kim & Paik [84] apply. We also compare to LinUCB [3, 13], which does not directly deal with the confounding. For each repetition we sample $k = 20$ actions uniformly on the $d = 4$ dimensional unit sphere. We consider three different types of confounding: *a)* no confounding; *b)* the adversary repeats the last observation with a minus sign, $b_t = -y_{t-1}$; *c)* a continues drift, $b_t = -0.1t$. Since the drift results in an unbounded bias term, all methods except for IDS-2P suffer linear regret (plot not shown). An immediate fix is to compensate the drift by adding the last observation,

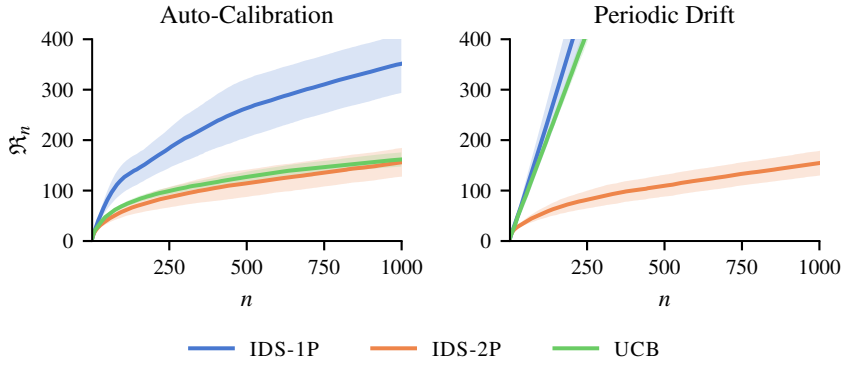


FIGURE 8.3: Results in the kernelized setting. UCB and IDS-2P are competitive despite the confounding by a simulated re-calibration process (left). With a bias that adds a linear and a period function of the time step, only IDS-2P achieves sublinear regret, since IDS-1P requires the bias to be bounded (right).

thereby making the range of the bias term bounded. The results are shown in Fig. 8.2.

CAMELBACK The second experiment is in the kernelized setting. As benchmark we use the camelback function on the domain $[-2, 2] \times [-1, 1]$:

$$f(x_1, x_2) = -\min\left(x_1^2(4 - 2.1x_1^2 + \tfrac{x_1^4}{3.}) + x_1x_2 + x_2^2(4x_2^2 - 4), 2.5\right).$$

We discretize the input space using 30 points per dimension. The only direct competitor is the method of Bogunovic *et al.* [24]. This method is equivalent to GP-UCB [152] with an up-scaled confidence coefficient. This suggests that the UCB approach is inherently robust up to a certain degree of corruption, which is also visible in our experiments. We use two types of confounding that we expect is relevant in applications: *a)* periodic drift of the objective, $b_t = \sin(0.2t) - 0.1t$; and *b)* a calibration process, which monitors a moving average over the last 10 observations and adjusts the output range to $[-0.1, 0.1]$ whenever the average is no longer in this range. Results are shown in Figure 8.3.

8.2.5 *Example: Gradient-Only Global Optimization*

Bayesian optimization [119, 145, 152] is typically phrased as a zero-order global optimization method with noisy evaluations and is closely related to

the kernelized bandit setting. Previous work also incorporates gradient and hessian information where it is available [173, 174]. Since the gradient is a linear operator, we can directly apply IDS in this setting. Assuming that the learner observes the function evaluation *and* the gradient feedback, we immediately obtain the same bounds as in the bandit setting. While our worst-case regret bounds do not reflect any improvement from the additional information, we note that IDS explicitly incorporates the anticipated gradient feedback in the information gain.

We analyze a different setting, where the learner observes *only* the gradient, which can be understood as a type of dueling bandit. Arguably, few optimization settings exist where gradient information is available while function evaluations are not. As such, the example serves mainly as an illustration. We describe a link to adversarial regret minimization at the end of the section.

In the following, we let $\mathcal{A} \subset \mathbb{R}^m$ be a compact and connected set, and $\mathcal{H}$ a RKHS over $\mathcal{A}$ with differentiable kernel $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. The action features are set to the kernel functionals, $\phi_a = k_a = k(a, \cdot)$. We compute the covariance for the observation operator $M_a = \nabla_x : \mathcal{H} \to \mathbb{R}^m$ next. For $a, b \in \mathcal{A}$ and $i, j = 1, \dots, m$, we get

$$[M_a M_b^*]_{ij} = \langle e_i, M_a M_b^* e_j \rangle = [\nabla_a \langle k_a, M_b^* e_j \rangle_{\mathcal{H}}]_i = \frac{\partial}{\partial a_i} \frac{\partial}{\partial b_j} k(a, b)\,,$$

$$[k_a M_b^*]_i = \langle k_a, M_b^* e_i \rangle_{\mathcal{H}} = \langle \nabla_b k_a, e_i \rangle = \frac{\partial}{\partial b_i} k(a, b)\,.$$

Through the gradient observations, the kernel regression results in *Bayesian quadrature* [51, 123]. In particular, any gap difference can be estimated by integrating the a path connecting the the inputs, which means the game is globally observable. Formally, we have to show that for all $a, b \in \mathcal{A}$, $k_a - k_b \in \operatorname{span}(\operatorname{im}(M_c^*) : c \in \mathcal{A})$. To this end, let $\tau : [0, 1] \to \mathcal{A}$ be a differentiable path with $\tau(0) = a$, $\tau(1) = b$ and constant velocity $\|\dot{\tau}\| = C$. We claim that

$$k_a - k_b = \int_0^1 M_{\tau(t)}^* \dot{\tau}(t) dt\,.$$

The claim follows by the fundamental theorem of calculus. For any $f \in \mathcal{H}$,

$$\begin{aligned}\left\langle \int_0^1 M^*_{\tau(t)} \dot{\tau}(t) dt, f \right\rangle &= \int_0^1 \langle M^*_{\tau(t)} \dot{\tau}(t), f \rangle dt \\ &= \int_0^1 \langle \dot{\tau}(t), M_a f \rangle dt \\ &= \int_0^1 \langle \dot{\tau}(t), \nabla_a f \rangle dt \\ &= f(a) - f(b) = \langle k_a - k_b, f \rangle .\end{aligned}$$

The alignment constant (Eq. (6.4)) can be bounded using Lemma 6.1 by $\alpha(\mathcal{M}) \leq (\int_0^1 \|\dot{\tau}(t)\| dt)^2 \leq C^2 \operatorname{diam}(\mathcal{A})^2$. Hence, for functions f with $\|f\|_{\mathcal{H}} \leq B$, Theorem 6.1 guarantees that IDS has worst-case regret at at most $\mathfrak{R}_n \leq (n^{2/3}(\operatorname{diam}(\mathcal{A}) B \gamma_n \beta_n)^{1/3})$. For standard kernel function such as the RBF kernel, the regret bound sublinear, which implies global convergence. In general, the learner has to evaluate suboptimal actions to estimate the reward difference between two plausible (local) optima, therefore the game is not locally observable.

The model can be applied in a semi-adversarial bandit setting, similar to the model discussed in Section 8.2.4. Concretely, we consider the case where feedback is subject to a time-dependent drift,

$$y_t = f(a_t) + \kappa(t) + \epsilon_t .$$

The drift function $\kappa(t)$ depends only on the evaluation step and $a_t \in \mathbb{R}^m$ is a continuous input parameter. Since the drift is independent of the parameter a_t, the gradient of the feedback with respect to the input parameter is not affected by the drift, but the challenge is to construct a gradient estimator from a single observation.

One-point gradient estimates are common in the literature on online bandit convex optimization [58, 75] and previously have been explored in conjunction with dueling bandits [176]. In the simplest formulation, the learner requests gradient feedback for a point $a_t \in \mathcal{A}$. The system is evaluated at $a_t + \epsilon u_t$, where $\epsilon > 0$ is a step-size and $u_t \sim \text{Uniform}(\mathbb{S}^m)$ is an independently sample from the unit sphere $\mathbb{S}^m \triangleq \{u \in \mathbb{R}^m : \|u\| = 1\}$, and the gradient estimate is defined as $\hat{g}_t \triangleq \frac{m}{\epsilon} y_t$. This estimate is understood as an unbiased gradient observation of a smoothed reward,

$$\mathbb{E}_{u_t}[\hat{g}_t] = \nabla_a \mathbb{E}_{u \sim \text{Uniform}(\mathbb{S}^m)}[f(a_t + \epsilon u)] ,$$

and the bias can be controlled assuming smoothness [75, Lemma 5 & 7]. However, compared to the dueling bandit reduction from Section 8.2.4,

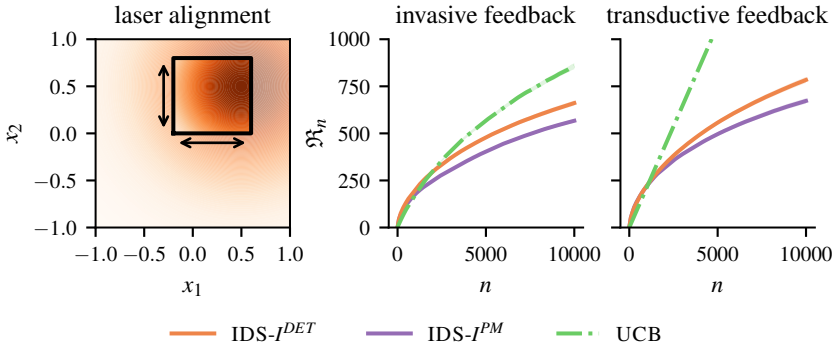


FIGURE 8.4: A demonstration of the stylized laser example with $I(a^1, a^2) = \exp\big(-((a^1 - 0.5)^2 + (a^2 - 0.5)^2)\big)$. The left plot shows the intensity function of the laser on the two dimensional plane. The objective is to shift the square target such that the integrated intensity within the square is maximized. The learner chooses actions to either observe a noisy measurement of the intensity, or alternatively, the energy function directly, evaluated on a measurement grid within the square (invasive feedback). The latter feedback is obtained from a screen that is put in the line of the laser, which blocks the beam and voids the reward signal. In the second variant (transductive feedback), the learner obtains information *only* through the invasive measurements. To solve the task, the learner needs to estimate the function with invasive measurements, while keeping the unobstructed target in a position with maximum integrated intensity sufficiently often. The plots on the right show the regret of IDS (with directed and undirected information gain) compared to the UCB algorithm. Note that UCB *never* chooses the informative actions and therefore suffers linear regret on the second variant.

using just a gradient oracle leads to a slower regret rate due to the local nature of the gradient feedback.

