Thompson Sampling for Bandit Convex Optimization

by

Alireza Bakhtiari

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Thompson sampling (TS) is a popular and empirically successful algorithm for online decision-making problems. This thesis advances our understanding of TS when applied to bandit convex optimization (BCO) problems, by providing new theoretical guarantees and characterizing its limitations.

First, we analyze 1-dimensional BCO and show that TS achieves a near-optimal Bayesian regret of at most $\tilde{O}(\sqrt{n})$, where n is the time horizon. This result holds without strong assumptions on the loss functions, requiring only convexity, boundedness, and a mild Lipschitz condition. In sharp contrast, we demonstrate that for general high-dimensional problems, TS can fail catastrophically.

More positively, we establish a Bayesian regret bound of $\tilde{O}(d^{2.5}\sqrt{n})$ for TS in generalized linear bandits, even when the convex monotone link function is unknown. Finally, we prove a fundamental limitation of current analysis techniques: we show that the standard information-theoretic machinery can never yield a regret bound better than the existing $\tilde{O}(d^{1.5}\sqrt{n})$ in the general case.

Preface

The first part of this thesis has been published as Alireza Bakhtiari, Tor Lattimore, and Csaba Szepesvári. Thompson Sampling for Bandit Convex Optimization. In International Conference on Learning Theory, 2025. The second part of this thesis has been published as Johannes Kirschner, Alireza Bakhtiari, Kushagra Chandak, Volodymyr Tkachuk, Csaba Szepesvári. Regret Minimization via Saddle Point Optimization. In Advances in Neural Information Processing Systems, 2023.

Acknowledgements

I would like to sincerely thank my supervisor, Csaba, for his time, mentorship, guidance, and patience throughout my studies. Csaba gave me the opportunity to pursue what I love most and provided me with unconditional intellectual freedom. His passion for research and honest approach to science have been a true inspiration. I also extend my gratitude to Tor Lattimore for his time and for teaching me the ways of research; his insights have profoundly shaped my thinking. Special thanks go to Dave Janz and Johannes Krishner for their close guidance and constructive feedback in shaping my research. Moreover, I am deeply grateful to my labmates—listed in alphabetical order as Alex prefers—Alex Ayoub, Kushagra Chandak, Shuai Liu, Samuel Robertson, and Vlad Tkachuk, for patiently listening to my ideas, providing valuable feedback, and assisting me in my research. This journey would have been far less enjoyable without their presence. I am also thankful to my friends for always being there when needed, and to my partner, Zahra, for her unwavering patience, support, and encouragement. Lastly, I feel incredibly fortunate to have my parents, Pooneh and Majid, whose although been far away, have always kept their love and care present in my life, and my brother, Amin, whose presence makes my life much more colorful and vibrant.

Contents

List of Tables

List of Figures

List of Algorithms

Part I

Thompson Sampling for Bandit Convex Optimization

Chapter 1

Introduction

Convexity is a common assumption in optimization problems [BV04]. Bandit convex optimization (BCO) addresses the fundamental problem of minimizing a convex function over a convex set when only noisy evaluations of the function are available at selected points. This setting naturally arises in scenarios where:

(i) Limited Access: The algorithm can only observe noisy evaluations of the objective function.

(ii) Cumulative Cost: The goal is to minimize the cumulative cost of these evaluations over time, rather than to simply identify the function's minimizer.

In classical optimization, it is typically assumed that the optimizer has access to the full function, or at least to its value and gradient at any point in the domain. However, this assumption often fails in practice. A representative example is dynamic pricing, where a seller selects a price (the input) and observes the resulting profit (the output), which is often modeled as a concave function of the price. The seller cannot directly observe the profit function itself, but only noisy feedback through customer purchases at chosen prices. Each evaluation corresponds to a real transaction and thus incurs a potentially significant cost. In such cases, the objective is not to eventually find the best price at any expense, but rather to make pricing decisions that yield high cumulative profit over time.

This thesis focuses on this *zeroth-order* or *bandit feedback* setting, which is prevalent in domains where gradients are unavailable or costly to compute, and where each function evaluation carries a tangible cost. While convexity is an idealization, it is a broadly applicable assumption that facilitates principled algorithm design and analysis. Unlike traditional optimization settings—where function evaluations are often treated as free abstractions—many real-world problems demand an approach that explicitly accounts for the cumulative cost incurred during learning.

Formally, the goal is to approximately solve the optimization problem

$$\arg\min_{x \in \mathcal{K}} f(x)\,, \tag{1.1}$$

where $\mathcal{K} \subset \mathbb{R}^d$ is a convex set (typically compact with non-empty interior) and $f : \mathcal{K} \to \mathbb{R}$ is a convex function. In the BCO setting, the learner does not observe gradients or even function values at arbitrary points; instead, in each round of interaction it selects a point $x_t \in \mathcal{K}$ and receives noisy feedback $y_t = f(x_t) + \varepsilon_t$, where ε_t models the noise. While the convexity assumption may seem restrictive, it captures a rich class of problems and provides a tractable framework to develop principled algorithms with provable guarantees.

The learner aims to minimize the cumulative loss over T rounds, which leads to the online variant of the optimization problem, where the learner's goal is to minimize cumulative loss compared to the best action in hindsight, which is

$$\text{Regret}(T) = \sum_{t=1}^{T} f(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f(x)\,, \tag{1.2}$$

often referred to as *bandit regret*. This perspective connects BCO to the broader literature on online learning and multi-armed bandits. While BCO is challenging due to the simultaneous absence of gradient information and the need for exploration, it offers a powerful abstraction for decision-making under uncertainty with limited feedback. The convexity assumption, though idealized, provides a tractable and theoretically rich framework for developing and analyzing algorithms that strike a balance between exploration and exploitation.

1.1 Examples of BCO

This setting naturally arises in a variety of applications where gradient information is unavailable, unreliable, or too expensive to compute, such as

Manufacturing: Consider a cheese factory that aims to optimize its recipe by adjusting the temperature and humidity of its warehouse. The quality of the final product can be modeled as a convex function of these parameters. However, each measurement becomes available only after the batch is complete, and producing a batch incurs a financial cost. The factory's goal is to iteratively improve product quality based on customer feedback, but it cannot afford to ruin too many batches in the process. This situation is even more pronounced for expensive or custom products—like cars, airplanes, or musical instruments—where each failed experiment is prohibitively costly and feedback may only be available post-sale.

Dynamic Pricing: In dynamic pricing, a retailer interacts sequentially with an uncertain market. At each round, they select a price $X_t \in \mathcal{K} \subset \mathbb{R}$, and the associated loss $f(X_t)$ represents the negative of expected profit. Prices that are too high may deter purchases, while prices that are too low leave revenue on the table. The profit function f is unknown in advance and customer behavior introduces noise into observations. The goal is to adjust prices over time to maximize cumulative profit, not just to identify an optimal price after at any cost.

Service Personalization: Large Language Models (LLMs) often personalize their responses based on user preferences. At each interaction, the system selects a response style $X_t \in \mathcal{K} \subset \mathbb{R}^d$ — e.g., controlling tone, formality, or humor. The user's satisfaction is captured by a loss function $f(X_t)$, which reflects poor alignment with their preferences. This function is unknown, subjective, and observed only through noisy feedback such as click-through rates or engagement metrics. The system must learn and adapt over time to minimize dissatisfaction, making this a natural fit for the bandit convex optimization setting.

Resource Allocation: Many decision-making problems involve allocating limited resources—like budget or bandwidth—across competing options. For example, a company might distribute marketing funds across various channels. The return on investment typically exhibits diminishing returns and can be modeled by a convex utility function. The utility is not directly observable but can be estimated after committing to a specific allocation. Since each evaluation carries cost, the company seeks to minimize cumulative regret over time by carefully balancing exploration and exploitation.

Online Advertising—$\mathbb{R}$-valued parameters: Online advertisement is a classical application of bandit algorithms. Often, advertisers can choose among nearly-continuous design parameters—like font size and color. These decisions affect user engagement in a way that can be modeled by a convex function over the parameter space. The feedback (e.g., click-through rates) is noisy and delayed, and testing each variation has opportunity cost. Bandit convex optimization provides a principled framework to navigate this space efficiently and improve ad performance over time.

Efficiency Tuning: In commercial aviation, dispatchers must decide the cruise altitude and Mach number for each flight, represented as $X_t = (\text{Mach}, \text{Altitude}) \in \mathcal{K} \subset \mathbb{R}^2$. The loss $f(X_t)$ is the fuel burned per seat-kilometre—a quantity well-approximated by a convex function due to the trade-offs between speed, altitude, and drag. Crucially, the true fuel burn is revealed only after the flight, and is confounded by weather, routing, and payload variability. Since each evaluation corresponds to a real flight with significant operational cost, the airline cannot afford extensive trial-and-error. Instead, it must adapt its cruise settings sequentially, improving fuel efficiency over time while minimizing total cost—an ideal use case for bandit convex optimization.

1.2 Thompson Sampling and Bayesian Bandits

Thompson sampling (TS) is a simple and often practical algorithm for interactive decision-making with a long history [Tho33a, RVK+18]. Our interest is in its application to Bayesian bandit convex optimization [Lat24].

A Bayesian bandit problem is a sequential decision-making problem where the learner has access to a prior distribution over the unknown objective function. The learner interacts with the environment by selecting actions and observing noisy feedback, which is modeled as a realization of a random variable whose distribution depends on the unknown objective function. This prior distribution captures the learner or the domain expert's beliefs about the objective function before any interaction.

At its core, Thompson Sampling is a Bayesian algorithm that maintains a posterior distribution over the unknown objective (or cost) function. In each round, it samples a function from this posterior and selects the action that minimizes the sampled function. The elegance of TS lies in its simplicity and flexibility—it requires no explicit exploration bonus or confidence bounds and can be implemented in a wide range of settings, provided posterior sampling is computationally feasible.

Chapter 2

Related work

BCO in the regret setting was first studied by [FKM05] and [Kle05]. Since then the field has grown considerably as summarized in the recent monograph by [Lat24]. Our focus is on the Bayesian version of the problem, which has seen only limited attention. [BDKP15] consider the adversarial version of the Bayesian regret and show that a (heavy) modification of TS enjoys a Bayesian regret of $\tilde{O}(\sqrt{n})$ when $d = 1$. Interestingly, they argue that TS without modification is not amenable to analysis via the information-theoretic machinery, but this argument only holds in the adversarial setting as our analysis shows. [BE18] and [Lat20] generalized the information-theoretic machinery used by [BDKP15] to higher dimensions, also in the adversarial setting. These works focus on proving bounds for information-directed sampling (IDS), which is a conceptually simple but computationally more complicated algorithm introduced by [RV14]. Nevertheless, we borrow certain techniques from these papers. Convex ridge functions have been studied before by [Lat21], who showed that IDS has a Bayesian regret in the stochastic setting of $\tilde{O}(d\sqrt{n})$, which matches the lower bound provided by linear bandits [DHK08]. Regrettably, however, this algorithm is not practically implementable, even under the assumption that you can sample efficiently from the posterior. [SNNJ21] also study a variation on the problem where the losses have the form $f(g(x))$ with $g : \mathbb{R}^d \to \mathbb{R}$ a *known* function and $f : \mathbb{R} \to \mathbb{R}$ an unknown convex function. When g is linear, then $f \circ g$ is a convex ridge function. The assumption that g is known dramatically changes the setting, however. The best known bound for an efficient algorithm in the monotone convex ridge function setting is $\tilde{O}(d^{1.5}\sqrt{n})$, which also holds for general convex functions, even in the frequentist setting [FvdHLM24b]. Convex ridge functions can also be viewed as a special case of the generalized linear model, which has been studied extensively as a reward model for stochastic bandits [FCGS10, and many more]. TS and other randomized algorithms have been studied with generalized linear models in Bayesian and frequentist settings [AL17, DMR19, KZS$^+$20]. None of these papers assume convexity (concavity for rewards) and consequentially suffer a regret that depends on other properties of the link function that can be arbitrarily large. Moreover, in generalized linear bandits it is standard to assume the link function is known.

Chapter 3

Bayesian BCO Problem and the TS Algorithm

Let K be a convex body in $\mathbb{R}^d$ and $\mathscr{F}$ be a set of convex functions from K to $[0, 1]$. We assume there is a known (prior) probability measure ξ on $\mathscr{F}$. The interaction between the learner and environment lasts for n rounds. At the beginning the environment secretly samples f from the prior ξ. Subsequently, the learner and environment interact sequentially. In round t the learner plays an action $X_t \in K$ and observes $Y_t \in \{0, 1\}$ for which $\mathbb{E}[Y_t|X_1, Y_1, \dots, X_t, f] = f(X_t)$. The assumption that the noise is Bernoulli is for convenience only. Our analysis would be unchanged with any bounded noise model and would continue to hold for sub-gaussian noise with minor modifications. A learner $\mathscr{A}$ is a (possibly random) mapping from sequences of action/loss pairs to actions and its Bayesian regret with respect to prior ξ is

$$\mathrm{BReg}_n(\mathscr{A}, \xi) = \mathbb{E}\left[\sup_{x \in K} \sum_{t=1}^{n} (f(X_t) - f(x))\right] .$$

Note that both f and the iterates (X_t) are random elements. Moreover, in the Bayesian setting the learner $\mathscr{A}$ is allowed to depend on the prior ξ. The main quantity of interest is

$$\sup_{\xi \in \mathscr{P}(\mathscr{F})} \mathrm{BReg}_n(\mathrm{TS}, \xi) \,, \tag{3.1}$$

where $\mathscr{P}(\mathscr{F})$ is the space of probability measures on $\mathscr{F}$ (with a suitable σ-algebra) and TS is Thompson sampling (Algorithm 1) with prior ξ (the dependence on the prior is always omitted from the notation). The quantity in Eq. (3.1) depends on the function class $\mathscr{F}$. Our analysis explores this dependence for various natural classes of convex functions. TS (Algorithm 1) is theoretically near-trivial. In every round it samples f_t from the posterior and plays X_t as the minimizer of f_t.

```
args: prior $\xi$
for $t = 1$ to $\infty$:
    sample $f_t$ from $\mathbb{P}(f = \cdot | X_1, Y_1, \dots, X_{t-1}, Y_{t-1})$
    play $X_t = x_{f_t}$ and observe $Y_t$
```

Algorithm 1: Thompson sampling

3.1 Thompson Sampling for Bandit Convex Optimization

With these definitions in place, we can now summarize our results:

- When $d = 1$, $\mathrm{BReg}_n(\mathrm{TS}, \xi) = \tilde{O}(\sqrt{n})$ for all priors (Theorem 15).
- A convex function f is called a monotone ridge function if there exists a convex monotone function $\ell : \mathbb{R} \to \mathbb{R}$ and $\theta \in \mathbb{R}^d$ such that $f(x) = \ell(\langle x, \theta \rangle)$. Theorem 17 shows when ξ is supported on monotone ridge functions, then $\mathrm{BReg}_n(\mathrm{TS}, \xi) = \tilde{O}(d^{2.5}\sqrt{n})$.
- In general, the Bayesian regret of TS can be exponential in the dimension (Theorem 22).
- The classical information-theoretic machinery used by [BE18] an [Lat20] cannot improve the regret for `BCO` beyond the best known upper bound of $\tilde{O}(d^{1.5}\sqrt{n})$.

Although the regret bounds are known already in the frequentist setting for different algorithms, there is still value in studying Bayesian algorithms and especially TS. Most notably, none of the frequentist algorithms can make use of prior information about the loss functions and adapting them to exploit such information is often painstaking and ad-hoc. TS, on the other hand, automatically exploits prior information. Our bounds for ridge functions can be viewed as a Bayesian regret bound for a kind of generalized linear bandit where the link function is unknown and assumed to be convex and monotone increasing.

Many problems are reasonably modelled as 1-dimensional convex bandits, with the classical example being dynamic pricing where K is a set of prices and convexity is a reasonable assumption based on the response of demand to price. The monotone ridge function class is a natural model for resource allocation problems where a single resource (e.g., money) is allocated to d locations. The success of some global task increases as more resources are allocated, but with diminishing returns. Problems like this can reasonably be modelled by convex monotone ridge functions with $K = \{x \geq \mathbf{0} : \|x\|_1 \leq 1\}$.

Our lower bounds show that TS does not behave well in the general BCO unless possibly the dimension is quite small. Perhaps more importantly, we show that the classical information-theoretic machinery used by [BE18] and [Lat20] cannot be used to improve the current best dimension dependence of the regret for BCO. Combining this with the duality between exploration-by-optimisation and information-directed sampling shows that exploration-by-optimisation (with negentropy potential) also cannot naively improve on the best known $\tilde{O}(d^{1.5}\sqrt{n})$ upper bound [ZL19, LG23]. We note that this does not imply a lower bound for BCO. The construction in the lower bound is likely amenable to methods for learning a direction based on the power method [LH21, HHK$^+$21]. The point is that the information ratio bound characterises the signal-to-noise ratio for the prior, but does not prove the signal-to-noise ratio does not increase as the learner gains information.

3.2 Notation

Let $\|\cdot\|$ be the standard euclidean norm on $\mathbb{R}^d$. For natural number k let $[k] = \{1, \ldots, k\}$. Define $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$ for positive definite Σ. Given a function $f : K \to \mathbb{R}$, let $\|f\|_\infty = \sup_{x \in K} |f(x)|$. The centered euclidean ball of radius $r > 0$ is $\mathbb{B}_r = \{x \in \mathbb{R}^d : \|x\| \leq r\}$ and the sphere is $\mathbb{S}_r = \{x \in \mathbb{R}^d : \|x\| = r\}$. We also let $\mathbb{B}_r(x) = \{y \in \mathbb{R}^d : \|x - y\| \leq r\}$. We let $H(x, \eta) = \{y : \langle y, \eta \rangle \geq \langle x, \eta \rangle\}$, which is a half-space with inward-facing normal η. Given a finite set $\mathcal{C}$ let $\textsc{pair}(\mathcal{C}) = \{(x, y) \in \mathcal{C} : x \neq y\}$ be the set of all distinct ordered pairs and abbreviate $\textsc{pair}(k) = \textsc{pair}([k])$. The convex hull of a subset A of a linear space is $\operatorname{conv}(A)$. The space of probability measures on K with respect to the Borel σ-algebra is $\mathscr{P}(K)$. Similarly, $\mathscr{P}(\mathscr{F})$ is a space of probability measures on $\mathscr{F}$ with some unspecified σ-algebra ensuring that $f \mapsto f(x)$ is measurable for all $x \in K$. Given a convex function $f : K \to \mathbb{R}$ we define $\operatorname{Lip}_K(f) = \sup_{x \neq y \in K}(f(x) - f(y))/\|x - y\|$ and $f_\star = \inf_{x \in K} f(x)$ and $x_f = \arg\min_{x \in K} f(x)$ where ties are broken in an arbitrary measurable fashion. Such a mapping exists and $f \mapsto f_\star$ is also measurable; [Nie92] showed that such a mapping always exists. Of course it follows that $f \mapsto f_\star = f(x_f)$ is also measurable. $\mathbb{P}_t = \mathbb{P}(\cdot | X_1, Y_1, \ldots, X_t, Y_t)$ and $\mathbb{E}_t$ be the expectation operator with respect to $\mathbb{P}_t$. The following assumption on $\mathcal{K}$ is considered global throughout:

Assumption 1. $\mathcal{K}$ is a convex body (compact, convex with non-empty interior) and $\mathbf{0} \in \mathcal{K}$.

3.2.1 Spaces of Convex Functions

A function $f : K \to \mathbb{R}$ is called a convex ridge function if there exists a convex $\ell : \mathbb{R} \to \mathbb{R}$ and $\theta \in \mathbb{R}^d$ such that $f(x) = \ell(\langle x, \theta \rangle)$. Moreover, f is called a monotone convex ridge function if it is a convex ridge function and ℓ is monotone increasing. We are interested in the following classes of convex functions: (a) $\mathscr{F}_{\texttt{b}}$ is the space of all bounded convex functions $f : K \to [0, 1]$. (b) $\mathscr{F}_{\texttt{l}}$ is the space of convex functions $f : K \to \mathbb{R}$ with $\operatorname{Lip}(f) \leq 1$. (c) $\mathscr{F}_{\texttt{r}}$ is the space of all convex ridge functions. (d) $\mathscr{F}_{\texttt{rm}}$ is the space of all monotone convex ridge functions. Intersections are represented as you might expect: $\mathscr{F}_{\texttt{bl}} = \mathscr{F}_{\texttt{b}} \cap \mathscr{F}_{\texttt{l}}$ and similarly for other combinations. The set $\mathscr{F}$ refers to a class of convex functions, which will always be either $\mathscr{F}_{\texttt{bl}}$ or $\mathscr{F}_{\texttt{blrm}}$.

The representation of f as a ridge convex function is not unique, meaning that there could be (are) two sets (θ_1, ℓ_1) and (θ_2, ℓ_2) such that $f = \ell_1(\langle x, \theta_1 \rangle)$ and $f = \ell_2(\langle x, \theta_2 \rangle)$. The following lemma ensures that the *link function* ℓ can be chosen in a way that the Lipschitzness of the original function f is preserved.

Lemma 2. *Suppose that K is a convex body and $f \in \mathscr{F}_{\texttt{lr}}$ is a Lipschitz convex ridge function. Then there exists a $\theta \in \mathbb{S}_1$ and a convex $\ell : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \ell(\langle x, \theta \rangle)$ and $\operatorname{Lip}(\ell) \leq \operatorname{Lip}_K(f)$.*

Proof. By assumption there exists a convex function $\ell : \mathbb{R} \to \mathbb{R}$ and $\theta \in \mathbb{S}_1$ such that $f(x) = \ell(\langle x, \theta \rangle)$. It remains to show that ℓ can be chosen so that $\operatorname{Lip}(\ell) \leq \operatorname{Lip}_K(f)$. Let h_K be the support function associated with K, given by $h_K(v) = \sup_{x \in K} \langle v, x \rangle$. Therefore ℓ is uniquely defined on $I = [-h_K(-\theta), h_K(\theta)]$ and can be defined in any way that preserves convexity outside. Let $Dg(x)[v]$ be the directional derivative of g at x in direction v, which for convex g exists for all x in the interior of the domain of g. Then

$$\begin{aligned}
\operatorname{Lip}_K(f) &\geq \sup_{x \in \operatorname{int}(K)} \max(Df(x)[\theta], Df(x)[-\theta]) \\
&= \sup_{x \in \operatorname{int}(K)} \max(D\ell(\langle x, \theta \rangle)[1], D\ell(\langle x, \theta \rangle)[-1]) \\
&= \sup_{x \in \operatorname{int}(I)} \max(|D\ell(x)[1]|, |D\ell(x)[-1]|) \\
&= \operatorname{Lip}_{\operatorname{int}(I)}(\ell) \\
&= \operatorname{Lip}_I(\ell)\,.
\end{aligned}$$

Then define ℓ on all of $\mathbb{R}$ via the classical extension [Lat24, Proposition 3.18, for example]. $\square$

Chapter 4

Generalized Information Ratio

The main theoretical tool is a version of the information ratio, which was introduced by [RV16] as a means to bound the Bayesian regret of TS for finite-armed and linear bandits. Given a distribution $\xi \in \mathscr{P}(\mathscr{F})$ and policy $\pi \in \mathscr{P}(K)$, let (X, f) have law $\pi \otimes \xi$ and with $\bar{f} = \mathbb{E}[f]$ define

$$\Delta(\pi, \xi) = \mathbb{E}\left[\bar{f}(X) - f_\star\right] \quad \text{and} \quad \mathcal{I}(\pi, \xi) = \mathbb{E}\left[(f(X) - \bar{f}(X))^2\right] ,$$

which are both non-negative. The quantity $\Delta(\pi, \xi)$ is the regret suffered by π when the loss function is sampled from ξ, while $\mathcal{I}(\pi, \xi)$ is a measure of the observed variation of the loss function. Intuitively, if $\mathcal{I}(\pi, \xi)$ is large, then the learner is gaining information. A classical version of the information ratio is $\Delta(\pi, \xi)^2/\mathcal{I}(\pi, \xi)$, though the most standard version replaces $\mathcal{I}(\pi, \xi)$ with an expected relative entropy term that is never smaller than $\mathcal{I}(\pi, \xi)$ [RV16]. Given a distribution ξ and a random function f with law ξ, we let $\pi^\xi_{\text{TS}} \in \mathscr{P}(K)$ be the law of x_f, which is the minimiser of f. The minimax generalised information ratio associated with TS on class of loss functions $\mathscr{F}$ is

$$\text{IR}(\mathscr{F}) = \left\{(\alpha, \beta) \in \mathbb{R}^2_+ : \sup_{\xi \in \mathscr{P}(\mathscr{F})} \left[\Delta(\pi^\xi_{\text{TS}}, \xi) - \alpha - \sqrt{\beta \mathcal{I}(\pi^\xi_{\text{TS}}, \xi)}\right] \leq 0\right\} .$$

Note that $(0, \beta) \in \text{IR}(\mathscr{F})$ is equivalent to $\Delta(\pi^\xi_{\text{TS}}, \xi)^2/\mathcal{I}(\pi^\xi_{\text{TS}}, \xi) \leq \beta$ for all $\xi \in \mathscr{P}(\mathscr{F})$. The α term is used to allow a small amount of slack that eases analysis and may even be essential in non-parametric and/or infinite-action settings.

Theorem 3. *Suppose that $\mathscr{F} \in \{\mathscr{F}_{\texttt{bl}}, \mathscr{F}_{\texttt{blrm}}\}$ and $(\alpha, \beta) \in \text{IR}(\mathscr{F})$. Then, for any prior $\xi \in \mathscr{P}(\mathscr{F})$,* *the regret of* TS *(Algorithm 1) is at most*

$$\mathit{BReg}_n(\mathrm{TS}, \xi) \le n\alpha + O\left(\sqrt{\beta n d \log(n \operatorname{diam}(K))}\right) ,$$

where the Big-O hides only a universal constant.

This theorem is a direct consequence of Theorem 13, so we defer the proof to Section 5.3. At a high level the argument is based on similar results by [BDKP15] and [BE18]. Also note that the space of ridge functions is not closed under convex combinations, which introduces certain challenges also noticed by [Lat21]. To address this last issue, we introduce a cover in Section 5.1 that let us work with subsets of $\mathscr{F}$ that are closed under convex combinations, and also satisfy some other properties.

4.1 Decomposition Lemma

We also introduce a new mechanism for deriving information ratio bounds specially for TS. The rough idea is to partition the function class $\mathscr{F}$ into disjoint subsets $\mathscr{F}_i$ such that an inequality similar to that of general information ratio holds for each partition.