8.2.6 *Example: Invasive Measurements and Bayesian Quadrature*

We consider a simplistic setup where the experimenter wishes to align a rectangular plate with a laser in a way that maximizes the brightness on the probe. The design parameters are $(a^1, a^2) \in \mathcal{A} = [-1, 1]^2$ and correspond to a vertical and horizontal shift of the target relative to the origin. The intensity of the laser on the two-dimensional plane at the target is given by an initially unknown function $I : \mathbb{R}^2 \to \mathbb{R}^2$. The setup is illustrated in

Fig. 8.4. The objective is to maximize the total power on the probe of size 1×1 that is centered at (a^1, a^2),

$$f(a^1, a^2) = \int_{a^1-0.5}^{a^1+0.5} \int_{a^2-0.5}^{a^2+0.5} I(a^1, a^2) da^1 da^2 \,,$$

In any round, the learner has the choice between a *direct* measurement of the objective and the *invasive* measurement, that is more informative but yields no reward. For the direct measurement, the experimenter chooses a design $a_t = (a_t^1, a_t^2)$ and observes the corresponding integrated reward $y_t = f(a_t^1, a_t^2) + \epsilon_t$. This action has standard bandit feedback. Alternatively, the experimenter can replace the target plate with a high-resolution screen centered at (a^1, a^2), that measures the intensity directly on $m \times m$ pixels. Specifically, the learner obtains m^2 measurements $\{I(b_i^1, b_j^2)\}_{i,j=1}^m$, plausibly at a much lower noise level than the integrated intensity measurement. As the screen blocks off the beam, there is no reward in such rounds.

Returning to the previously mentioned example of tuning particle accelerators [92, 93], both integrated signals and invasive measurements are common feedback mechanisms, that are not always easy to integrate in automated tuning methods. In a *transductive* variant of the same setting, the signal is observed *only* through the invasive measurements, and this case, the game is no longer locally observable.

We present a numerical simulation of this setup in Figure 8.4. Our set $\mathcal{A}$ is discrete with 9 actions corresponding to a unit shift in any direction (or no shift). We use 25-dimensional features computed from a radial basis function kernel. In the setup where the reward signal can be observed directly, UCB outperforms IDS for the first ~ 1000 steps; but then IDS gains an advantage from choosing the more informative measurements from time to time. UCB has linear regret in the transductive setting because the UCB action is not informative. On the other hand, IDS trades off the informative measurements with greedy parameter settings that yield reward.

More generally, our example can be understood in the context of *Bayesian quadrature* [51, 123] and the related regret setting, which was previously studied by Toscano-Palmerin & Frazier [161]. Here, the goal is to optimize a design $i \in \mathcal{I}$ with the following objective,

$$f(i) = \int_{\mathcal{Z}} g(i, z) dz \,.$$

The function $g : \mathcal{I} \times \mathcal{Z} \to \mathbb{R}$ known, but the integral is expensive to evaluate. By extending the action space to $\mathcal{A} = \mathcal{I} \times \mathcal{Z}$, the learner gains

direct access to evaluations of $g(a, b)$. Assuming that g is a function in a RKHS over $\mathcal{A}$ with kernel evaluation features $k_{i,z} \in \mathcal{H}$, the feedback maps are set to bandit evaluations, $M_{i,z} = k_{i,z}$. The reward features are set to $\phi_i = \int_{\mathcal{B}} k_{i,z'} dz'$, independent of z. Several applications with integrated objective functions are summarized in [161].

8.3 CONTRIBUTIONS AND RELATED WORK

The contextual and kernelized partial monitoring settings presented in this chapter is based on the following publication:

- Kirschner, J., Lattimore, T. & Krause, A. *Information Directed Sampling for Linear Partial Monitoring* in *Proc. International Conference on Learning Theory (COLT)* (July 2020)

The presentation here simplifies previous definitions and analysis, and provides additional details in the kernelized setting.

In a basic form, the contextual bandit setting goes back to the work by Woodroofe [172], but the modern formulation is due to Langford & Zhang [100]. The only work we are ware of that analyses partial monitoring in a contextual setting is by Bartók & Szepesvári [19]. Their result is for *finite* partial monitoring, and requires a local observability condition. In contrast, our work uses the linear partial monitoring formulation, and the analysis covers all game categories.

The use of prior information was investigated by [167], but only the locally observable case was analyzed. To the best of our knowledge, kernelized methods have not yet been investigated in the partial monitoring setting.

The dueling bandit and robust regret minimization examples are in:

- Kirschner, J. & Krause, A. *Bias-Robust Bayesian Optimization via Dueling Bandits* in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)* (July 2021)

Kernelized dueling bandits have been studied in the literature [67, 154, 156], as well as extensions with multi-point comparisons [155]. The linear model has been studied recently by [140]. All previous work explicitly considers the binary feedback, although in some cases the extension to the more general sub-Gaussian likelihood is possible. To the best of our knowledge, none of the previous works provides bounds on the cumulative regret in the kernelized setting.

9

CONCLUSION

In this thesis, we have developed a frequentist theory of information-directed sampling. Our results show that the IDS principle yields practical and effective algorithms for regret minimization in a wide range of settings. The primal-dual interpretation establishes a link between IDS and the asymptotic regret lower bound. The connection provides new insights into the IDS exploration mechanism and can be used to design new algorithms. Moreover, we analyzed IDS in the linear partial monitoring framework, thereby making it applicable to many models beyond bandit feedback.

Compared to bandit algorithms, partial monitoring has received much less attention in the literature, likely due to the complexity that comes with the additional generality. On the other hand, IDS is straightforward to implement and does not directly use the finer geometric structure of partial monitoring, apart from constraints on the parameter set. We illustrated the setting with many example applications, hoping that the IDS framework can serve as starting point to address new applications and guide the analysis where needed. We conclude with a short comparison between the Bayesian and frequentist IDS frameworks and a list of open questions for the future.

9.1 BAYESIAN AND FREQUENTIST IDS FRAMEWORKS

The Bayesian and frequentist IDS frameworks are based on the same principle of sampling actions from a distribution that minimizes the ratio of squared expected regret and information gain. The difference is how the gap is estimated and the information gain is defined. The Bayesian learner has access to the prior distribution, which is used to define the gaps and the information gain. The frequentist IDS framework replaces both quantities with suitable worst-case notions derived from the confidence set. We briefly compare Bayesian and frequentist framework along the three dimensions *analysis*, *computation* and *empirical performance*.

ANALYSIS The Bayesian analysis uses Bayes' rule, thereby providing results for a wide range of prior distributions and observation likelihoods. The proof is based on standard tools from information theory, and is short

and elegant. The analysis bounds regret in expectation over the prior, and the worst-case regret bound scales with the prior entropy of the optimal action. A more robust analysis of the Bayesian IDS algorithm, such as frequentist regret or prior misspecification, is still an open problem. Also, so far, only finite action sets have been analyzed and instance-dependent bounds are not known. Some progress of analyzing Bayesian IDS for partial monitoring is by Lattimore & Szepesvári [105].

The frequentist analysis of IDS in this thesis is specialized to linear models and least-squares estimation. The sub-Gaussian tail assumption on the observation likelihood provides flexibility since it covers many distributions beyond the normal distribution. The proofs for the worst-case regret are relatively simple, and readily apply to continuous action sets and partial monitoring. On the other hand, the asymptotic analysis is more involved, and a suitable choice of the information gain function is less obvious. As it is the case for most frequentist bandit algorithms, adaptations beyond the Gaussian-linear case require to derive specialized confidence sets. The choice of information gain function is less clear for non-Gaussian settings.

COMPUTATION Any Bayesian learner faces the challenge of computing the posterior distribution, which is easy for conjugate priors, but rather costly in general. For the Gaussian-linear models that we analyzed, the posterior distribution of the parameter is not harder to compute than the least-squares estimate. However, the Bayesian IDS algorithm further requires computing the posterior distribution of the optimal action and the mutual information. In many cases, this requires the computation of high-dimensional integrals. Russo & Van Roy [135] show that sample-based approximations and surrogates of the information gain function are effective, partly alleviating the limitations above.

In the frequentist IDS version, the computation of gap estimates and information gain function only requires basic linear algebra operators. In most cases, this step is efficient for finite action sets, and not more expensive than computing upper-confidence bound scores. For continuous action sets, so far, only heuristic implementations are known. In both formulations, we need to sample from the IDS distribution, which for finite action sets of size k can be done in $\mathcal{O}(k^2)$ steps exactly, or approximately in $\mathcal{O}(k)$ steps.

PERFORMANCE The Bayesian IDS algorithm is reported to achieve superior performance on standard benchmarks [135], in many cases outper-

forming Thompson sampling, which is known for its excellent performance. The effectiveness of frequentist methods heavily depends on the tightness of the confidence bounds, which are often conservative in practice. Without tuning, frequentist IDS is often competitive with UCB and standard methods in partial monitoring, but not necessarily with Bayesian algorithms, including Thompson sampling and Bayesian IDS.

On a large horizon, the asymptotic guarantees for IDS are visible in numerical simulations, where it clearly outperforms UCB and Thompson sampling on some instances of the linear bandit problem. However, the regime transition depends inversely on the minimum gap squared, arguably pushing it out of reach in many practical applications.

9.2 OPEN QUESTIONS

We close with a list of exciting directions for the future. Naturally, our focus is on open questions within the IDS framework, but more generally, the exploration-exploitation trade-off in models with structured feedback is not yet fully understood.

9.2.1 *First-Principles Derivation*

We have seen two perspectives that motivate the IDS principle: The first is the worst-case upper bound that IDS optimizes greedily. The second is the asymptotic lower bound and the primal-dual interpretation. While in both cases, the analysis yields optimal or near-optimal regret bounds for the linear setting, neither leads to the definition of the information ratio from first-principles considerations. Therefore, it is natural to ask if the IDS algorithm is a consequence of implicit assumptions on the class of policies that minimize regret. For example, one can require that the sampling distribution is a function of sufficient statistics associated with the linear least-squares estimator. Provided that the policy is also worst-case and asymptotically optimal, it appears likely that one has to randomize in a way similar to IDS. Progress in this direction could provide a better understanding of the scope and limitations of the IDS principle.