Lemma 4. *Suppose there exist natural numbers k and m such that for all $\bar{f} \in \operatorname{conv}(\mathscr{F})$ there exists a disjoint union $\mathscr{F} = \cup_{i=1}^m \mathscr{F}_i$ of measurable sets for which*

$$\max_{i\in[m]} \left[\sup_{f\in\mathscr{F}_i} (\bar{f}(x_f) - f_\star) - \alpha - \sqrt{\beta \inf_{f_1,\dots,f_k\in\mathscr{F}_i} \sum_{j,l\in\mathrm{PAIR}(k)} (f_j(x_{f_l}) - \bar{f}(x_{f_l}))^2} \right] \le 0 \,.$$

Then $(\alpha, k(k-1)m\beta) \in \mathrm{IR}(\mathscr{F})$.

Let us pause for a moment to provide some intuition. The supremum term is the worst possible regret within $\mathscr{F}_i$ while the infimum represents a kind of bound on the minimum amount of information obtained by TS. In particular, TS plays the optimal action for some sampled loss and gains information when there is variation of the losses at that point. The appearance of m in the information ratio bound arises from a Cauchy-Schwarz (what else?) that is somehow the 'same' Cauchy-Schwarz used in the analysis of the information ratio for finite-armed bandits [RV14] and in the information ratio decomposition by [Lat20].

Proof of Lemma 4. Let $\xi \in \mathscr{P}(\mathscr{F})$ and $\bar{f} = \mathbb{E}[f]$ and $\mathscr{F}_1, \ldots, \mathscr{F}_m$ be disjoint subsets of $\mathscr{F}$ such that $\mathscr{F} = \cup_{i=1}^m \mathscr{F}_i$ and

$$\max_{i \in [m]} \left[\sup_{f \in \mathscr{F}_i} \left(\bar{f}(x_f) - f_\star\right) - \alpha - \sqrt{\beta \inf_{f_1,\ldots,f_k \in \mathscr{F}_i} \sum_{j,l \in \text{PAIR}(k)} (\bar{f}_j(x_{f_l}) - f(x_{f_l}))^2}\right] \leq 0\,, \tag{4.1}$$

which exists by the assumptions in the lemma. When $\xi(\mathscr{F}_i) = 0$ define ν_i as an arbitrary probability measure on $\mathscr{F}$ and otherwise let $\nu_i(\cdot) = \xi(\cdot \cap \mathscr{F}_i)/\xi(\mathscr{F}_i)$ and $w_i = \xi(\mathscr{F}_i)$. Therefore

$$\begin{aligned}
\Delta(\pi, \xi) &= \int_{\mathscr{F}} \left(\bar{f}(x_f) - f_\star\right) \mathrm{d}\xi(f) \\
&= \sum_{i=1}^m w_i \int_{\mathscr{F}} \left(\bar{f}(x_f) - f_\star\right) \mathrm{d}\nu_i(f) \\
&\overset{\texttt{(a)}}{\leq} \sum_{i=1}^m w_i \sup_{f \in \mathscr{F}_i} (\bar{f}(x_f) - f_\star) \\
&\overset{\texttt{(b)}}{\leq} \alpha + \sum_{i=1}^m w_i \sqrt{\beta \inf_{f_1,\ldots,f_k \subset \mathscr{F}_i} \sum_{j,l \in \text{PAIR}(k)} (f_j(x_{f_l}) - \bar{f}(x_{f_l}))^2} \\
&\overset{\texttt{(c)}}{\leq} \alpha + \sum_{i=1}^m w_i \sqrt{\beta k(k-1) \int_{\mathscr{F}} \int_{\mathscr{F}} (\bar{f}(x_g) - f(x_g))^2 \,\mathrm{d}\nu_i(f)\,\mathrm{d}\nu_i(g)} \\
&\overset{\texttt{(d)}}{\leq} \alpha + \sqrt{\beta m k(k-1) \sum_{i=1}^m w_i^2 \int_{\mathscr{F}} \int_{\mathscr{F}} (\bar{f}(x_g) - f(x_g))^2 \,\mathrm{d}\nu_i(f)\,\mathrm{d}\nu_i(g)} \\
&\overset{\texttt{(e)}}{\leq} \alpha + \sqrt{\beta m k(k-1) \sum_{i=1}^m \sum_{j=1}^m w_i w_j \int_{\mathscr{F}} \int_{\mathscr{F}} (\bar{f}(x_g) - f(x_g))^2 \,\mathrm{d}\nu_i(f)\,\mathrm{d}\nu_j(g)} \\
&= \alpha + \sqrt{\beta m k(k-1) \int_{\mathscr{F}} \int_{\mathscr{F}} (\bar{f}(x_g) - f(x_g))^2 \,\mathrm{d}\xi(f)\,\mathrm{d}\xi(g)} \\
&= \alpha + \sqrt{\beta m k(k-1) \mathcal{I}(\pi, \xi)}\,,
\end{aligned} \tag{4.2}$$

where `(a)` is immediate from the definition of the integral. `(b)` follows from Eq. (4.1). `(c)` is true because if $f_1, \ldots, f_k$ are sampled independently from ν_i, then

$$\begin{aligned}\int_{\mathscr{F}} \int_{\mathscr{F}} (\bar{f}(x_g) - f(x_g))\, \mathrm{d}\nu_i(f)\, \mathrm{d}\nu_i(g) &= \frac{1}{k(k-1)} \mathbb{E}\left[\sum_{j,l \in \text{PAIR}(k)} (\bar{f}(x_{f_l}) - f_j(x_{f_l}))^2\right] \\ &\geq \frac{1}{k(k-1)} \inf_{f_1,\ldots,f_k \subset \mathscr{F}_i} \sum_{j,l \in \text{PAIR}(k)} (\bar{f}(x_{f_l}) - f_j(x_{f_l}))^2 \,.\end{aligned}$$

`(d)` follows from Cauchy-Schwarz and `(e)` by introducing additional non-negative terms. Since Eq. (4.2) holds for all $\xi \in \mathscr{P}(\mathscr{F})$ it follows that

$$\sup_{\xi \in \mathscr{P}(\mathscr{F})} \left[\Delta(\pi, \xi) - \alpha - \sqrt{\beta m k(k-1) \mathcal{I}(\pi, \xi)}\right] \leq 0$$

and therefore $(\alpha, \beta m k(k-1)) \in \text{IR}(\xi)$. □

Chapter 5

Approximate Thompson Sampling

Often *exact* minimization a convex function is computationally expensive. Therefore, it is natural to consider approximate minimization. In fact, we analyze this approximate version of TS, which is a strict generalization of the exact version. Later, we specialize the analysis of this approximate version, which we call approximate Thompson sampling (ATS), to get regret bounds for the exact version of TS. ATS is defined in Algorithm 2 and is similar to TS except that it only approximately minimizes the sampled loss function, i.e. X_t only needs to approximately minimize f_t. The analysis of this algorithm is surprisingly subtle, and indeed, we were only able to analyze an approximate version of TS that uses a small amount of regularization.

```
args: prior $\xi$
for $t = 1$ to $\infty$:
    sample $f_t$ from $\mathbb{P}(f = \cdot | X_1, Y_1, \ldots, X_{t-1}, Y_{t-1})$
    play $X_t \in \bar{x}_{f_t}$
    observe $Y_t$
```

Algorithm 2: Approximate Thompson sampling

Definition 5. Let $\epsilon_O \leq \epsilon_R$ be non-negative constants called the optimization accuracy and regularization parameter, respectively. Given $f \in \mathscr{F}_1$ let $\tilde{f}(x) = f(x) + \frac{\epsilon_R}{2} \|x\|^2$ when $\epsilon_R > 0$ define

$$\tilde{x}_f = \operatorname*{arg\,min}_{x \in K} \tilde{f}(x) \qquad\qquad \bar{x}_f = \left\{ x : \tilde{f}(x) \leq \min_{y \in K} \tilde{f}(y) + \epsilon_O \right\} .$$

When the regularization parameter $\epsilon_R = 0$, define $\tilde{x}_f = x_f$ and $\bar{x}_f = \{x_f\}$.

When $\epsilon_R = \epsilon_O = 0$, then ATS and TS are equivalent, though we note the importance in our analysis that the ties in TS are broken in a consistent fashion. The regularization in the definition of $\tilde{f}$ ensures that all points in $\bar{x}_f$ are reasonably close to $\tilde{x}_f$ and introduces a degree of stability into ATS. An obvious question is whether or not you could do away with the regularization and define $\bar{x}_f$ by $\{x : f_t(x) \leq f_{t\star} + \epsilon\}$ for suitably small $\epsilon \geq 0$. We suspect the answer is yes but do not currently have a proof. The regularization ensures that $\bar{x}_f$ has small diameter, which need not be true in general for $\{x : f(x) \leq f_\star + \epsilon\}$, even if ϵ is arbitrarily small.

Remark 6. It's also important to note that x_f is not $\bar{x}_f$ necessarily.

5.1 A Convex Cover

We start by defining a kind of cover of a set of convex functions $\mathscr{F}$. In the standard analysis introduced by [BDKP15] and [BE18], this cover was defined purely in terms of the optimal action. As noticed by [Lat21], this argument relies on $\mathscr{F}$ being closed under convex combinations, which is not true for the space of ridge functions. Here we introduce a new notion of cover for function classes $\mathscr{F}$ that are not closed under convex combinations.

Definition 7. Let $\mathscr{F}$ be a set of convex functions from $\mathcal{K}$ to $\mathbb{R}$ and $\epsilon > 0$. Define $N(\mathscr{F}, \epsilon)$ to be the smallest number N such that there exists $\{\mathscr{F}_1, \ldots, \mathscr{F}_N\}$ such that for all $k \in [N]$:

- *Closure:* $\mathscr{F}_k$ is a subset of $\mathscr{F}$ and $\operatorname{conv}(\mathscr{F}_k) \subset \mathscr{F}$.
- *Common near-minimiser:* There exists an $x_k \in K$ such that $\|\tilde{x}_f - x_k\| \leq \epsilon$ for all $f \in \mathscr{F}_k$.

Moreover:

- *Approximation:* For all $f \in \mathscr{F}$ there exists a $k \in [N]$ and $g \in \mathscr{F}_k$ such that $\|f - g\|_\infty \leq \epsilon$ and $\|\tilde{x}_f - x_k\| \leq \epsilon$.

We now bound the covering number $N(\mathscr{F}, \epsilon)$ for function classes $\mathscr{F}_{\texttt{bl}}$ and $\mathscr{F}_{\texttt{blrm}}$. The former class is closed under convex combinations, which somewhat simplifies the situation.

Proposition 8. *Suppose that* $\mathscr{F} = \mathscr{F}_{\texttt{bl}}$. *Then* $\log N(\mathscr{F}, \epsilon) = O\left(d \log\left(\frac{\operatorname{diam}(K)}{\epsilon}\right)\right)$.

Proof. Let $\mathcal{C}_K$ be a finite subset of K such that for all $x \in K$ there exists a $y \in \mathcal{C}_K$ with $\|x - y\| \leq \epsilon$. Standard bounds on covering numbers [AAGM15, §4] show that $\mathcal{C}_K$ can be chosen so that

$$|\mathcal{C}_K| \leq \left(1 + \frac{2\operatorname{diam}(K)}{\epsilon}\right)^d .$$

Given $x \in \mathcal{C}_K$ define $\mathscr{F}_x = \{f \in \mathscr{F} : \|\tilde{x}_f - x\| \leq \epsilon\}$. Since $\operatorname{conv}(\mathscr{F}) = \mathscr{F}$ it follows trivially that $\operatorname{conv}(\mathscr{F}_x) \subset \operatorname{conv}(\mathscr{F}) = \mathscr{F}$. The common near minimiser property is satisfied automatically by definition. Suppose that $f \in \mathscr{F}$ is arbitrary and let $x \in \mathcal{C}_K$ be such that $\|x - \tilde{x}_f\| \leq \epsilon$, which exists by construction. Therefore $f \in \mathscr{F}_x$ and the approximation property also holds. □

Proposition 9. *Suppose that* $\mathscr{F} = \mathscr{F}_{\mathtt{blrm}}$. *Then* $\log N(\mathscr{F}, \epsilon) = O\left(d \log\left(\frac{\operatorname{diam}(K)}{\epsilon}\right)\right)$.

Proof. To begin, define $\epsilon_{\mathbb{S}} = \epsilon / \operatorname{diam}(K)$. Given a ridge function $f \in \mathscr{F}$, let $\theta_f \in \mathbb{S}_1$ be a direction such that $f(\cdot) = u(\langle \theta, \cdot \rangle)$ for some convex function u. Given $x \in K$ and $\theta \in \mathbb{S}_1$ let

$$\mathscr{F}_{x,\theta} = \{f \in \mathscr{F} : \|\tilde{x}_f - x\| \leq \epsilon \text{ and } \theta_f = \theta\} .$$

Note that $\{f \in \mathscr{F} : f_\theta = \theta\}$ is convex and hence $\operatorname{conv}(\mathscr{F}_{x,\theta}) \subset \mathscr{F}$ holds. Let $\mathcal{C}_{\mathbb{S}}$ be a finite subset of $\mathbb{S}_1$ such that for all $\theta \in \mathbb{S}_1$ there exists an $\eta \in \mathcal{C}_{\mathbb{S}}$ for which $\|\theta - \eta\| \leq \epsilon_{\mathbb{S}}$. Similarly, let $\mathcal{C}_K$ be a finite subset of K such that for all $x \in K$ there exists a $y \in \mathcal{C}_K$ with $\|x - y\| \leq \epsilon$. Classical covering number results [AAGM15, §4] show that $\mathcal{C}_{\mathbb{S}}$ and $\mathcal{C}_K$ can be chosen so that

$$|\mathcal{C}_{\mathbb{S}}| \leq \left(1 + \frac{4}{\epsilon_{\mathbb{S}}}\right)^d \qquad\qquad |\mathcal{C}_K| \leq \left(1 + \frac{2\operatorname{diam}(K)}{\epsilon}\right)^d .$$

Consider the collection $\{\mathscr{F}_{x,\theta} : x \in \mathcal{C}_K, \theta \in \mathcal{C}_{\mathbb{S}}\}$, which has size $N = |\mathcal{C}_K||\mathcal{C}_{\mathbb{S}}|$. Let $f \in \mathscr{F}$ be arbitrary and let $\theta \in \mathcal{C}_{\mathbb{S}}$ and $x \in \mathcal{C}_K$ be such that $\|\theta - \theta_f\| \leq \delta$ and $\|x - \tilde{x}_f\| \leq \epsilon$. Then define $g = u_f(\langle \cdot, \theta \rangle) \in \mathscr{F}$, which satisfies

$$\|f - g\|_\infty = \sup_{x \in K} |u_f(\langle x, \theta \rangle) - u_f(\langle x, \theta_f \rangle)| \leq \sup_{x \in K} |\langle x, \theta - \theta_f \rangle| \leq \epsilon_{\mathbb{S}} \operatorname{diam}(K) \leq \epsilon .$$

Therefore the approximation property holds. □

5.2 Continuity of Regret and Information Gain

In order to find a pair (α, β) in $\mathrm{IR}(\mathscr{F})$, we need to bound the regret $\Delta(\pi, \xi)$ in terms of the information gain $I(\pi, \xi)$ for a prior $\xi \in \mathscr{P}(\mathscr{F})$ and policy $\pi \in \mathscr{P}(K)$. It turns out useful to prove continuity (Lipschitzness) properties of these quantities in terms of the distance between two different priors $\xi, \nu \in \mathscr{P}(\mathscr{F})$ and the distance between two different policies $\pi, \rho \in \mathscr{P}(K)$. Of course, the proper *distance* metric on $\mathscr{P}(\mathscr{F})$ and $\mathscr{P}(K)$ needs to be specified to make this precise.

Lemma 10. *Suppose f and g are random elements in $\mathscr{F}$ with laws ξ and ν and that $X, Y \in K$ are independent of f and g and have laws π and ρ. Suppose that $\|f - g\|_\infty \leq \epsilon$ almost surely and $\|X - Y\| \leq \epsilon$ almost surely. Then*

(a) $I(\pi, \nu)^{1/2} \leq I(\pi, \xi)^{1/2} + \epsilon.$

(b) $I(\pi, \nu)^{1/2} \leq I(\rho, \nu)^{1/2} + \epsilon.$

Proof. For random variable X let $\|X\|_{\mathrm{L2}} = \mathbb{E}[X^2]^{1/2}$, which is a norm on the space of square integrable random variables on some probability space with suitable a.s. identification. Let $\bar{f} = \mathbb{E}[f]$ and $\bar{g} = \mathbb{E}[g]$. By definition $I(\pi, \xi)^{1/2} = \|f(X) - \bar{f}(X)\|_{\mathrm{L2}}$ and $I(\pi, \nu)^{1/2} = \|g(X) - \bar{g}(X)\|_{\mathrm{L2}}$. The first claim follows since $\|\cdot\|_{\mathrm{L2}}$ is a norm. The second claim follows in the same manner and using the fact that $f, g, \bar{f}, \bar{g}$ are Lipschitz. □

Lemma 11. *Suppose that $\alpha, \beta \in \mathrm{IR}(\mathscr{F})$. Suppose that X and f are (possibly dependent) random elements with laws $\pi \in \mathscr{P}(K)$ and $\nu \in \mathscr{P}(\mathscr{F})$ and such that $f(X) \leq f_\star + \epsilon$ almost surely. Then*

$$\Delta(\pi, \nu) \leq \alpha + \sqrt{\beta I(\pi, \nu)} + \epsilon \left[1 + \sqrt{\beta}\right] .$$

Proof. Let $g(x) = \max(f(x), f(X))$ and ξ be the law of g, which means that $\pi \in \mathrm{TS}(\xi)$. As usual, let $\bar{f} = \mathbb{E}[f]$ and $\bar{g} = \mathbb{E}[g]$. By construction

$$\begin{aligned}
\|f - g\|_\infty &= \sup_{x \in \mathcal{K}} |\max(f(x), f(X)) - f(x)| \\
&= \sup_{x \in \mathcal{K}} \max(f(x), f(X)) - f(x) \\
&\leq \sup_{x \in \mathcal{K}} \max(f(x), f(x) + \epsilon) - f(x) \\
&\leq \epsilon \,,
\end{aligned}$$

almost surely. Therefore

$$\Delta(\pi, \nu) = \mathbb{E}[\bar{f}(X) - f_\star] \leq \mathbb{E}[\bar{g}(X) - g_\star] + \epsilon = \Delta(\pi, \xi) + \epsilon \leq \alpha + \sqrt{\beta I(\pi, \xi)} + \epsilon\,.$$

The result now follows from Lemma 10. □

The next lemma establishes basic properties of the regularised minimisers $\tilde{x}_f$ and $\bar{x}_f$, which are defined in Definition 5. Remember that ϵ_R is the amount of regularisation. Larger values make $\tilde{x}_f$ more stable but also a worse approximation of x_f. The approximation error in the definition of $\bar{x}_f$ is ϵ_O, which can be chosen extremely small.

Lemma 12. *Suppose that $f, g \in \mathscr{F}$. Then*

(a) $\sup\{\|\tilde{x}_f - y\| : y \in \bar{x}_f\} \leq \sqrt{2\epsilon_O/\epsilon_R}$ *with* $0/0 \triangleq 0$.

(b) $f(\tilde{x}_f) \leq f_\star + \frac{\epsilon_R}{2}\operatorname{diam}(K)^2$.

Proof. Note the special case that $\epsilon_R = \epsilon_O = 0$, then $\bar{x}_f = \{\tilde{x}_f\}$ by definition and (a) is immediate. Otherwise let $\tilde{f}(x) = f(x) + \frac{\epsilon_R}{2}\|x\|^2$ and $x = \tilde{x}_f$ and $y \in \bar{x}_f$. Then

$$\tilde{f}(\tilde{x}_f) + \epsilon_O \geq \tilde{f}(y) \geq \tilde{f}(\tilde{x}_f) + D\tilde{f}(\tilde{x}_f)[y - \tilde{x}_f] + \frac{\epsilon_R}{2}\|\tilde{x}_f - y\|^2 \geq \tilde{f}(\tilde{x}_f) + \frac{\epsilon_R}{2}\|\tilde{x}_f - y\|^2\,.$$

Rearranging completes the proof of the first part. For (b), let $y \in K$ be arbitrary

$$f(\tilde{x}_f) + \frac{\epsilon_R}{2}\|x\|^2 \leq f(y) + \frac{\epsilon_R}{2}\|y\|^2\,.$$

And the result follows since $\|y\|^2 - \|x\|^2 \leq \operatorname{diam}(K)^2$. □

5.3 A Regret Bound in Terms of the Information Ratio

We can now state a general theorem from which Theorem 3 follows.

Theorem 13. *Suppose that $\epsilon \in (0, 1)$ and $\frac{1}{2}\epsilon_R \operatorname{diam}(K)^2 \leq \epsilon$ and $2\epsilon_O/\epsilon_R \leq \epsilon^2$ and let $\mathscr{F}$ be a set of convex functions from K to $[0, 1]$ and $(\alpha, \beta) \in \text{IR}(\mathscr{F})$. Then the Bayesian regret of* ATS *for any prior ξ is at most*

$$\mathit{BReg}_n(\text{ATS}, \xi) \leq n\alpha + 3n\epsilon[1 + \sqrt{\beta}] + \sqrt{\frac{\beta n}{2}\log\left(N\left(\mathscr{F}, 1/\epsilon\right)\right)}\,.$$

Theorem 3 follows by choosing $\epsilon = 1/n$ and $\epsilon_R = \epsilon_O = 0$ and by Proposition 8 and Proposition 9 to bound the covering numbers for the relevant classes.

Proof. Let $N = N(\mathscr{F}, \epsilon)$ and $\mathscr{F}_1, \ldots, \mathscr{F}_N$ be a collection of subset of $\mathscr{F}$ satisfying the conditions of Definition 7. Hence, there exists a sequence $x_1, \ldots, x_N$ such that for all $k \in [N]$ and $f \in \mathscr{F}_k$,

$$\|\tilde{x}_f - x_k\| \leq \epsilon\,.$$

Let $\xi \in \mathscr{P}(\mathscr{F})$ be any prior. Suppose that f is sampled from ξ and $X_\star$ be a random element in K such that $X_\star \in \bar{x}_f$ and let X be an independent copy of $X_\star$ and Y be a random variable such that $\mathbb{E}[Y|f, X] = f(X)$ and $Y \in [0, 1]$ almost surely. By Definition 7 there exists an $[N]$-valued random variable κ such that:

(i) There exists a random function $f_\kappa \in \mathscr{F}_\kappa$ with $\|f - f_\kappa\|_\infty \leq \epsilon$; and

(ii) $\|\tilde{x}_f - x_\kappa\| \leq \epsilon$.

Define π as the law of X, which is a (approximate) Thompson sampling policy for ξ.

Lemma 14. *The following holds:*

$$\Delta(\pi, \xi) \leq \alpha + \sqrt{\beta I(\kappa; X, Y)} + 3\epsilon[1 + \sqrt{\beta}]\,,$$

where $I(\kappa; X, Y)$ is the mutual information between κ and the pair (X, Y).

Proof. Let ν be the law of $\mathbb{E}[f|\kappa]$ and ν_κ be the law of $\mathbb{E}[f_\kappa|\kappa]$ and π_κ be the law of x_κ. Let

$\bar{f}_\kappa = \mathbb{E}[f_\kappa]$. Then

$$\begin{aligned}
\Delta(\pi, \xi) &= \mathbb{E}[\bar{f}(X) - f_\star] \\
&\overset{\texttt{(a)}}{=} \mathbb{E}[\bar{f}(X) - \mathbb{E}[f_\star|\kappa]] \\
&\overset{\texttt{(b)}}{\leq} \mathbb{E}[\bar{f}_\kappa(X) - \mathbb{E}[f_\star|\kappa]] + \epsilon \\
&\overset{\texttt{(c)}}{\leq} \mathbb{E}[\bar{f}_\kappa(x_\kappa) - \mathbb{E}[f_\kappa|\kappa]_\star] + 3\epsilon \\
&\overset{\texttt{(d)}}{=} \Delta(\pi_\kappa, \nu_\kappa) + 3\epsilon \\
&\overset{\texttt{(e)}}{\leq} \alpha + \sqrt{\beta I(\pi_\kappa, \nu_\kappa)} + 3\epsilon \\
&\overset{\texttt{(f)}}{\leq} \alpha + \sqrt{\beta I(\pi_\kappa, \nu)} + \epsilon[3 + \sqrt{\beta}] \\
&\overset{\texttt{(g)}}{\leq} \alpha + \sqrt{\beta I(\pi, \nu)} + 3\epsilon[1 + \sqrt{\beta}] \\
&\overset{\texttt{(h)}}{\leq} \alpha + \sqrt{\frac{\beta I(\kappa; X, Y)}{2}} + 3\epsilon[1 + \sqrt{\beta}]\,,
\end{aligned}$$

where

(a) by the tower rule.

(b) follows since $\|f_\kappa - f\|_\infty \leq \epsilon$ by definition and by convexity of $\|\cdot\|_\infty$,

$$\|\bar{f} - \bar{f}_\kappa\|_\infty = \|\mathbb{E}[f] - \mathbb{E}[f_\kappa]\|_\infty \leq \mathbb{E}[\|f - f_\kappa\|_\infty] \leq \epsilon\,.$$

(c) follows because $\|f - f_\kappa\|_\infty \leq \epsilon$ and $\|\tilde{x}_f - x_\kappa\| \leq \epsilon$ so that

$$\begin{aligned}
\mathbb{E}[f_\star|\kappa] &= \mathbb{E}[f(x_f)|\kappa] \\
&\geq \mathbb{E}[f(\tilde{x}_f)|\kappa] - \frac{\epsilon_R}{2}\operatorname{diam}(K)^2 && \text{by Lemma 12(b)} \\
&\geq \mathbb{E}[f(x_\kappa)|\kappa] - 2\epsilon && \text{by the assumption on } \epsilon_R \text{ and (ii)} \\
&\geq \mathbb{E}[f_\kappa(x_\kappa)|\kappa] - 3\epsilon && \text{by (i)} \\
&= \mathbb{E}[f_\kappa|\kappa]_\star - 3\epsilon\,.
\end{aligned}$$

And because by the triangle inequality, the definition of $X_\star$ and Lemma 12(a),

$$\|X_\star - x_\kappa\| \leq \|X_\star - \tilde{x}_f\| + \|\tilde{x}_f - x_\kappa\| \leq \sqrt{2\epsilon_O/\epsilon_R} + \epsilon \leq 2\epsilon\,, \tag{5.1}$$

which implies that $\mathbb{E}[\bar{f}_\kappa(X)] = \mathbb{E}[\bar{f}_\kappa(X_\star)] \leq \mathbb{E}[\bar{f}_\kappa(x_\kappa)] + 2\epsilon$.

`(d)` follows by definition.

`(e)` follows by $(\alpha, \beta) \in \text{IR}(\mathcal{F})$.