9.2.2 *Asymptotic and Instance-Dependent Regret*

The asymptotic analysis can be generalized in several directions, including the more generic structured bandit setting, linear partial monitoring, and

the contextual case. While it seems plausible that similar techniques as presented here will be helpful along the way, new progress is most valuable if, at the same time, the proofs can be further simplified. Our asymptotic analysis follows a relatively clean and modular structure and raises the hope that there exists a *really* simple proof. Finding an information gain that preserves the guarantees and telescopes more easily could be a first step towards this end. Lastly, optimizing instance-dependent regret in finite time is largely an open problem.

9.2.3 *Partial Monitoring*

There are many open questions left in stochastic partial monitoring. First, the classification of linear partial monitoring with arbitrary constraints on the parameter set is not yet completed, since we only provided upper bounds. One can also ask to classify the regret rate on continuous action sets. There is an indication that such a result depends on finer properties of the action set, such as curvature [90]. Finding an oracle-efficient approximation of the IDS principle is another practically relevant question.

9.2.4 *Other Information Trade-Offs*

We derived the information gain functions for a sub-Gaussian noise likelihood, capturing a larger class of light-tailed distributions. We point out that the asymptotic information gain can be interpreted as a log-likelihood ratio test with alternatives $\nu_1, \dots, \nu_l \in \mathcal{M}$ and weights $q_t \in \mathscr{P}([l])$:

$$I_t^{\mathrm{LR}}(x) = \sum_{i=1}^{l} q_t(i) \mathbb{E}_y \left[\log \frac{\mathcal{L}(\hat{\theta}_t; x_t, y)}{\mathcal{L}(\nu_i; x_t, y)} \right].$$

This could serve as a starting point to derive an information gain that preserves the guarantees for a larger class of likelihood functions, for instance, heavy-tailed noise. There are also other information-regret trade-offs than cumulative regret minimization. For instance, sparsity of the linear parameter vector is an important assumption in high-dimensional settings and leads to complex information trade-offs [73]. Simple regret is another well-studied objective, where the learner strives to minimize the prediction error of the optimum with a minimal number of steps. Naturally, the learner can optimize the information gain directly, which indicates that information gain functions can be studied outside the cumulative regret minimization framework.

9.2.5 *Reinforcement Learning*

Departing from the stateless world of bandit feedback, in reinforcement learning, the learner's actions affect the options available in the future. Consequently, the learner faces a planning problem. Both upper-confidence bound algorithms and Thompson sampling have been proven useful in reinforcement learning, but information-theoretic tools are far less developed, with some progress made by Lu & Van Roy [112], Lu *et al.* [113], and Zanette & Sarkar [177], and by the author and collaborators [122]. From a theoretical perspective, we can view policy optimization in the episodic setting as a bandit problem. Each policy corresponds to a single action, and the feedback stems from the episodic role-out. Another possibility is to define the regret estimate and information gain for several steps into the future.

A

CONCENTRATION INEQUALITIES

We first state a consequence of Freedman's inequality.

Lemma A.1 ([81, Lemma 3])**.** *Let $X_1, \dots, X_n$ be a martingale difference sequence on a filtration $\mathcal{F}_t$ such that $X_t \le B$ holds for all $t = 1, \dots, n$. Denote the corresponding martingale by $M_n = \sum_{t=1}^n X_t$ and the sum of conditional variances by $V_n = \sum_{t=1}^n \mathbb{E}\left[X_t^2 | \mathcal{F}_{t-1}\right]$. Then, for any $\beta > 1$, $l = \left\lceil \frac{\log(n\lambda^{-2})}{\log \beta} \right\rceil$, $\lambda \ge 0$,*

$$\mathbb{P}[M_n \ge \lambda \max\{\lambda B, \sqrt{V_n}\}] \le (l+1) \exp\left(-\frac{\lambda^2}{2\beta + \frac{2}{3}}\right).$$

Moreover, with probability at least $1 - \delta$,

$$M_n \le \max\left\{4B \log \frac{2n+2}{\delta}, 2\sqrt{V_n \log \frac{2n+2}{\delta}}\right\}.$$

Proof. Our proof is a refined version of [81, Lemma 3]. Define $l = \left\lceil \frac{\log n\lambda^{-2}}{\log \beta} \right\rceil$, and $\alpha_i = \lambda^2 B^2 \beta^i$ for $i = 0, \dots, l$, further set $\alpha_{-1} = 0$ for notational convenience. Note that since $V_n \le nB^2$, our choice of l implies $\alpha_l \ge V_n$. Then,

$$\begin{aligned}
&\mathbb{P}[M_n \ge \lambda \max\{\lambda B, \sqrt{V_n}\}] \\
&= \mathbb{P}[M_n \ge \lambda \max\{\lambda B, \sqrt{V_n}\},\ \alpha_{i-1} \le V_n \le \alpha_i \text{ for } i = 0, \dots, l] \\
&\le \sum_{i=0}^{l} \mathbb{P}[M_n \ge \lambda \max\{\lambda B, \sqrt{V_n}\},\ \alpha_{i-1} \le V_n \le \alpha_i] \\
&\le \sum_{i=0}^{l} \mathbb{P}[M_n \ge \lambda \max\{\lambda B, \sqrt{\alpha_{i-1}}\},\ V_n \le \alpha_i] \\
&\overset{\text{(i)}}{\le} \sum_{i=0}^{l} \exp\left(-\frac{\lambda^2 \max\{\lambda^2 B^2, \alpha_{i-1}\}}{2\alpha_i + \frac{2}{3} B\lambda \max\{\lambda b, \sqrt{\alpha_{i-1}}\}}\right) \\
&= \exp\left(-\frac{\lambda^2}{2 + \frac{2}{3}}\right) + \sum_{i=1}^{l} \exp\left(-\frac{\lambda^2}{2\beta + \frac{2}{3}\sqrt{\frac{1}{\beta^{i-1}}}}\right) \\
&\overset{\text{(ii)}}{\le} (l+1) \exp\left(-\frac{\lambda^2}{2\beta + \frac{2}{3}}\right).
\end{aligned}$$

Here, (i) is Freedman's inequality [61], and (ii) uses that $\beta > 1$. This shows the first part of the lemma. For the second part, note that for $n = 1$, the claim is trivially true by $X_1 \leq B$, hence we assume $n \geq 2$. If also $\lambda \geq 1$, we can upper bound $l \leq \left\lceil \frac{\log n}{\log \beta} \right\rceil =: l'$. We choose $\beta = 5/3$, and set $\lambda = \sqrt{\log \frac{l'+1}{\delta}(2\beta + \frac{2}{3})} = 2\sqrt{\log \frac{l'+1}{\delta}}$, such that indeed $\lambda \geq 2\sqrt{\log\left(\frac{\log(2)}{\log(5/3)} + 1\right)} \geq 1$ for $n \geq 2$, hence proving the claim. □

The next concentration inequality for supermartingales is an anytime variant of Corollary 2.7 by [55], and might be of independent interest.

Theorem A.1. *Let $M_n = \sum_{t=1}^n X_t$ be a sum of supermartingale differences X_t on a filtration $\mathcal{F}_t$. Further, let U_t be a non-negative predictable process, such that $X_t \leq U_t$ holds for all $t \geq 1$. Define*

$$C_t^2 = \begin{cases} \mathbb{E}\left[X_t^2 | \mathcal{F}_{t-1}\right], & \text{if } \mathbb{E}\left[X_t^2 | \mathcal{F}_{t-1}\right] \geq U_t^2\,, \\ \frac{1}{4}\left(U_t + \frac{\mathbb{E}\left[X_t^2 | \mathcal{F}_{t-1}\right]}{U_t}\right)^2 & \text{otherwise.} \end{cases}$$

Further denote $A_n = \sum_{t=1}^n C_t^2$. Then, for any fixed positive sequence $(L_t)_{t=1}^n$, with probability at least $1 - \delta$,

$$\forall\, n \geq 1, \quad M_n \leq \sqrt{2(A_n + L_n)\log\left(\frac{1}{\delta}\frac{(A_n + L_n)^{1/2}}{L_n^{1/2}}\right)}\,.$$

The proof combines Corollary 2.7 in [55], with the *method of mixtures* [127], see also [3, Theorem 1]. We start with the following lemma.

Lemma A.2. *Let X_t, C_t as in Theorem A.1, and define for $\lambda \geq 0$, $t \geq 1$,*

$$M_t^\lambda = \exp\left(\sum_{s=1}^t \lambda X_s - \frac{\lambda^2}{2} C_s^2\right). \tag{A.1}$$

Further, let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}$. Then M_t^λ is a supermartingale, M_τ^λ is almost surely well-defined, and $\mathbb{E}\left[M_\tau^\lambda\right] \leq 1$.

Proof. The proof is along the lines of Lemma 8 in [3], where we replace the subgaussian condition by the suitable analog to showing that M_t^λ is a supermartingale. Let

$$D_s^\lambda = \exp\left(\lambda X_s - \frac{\lambda}{2} C_s^2\right).$$

By Corollary 2.6 of [55], we have that $\mathbb{E}\left[e^{\lambda X_s}|\mathcal{F}_{s-1}\right] \leq \exp\left(\frac{\lambda^2}{2}C_s^2\right)$ for all $\lambda > 0$, and consequently, $\mathbb{E}[D_s|\mathcal{F}_{s-1}] \leq 1$. Therefore, $\mathbb{E}\left[M_t^\lambda|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[D_t^\lambda|\mathcal{F}_{t-1}\right]M_{t-1}^\lambda \leq M_{t-1}^\lambda$, which shows that M_t^λ is a supermartingale such that $\mathbb{E}\left[M_t^\lambda\right] \leq 1$ for all $t \geq 1$. The rest of the argument is standard [53, Theorem 5.7.6]; by the convergence theorem for non-negative supermartingales, M_τ^λ is well defined for any stopping time $\tau \leq \infty$, then using Fatou's lemma it follows that $M_\tau^\lambda \leq \liminf_{t\to\infty} M_{\tau\wedge t}^\lambda \leq 1$. □

To prove Theorem A.1, we use the *method of mixtures*, similar to Theorem 1 in [3]. The main difference is that the supermartingale M_t^λ from the the previous lemma is only defined for $\lambda \geq 0$, which requires to choose a mixing density supported on $[0,\infty)$.