`(f)` follows from Lemma 10 and because

$$\|\mathbb{E}[f_\kappa|\kappa] - \mathbb{E}[f|\kappa]\|_\infty \leq \mathbb{E}[\|f_\kappa - f\|_\infty] \leq \epsilon\,.$$

`(g)` follows from Lemma 10 and Eq. (5.1).

`(h)` follows from Pinsker's inequality. Let $\mathrm{KL}(\cdot, \cdot)$ be the relative entropy. Then

$$\begin{aligned}\mathcal{I}(\pi, \nu) &= \mathbb{E}[(\mathbb{E}[f(X)|X] - \mathbb{E}[f(X)|\kappa, X])^2] \\ &= \mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y|\kappa, X])^2] \\ &\leq \frac{1}{2}\mathbb{E}[\mathrm{KL}(\mathbb{P}_{Y|X}, \mathbb{P}_{Y|X,\kappa})] \\ &= \frac{1}{2}I(\kappa; X, Y)\end{aligned}$$

This concludes the explanation of the steps and so the proof of the lemma. □

We are now in a position to prove Theorem 13. Let π_t be the law of X_t under $\mathbb{P}_{t-1}$ and ξ_t be the law of f under $\mathbb{P}_{t-1}$. By Lemma 14,

$$\Delta(\pi_t, \xi_t) \leq \alpha + \sqrt{\beta I_t(\kappa; X_t, Y_t)} + 3\epsilon[1 + \sqrt{\beta}]\,.$$

Hence, letting I_t be the mutual information with respect to probability measure $\mathbb{P}_t$,

$$\begin{aligned}\mathrm{BReg}_n(\mathrm{ATS}, \xi) &= \mathbb{E}\left[\sum_{t=1}^n \Delta(\pi_t, \xi_t)\right] \\ &\leq n\alpha + \mathbb{E}\left[\sum_{t=1}^n \sqrt{\beta I_{t-1}(\kappa; X_t, Y_t)}\right] + 3n\epsilon[1 + \sqrt{\beta}] \\ &\leq n\alpha + \sqrt{\beta n \mathbb{E}\left[\sum_{t=1}^n I_{t-1}(\kappa; X_t, Y_t)\right]} + 3n\epsilon[1 + \sqrt{\beta}] \\ &\leq n\alpha + \sqrt{\beta n \log(N)} + 3n\epsilon[1 + \sqrt{\beta}]\,,\end{aligned}$$

where the final inequality holds by the chain rule for the mutual information and because $\kappa \in [N]$ and hence its entropy is at most $\log(N)$. □

Chapter 6

Thompson Sampling in 1-dimension

Our first main theorem shows that TS is statistically efficient when the loss is bounded and Lipschitz and $d = 1$.

Theorem 15. *When* $d = 1$, $\sup_{\xi \in \mathscr{P}(\mathscr{F}_{b1})} \mathit{BReg}_n(\mathrm{TS}, \xi) = O\left(\sqrt{n \log(n) \log(n \operatorname{diam}(K))}\right)$.

Theorem 15 is established by combining the following bound on the information ratio and $\alpha = 1/n$ with Theorem 3.

Theorem 16. *Suppose that* $d = 1$ *and* $\alpha \in (0, 1)$*. Then* $(\alpha, 10^4 \lceil \log(1/\alpha) \rceil) \in \mathrm{IR}(\mathscr{F}_{b1})$.

Proof. Let $\bar{f} \in \operatorname{conv}(\mathscr{F}_{\text{b1}})$ and for integer i define

$$\mathscr{F}_i = \begin{cases} \{f \in \mathscr{F}_{\text{b1}} : \bar{f}(x_f) - f_\star \in (\alpha 2^{|i|-1}, \alpha 2^{|i|}],\, x_f \geq x_{\bar{f}}\} & \text{if } i > 0 \\ \{f \in \mathscr{F}_{\text{b1}} : \bar{f}(x_f) - f_\star \in (\alpha 2^{|i|-1}, \alpha 2^{|i|}],\, x_f < x_{\bar{f}}\} & \text{if } i < 0 \\ \{f \in \mathscr{F}_{\text{b1}} : \bar{f}(x_f) - f_\star \leq \alpha\} & \text{if } i = 0\,. \end{cases} \tag{6.1}$$

Since the losses in $\mathscr{F}_{\text{b1}}$ are bounded by assumption, for $|i| > m = \lceil \log_2(1/\alpha) \rceil$, $\mathscr{F}_i = \emptyset$ so that $\mathscr{F}_{\text{b1}} = \cup_{i=-m}^m \mathscr{F}_i$. In a moment we will show that with $k = 4$ and $-m \leq i \leq m$ and $\epsilon = \alpha 2^{|i|}$ that

$$\sup_{f \in \mathscr{F}_i} \left(\bar{f}(x_f) - f_\star\right) \leq \epsilon \leq \alpha + \sqrt{230 \inf_{f_1, \ldots, f_k \in \mathscr{F}_i} \sum_{j,l \in \mathrm{PAIR}(k)} (f_j(x_{f_l}) - \bar{f}(x_{f_l}))^2}\,. \tag{6.2}$$

Hence, by Lemma 4 and naive simplification of constants $(\alpha, 10^4 \lceil \log(1/\alpha) \rceil) \in \mathrm{IR}(\mathscr{F}_{\text{b1}})$ as desired. The first inequality in Eq. (6.2) is an immediate consequence of the definition of $\mathscr{F}_i$ and

ϵ. The second is also immediate when $i = 0$. The situation when $i < 0$ and $i > 0$ is symmetric, so for the remainder we prove that the second inequality in Eq. (6.2) holds for any $i > 0$. Let $f_1, \ldots, f_k \in \mathscr{F}_i$ and for $j \in [4]$ let $x_j = x_{f_j}$ and assume without loss of generality that $x_1 \leq x_2 \leq x_3 \leq x_4$. Suppose that

$$\sum_{j,l \in \textsc{pair}(4)} (f_j(x_{f_l}) - \bar{f}(x_{f_l}))^2 < c^2\epsilon^2 \qquad \text{with} \qquad c = \frac{\sqrt{65} - 7}{16} . \tag{6.3}$$

Let us now establish a contradiction, which we do in three steps. The main argument in each step is illustrated in Figure 6.1.

STEP 1 We start by showing that x_2 must be somewhat closer to x_3 than to x_1.

$$f_1(x_3) \overset{\text{(a)}}{\leq} \bar{f}(x_3) + \epsilon c \overset{\text{(b)}}{\leq} f_3(x_3) + \epsilon[c+1] \leq f_3(x_1) + \epsilon[c+1] \overset{\text{(c)}}{\leq} \bar{f}(x_1) + \epsilon[2c+1] ,$$

where `(a)` follows from Eq. (6.3), `(b)` since for $f \in \mathscr{F}_i$, $f(x_f) \geq \bar{f}(x_f) - \epsilon$ and `(c)` from Eq. (6.3) again. Hence, with $p \in [0, 1]$ such that $x_2 = (1-p)x_1 + px_3$,

$$\begin{aligned} \bar{f}(x_1) \overset{\text{(a)}}{\leq} \bar{f}(x_2) \overset{\text{(b)}}{\leq} f_1(x_2) + c\epsilon &\overset{\text{(c)}}{\leq} (1-p)f_1(x_1) + pf_1(x_3) + c\epsilon \\ &\overset{\text{(d)}}{\leq} (1-p)(\bar{f}_1(x_1) - \epsilon/2) + pf_1(x_3) + c\epsilon \\ &\overset{\text{(e)}}{\leq} \bar{f}(x_1) + \epsilon\left[c + p[2c + 3/2] - 1/2\right] , \end{aligned}$$

where `(a)` follows because $\bar{f}$ is non-decreasing on $[x_1, x_4]$ by the definition of $\mathscr{F}_i$ and $i > 0$, `(b)` from Eq. (6.3), `(c)` by convexity and the definition of p, `(d)` since $f_1(x_1) \leq \bar{f}_1(x_1) - \epsilon/2$ by the definition of $\mathscr{F}_i$ and `(e)` is true by the previous display. Therefore $p \geq (1/2 - c)/(2c + 3/2) \approx 0.27$.

STEP 2 Having shown that x_2 is close to x_3, we now show that $f_3(x_3)$ is not much smaller than $\bar{f}(x_1)$. Indeed,

$$\begin{aligned} \bar{f}(x_1) \overset{\text{(a)}}{\leq} \bar{f}(x_2) &\overset{\text{(b)}}{\leq} f_3(x_2) + c\epsilon \\ &\overset{\text{(c)}}{\leq} (1-p)f_3(x_1) + pf_3(x_3) + c\epsilon \\ &\overset{\text{(d)}}{\leq} (1-p)\bar{f}(x_1) + pf_3(x_3) + 2c\epsilon , \end{aligned}$$

`(a)`-`(c)` follows as above in **STEP 1** and `(d)` from Eq. (6.3). Rearranging shows that $f_3(x_3) \geq \bar{f}(x_1) - \frac{2c\epsilon}{p}$.

STEP 3 Lastly we derive a contradiction using **STEP 1** and **STEP 2** since

$$\begin{aligned} f_4(x_3) &\overset{\text{(a)}}{\leq} f_4(x_1) \overset{\text{(b)}}{\leq} \bar{f}(x_1) + c\epsilon \\ &\overset{\text{(c)}}{\leq} f_3(x_3) + \epsilon\left[c + \frac{2c}{p}\right] \\ &\overset{\text{(d)}}{\leq} \bar{f}(x_3) + \epsilon\left[c + \frac{2c}{p} - \frac{1}{2}\right] \\ &\overset{\text{(e)}}{\leq} \bar{f}(x_3) - c\epsilon\,, \end{aligned}$$

where (a) follows by convexity and because f_4 is minimised at x_4, (b) from Eq. (6.3), (c) from **STEP 2**, (d) since $f_3(x_3) \leq \bar{f}(x_3) - \epsilon/2$ by the definition of $\mathscr{F}_i$ and (e) from the bound on p in **STEP 1** and the definition of c. But this contradicts Eq. (6.3). Hence Eq. (6.3) does not hold. And since $c^2 \geq \frac{1}{230}$ it follows that Eq. (6.2) holds. □

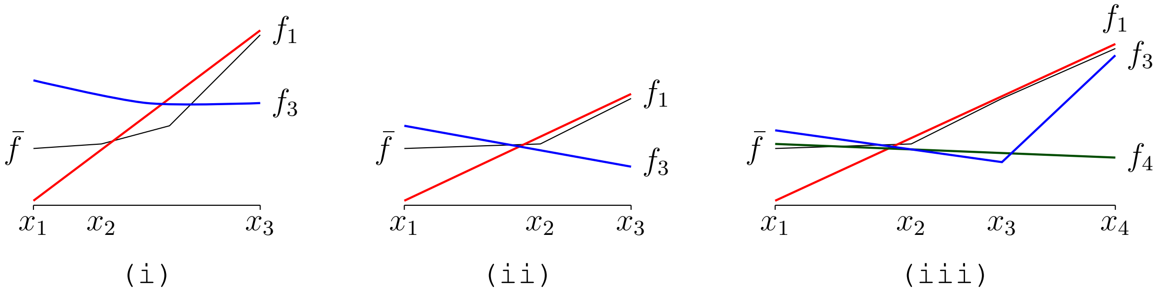


Figure 6.1: (i) shows that if x_2 is too close to x_1, then $f_1(x_3)$ must be large, which implies that $f_3(x_3)$ must be large and so too must $f_3(x_1)$, which shows that $f_3(x_1) - \bar{f}(x_1)$ is large. (ii) shows what happens if $f_3(x_3)$ is too far below $\bar{f}(x_1)$, which is that $f_3(x_1)$ must be much larger than $\bar{f}(x_1)$. (iii) shows that $f_4(x_3)$ cannot be much larger than $f_3(x_3)$ and therefore $\bar{f}(x_3) - f_4(x_3)$ must be large.

Chapter 7

Thompson Sampling for Ridge Functions

We now consider the multi-dimensional convex monotone ridge function setting where $\mathscr{F} = \mathscr{F}_{\texttt{blrm}}$

Theorem 17. $\sup_{\xi \in \mathscr{P}(\mathscr{F}_{\texttt{blrm}})} \mathit{BReg}_n(\mathrm{TS}, \xi) = O\left(d^{2.5}\sqrt{n}\log(nd\,\mathrm{diam}(K))^2\right)$.

[RV16] used information-theoretic means to show that for linear bandits the regret is at most $\tilde{O}(d\sqrt{n})$. [Lat21] showed that for (possibly non-monotone) convex ridge functions a version of IDS has Bayesian regret at most $\tilde{O}(d\sqrt{n})$. The downside is that IDS is barely implementable in practice, even given efficient access to posterior samples. Like Theorem 15, Theorem 17 is established by combining a bound on the information ratio with Theorem 3.

Theorem 18. $(\alpha, \beta\left\lceil\log(1/\alpha)\right\rceil) \in \mathrm{IR}(\mathscr{F}_{\texttt{blrm}})$ *whenever* $\alpha \in (0,1)$ *and*

$$\beta = O\left(d^4 \log\left(\frac{d\,\mathrm{diam}(K)}{\alpha}\right)^2\right).$$

with the Big-O hiding only a universal constant.

Proof. Abbreviate $\mathscr{F} = \mathscr{F}_{\texttt{blmr}}$. The high-level argument follows the proof of Theorem 16. The main challenge is lower bounding the quadratic (information gain) term that appears in Lemma 4, which uses an argument based on the method of inscribed ellipsoid for optimisation [TKE88]. Let $\bar{f} \in \mathrm{conv}(\mathscr{F})$. Given a nonempty finite set $\mathcal{C} \subset \mathscr{F}$ and $\delta > 0$, let $J_\delta(\mathcal{C}) = \mathrm{conv}\left(\cup_{g \in \mathcal{C}} \mathbb{B}_\delta(x_g)\right)$. Moreover, let $E_\delta(\mathcal{C})$ be the ellipsoid of maximum volume enclosed in $J_\delta(\mathcal{C})$, which is called John's ellipsoid. We now need two lemmas. The first shows that for a suitable subset $\mathcal{C}$ of loss functions either the information gain is reasonably large or some function f can be removed from $\mathcal{C}$ in such a way that $E_\delta(\mathcal{C} \setminus \{f\})$ is considerably smaller than $E_\delta(\mathcal{C})$.

Lemma 19. *Let $\epsilon > 0$, $\delta = \frac{\epsilon}{12(d+1)}$ and $\mathcal{C} \subset \mathscr{F}$ be a nonempty finite set such that for all $f, g \in \mathcal{C}$, $\bar{f}(x_f) - f_\star \in [\epsilon/2, \epsilon]$ and $|\bar{f}(x_f) - \bar{f}(x_g)| \leq \delta$. Then at least one of the following holds:*

`(i)` *There exists an $f \in \mathcal{C}$ such that $\operatorname{vol}(E_\delta(\mathcal{C} \setminus \{f\})) \leq 0.85 \operatorname{vol}(E_\delta(\mathcal{C}))$.*

`(ii)` *There exists a pair $f, g \in \text{PAIR}(\mathcal{C})$ such that $(f(x_g) - \bar{f}(x_g))^2 \geq \delta^2$.*

Proof. Let $\mu \in \mathbb{R}^d$ and Σ be positive definite such that $E_\delta(\mathcal{C}) = \{x : \|x - \mu\|_{\Sigma^{-1}} \leq 1\}$ and for $r > 0$, let $E_{\delta,r}(\mathcal{C}) = \{x : \|x - \mu\|_{\Sigma^{-1}} \leq r\}$. By John's theorem [AAGM15, Remark 2.1.17], $E_\delta(\mathcal{C}) \subset J_\delta(\mathcal{C}) \subset E_{\delta,d}(\mathcal{C})$. Let $f \in \mathcal{C}$ be arbitrary. By assumption there exists a convex monotone increasing ℓ and $\theta \in \mathbb{S}_1$ such that $f = \ell(\langle \cdot, \theta\rangle)$. Let $H = H(\mu, \theta)$, which is the half-space passing through the center of John's ellipsoid $E_\delta(\mathcal{C})$ with inward-facing normal θ. Consider the following cases, illustrated in Figure 7.1:

CASE 1 $\langle x_g, \theta\rangle \geq \langle \mu, \theta\rangle + \delta$ for all $g \in \mathcal{C} \setminus \{f\}$. In this case $J_\delta(\mathcal{C} \setminus \{f\}) \subset H \cap J_\delta(\mathcal{C})$ and therefore the inequality of [Kha90] shows that $\operatorname{vol}(E_\delta(\mathcal{C} \setminus \{f\}))) \leq 0.85 \operatorname{vol}(E_\delta(\mathcal{C}))$.

CASE 2 There exists a $g \in \mathcal{C} \setminus \{f\}$ such that such that $\langle x_g, \theta\rangle < \langle \mu, \theta\rangle + \delta$. Since $E_\delta(\mathcal{C}) \subset J_\delta(\mathcal{C})$, there exists an $x \in J_\delta(\mathcal{C})$ such that $\langle x, \theta\rangle \geq \langle \mu, \theta\rangle + \|\theta\|_\Sigma$. By the definition of $J_\delta(\mathcal{C})$ it follows that there there exists a $h \in \mathcal{C}$ such that $\langle x_h, \theta\rangle \geq \langle \mu, \theta\rangle + \|\theta\|_\Sigma - \delta$. Let $x'_h = x_h + \delta\theta$ and $x'_g = x_g - \delta\theta$ and $x'_f = x_f - \delta\theta$. Collecting the above facts, we have

$$\langle x'_f, \theta\rangle \geq \langle \mu, \theta\rangle - d\,\|\theta\|_\Sigma \qquad \text{(since } J_\delta(\mathcal{C}) \subset E_{\delta,d}(\mathcal{C})\text{)} \tag{7.1}$$

$$\langle x'_g, \theta\rangle \leq \langle \mu, \theta\rangle \tag{7.2}$$

$$\langle x'_h, \theta\rangle \geq \langle \mu, \theta\rangle + \|\theta\|_\Sigma\,. \qquad \text{(since } E_\delta(\mathcal{C}) \subset J_\delta(\mathcal{C})\text{)} \tag{7.3}$$

Since ℓ is nondecreasing and $f(x_f) \leq f(x_g)$ it follows that $\langle x'_f, \theta\rangle \leq \langle x'_g, \theta\rangle \leq \langle x'_h, \theta\rangle$. Therefore,

$$\begin{aligned}
f(x_g) &\overset{\text{(a)}}{\leq} \delta + f(x'_g) \overset{\text{(b)}}{=} \delta + f\left(\frac{\langle x'_g - x'_f, \theta\rangle}{\langle x'_h - x'_f, \theta\rangle} x'_h + \frac{\langle x'_h - x'_g, \theta\rangle}{\langle x'_h - x'_f, \theta\rangle} x'_f\right) \\
&\overset{\text{(c)}}{\leq} \delta + f(x'_h) + \frac{\langle x'_h - x'_g, \theta\rangle}{\langle x'_h - x'_f, \theta\rangle}(f(x'_f) - f(x'_h)) \\
&\overset{\text{(d)}}{\leq} \delta + f(x'_h) + \frac{1}{d+1}(f(x'_f) - f(x'_h)) \overset{\text{(e)}}{\leq} 3\delta + f(x_h) + \frac{1}{d+1}(f(x_f) - f(x_h))\,,
\end{aligned} \tag{7.4}$$

where `(a)` follows because f is a Lipschitz ridge function and using Lemma 2 and the definition of x'_g so that $|\langle x_g - x'_g, \theta\rangle| = \delta$. `(b)` by definitions and the fact that $f(\cdot) = \ell(\langle \cdot, \theta\rangle)$, `(c)` by convexity of f, `(d)` from Eq. (7.3) and because $f(x'_f) \leq f(x'_h)$. `(e)` uses again that f is a

Lipschitz ridge function and Lemma 2 and that $|\langle x'_h - x_h, \theta\rangle| = \delta$. Suppose that $f(x_h) \geq \bar{f}(x_h) + \delta$, then $(f(x_h) - \bar{f}(x_h))^2 \geq \delta^2$ and (ii) holds. Otherwise $f(x_h) < \bar{f}(x_h) + \delta$ and so

$$\begin{aligned} f(x_g) &\overset{\text{(a)}}{\leq} 3\delta + \frac{d}{d+1} f(x_h) + \frac{1}{d+1} f(x_f) \overset{\text{(b)}}{\leq} 3\delta + \frac{d}{d+1}(\bar{f}(x_h) + \delta) + \frac{1}{d+1}(\bar{f}(x_h) + \delta - \epsilon/2) \\ &\overset{\text{(c)}}{\leq} 4\delta + \bar{f}(x_h) - \frac{\epsilon}{2(d+1)} \overset{\text{(d)}}{=} \bar{f}(x_h) - 2\delta \overset{\text{(e)}}{\leq} \bar{f}(x_g) - \delta \,, \end{aligned}$$

where (a) follows from Eq. (7.4), (b) follows from the assumption in the lemma statement that $f(x_f) \leq \bar{f}(x_f) - \epsilon/2 \leq \bar{f}(x_h) + \delta - \epsilon/2$ and $f(x_h) < \bar{f}(x_h) + \delta$. (c) by naive simplification, (d) by the definition of ϵ and δ and (e) by the assumptions in the lemma. Therefore $f(x_g) \leq \bar{f}(x_g) - \delta$, which implies that $(f(x_g) - \bar{f}(x_g))^2 \geq \delta^2$ and again (ii) holds.

Summarising, in **CASE 1**, (i) holds while in **CASE 2**, (ii) holds. □

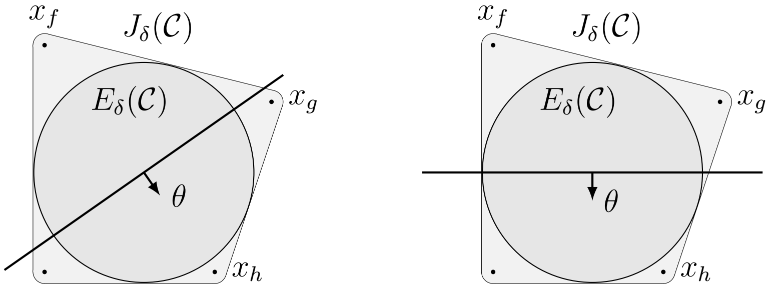


Figure 7.1: The two cases considered in the proof of Lemma 19. In the left figure, the situation is such that $E_\delta(\mathcal{C} \setminus \{f\})$ is a constant fraction less volume than $E_\delta(\mathcal{C})$. On the other hand, in the figure on the right one of $(f(x_h) - \bar{f}(x_h))^2$ or $(f(x_g) - \bar{f}(x_g))^2$ must be reasonably large.

The next lemma uses an inductive argument to show that any suitably large set $\mathcal{C}$ satisfying the conditions of the previous lemma necessarily yields a large information gain.

Lemma 20. *Suppose that $\mathcal{C}$ satisfies the conditions of Lemma 19 for some $\epsilon \geq \alpha$ and $|\mathcal{C}| = 1 + 2d + 8d \left\lceil \log\left(\frac{24d(d+1)\operatorname{diam}(K)}{\alpha}\right)\right\rceil$. Then*

$$\sum_{f,g \in \text{PAIR}(\mathcal{C})} (f(x_g) - \bar{f}(x_g))^2 \geq d\delta^2 \,.$$

Proof. Define a sequence $(\mathcal{C}_k)$ of sets as follows. Let $\mathcal{C}_1 = \mathcal{C}$ and $2m - 1 \triangleq |\mathcal{C}|$. Then, given $\mathcal{C}_k$, define $\mathcal{C}_{k+1} \subset \mathcal{C}_k$ as a set such that one of two properties hold:

(i) $|\mathcal{C}_{k+1}| = |\mathcal{C}_k| - 1$ and $\operatorname{vol}(E_\delta(\mathcal{C}_{k+1})) \leq 0.85 \operatorname{vol}(E_\delta(\mathcal{C}_k))$; or

(ii) $\mathcal{C}_{k+1} = \mathcal{C}_k \setminus \{f, g\}$ for some $f, g \in \text{PAIR}(\mathcal{C}_k)$ and $(f(x_g) - \bar{f}(x_g))^2 \geq \delta^2$.

Such a sequence exists by Lemma 19. By definition $|\mathcal{C}_1| = 2m - 1$ and since $|\mathcal{C}_{k+1}| \geq |\mathcal{C}_k| - 2$, $|\mathcal{C}_m| \geq |\mathcal{C}_1| - 2(m-1) = 1$. Recall that by John's theorem $E_\delta(\mathcal{C}_m) \subset J_\delta(\mathcal{C}_m) \subset E_{\delta,d}(\mathcal{C}_m)$, which means that

$$\operatorname{vol}(E_\delta(\mathcal{C}_m)) = \left(\frac{1}{d}\right)^d \operatorname{vol}(E_{\delta,d}(\mathcal{C}_m)) \geq \left(\frac{1}{d}\right)^d \operatorname{vol}(J_\delta(\mathcal{C}_m)) \geq \left(\frac{1}{d}\right)^d \operatorname{vol}(\mathbb{B}_\delta)\,.$$

Furthermore, $E_\delta(\mathcal{C}_1) \subset K + \mathbb{B}_\delta \subset \mathbb{B}_{\operatorname{diam}(K)+\delta}$. Let τ be the number of times (i) occurs. Then

$$\left(\frac{1}{d}\right)^d \operatorname{vol}(\mathbb{B}_\delta) \leq \operatorname{vol}(E_\delta(\mathcal{C}_m)) \leq (0.85)^\tau \operatorname{vol}(E_\delta(\mathcal{C}_1)) \leq (0.85)^\tau \operatorname{vol}(\mathbb{B}_{\operatorname{diam}(K)+\delta})\,.$$

Therefore $(0.85)^\tau \geq \left(\frac{\delta}{d(\operatorname{diam}(K)+\delta)}\right)^d$, which shows that

$$\tau \leq \frac{d \log\left(\frac{\delta}{d(\operatorname{diam}(K)+\delta)}\right)}{\log(0.85)} \leq 8d \log\left(\frac{2d\operatorname{diam}(K)}{\delta}\right) \leq |\mathcal{C}| - 2d - 1\,.$$

Therefore (ii) happens at least d times and the claim follows by the definition of (ii). □

The last step is to introduce a decomposition of the space of loss functions and show how to obtain finite sets satisfying the conditions of Lemma 20. Let

$$k = (25d + 24)\left[1 + 2d + 8d\left\lceil \log\left(\frac{24d(d+1)\operatorname{diam}(K)}{\alpha}\right)\right\rceil\right] = O\left(d^2 \log\left(\frac{d\operatorname{diam}(K)}{\alpha}\right)\right)\,.$$