Proof of Theorem A.1. Remember that $S_t = \sum_{s=1}^t X_s$, $A_t = \sum_{s=1}^t C_s^2$ and $M_t^\lambda = \exp\left(\lambda S_t - \frac{\lambda^2}{2}A_t\right)$. Further, let $\Lambda = (\Lambda_t)_{t\geq 1}$ be a sequence of independent Gaussian random variable truncated to $[0,\infty)$ with densities $f_{\Lambda_t}(\lambda) = c(L_t)\exp\left(-\frac{1}{2}\lambda^2 L_t\right)\mathbb{1}\{\lambda \geq 0\}$ where $c(A) = \sqrt{\frac{2A}{\pi}}$ is a normalizing constant. Using Λ as a *mixing distribution* we define

$$M_t = \mathbb{E}\left[M_t^{\Lambda_t}|\mathcal{F}_\infty\right], \tag{A.2}$$

where $\mathcal{F}_\infty = \sigma\left(\cup_{t=1}^\infty \mathcal{F}_t\right)$ is the tail σ-algebra of the filtration $\mathcal{F}_t$. In particular, using Fubini's theorem, we still get $\mathbb{E}[M_\tau] = \mathbb{E}\left[\mathbb{E}\left[M_\tau^{\Lambda_t}|\Lambda\right]\right] \leq 1$. In the next step, we explicitly calculate M_t for any $t \geq 1$,

$$\begin{aligned}
M_t &= \int_{\mathbb{R}^+} \exp\left(\lambda S_t - \tfrac{\lambda^2}{2}A_t\right) f_{\Lambda_t}(\lambda)\, d\lambda \\
&= \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left(\lambda - \frac{S_t}{A_t}\right)^2 A_t + \frac{1}{2}\frac{S_t^2}{A_t}\right) f_{\Lambda_t}(\lambda)\, d\lambda \\
&= \exp\left(\frac{1}{2}\frac{S_t^2}{A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left(\lambda - \frac{S_t}{A_t}\right)^2 A_t\right) f_{\Lambda_t}(\lambda)\, d\lambda \\
&= c(L_t)\exp\left(\frac{1}{2}\frac{S_t^2}{A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left((\lambda - S_t/A_t)^2 A_t + \lambda^2 L_t\right)\right) d\lambda\,.
\end{aligned}$$

Completing the square yields

$$\left(\lambda - \frac{S_t}{A_t}\right)^2 A_t + \lambda^2 A = \left(\lambda - \frac{S_t}{L_t + A_t}\right)^2 (L_t + A_t) + \frac{S_t^2}{A_t} - \frac{S_t^2}{L_t + A_t}\,,$$

and with the previous equation,

$$
\begin{aligned}
M_t &= c(L_t)\exp\left(\tfrac{1}{2}\tfrac{S_t^2}{L_t+A_t}\right)\int_{\mathbb{R}^+}\exp\left(-\tfrac{1}{2}\left(\lambda-\tfrac{S_t}{L_t+A_t}\right)^2(L_t+A_t)\right)d\lambda \\
&\overset{\text{(i)}}{\geq} \mathbb{1}(S_t\geq 0)c(L_t)\exp\left(\tfrac{1}{2}\tfrac{S_t^2}{L_t+A_t}\right)\int_{\mathbb{R}^+}\exp\left(-\tfrac{1}{2}\left(\lambda-\tfrac{S_t}{L_t+A_t}\right)^2(L_t+A_t)\right)d\lambda \\
&\overset{\text{(ii)}}{\geq} \mathbb{1}(S_t\geq 0)c(L_t)\exp\left(\tfrac{1}{2}\tfrac{S_t^2}{L_t+A_t}\right)\int_{\mathbb{R}^+}\exp\left(-\tfrac{1}{2}\left(\lambda^2(L_t+A_t)\right)\right)d\lambda \\
&= \mathbb{1}(S_t\geq 0)\tfrac{c(L_t)}{c(L_t+A_t)}\exp\left(\tfrac{1}{2}\tfrac{S_t^2}{L_t+A_t}\right).
\end{aligned}
$$

In (i) we introduced an indicator function and used that all other terms are positive. To get (ii), we first applied a change of variables $\lambda' = \lambda - S_t/(L_t + A_t)$ and then made use of $S_t \geq 0$ to reduce the integration range (and again that the integrand is positive).

A final application of Markov's inequality yields

$$
\begin{aligned}
&\mathbb{P}\left[S_\tau \geq \sqrt{2(L_\tau+A_\tau)\log\left(\frac{1}{\delta}\frac{(L_\tau+A_\tau)^{1/2}}{L_\tau^{1/2}}\right)}\right] \\
&= \mathbb{P}\left[\frac{c(L_\tau)}{c(L_\tau+A_\tau)}\exp\left(\frac{1}{2}\frac{S_\tau^2}{L_\tau+A_\tau}\right)\geq\frac{1}{\delta}, S_\tau\geq 0\right] \\
&\leq \delta\cdot\mathbb{E}\left[\mathbb{1}\{S_\tau\geq 0\}\frac{c(L_\tau)}{c(L_\tau+A_\tau)}\exp\left(\frac{1}{2}\frac{S_\tau^2}{L_\tau+A_\tau}\right)\right] \\
&\overset{\text{(i)}}{\leq} \delta\cdot\mathbb{E}[M_\tau]\overset{\text{(ii)}}{\leq}\delta,
\end{aligned}
$$

where (i) uses the inequality for M_t derived above, and (ii) follows from Lemma A.2.

To get the anytime result as stated in the Theorem, we use the same argument as in [3] on the stopping time

$$
\tau = \min\left\{t\geq 1 \mid S_t \geq \sqrt{2(L_t+A_t)\log\left(\frac{1}{\delta}\frac{(L_t+A_t)^{1/2}}{L_t^{1/2}}\right)}\right\}.
$$

Expressing the quantity of interest in terms of τ, and applying the previous inequality yields

$$\begin{aligned}
&\mathbb{P}\left[S_t \geq \sqrt{2(L_t + A_t)\log\left(\frac{1}{\delta}\frac{(L_t + A_t)^{1/2}}{L_t^{1/2}}\right)} \text{ for any } t \geq 1\right] \\
&= \mathbb{P}\left[\tau < \infty,\, S_\tau \geq \sqrt{2(L_\tau + A_\tau)\log\left(\frac{1}{\delta}\frac{(L_\tau + A_\tau)^{1/2}}{L_\tau^{1/2}}\right)}\right] \\
&\leq \mathbb{P}\left[S_\tau \geq \sqrt{2(L_\tau + A_\tau)\log\left(\frac{1}{\delta}\frac{(L_\tau + A_\tau)^{1/2}}{L_\tau^{1/2}}\right)}\right] \\
&\leq \delta\,.
\end{aligned}$$

This completes the proof. □

As a consequence of the previous result, we have the following lemma.

Lemma A.3. *Let X_t be a non-negative stochastic process adapted to a filtration $\mathcal{F}_t$. Assume that $X_t \leq B_t$ for a fixed, non-decreasing sequence $(B_t)_{t=1}^\infty$. Define $M_n = \sum_{t=1}^n X_t$ and $\bar{M}_n = \sum_{t=1}^n \mathbb{E}[X_t|\mathcal{F}_{t-1}]$, and let $(L_t)_{t=1}^\infty$ be any fixed, positive sequence. Then, with probability at least $1-\delta$,*

$$\forall\, n \geq 1, \quad \bar{M}_n - M_n \leq \sqrt{2(B_n\bar{M}_n + L_n)\log\left(\frac{1}{\delta}\frac{(B_n\bar{M}_n + L_n)^{1/2}}{L_n^{1/2}}\right)}\,.$$

Further, if $B_n \geq 1$, with probability at least $1-\delta$ for any $n \geq 1$ it holds that,

$$\begin{aligned}
\bar{M}_n &\leq 2M_n + 2\sqrt{4B_n^2\log\left(\frac{4B_n}{\delta^2}\right)\log\left(\frac{4B_n}{\delta^2}\log\left(\frac{4B_n}{\delta^2}\right)\right)} + e \\
&\leq 2M_n + \mathcal{O}\Big(B\log(\tfrac{B}{\delta})\Big)
\end{aligned}$$

Proof. Clearly, $\xi_t = \bar{X}_t - X_t$ is a martingale difference sequence such that $\xi_t \leq \bar{X}_t$ and $\bar{X}_t$ is a predictable process. Hence Theorem A.1 applies with

$$C_t^2 \triangleq \begin{cases} \mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right], & \text{if } \mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right] \geq \bar{X}_t^2\,, \\ \frac{1}{4}\left(\bar{X}_t + \frac{\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]}{\bar{X}_t}\right)^2 & \text{otherwise.} \end{cases}$$

In particular $C_t^2 \leq B_t\bar{X}_t$. To see this, note that the Bhatia-Davis inequality (Lemma A.4) implies $\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right] = \mathrm{Var}(X_t|\mathcal{F}_{t-1}) \leq \bar{X}_t(B_t - \bar{X}_t) \leq \bar{X}_t B_t$.

Further, if $\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right] \leq \bar{X}_t^2$, $\frac{1}{4}\left(\bar{X}_t + \frac{\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]}{\bar{X}_t}\right)^2 \leq \bar{X}_t^2 \leq \bar{X}_t B_t$. Consequently, $A_T = \sum_{t=1}^T C_t^2 \leq B_n \sum_{t=1}^T \bar{X}_t = B_n \bar{M}_n$. A direct application of Theorem A.1 shows the first inequality.

For the second claim, we set $L_t = eB_t$ where $e = \exp(1)$. With this choice,

$$\bar{M}_n - M_n \leq \sqrt{(B_n(\bar{M}_n + e)\log\left(\frac{\bar{M}_n + e}{\delta^2}\right)}.$$

We substitute $x = \bar{M}_n + e$. The claim follows by rearranging if we show that

$$\sqrt{B_n x \log\left(\frac{x}{\delta^2}\right)} \leq \frac{x}{2} + \sqrt{4B_n^2 \log\left(\frac{4B_n}{\delta^2}\right)\log\left(\frac{4B_n}{\delta^2}\log\left(\frac{4B_n}{\delta^2}\right)\right)}.$$

Clearly, the left side of the inequality is a concave function in x, hence it suffices to show that the derivative is smaller than $\frac{1}{2}$ for all $x \geq 4B_n \log(\frac{4B_n}{\delta^2})$. The derivative is

$$\frac{d}{dx}\sqrt{B_n x \log\left(\frac{x}{\delta^2}\right)} = \frac{\sqrt{B_n}\left(\log(\frac{x}{\delta^2}) + 1\right)}{2\sqrt{x\log(\frac{x}{\delta^2})}} \leq \sqrt{\frac{B_n \log(\frac{x}{\delta^2})}{x}}.$$

The inequality uses that $x \geq e$. Hence all we need to show is that

$$4B_n \log\left(\tfrac{x}{\delta^2}\right) \leq x\,,$$

which is true for $x \geq 4B_n \log(\frac{4B_n}{\delta^2})$.

□

The following lemma bounds the variance of a random variable supported on an interval.

Lemma A.4 (Bhatia & Davis [23])**.** *Let X be a real random variable supported in $[m, M]$. Then,*

$$\mathrm{Var}(X) \leq (M - \mathbb{E}[X])(\mathbb{E}[X] - m)\,,$$

and the bound is tight, if all mass is concentrated on the end-points of the interval.

B

ASYMPTOTIC INFORMATION GAIN: PROOFS

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

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

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

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

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

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

The equality follows from Lemma B.4. Next, we derive an upper bound to the term (B).

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

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

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

$$\begin{aligned}&|[L_s - L_{s+1}](a)|\\ &\leq 4|\epsilon|^2\|a_s\|_{V_s^{-1}}^2 + 2|\langle \hat{v}_s - \hat{\theta}_s, a_s\rangle||\epsilon_s| + 8\beta_s\|a_s\|_{V_s^{-1}}^2 + \langle \hat{v}_s - \hat{\theta}_s, a_s\rangle^2\end{aligned}$$

Proof. For one direction, we can reuse Lemma B.1,

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

For the other direction, we have

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

where for the last step we denote $u_s = \frac{\epsilon_s + \langle a_s, \theta - \hat{\theta}_s\rangle}{1 + \|a_s\|_{V_s^{-1}}^2}$ and use Lemma B.4. Further unwrapping the square gives

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

Combining both directions yields the claim. □

Lemma B.3. *Let s such that $24^2\eta_s\beta_s\|a_s\|^2_{V_s^{-1}} \le 1$ and $\beta_s\|a_s\|^2_{V_s^{-1}} \le 1$. Then*

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

Proof.

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

For the last step we used $(x+y+z)^i \le (3x)^i + (3y)^i + (3z)^i$ for $x,y,z \ge 0$. Further, ρ-subgaussian noise ϵ_s satisfies $\mathbb{E}\big[|\epsilon|^i\big] \le (2\rho^2)^{i/2}i\Gamma(i/2) \le (2\rho^2)^i i!$ and $\mathbb{E}\big[|\epsilon|^2 i\big] \le (2\rho^2)^i 2i!$ for all $i \in \mathbb{N}$ [Lemma 1.4, 131]. Hence,

$$\begin{aligned}
&\mathbb{E}_s\left[\frac{|\eta_s(L_{s+1}(x)-L_s(x))|^i}{i!}\right] \\
&\le \mathbb{E}_s\left[\frac{(12\eta_s|\epsilon|^2\|a_s\|^2_{V_s^{-1}})^i}{i!}\right] + \mathbb{E}_s\left[\frac{(6\eta_s|\langle\hat{v}_s-\hat{\theta}_s,a_s\rangle|\epsilon_s|)^i}{i!}\right] \\
&\quad + \frac{(24\eta_s\beta_s\|a_s\|^2_{V_s^{-1}} + 3\eta_s\langle\hat{v}_s-\hat{\theta}_s,a_s\rangle^2)^i}{i!}
\end{aligned}$$

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

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

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

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

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

Lemma B.4. *The one-step update to the least-squares estimator with data $y_s = \langle a_s, \theta\rangle + \epsilon_s$ is*

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

Proof. The difference can be computed with the Sherman-Morrison formula (Lemma D.2),

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

□

C

PARTIAL MONITORING: LOWER BOUNDS

The lower bounds complete the classification theorem. These results are almost implied by existing theorems from finite partial monitoring. The only difference is that here the outcome space is infinite, which does not change the structure of the proofs. We include here the key details and intuition. As expected, the key tool is Le Cam's method in combination with the Bretagnolle–Huber inequality [28] and an elementary calculation of the relative entropy between measures on interaction sequences induced by a fixed policy and for different environments. For the remainder of this section, we fix an arbitrary policy and finite game with actions $\mathcal{A}$ and feedback functions $(M_a)_{a\in\mathcal{A}}$. For simplicity, we assume the noise is Gaussian and $\mathcal{A}$ spans $\mathbb{R}^d$. Given a $\theta \in \mathbb{R}^d$ let $\mathbb{P}_\theta^n$ be the measure on action/observation sequences of length n when the learner interacts with the game for parameter θ. Before the theorems and proofs we need a little more notation. Let

$$V_n(\theta) = \mathbb{E}_\theta[V_n] = \mathbb{E}_\theta\left[\sum_{t=1}^n M_{a_t} M_{a_t}^\top\right].$$

Then define $E_n(\theta)$ as the binary random variable that the algorithm plays a suboptimal action at least $n/2$ times.

$$E_n(\theta) = \mathbb{1}\left(\sum_{t=1}^n \mathbb{1}(a_t \notin \mathcal{P}(\theta)) \geq n/2\right).$$

Notice that if θ, θ' are such that $\mathcal{P}(\theta) \cap \mathcal{P}(\theta') = \emptyset$, then $E_n(\theta') \geq 1 - E_n(\theta)$. In the following we let $\mathfrak{R}_n(\theta)$ be the expected regret of a learner that interacts with the environment determined by $\theta \in \mathcal{M}$, without specifying the policy.

Lemma C.1. *The relative entropy between $\mathbb{P}_\theta^n$ and $\mathbb{P}_{\theta'}^n$ satisfies*

$$D_{KL}(\mathbb{P}_\theta^n \| \mathbb{P}_{\theta'}^n) = \frac{1}{2}\left\|\theta - \theta'\right\|_{V_n(\theta)}^2.$$

For a proof refer to [103, Theorem 24.1].

Lemma C.2. *(Bretagnolle-Huber inequality) Let P and Q be probability measures on the same measurable space (F, Ω) and let $A \in F$ be an arbitrary event. Then*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D_{KL}(P\|Q)) .$$

Theorem C.1. *Suppose that* $\operatorname{span}(M_a : a \in \mathcal{A}) \neq \mathbb{R}^d$*, then there exists a game-dependent constant $C > 0$ such that for all $n \geq 1$ there exists a θ for which $\mathfrak{R}_n(\theta) \geq Cn$.*

Proof. Let $\theta \in \mathbb{R}^d$ be a non-zero vector such that $M_a\theta = 0$ for all $a \in \mathcal{A}$, which exists by the assumption that $\operatorname{span}(M_a : a \in \mathcal{A}) \neq \mathbb{R}^d$. Next, let $\theta' = -\theta$ and notice that by Lemma C.1,

$$D_{\mathrm{KL}}(\mathbb{P}_\theta^n \| \mathbb{P}_{\theta'}^n) = 0 .$$

By our choice, the optimal action for the environment determined by θ and θ' are different: $\mathcal{P}(\theta) \cap \mathcal{P}(\theta') = \emptyset$. The Bretagnolle-Huber inequality (Lemma C.2) implies that

$$\begin{aligned} \mathbb{P}_\theta^n(E_n(\theta)) + \mathbb{P}_{\theta'}^n(E_n(\theta')) &\geq \mathbb{P}_\theta^n(E_n(\theta)) + \mathbb{P}_{\theta'}^n(1 - E_n(\theta)) \\ &\geq \frac{1}{2} \exp(-D_{\mathrm{KL}}((\|\mathbb{P})_\theta^n, \mathbb{P}_{\theta'}^n)) \geq \frac{1}{2} . \end{aligned} \tag{C.1}$$

Furthermore, there exists an $\epsilon > 0$ such that $\mathfrak{R}_n(\theta) \geq \epsilon n \mathbb{P}_\theta^n(E_n(\theta))/2$. Hence, by Eq. (C.1), the regret is linear for either environment θ or θ'. □

The key lemma for proving the lower bound for globally observable games shows that in games that are not locally observable, there exists a pair of neighbouring Pareto optimal actions a, b and a parameter θ such that both actions a, b are optimal, but $\langle \phi_a - \phi_b, \theta \rangle$ can *not* be estimated by playing only actions from the neighborhood $\mathcal{N}_{ab}$.

Lemma C.3. *Suppose a game is not locally observable. Then there exists a pair a, b of neighbouring Pareto optimal actions and $\theta \in \operatorname{relint}(\mathcal{C}_a \cap \mathcal{C}_b)$ such that $\phi_a - \phi_b \notin \operatorname{span}\{M_c : c \in \mathcal{N}_{ab}\}$.*

Proof. The lemma follows from the definition of local observability its equivalent characterization provided in Lemma 6.6. □

Theorem C.2. *Suppose the game is globally observable, but not locally observable. Then there exists a game-dependent constant $C > 0$ and $\theta \in \mathbb{R}^d$ such that the regret is $\mathfrak{R}_n(\theta) \geq Cn^{2/3}$.*

Proof. By Lemma C.3, there exists a pair of neighboring Pareto optimal actions $a, b \in \mathrm{ext}(\mathrm{conv}(\mathcal{A}))$ and $\theta \in \mathrm{relint}(\mathcal{C}_a \cap \mathcal{C}_b)$ such that $\phi_a - \phi_b \notin \mathrm{span}(\{M_c : c \in \mathcal{P}(\theta)\}) = L$. Let $\phi_a - \phi_b = u + v$, where $u \in L$ and $v \in L^\perp$. Since $\phi_a - \phi_b \notin L$ it follows that

$$\langle \phi_a - \phi_b, v \rangle = \langle u + v, v \rangle = \|v\|^2 > 0\,.$$

In particular, for suitably small $\epsilon > 0$ it holds that $\theta + \epsilon v \in \mathcal{C}_a$ and $\theta - \epsilon v \in \mathcal{C}_b$. Define

$$\theta_n = \theta + n^{-1/3} v \quad \text{and} \quad \theta'_n = \theta - n^{-1/3} v$$

and let assume n is sufficiently large that $\theta_n \in \mathcal{C}_a$ and $\theta'_n \in \mathcal{C}_b$. Next, decompose $V_n(\theta)$ as $V_n(\theta) = U_n(\theta) + W_n(\theta)$, where

$$U_n(\theta) \triangleq \mathbb{E}_\theta \left[\sum_{t=1}^n \mathbb{1}(a_t \in \mathcal{P}(\theta)) M_{a_t} M_{a_t}^\top \right] \quad \text{and}$$
$$W_n(\theta) \triangleq \mathbb{E}_\theta \left[\sum_{t=1}^n \mathbb{1}(a_t \notin \mathcal{P}(\theta)) M_{a_t} M_{a_t}^\top \right].$$