For $0 \leq i \leq \lceil \log_2(1/\alpha) \rceil$ let

$$\mathscr{F}_i = \begin{cases} \{f \in \mathscr{F} : \bar{f}(x_f) - f_\star \in [\alpha 2^{|i|-1}, \alpha 2^{|i|})\} & \text{if } i > 0 \\ \{f \in \mathscr{F} : \bar{f}(x_f) - f_\star < \alpha\} & \text{if } i = 0\,. \end{cases}$$

In order to apply Lemma 4 we will show that for all $0 \leq i \leq \lceil \log_2(1/\alpha) \rceil$ and with $\epsilon = \alpha 2^i$ that for all $f_1, \ldots, f_k \in \mathscr{F}_i$,

$$\sup_{f \in \mathscr{F}_i} \left(\bar{f}(x_f) - f_\star\right) \leq \epsilon \leq \alpha + \sqrt{512 \sum_{j,l \in \text{PAIR}(k)} (f_j(x_{f_l}) - \bar{f}(x_{f_l}))^2}\,. \tag{7.5}$$

This implies that $(\alpha, 512k(k-1)(1+\lceil\log(1/\alpha)\rceil)) \in \text{IR}(\mathscr{F})$ as required. The first inequality in Eq. (7.5) follows immediately from the definition of ϵ and $\mathscr{F}_i$. The second is also immediate when $i = 0$ by the definition of $\mathscr{F}_i$. Suppose now that $i > 0$. Let $f_1, \ldots, f_k \in \mathscr{F}_i$ and assume without loss of generality that $j \mapsto \bar{f}(x_{f_j})$ is nonincreasing. The second inequality in Eq. (7.5) holds immediately if $f_j = f_l$ for some $j, l \in \text{PAIR}(k)$ since $f_j(x_{f_l}) = f_j(x_{f_j}) \leq \bar{f}(x_{f_j}) - \epsilon/2$. Suppose this is not the case and consider two cases:

CASE 1 $\bar{f}(x_{f_1}) \geq \bar{f}(x_{f_k}) + 2\epsilon$. In this case $f_1(x_{f_k}) \geq f_1(x_{f_1}) \geq \bar{f}(x_{f_1}) - \epsilon \geq \bar{f}(x_{f_k}) + \epsilon$, which shows that $(f_1(x_{f_k}) - \bar{f}(x_{f_k}))^2 \geq \epsilon^2$ and the second inequality Eq. (7.5) holds.

CASE 2 $\bar{f}(x_1) < \bar{f}(x_k) + 2\epsilon$. Let $\delta = \frac{\epsilon}{12(d+1)}$ as in Lemma 20. Let $b = 25d + 24$ and $\mathcal{C}_1, \ldots, \mathcal{C}_b$ be formed by dividing $\{f_1, \ldots, f_k\}$ in order into b blocks of equal size. Let $s_a = \max_{f,g \in \mathcal{C}_a} |\bar{f}(x_f) - \bar{f}(x_g)|$. Given the conditions of the case we have $2\epsilon > \sum_{a=1}^b s_a \geq \sum_{a=1}^b \delta \mathbf{1}(s_a > \delta)$, which means that $\sum_{a=1}^b \mathbf{1}(s_a > \delta) \leq 2\epsilon/\delta \leq 24(d+1) = b - d$. Hence there exist at least d blocks $\mathcal{C}_a$ for which $s_a \leq \delta$ and these blocks satisfy the conditions of Lemma 20 so that $\sum_{j,l \in \text{PAIR}(k)} (f_j(x_{f_l}) - \bar{f}(x_{f_l}))^2 \geq d^2\delta^2 \geq \epsilon^2/512$ and again the second inequality in Eq. (7.5) holds.

Hence Eq. (7.5) holds and the claim follows from Lemma 4 and the definitions of k and m. □

Chapter 8

TS Lower-Bound for general Convex Functions

We now prove that Thompson sampling has poor behaviour for general multi-dimensional convex functions and that the classical information-theoretic techniques cannot improve on the best known bound for general bandit convex optimisation of $\tilde{O}(d^{1.5}\sqrt{n})$. While these seem like quite different results, they are based on the same construction, which is based on finding a family of functions and prior that makes learning challenging.

Lemma 21. *Let $\epsilon \in [0, 1/2]$ and $\theta \in \mathbb{S}_1$ and define functions f and f_θ by*

$$f(x) = \epsilon + \frac{1}{2}\|x\|^2 \qquad f_\theta(x) = \begin{cases} f(x) & \textit{if } \|\theta - x\|^2 \geq 1 + 2\epsilon \\ \langle \theta, x\rangle - 1 + \sqrt{1+2\epsilon}\,\|\theta - x\| & \textit{otherwise}\,. \end{cases}$$

Then f_θ is convex and minimised at θ and $\operatorname{Lip}_{\mathbb{B}_1}(f_\theta) \leq \sqrt{2+2\epsilon}$.

The function f_θ arises naturally as the largest convex function for which both $f_\theta(\theta) = 0$ and $f_\theta(x) \leq f(x)$ for all $x \in \mathbb{R}^d$. Equivalently, its epigraph is the convex hull of the epigraphs of f and the convex indicator function: $\infty\mathbf{1}_\theta(\cdot)$.

Proof of Lemma 21. Recall that $\epsilon \in [0, 1/2]$ and

$$f(x) = \epsilon + \frac{1}{2}\|x\|^2 \qquad f_\theta(x) = \begin{cases} f(x) & \text{if } \|x - \theta\|^2 \geq 1 + 2\epsilon \\ \langle \theta, x\rangle - 1 + \sqrt{1+2\epsilon}\,\|\theta - x\| & \text{otherwise}\,. \end{cases}$$

We need to prove that f_θ on $\mathbb{B}_1$ is convex, Lipschitz and minimised at θ. Convexity follows because

$$\begin{aligned}
g_\theta(x) &\triangleq \sup_{y \in \mathbb{R}^d} \left\{ f(y) + \langle f'(y), x - y \rangle : f(y) + \langle f'(y), \theta - y \rangle \leq 0 \right\} \\
&= \sup_{\substack{y \in \mathbb{R}^d \\ r \leq 0}} \left\{ f(y) + \langle f'(y), x - y \rangle : f(y) + \langle f'(y), \theta - y \rangle = r \right\} \\
&= \sup_{\substack{y \in \mathbb{R}^d \\ r \leq 0}} \left\{ \langle f'(y), x - \theta \rangle + r : f(y) + \langle f'(y), \theta - y \rangle = r \right\} \\
&= \sup_{\substack{y \in \mathbb{R}^d \\ r \leq 0}} \left\{ \langle y, x - \theta \rangle + r : \|y - \theta\|^2 = 1 + 2\epsilon - 2r \right\} \\
&= \sup_{r \leq 0} \left\{ \langle \theta, x - \theta \rangle + r + \sqrt{1 + 2\epsilon - 2r}\, \|x - \theta\| \right\} \\
&= \sup_{r \leq 0} \left\{ \langle \theta, x \rangle - 1 + r + \sqrt{1 + 2\epsilon - 2r}\, \|x - \theta\| \right\} \qquad (8.1) \\
&= f_\theta(x)\,,
\end{aligned}$$

where in the final inequality we note that the maximising r is

$$r = \begin{cases} \frac{1}{2} + \epsilon - \frac{1}{2}\|x - \theta\|^2 & \text{if } \|x - \theta\|^2 \geq 1 + 2\epsilon \\ 0 & \text{otherwise }. \end{cases}$$

Therefore f_θ is the supremum over a set of linear functions and hence convex. That f_θ is minimised at θ follows directly from the first-order optimality conditions. Let $\eta \in \mathbb{S}_1$. Then

$$Df_\theta(\theta)[\eta] = \langle \theta, \eta \rangle + \sqrt{1 + 2\epsilon}\, \|\eta\| > 0\,,$$

where $Df_\theta(\theta)[\cdot]$ is the directional derivative operator (noting that f_θ is not differentiable at θ). Lastly, for Lipschitzness. Since f_θ is continuous, it suffices to bound $\|f'_\theta(\cdot)\|$ on $\operatorname{int}(\mathbb{B}_1)$ where f_θ is differentiable. When $\|x - \theta\|^2 \geq 1 + 2\epsilon$, then $\|f'_\theta(x)\| = \|f'(x)\| = \|x\| \leq 1$. On the other hand, if $\|x - \theta\|^2 < 1 + 2\epsilon$, then

$$\begin{aligned}
\|f'_\theta(x)\|^2 &= \left\| \theta + \sqrt{1 + 2\epsilon} \frac{x - \theta}{\|x - \theta\|} \right\|^2 \\
&= 2 + 2\epsilon + \sqrt{1 + 2\epsilon} \frac{\langle \theta, x - \theta \rangle}{\|x - \theta\|} \\
&\leq 2 + 2\epsilon\,.
\end{aligned}$$

Therefore $\mathrm{Lip}_{\mathbb{B}_1}(f) \leq \sqrt{2+2\epsilon}$. □

Theorem 22. *When $K = \mathbb{B}_1$ is the standard euclidean ball. There exists a prior ξ on $\mathscr{F}_{b1}$ such that*

$$\mathit{BReg}_n(\mathrm{TS}, \xi) \geq \frac{1}{2} \min\left(n, \left\lfloor \frac{1}{4}\exp(d/32)\right\rfloor\right) .$$

sketch. The idea is to construct a prior such that with high probability TS obtains limited information while suffering high regret. We assume there is no noise and let f and f_θ be defined as in Lemma 21 with $\epsilon = 1/4$. Let σ be the uniform probability measure on $\mathbb{S}_1$ and the prior ξ be the law of f_θ when θ is sampled from σ. By the definition of f_θ and the fact that $\epsilon = 1/4$, for any $x \in \mathbb{S}_1$ $(f(x) = f_\theta(x)) \Leftrightarrow \langle x, \theta\rangle \leq \frac{1}{4}$. Since TS plays the minimiser of some f_θ in every round, it follows that TS always plays in $\mathbb{S}_1$. Let $\mathcal{C}_\theta = \{x \in \mathbb{S}_1 : \langle x, \theta\rangle > \frac{1}{4}\}$ and $\delta = \sigma(\mathcal{C}_\theta)$. Let $f_{\theta_\star}$ be the true loss function sampled from ξ. Suppose that $X_1, \ldots, X_t \in \mathbb{S}_1 \setminus \mathcal{C}_{\theta_\star}$, which means that $Y_s = \frac{3}{4}$ for $1 \leq s \leq t$ and the posterior is the uniform distribution on $\Theta_{t+1} = \mathbb{S}_1 \setminus \cup_{s=1}^t \mathcal{C}_{X_s}$. Provided that $t\delta \leq \frac{1}{2}$,

$$\mathbb{P}(X_{t+1} \in \mathcal{C}_{\theta_\star} | X_1, \ldots, X_t \notin \mathcal{C}_{\theta_\star}) = \frac{\sigma(\mathcal{C}_{\theta_\star} \cap \Theta_{t+1})}{\sigma(\Theta_{t+1})} \leq \frac{\delta}{1 - t\delta} \leq 2\delta .$$

Hence, with $n_0 = \min(n, \lfloor 1/(4\delta)\rfloor)$,

$$\mathrm{BReg}_n(\mathrm{TS}, \xi) \geq \mathrm{BReg}_{n_0}(\mathrm{TS}, \xi) \geq \frac{3n_0}{4}\mathbb{P}(X_1, \ldots, X_{n_0} \notin \mathcal{C}_{\theta_\star}) \geq \frac{3n_0}{4}(1 - 2n_0\delta) \geq \frac{3n_0}{8} .$$

The result is completed since $\delta \leq \exp(-d/32)$ follows from concentration of measure on the sphere [Tko18, Theorem B.1]. □

Definition 23. For $\theta \in \mathbb{S}_1$ and $\epsilon \in [0, 1/2]$, define

$$\mathcal{C}_{\theta,\epsilon} = \left\{x \in \mathbb{S}_1 : \|\theta - x\|^2 < 1 + 2\epsilon\right\}$$

and $\bar{\mathcal{C}}_{\theta,\epsilon} = \mathbb{S}_1 \setminus \mathcal{C}_{\theta,\epsilon}$.

Definition 24. Let $\sigma(\cdot)$ be the uniform distribution over the unit sphere $\mathbb{S}_1$. Moreover, let $\sigma_S(\cdot)$ be the uniform distribution over a set $S \subseteq \mathbb{S}_1$ defined as $\sigma_S(\cdot) = \frac{\sigma(\cdot \cap S)}{\sigma(S)}$.

We use the following theorem from [Tko18] to bound the surface area of spherical caps.