Let $T_n(\mathcal{Y}) = \sum_{t=1}^n \mathbb{1}(a_t \in \mathcal{Y})$ be the number of times an action in $\mathcal{Y} \subset \mathcal{A}$ is played. Notice, since $v \in L^\perp$, that

$$\begin{aligned}
\frac{1}{2}\|\theta_n - \theta'_n\|^2_{V_n(\theta_n)} &= 2n^{-2/3}\|v\|^2_{V_n(\theta_n)} \\
&= 2n^{-2/3}\|v\|^2_{W_n(\theta_n)} \le 2n^{-2/3}\mathbb{E}_{\theta_n}[T_n(\mathcal{P}(\theta)^c)]\|v\|^2_G\,,
\end{aligned}$$

where $G = \sum_{a \in \mathcal{A}} M_a M_a^\top$. Now, there exists a game-dependent constant $\epsilon > 0$ such that

$$\mathfrak{R}_n(\theta_n) \ge \epsilon \mathbb{E}_{\theta_n}[T_n(\mathcal{P}(\theta)^c)]\,.$$

Hence if $\mathbb{E}_{\theta_n}[T_n(\mathcal{P}(\theta)^c)] \ge n^{2/3}$, then $\mathfrak{R}_n(\theta_n) \ge \epsilon n^{2/3}$. Assume that $\mathbb{E}_{\theta_n}[T_n(\mathcal{P}(\theta)^c)] \le n^{2/3}$. By the Bretagnolle-Huber inequality (Lemma C.2), there exists another game-dependent constant $\epsilon' > 0$ such that

$$\begin{aligned}
\mathfrak{R}_n(\theta_n) + \mathfrak{R}_n(\theta'_n) &\ge n^{2/3}\epsilon' \exp\Big(-2n^{-2/3}\mathbb{E}_{\theta_n}[T_n(\mathcal{P}(\theta)^c)]\|v\|^2_G\Big) \\
&\ge n^{2/3}\epsilon' \exp(-2\|v\|^2_G)\,.
\end{aligned}$$

Combining the last two displays completes the proof. □

Theorem C.3. *Suppose the game is locally observable, then there exists a constant $C > 0$ such that for all n there is a θ for which $\mathfrak{R}_n(\theta) \geq Cn^{1/2}$.*

Proof. Let $\theta \in \mathbb{R}^d$ be arbitrary and $\theta_n = n^{-1/2}\theta$ and $\theta'_n = -\theta_n$. By the assumption that $\mathcal{A}$ spans $\mathbb{R}^d$, it follows that $\mathcal{P}(\theta_n) \cap \mathcal{P}(\theta'_n) = \emptyset$. By Lemma C.1,

$$D_{\mathrm{KL}}(\mathbb{P}^n_{\theta_n} \| \mathbb{P}^n_{\theta'_n}) = \frac{1}{2}\|\theta'_n - \theta_n\|^2_{V_n(\theta_n)} = \frac{1}{2}\|\theta\|^2_{V_n(\theta_n)/n}\,.$$

Clearly, $G = \sum_{a \in \mathcal{A}} M_a M_a^\top \succ V_n(\theta_n)/n$. Hence, there exists a constant $C > 0$ such that for all $n \geq 1$,

$$D_{\mathrm{KL}}(\mathbb{P}^n_{\theta_n} \| \mathbb{P}^n_{\theta'_n}) \leq C\,.$$

Then, using the same argument as in the proof of Theorem C.1, we have

$$\begin{aligned}\mathbb{P}^n_{\theta_n}(E_n(\theta_n)) + \mathbb{P}^n_{\theta'_n}(E_n(\theta'_n)) &\geq \mathbb{P}^n_{\theta_n}(E_n(\theta_n)) + \mathbb{P}^n_{\theta'_n}(1 - E_n(\theta_n)) \\ &\geq \frac{1}{2}\exp(-C)\,.\end{aligned}$$

The result follows because there exists an $\epsilon > 0$ such that $\mathfrak{R}_n(\theta) \geq \mathbb{P}^n_\theta(E_n(\theta))\epsilon\sqrt{n}/2$. □

D

LINEAR ALGEBRA

Standard results from linear algebra are collected here for reference.

Lemma D.1 (Matrix determinant lemma)**.** *Let $V \in \mathbb{R}^{d\times d}$ be an invertible matrix and $a, b \in \mathbb{R}^d$. Then*

$$\det(V + ab^\top) = (1 + aV^{-1}b)\det(V)\,.$$

More generally, for let $A, B \in \mathbb{R}^{d\times m}$,

$$\det(V + AB^\top) = \det(\mathbf{1}_m + A^\top V^{-1}B)\det(V)\,.$$

Lemma D.2 (Sherman-Morrison-Woodbury formula [147, 171])**.** *Let $V \in \mathbb{R}^{d\times d}$ be an invertible matrix and $A \in \mathbb{R}^{d\times m}$,*

$$(V + AA^\top)^{-1} = V^{-1} - V^{-1}A(\mathbf{1}_m + A^\top V^{-1}A)^{-1}A^\top V^{-1}\,.$$

BIBLIOGRAPHY

1. Abbasi-Yadkori, Y. *Online Learning for Linearly Parametrized Control Problems* PhD thesis (2012).
2. Abbasi-Yadkori, Y., Antos, A. & Szepesvári, C. *Forced-exploration based algorithms for playing in stochastic linear bandits* in *COLT Workshop on On-line Learning with Limited Feedback* (2009).
3. Abbasi-Yadkori, Y., Pál, D. & Szepesvári, C. *Improved algorithms for linear stochastic bandits* in *Advances in Neural Information Processing Systems* (2011), 2312.
4. Abe, N. & Long, P. M. *Associative Reinforcement Learning Using Linear Probabilistic Concepts* in *Proceedings of the Sixteenth International Conference on Machine Learning* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999), 3.
5. Abeille, M. & Lazaric, A. *Linear Thompson Sampling Revisited* in *Artificial Intelligence and Statistics* (2017), 176.
6. Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., Sen, S. & Slivkins, A. Making Contextual Decisions with Low Technical Debt. *arXiv: Learning* (2016).
7. Agrawal, S. & Goyal, N. *Thompson sampling for contextual bandits with linear payoffs* in *International Conference on Machine Learning* (2013), 127.
8. Aitken, A. C. IV.—On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* **55**, 42 (1936).
9. Antos, A., Bartók, G., Pál, D. & Szepesvári, C. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science* **473**, 77 (2013).
10. Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society* **68**, 337 (1950).
11. Arora, S., Hazan, E. & Kale, S. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing* **8**, 121 (2012).

12. Audibert, J.-Y. & Bubeck, S. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research* **11**, 2785 (2010).

13. Auer, P. Using Confidence Bounds for Exploitation-exploration Tradeoffs. *J. Mach. Learn. Res.* **3**, 397 (2003).

14. Auer, P., Cesa-Bianchi, N. & Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning* **47**, 235 (2002).

15. Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R. E. *Gambling in a rigged casino: The adversarial multi-armed bandit problem* in *Proceedings of IEEE 36th Annual Foundations of Computer Science* (1995), 322.

16. Bartók, G., Foster, D. P., Pál, D., Rakhlin, A. & Szepesvári, C. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research* **39**, 967 (2014).

17. Bartók, G., Pál, D. & Szepesvári, C. *Minimax regret of finite partial-monitoring games in stochastic environments* in *Proceedings of the 24th Annual Conference on Learning Theory* (2011), 133.

18. Bartók, G., Zolghadr, N. & Szepesvári, C. *An Adaptive Algorithm for Finite Stochastic Partial Monitoring* in *Proceedings of the 29th International Coference on International Conference on Machine Learning* (Omnipress, Edinburgh, Scotland, 2012), 1779.

19. Bartók, G. & Szepesvári, C. *Partial monitoring with side information* in *International Conference on Algorithmic Learning Theory* (2012), 305.

20. Bastani, H., Bayati, M. & Khosravi, K. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011* (2017).

21. Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* **47**, 253 (2013).

22. Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A. & Hüllermeier, E. Preference-based Online Learning with Dueling Bandits: A Survey. *Journal of Machine Learning Research* **22**, 1 (2021).

23. Bhatia, R. & Davis, C. A Better Bound on the Variance. *The American Mathematical Monthly* **107**, 353 (2000).

24. Bogunovic, I., Losalka, A., Krause, A. & Scarlett, J. Stochastic linear bandits robust to adversarial attacks. *arXiv preprint arXiv:2007.03285* (2020).

25. Bogunovic, I., Scarlett, J., Jegelka, S. & Cevher, V. Adversarially robust optimization with gaussian processes. *arXiv preprint arXiv:1810.10775* (2018).

26. Bouneffouf, D., Rish, I. & Aggarwal, C. *Survey on Applications of Multi-Armed and Contextual Bandits* in *2020 IEEE Congress on Evolutionary Computation (CEC)* (2020), 1.

27. Boyd, S., Boyd, S. P. & Vandenberghe, L. *Convex optimization* (Cambridge university press, 2004).

28. Bretagnolle, J. & Huber, C. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**, 119 (1979).

29. Bubeck, S. & Eldan, R. *Multi-scale exploration of convex functions and bandit convex optimization* in *Conference on Learning Theory* (2016), 583.

30. Bubeck, S. & Liu, C.-Y. *Prior-free and prior-dependent regret bounds for thompson sampling* in *2014 48th Annual Conference on Information Sciences and Systems (CISS)* (2014), 1.

31. Burnetas, A. N. & Katehakis, M. N. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* **17**, 122 (1996).

32. Bush, R. R. & Mosteller, F. A stochastic model with applications to learning. *The Annals of Mathematical Statistics*, 559 (1953).

33. Carmeli, C., De Vito, E. & Toigo, A. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications* **4**, 377 (2006).

34. Caron, S., Kveton, B., Lelarge, M. & Bhagat, S. *Leveraging side observations in stochastic bandits* in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (2012), 142.

35. Cesa-Bianchi, N., Lugosi, G. & Stoltz, G. Regret minimization under partial monitoring. *Mathematics of Operations Research* **31**, 562 (2006).

36. Cesa-Bianchi, N. & Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences* **78**, 1404 (2012).

37. Chaloner, K. & Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 273 (1995).

38. Chapelle, O. & Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems* **24**, 2249 (2011).

39. Chaudhuri, K., Jain, P. & Natarajan, N. *Active heteroscedastic regression* in *International Conference on Machine Learning* (2017), 694.

40. Chaudhuri, S. & Tewari, A. *Phased exploration with greedy exploitation in stochastic combinatorial partial monitoring games* in *Advances in Neural Information Processing Systems* (2016), 2433.

41. Chernoff, H. Sequential design of experiments. *The Annals of Mathematical Statistics* **30**, 755 (1959).

42. Chowdhury, S. R. & Gopalan, A. *On Kernelized Multi-armed Bandits* in *International Conference on Machine Learning* (2017).

43. Combes, R., Magureanu, S. & Proutiere, A. *Minimal exploration in structured stochastic bandits* in *Advances in Neural Information Processing Systems* (2017), 1763.

44. Cowan, W., Honda, J. & Katehakis, M. N. Normal bandits of unknown means and variances. *The Journal of Machine Learning Research* **18**, 5638 (2017).

45. Cowen-Rivers, A. I., Lyu, W., Tutunov, R., Wang, Z., Grosnit, A., Griffiths, R. R., Jianye, H., Wang, J. & Ammar, H. B. An Empirical Study of Assumptions in Bayesian Optimisation. *arXiv preprint arXiv:2012.03826* (2020).

46. Dani, V., Hayes, T. P. & Kakade, S. M. *Stochastic Linear Optimization under Bandit Feedback* in *COLT* (Omnipress, 2008), 355.

47. De Rooij, S., Van Erven, T., Grünwald, P. D. & Koolen, W. M. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research* **15**, 1281 (2014).

48. Degenne, R., Koolen, W. M. & Ménard, P. *Non-asymptotic pure exploration by solving games* in *Advances in Neural Information Processing Systems* (2019), 14492.

49. Degenne, R. & Perchet, V. *Anytime optimal algorithms in stochastic multi-armed bandits* in *International Conference on Machine Learning* (2016), 1587.

50. Degenne, R., Shao, H. & Koolen, W. Structure Adaptive Algorithms for Stochastic Bandits (2020).

51. Diaconis, P. Bayesian numerical analysis. *Statistical decision theory and related topics IV* **1**, 163 (1988).

52. Dong, S. & Roy, B. V. *An information-theoretic analysis for thompson sampling with many actions* in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (2018), 4161.

53. Durrett, R. *Probability: Theory and Examples* (Cambridge university press, 2010).

54. Erven, T., Koolen, W. M., Rooij, S. & Grünwald, P. Adaptive hedge. *Advances in Neural Information Processing Systems* **24**, 1656 (2011).

55. Fan, X., Grama, I., Liu, Q., *et al.* Exponential inequalities for martingales with applications. *Electronic Journal of Probability* **20** (2015).

56. Fiez, T., Jain, L., Jamieson, K. G. & Ratliff, L. in *Advances in Neural Information Processing Systems 32* 10666 (Curran Associates, Inc., 2019).

57. Fisher, R. A. *et al.* The design of experiments. *The design of experiments.* (1937).

58. Flaxman, A. D., Kalai, A. T. & McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. *arxiv* (2004).

59. Frank, M., Wolfe, P., *et al.* An algorithm for quadratic programming. *Naval research logistics quarterly* **3**, 95 (1956).

60. Frazier, P. I. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811* (2018).

61. Freedman, D. A. On Tail Probabilities for Martingales. *the Annals of Probability*, 100 (1975).

62. Gajane, P. & Urvoy, T. Utility-based dueling bandits as a partial monitoring game. *arXiv preprint arXiv:1507.02750* (2015).

63. Garg, N. & Koenemann, J. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. *SIAM Journal on Computing* **37**, 630 (2007).

64. Girosi, F. An equivalence between sparse approximation and support vector machines. *Neural computation* **10**, 1455 (1998).

65. Gittins, J. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, 241 (1974).

66. Glowacka, D. Bandit Algorithms in Information Retrieval. *Foundations and Trends® in Information Retrieval* **13**, 299 (2019).

67. González, J., Dai, Z., Damianou, A. & Lawrence, N. D. *Preferential bayesian optimization* in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), 1282.

68. Graves, T. L. & Lai, T. L. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization* **35**, 715 (1997).

69. Greenewald, K., Tewari, A., Murphy, S. & Klasnja, P. *Action centered contextual bandits* in *Advances in neural information processing systems* (2017), 5977.

70. Hamidi, N. & Bayati, M. On Worst-case Regret of Linear Thompson Sampling. *arXiv preprint arXiv:2006.06790* (2020).

71. Hamidi, N. & Bayati, M. The Randomized Elliptical Potential Lemma with an Application to Linear Thompson Sampling. *arXiv preprint arXiv:2102.07987* (2021).

72. Hao, B., Lattimore, T. & Szepesvari, C. Adaptive Exploration in Linear Contextual Bandit. *arXiv preprint arXiv:1910.06996* (2019).

73. Hao, B., Lattimore, T. & Wang, M. *High-Dimensional Sparse Linear Bandits* in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) **33** (Curran Associates, Inc., 2020), 10753.

74. Hazan, E. *et al.* Introduction to Online Convex Optimization. *Foundations and Trends® in Optimization* **2**, 157 (2016).

75. Hazan, E. & Levy, K. *Bandit Convex Optimization: Towards Tight Bounds* in *Advances in Neural Information Processing Systems* (eds Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q.) **27** (Curran Associates, Inc., 2014).

76. Jaggi, M. *Revisiting Frank-Wolfe: Projection-free sparse convex optimization* in *International Conference on Machine Learning* (2013), 427.

77. Jaksch, T., Ortner, R. & Auer, P. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research* **11** (2010).

78. Jourdan, M., Mutný, M., Kirschner, J. & Krause, A. *Efficient Pure Exploration for Combinatorial Bandits with Semi-Bandit Feedback* in *Algorithmic Learning Theory* (2021), 805.

79. Jun, K.-S. & Zhang, C. Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality (2020).

80. Juranić, P., Rehanek, J., Arrell, C. A., Pradervand, C., Ischebeck, R., Erny, C., Heimgartner, P., Gorgisyan, I., Thominet, V., Tiedtke, K., *et al.* SwissFEL Aramis beamline photon diagnostics. *Journal of synchrotron radiation* **25**, 1238 (2018).

81. Kakade, S. M. & Tewari, A. in *Advances in Neural Information Processing Systems 21* 801 (2009).

82. Kalkanlı, C. & Özgür, A. *An Improved Regret Bound for Thompson Sampling in the Gaussian Linear Bandit Setting* in *2020 IEEE International Symposium on Information Theory (ISIT)* (2020), 2783.

83. Kanagawa, M., Hennig, P., Sejdinovic, D. & Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582* (2018).

84. Kim, G.-S. & Paik, M. C. *Contextual Multi-armed Bandit Algorithm for Semiparametric Reward Model* in *International Conference on Machine Learning* (2019), 3389.

85. Kimeldorf, G. S. & Wahba, G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41**, 495 (1970).

86. Kirschner, J., Bogunovic, I., Jegelka, S. & Krause, A. *Distributionally Robust Bayesian Optimization* in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)* (2020).

87. Kirschner, J. & Krause, A. *Information Directed Sampling and Bandits with Heteroscedastic Noise* in *Proc. International Conference on Learning Theory (COLT)* (2018).

88. Kirschner, J. & Krause, A. *Stochastic Bandits with Context Distributions* in *Proc. Neural Information Processing Systems (NeurIPS)* (2019).

89. Kirschner, J. & Krause, A. *Bias-Robust Bayesian Optimization via Dueling Bandits* in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)* (2021).

90. Kirschner, J., Lattimore, T. & Krause, A. *Information Directed Sampling for Linear Partial Monitoring* in *Proc. International Conference on Learning Theory (COLT)* (2020).

91. Kirschner, J., Lattimore, T., Vernade, C. & Szepesvári, C. *Asymptotically Optimal Information-Directed Sampling* in *Proc. International Conference on Learning Theory (COLT)* (2021).

92. Kirschner, J., Mutný, M., Hiller, N., Ischebeck, R. & Krause, A. *Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional Subspaces* in *Proc. International Conference for Machine Learning (ICML)* (2019).

93. Kirschner, J., Nonnenmacher, M., Mutný, M., Hiller, N., Adelmann, A., Ischebeck, R. & Krause, A. *Bayesian Optimization for Fast and Safe Parameter Tuning of SwissFEL* in *Proc. International Free-Electron Laser Conference (FEL2019)* (2019).

94. Kleinberg, R. & Leighton, T. *The value of knowing a demand curve: Bounds on regret for online posted-price auctions* in *44th Annual IEEE Symposium on Foundations of Computer Science.* (2003), 594.

95. Kocsis, L. & Szepesvári, C. *Bandit based monte-carlo planning* in *European conference on machine learning* (2006), 282.

96. Komiyama, J., Honda, J. & Nakagawa, H. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 1792 (Curran Associates, Inc., 2015).

97. Krishnamurthy, A., Wu, Z. S. & Syrgkanis, V. *Semiparametric Contextual Bandits* in *International Conference on Machine Learning* (2018), 2776.

98. Kveton, B., Szepesvari, C., Wen, Z. & Ashkan, A. *Cascading bandits: Learning to rank in the cascade model* in *International Conference on Machine Learning* (2015), 767.

99. Lai, T. L. & Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6**, 4 (1985).

100. Langford, J. & Zhang, T. in *Advances in Neural Information Processing Systems 20* (eds Platt, J. C., Koller, D., Singer, Y. & Roweis, S. T.) 817 (Curran Associates, Inc., 2008).

101. Lattimore, T. Improved regret for zeroth-order adversarial bandit convex optimisation. *arXiv preprint arXiv:2006.00475* (2020).

102. Lattimore, T. & György, A. Mirror Descent and the Information Ratio. *arXiv preprint arXiv:2009.12228* (2020).

103. Lattimore, T. & Szepesvari, C. *Bandit Algorithms* (Cambridge University Press, 2020).

104. Lattimore, T. & Szepesvári, C. *The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits* in *Artificial Intelligence and Statistics* (2017), 728.

105. Lattimore, T. & Szepesvári, C. An Information-Theoretic Approach to Minimax Regret in Partial Monitoring. *arXiv preprint arXiv:1902.00470* (2019).

106. Lattimore, T. & Szepesvári, C. *Cleaning up the neighborhood: A full classification for adversarial partial monitoring* in *Algorithmic Learning Theory* (2019), 529.

107. Li, L., Chu, W., Langford, J. & Schapire, R. E. *A contextual-bandit approach to personalized news article recommendation* in *Proceedings of the 19th international conference on World wide web* (2010), 661.

108. Li, Y., Lou, E. Y. & Shan, L. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109* (2019).

109. Lin, T., Abrahao, B., Kleinberg, R., Lui, J. & Chen, W. *Combinatorial partial monitoring game with linear feedback and its applications* in *International Conference on Machine Learning* (2014), 901.

110. Littlestone, N. & Warmuth, M. K. The weighted majority algorithm. *Information and computation* **108**, 212 (1994).

111. Liu, F., Buccapatnam, S. & Shroff, N. *Information directed sampling for stochastic bandits with graph feedback* in *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

112. Lu, X. & Van Roy, B. *Information-Theoretic Confidence Bounds for Reinforcement Learning* in *Advances in Neural Information Processing Systems* (eds Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R.) **32** (Curran Associates, Inc., 2019).

113. Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I. & Wen, Z. Reinforcement Learning, Bit by Bit. *arXiv preprint arXiv:2103.04047* (2021).

114. Lykouris, T., Mirrokni, V. & Paes Leme, R. *Stochastic bandits robust to adversarial corruptions* in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (2018), 114.

115. Mannor, S. & Shamir, O. *From Bandits to Experts: On the Value of Side-Observations* in *Advances in Neural Information Processing Systems* (eds Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F. & Weinberger, K. Q.) **24** (Curran Associates, Inc., 2011).

116. McIntire, M., Cope, T., Ermon, S. & Ratner, D. *Bayesian Optimization of FEL Performance at LCLS* in *7th Int. Particle Accelerator Conf.(IPAC'16), Busan, Korea, May 8-13, 2016* (2016), 2972.

117. Ménard, P. & Garivier, A. *A minimax and asymptotically optimal algorithm for stochastic bandits* in *International Conference on Algorithmic Learning Theory* (2017), 223.

118. Milne, C. J. *et al.* SwissFEL: The Swiss X-ray Free Electron Laser. *Applied Sciences* **7** (2017).

119. Mockus, J. The Bayesian approach to global optimization. *System Modeling and Optimization*, 473 (1982).

120. Mutný, M., Kirschner, J. & Krause, A. *Experimental Design for Optimization of Orthogonal Projection Pursuit Models* in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)* (2020).

121. Mutný, M. & Krause, A. *Efficient High Dimensional Bayesian Optimization with Additivity and Quadrature Fourier Features* in *Neural and Information Processing Systems (NeurIPS)* (2018).

122. Nikolov, N., Kirschner, J., Berkenkamp, F. & Krause, A. *Information-Directed Exploration for Deep Reinforcement Learning* in *Proc. International Conference on Learning Representations (ICLR)* (2019).

123. O'Hagan, A. Bayes–hermite quadrature. *Journal of statistical planning and inference* **29**, 245 (1991).

124. Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213* (2019).

125. Osband, I., Russo, D. & Van Roy, B. (More) efficient reinforcement learning via posterior sampling. *arXiv preprint arXiv:1306.0940* (2013).

126. Pedrick, G. *Theory of reproducing kernels for Hilbert spaces of vector valued functions* PhD thesis (University of Kansas, 1957).

127. Peña, V. H., Lai, T. L. & Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications* (Springer Science & Business Media, 2008).

128. Piccolboni, A. & Schindelhauer, C. *Discrete prediction games with arbitrary feedback and loss* in *International Conference on Computational Learning Theory* (2001), 208.

129. Prokhorov, Y. V. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications* **1**, 157 (1956).

130. Rasmussen, C. E. in *Advanced lectures on machine learning* 63 (Springer, 2004).

131. Rigollet, P. 18. S997: High dimensional statistics (2015).

132. Rusmevichientong, P. & Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research* **35**, 395 (2010).

133. Russo, D. Simple Bayesian Algorithms for Best-Arm Identification. *Operations Research* (2020).

134. Russo, D. A note on the equivalence of upper confidence bounds and gittins indices for patient agents. *Operations Research* **69**, 273 (2021).

135. Russo, D. & Van Roy, B. *Learning to optimize via information-directed sampling* in *Advances in Neural Information Processing Systems* (2014), 1583.

136. Russo, D. & Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research* **39**, 1221 (2014).

137. Russo, D. & Van Roy, B. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research* **17**, 2442 (2016).

138. Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I. & Wen, Z. A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning* **11**, 1 (2018).

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

140. Saha, A. & Gopalan, A. Regret Minimization in Stochastic Contextual Dueling Bandits. *arXiv preprint arXiv:2002.08583* (2020).

141. Scheinker, A., Bohler, D., Tomin, S., Kammering, R., Zagorodnov, I., Schlarb, H., Scholz, M., Beutner, B. & Decking, W. Model-independent tuning for maximizing free electron laser pulse energy. *Physical Review Accelerators and Beams* **22**, 082802 (2019).

142. Schölkopf, B., Herbrich, R. & Smola, A. J. *A generalized representer theorem* in *International conference on computational learning theory* (2001), 416.

143. Schölkopf, B., Smola, A. J., Bach, F., *et al. Learning with kernels: support vector machines, regularization, optimization, and beyond* (2002).

144. Seidel, M., Adam, S., Adelmann, A., Baumgarten, C., Bi, Y., Doelling, R., Fitze, H., Fuchs, A., Humbel, M., Grillenberger, J., Kiselev, D., Mezger, A., Reggiani, D., Schneider, M., Yang, J., Zhang, H. & Zhang, T. *Production of A 1.3 MW proton beam at PSI* in (2010).

145. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* **104**, 148 (2016).

146. Shalev-Shwartz, S. & Singer, Y. *Online learning: Theory, algorithms, and applications* (2007).

147. Sherman, J. & Morrison, W. J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* **21**, 124 (1950).

148. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **529**, 484 (2016).

149. Slivkins, A. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning* **12**, 1 (2019).

150. Soare, M., Lazaric, A. & Munos, R. *Best-arm identification in linear bandits* in *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1* (2014), 828.

151. Sorokin, A. A., Bican, Y., Bonfigt, S., Brachmanski, M., Braune, M., Jastrow, U. F., Gottwald, A., Kaser, H., Richter, M. & Tiedtke, K. An X-ray gas monitor for free-electron lasers. *Journal of synchrotron radiation* **26**, 1092 (2019).

152. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *International Conference on Machine Learning* (2010).

153. Strens, M. J. *A Bayesian Framework for Reinforcement Learning* in *Proceedings of the Seventeenth International Conference on Machine Learning* (2000), 943.

154. Sui, Y., Yue, Y. & Burdick, J. W. Correlational dueling bandits with application to clinical treatment in large decision spaces. *arXiv preprint arXiv:1707.02375* (2017).

155. Sui, Y., Zhuang, V., Burdick, J. W. & Yue, Y. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253* (2017).

156. Sui, Y., Zoghi, M., Hofmann, K. & Yue, Y. *Advancements in Dueling Bandits.* in *IJCAI* (2018), 5502.

157. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (2018).

158. Tang, L., Rosales, R., Singh, A. & Agarwal, D. *Automatic ad format selection via contextual bandits* in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (2013), 1587.

159. Thompson, W. R. On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **25**, 285 (1933).

160. Tirinzoni, A., Pirotta, M., Restelli, M. & Lazaric, A. An Asymptotically Optimal Primal-Dual Incremental Algorithm for Contextual Linear Bandits. *Advances in Neural Information Processing Systems* **33** (2020).

161. Toscano-Palmerin, S. & Frazier, P. I. Bayesian optimization with expensive integrands. *arXiv preprint arXiv:1803.08661* (2018).

162. Tsuchiya, T., Honda, J. & Sugiyama, M. *Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring* in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) **33** (Curran Associates, Inc., 2020), 8861.

163. Vakili, S., Khezeli, K. & Picheny, V. On Information Gain and Regret Bounds in Gaussian Process Bandits. *arXiv preprint arXiv:2009.06966* (2020).

164. Valko, M., Korda, N., Munos, R., Flaounas, I. & Cristianini, N. Finite-Time Analysis of Kernelised Contextual Bandits. *arXiv:1309.6869 [cs, stat]* (2013).

165. Valko, M., Munos, R., Kveton, B. & Kocák, T. *Spectral bandits for smooth graph functions* in *International Conference on Machine Learning* (2014), 46.

166. Van Parys, B. P. & Golrezaei, N. Optimal Learning for Structured Bandits. *arXiv preprint arXiv:2007.07302* (2020).

167. Vanchinathan, H., Bartók, G. & Krause, A. *Efficient Partial Monitoring with Prior Information* in *Neural Information Processing Systems (NIPS)* (2014).

168. Vovk, V. G. Aggregating strategies. *Proc. of Computational Learning Theory, 1990* (1990).

169. Wagenmaker, A., Katz-Samuels, J. & Jamieson, K. Experimental Design for Regret Minimization in Linear Bandits. *arXiv preprint arXiv:2011.00576* (2020).

170. Wang, Z., Zhou, B. & Jegelka, S. *Optimization as estimation with Gaussian processes in bandit settings* in *Artificial Intelligence and Statistics* (2016), 1022.

171. Woodbury, M. A. *Inverting modified matrices* (Statistical Research Group, 1950).

172. Woodroofe, M. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association* **74**, 799 (1979).

173. Wu, A., Aoi, M. C. & Pillow, J. W. Exploiting gradients and hessians in bayesian optimization and bayesian quadrature. *arXiv preprint arXiv:1704.00060* (2017).

174. Wu, J., Poloczek, M., Wilson, A. G. & Frazier, P. *Bayesian optimization with gradients* in *Advances in Neural Information Processing Systems* (2017), 5267.

175. Yue, Y., Broder, J., Kleinberg, R. & Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences* **78**, 1538 (2012).

176. Yue, Y. & Joachims, T. *Interactively optimizing information retrieval systems as a dueling bandits problem* in *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), 1201.

177. Zanette, A. & Sarkar, R. *Information directed reinforcement learning* tech. rep. (Technical report, Technical report, 2017).