Theorem 25. *[Tko18, Theorem B.1] For all $\epsilon \in [0,1]$ and $\theta \sim \sigma$ we have*

$$\mathbb{P}\left(\langle \theta, e_1 \rangle \geq \epsilon\right) \leq \exp\left(-\frac{d\epsilon^2}{2}\right) .$$

Proof of Theorem 22. Define f and f_θ as in Lemma 21 with $\epsilon = 1/4$, which then using Definition 23 can be written as

$$f_\theta(x) = \begin{cases} f(x) & \text{if } x \in \bar{\mathcal{C}}_\theta, \\ \langle \theta, x \rangle - 1 + \sqrt{\frac{3}{2}}\, \|\theta - x\| & \text{if } x \in \mathcal{C}_\theta \,, \end{cases}$$

where we drop the ϵ from $\mathcal{C}_{\theta,\epsilon}$ and $\bar{\mathcal{C}}_{\theta,\epsilon}$ in the notation for simplicity. We define the bandit instance by setting the prior ξ_1 to be the law of f_θ when θ has law σ, and letting the observation noise to be zero, meaning that

$$Y_t = f_{\theta_*}(X_t) \,,$$

where X_t is the action played at round t, Y_t is the loss observed at round t, and f_{θ_*} is the true function that is secretly sampled from the prior ξ_1. Also, define the random sets $\Theta_t \subseteq \mathbb{S}_1$ as

$$\Theta_t = \left\{ \theta \in \mathbb{S}_1 : f_\theta(X_s) = \frac{3}{4}, \forall s \in [t-1] \right\} .$$

We also make extensive use of the fact that for any two $\theta_1, \theta_2 \in \mathbb{S}_1$, we have

$$f_{\theta_1}(\theta_2) = f_{\theta_2}(\theta_1) = \frac{3}{4}, \quad \text{if and only if} \quad \theta_1 \in \bar{\mathcal{C}}_{\theta_2} \;\Leftrightarrow\; \theta_2 \in \bar{\mathcal{C}}_{\theta_1} \,,$$

which follows from the definition of f_θ.

Step 1: First we show that if the posterior distribution ξ_t at round $t \in [T]$ is uniform over Θ_t, and the algorithm observes the loss $Y_t = \frac{3}{4}$ as a result of playing X_t, then the posterior distribution ξ_{t+1} at round $t+1$ is uniform over Θ_{t+1}, i.e., $\xi_{t+1} = \sigma_{\Theta_{t+1}}$. To this end, observe that if $Y_t = \frac{3}{4}$ then for any set $B \subseteq \Theta_t$,

$$\xi_{t+1}(B) = \mathbb{P}_{t-1}\left(\theta_\star \in B | Y_t = \frac{3}{4}, X_t\right) = \frac{\mathbb{P}_{t-1}(\theta_\star \in B, Y_t = \frac{3}{4}|X_t)}{\mathbb{P}_{t-1}(Y_t = \frac{3}{4}|X_t)}$$
$$= \frac{\mathbb{P}_{t-1}(Y_t = \frac{3}{4}|X_t, \theta_\star \in B)\mathbb{P}_{t-1}(\theta_\star \in B|X_t)}{\mathbb{P}_{t-1}(Y_t = \frac{3}{4}|X_t)}$$
$$= \frac{\mathbb{P}_{t-1}(f_{\theta_\star}(X_t) = \frac{3}{4}|X_t, \theta_\star \in B)\mathbb{P}_{t-1}(\theta_\star \in B)}{\mathbb{P}_{t-1}(f_{\theta_\star}(X_t) = \frac{3}{4}|X_t)} . \quad (8.2)$$

Note that TS samples f_{θ_t} from ξ_t, and then plays the minimizer of f_{θ_t}, which from Lemma 21 is θ_t, i.e. $X_t = \theta_t$. Consequently, continuing from Eq. (8.2) with the assumption of this step that $\theta_t, \theta_\star \sim \sigma_{\Theta_t}$ and the fact that $X_t = \theta_t$, we have

$$\xi_{t+1}(B) = \frac{\mathbb{P}_{t-1}(f_{\theta_\star}(X_t) = \frac{3}{4}|X_t, \theta_\star \in B)\mathbb{P}_{t-1}(\theta_\star \in B)}{\mathbb{P}_{t-1}(f_{\theta_\star}(X_t) = \frac{3}{4}|X_t)}$$
$$= \frac{\mathbb{P}_{t-1}(\theta_\star \in \bar{\mathcal{C}}_{\theta_t}|\theta_t, \theta_\star \in B)\mathbb{P}_{t-1}(\theta_\star \in B)}{\mathbb{P}_{t-1}(\theta_\star \in \bar{\mathcal{C}}_{\theta_t}|\theta_t)}$$
$$= \frac{\frac{\sigma(B\cap\bar{\mathcal{C}}_{\theta_t})}{\sigma(B)} \cdot \frac{\sigma(B)}{\sigma(\Theta_t)}}{\frac{\sigma(\bar{\mathcal{C}}_{\theta_t}\cap\Theta_t)}{\sigma(\Theta_t)}}$$
$$= \frac{\sigma(B \cap \bar{\mathcal{C}}_{\theta_t})}{\sigma(\Theta_t \cap \bar{\mathcal{C}}_{\theta_t})}$$

which implies that ξ_{t+1} is uniform over $\Theta_t \cap \bar{\mathcal{C}}_{\theta_t}$. Lastly, note that

$$\Theta_{t+1} = \left\{\theta \in \mathbb{S}_1 : f_\theta(X_s) = \frac{3}{4}, \forall s \in [t]\right\} = \Theta_t \cap \left\{\theta \in \mathbb{S}_1 : f_\theta(\theta_t) = \frac{3}{4}\right\} = \Theta_t \cap \bar{\mathcal{C}}_{\theta_t} ,$$

which means that ξ_{t+1} is uniform over Θ_{t+1}.

Step 2: Let $\delta = \sigma(\mathcal{C}_{\theta_\star})$, and note that $\sigma(\mathcal{C}_\theta) = \delta$ for all $\theta \in \mathbb{S}_1$ due to the shape of the $\mathcal{C}_\theta$ which is a spherical cap with a fixed radius. Consider the event $\mathcal{E}_t$ where $X_1, \ldots, X_t \in \bar{\mathcal{C}}_{\theta_\star}$, which implies both that $Y_1, \dots, Y_t = \frac{3}{4}$, and that ξ_{t+1} is uniform over Θ_{t+1}. Conditioned on $\mathcal{E}_t$, we have

$$\begin{aligned}
\sigma(\Theta_{t+1}) &= \sigma\left(\left\{\theta \in \mathbb{S}_1 : f_\theta(X_s) = \frac{3}{4}, \forall s \in [t]\right\}\right) \\
&= \sigma\left(\left\{\theta \in \mathbb{S}_1 : f_\theta(\theta_s) = \frac{3}{4}, \forall s \in [t]\right\}\right) \\
&= \sigma\left(\left\{\theta \in \mathbb{S}_1 : \theta \in \bar{\mathcal{C}}_{\theta_s}, \forall s \in [t]\right\}\right) \\
&= \sigma\left(\left\{\theta \in \mathbb{S}_1 : \theta \notin \cup_{s=1}^{t} \mathcal{C}_{\theta_t}\right\}\right) \\
&\geq 1 - t\delta\,.
\end{aligned}$$

Therefore, the probability of TS playing $X_{t+1} \in \mathcal{C}_{\theta_\star}$ is upper bounded by

$$\mathbb{P}(X_{t+1} \in \mathcal{C}_{\theta_\star} | \theta_\star, \mathcal{E}_t) = \frac{\sigma(\mathcal{C}_{\theta_\star} \cap \Theta_{t+1})}{\sigma(\Theta_{t+1})} \leq \frac{\delta}{1 - t\delta}\,,$$

which further implies that

$$\mathbb{P}(\mathcal{E}_{t+1} | \mathcal{E}_t) = \mathbb{P}\left(Y_{t+1} = \frac{3}{4} | \mathcal{E}_t\right) = \mathbb{P}(X_{t+1} \in \bar{\mathcal{C}}_{\theta_\star} | \mathcal{E}_t) \geq 1 - \frac{\delta}{1 - t\delta}\,,$$

and therefore

$$\mathbb{P}(\mathcal{E}_{t+1}) = \mathbb{P}\left(\mathcal{E}_t\right) \mathbb{P}\left(\mathcal{E}_{t+1} | \mathcal{E}_t\right) \geq \mathbb{P}\left(\mathcal{E}_t\right) \left(1 - \frac{\delta}{1 - t\delta}\right) \geq \mathbb{P}\left(\mathcal{E}_t\right) - \frac{\delta}{1 - t\delta}\,.$$

Let $n_0 = \min(\left\lfloor \frac{1}{4\delta} \right\rfloor, n)$, then

$$\mathbb{P}(\mathcal{E}_{n_0}) \geq \mathbb{P}(\mathcal{E}_{n_0 - 1}) - \frac{\delta}{1 - (n_0 - 1)\delta} \geq \mathbb{P}(\mathcal{E}_1) - \sum_{t=1}^{n_0 - 1} \frac{\delta}{1 - t\delta} = 1 - \sum_{t=0}^{n_0 - 1} \frac{\delta}{1 - t\delta}$$

where the last equality follows from $\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\theta_1 \in \bar{\mathcal{C}}_{\theta_\star}) = 1 - \delta$. Since $t\delta \leq n_0\delta \leq 1/4$ for all $t < n_0$, we have

$$\mathbb{P}(\mathcal{E}_{n_0}) \geq 1 - \sum_{t=0}^{n_0 - 1} \frac{\delta}{1 - 1/4} = 1 - \frac{4}{3} n_0 \delta \geq \frac{2}{3}.$$

Therefore, the expected regret of TS is lower bounded by

$$\text{BReg}_n(\text{TS}, \xi_1) \geq \text{BReg}_{n_0}(\text{TS}, \xi_1) \geq \frac{3}{4} n_0 \mathbb{P}(\mathcal{E}_{n_0}) \geq \frac{1}{2} n_0\,,$$

since the algorithm incurs maximum regret of $\frac{3}{4}$ in every round $s \in [n_0]$ given the event $\mathcal{E}_{n_0}$. Finally, using Theorem 25, we have

$$\delta \leq \exp(-d/32),$$

which implies that

$$\text{BReg}_n(\text{TS}, \xi_1) \geq \frac{1}{2} \min\left(n, \left\lfloor \frac{1}{4} \exp(d/32) \right\rfloor\right) .$$

□

Chapter 9

IR Lower-Bound for General Convex Functions

Theorem 22 shows that Thompson sampling has large regret for general bandit convex optimisation. The next theorem shows there exist priors for which the information ratio for any policy is at least $\Omega(d^2)$. At least naively, this means that the information-theoretic machinery will not yield a bound on the regret for general bandit convex optimisation that is better than $\tilde{O}(d^{1.5}\sqrt{n})$

Theorem 26. *Suppose that $K = \mathbb{B}_1$ and $d > 256$. Then there exists a prior ξ on $\mathscr{F}_{b1}$ such that for all probability measures π on K, $\Delta(\pi, \xi) \geq 2^{-19} \frac{d}{\log(d)} \sqrt{\mathcal{I}(\pi, \xi)}$.*

The prior ξ is the same as used in the proof of Theorem 22 but with $\epsilon = \tilde{\Theta}(1/d)$. The argument is based on proving that for any policy the regret is $\Omega(\epsilon)$ while the information gain is $\tilde{O}(\epsilon^2)$.

Throughout this section, we use the same construction as the one used in Chapter 8 except with $\epsilon = \frac{8\log(d)}{d}$. Therefore, we have

$$f(x) = \frac{1}{2}\|x\| + \frac{8\log(d)}{d}\,, \quad \text{and} \quad f_\theta(x) = \begin{cases} f(x) & \text{if } x \in \bar{\mathcal{C}}_\theta, \\ \langle \theta, x\rangle - 1 + \sqrt{1 + \frac{16\log(d)}{d}}\,\|\theta - x\| & \text{if } x \in \mathcal{C}_\theta\,. \end{cases}$$

Further, let ξ be the law of f_θ where $\theta \sim \sigma$, and for $x \in \mathbb{B}_1$ define

$$\Delta_x = \mathbb{E}\left[f_\theta(x) - f_\theta(\theta)\right] = \mathbb{E}\left[f_\theta(x)\right]\,, \quad \text{and} \quad \mathcal{I}_x = \mathbb{E}\left[\left(f_\theta(x) - \mathbb{E}\left[f_\theta(x)\right]\right)^2\right]\,,$$

which should be thought of as the expected loss and the expected information gain at x. Therefore, for any policy π we have

$$\Delta(\pi,\xi) = \mathbb{E}[\Delta_X] \qquad \text{and} \qquad \mathcal{I}(\pi,\xi) = \mathbb{E}[\mathcal{I}_X]\,,$$

where $X \sim \pi$. The basic idea is to prove that $\Delta_x = \Omega(\log(d)/d)$ for all $x \in \mathbb{B}_1$ and $\mathcal{I}_x = O(\log(d)^4 d^{-4})$ for all $x \in \mathbb{B}_1$, which implies that $\Delta(\pi,\xi) = \Omega(\log(d)/d)$ and $\mathcal{I}(\pi,\xi) = \tilde{O}(\log(d)^4 d^{-4})$ for any policy π, and hence the claimed lower bound.

Additional to $\epsilon = 8\log(d)/d$, we fix $\tau = \sqrt{8\epsilon} = 8\sqrt{\log(d)/d}$ in the rest of this section.

Lemma 27. *For $\tau \le r$, and $x \in \mathbb{B}_1$ with $\|x\| = r$, $\mathbb{P}\left(f_\theta(x) = f(x)\right) \ge 1 - \dfrac{1}{d^4}$.*

Proof. From Lemma 21, $f_\theta(x) = f(x)$ if $\|\theta - x\|^2 \ge 1 + 2\epsilon$. Moreover, for $x \in \mathbb{B}_1$ with $\|x\| = r$,

$$\|\theta - x\|^2 = \|\theta\|^2 + r^2 - 2\langle \theta, x\rangle = 1 + r^2 - 2\langle \theta, x\rangle\,,$$

which implies that $f(x) = f_\theta(x)$ if $\langle \theta, x\rangle \le \frac{r^2}{2} - \epsilon$. Since $\theta \sim \sigma$,

$$\begin{aligned}
\mathbb{P}\left(\langle \theta, x\rangle \le \frac{r^2}{2} - \epsilon\right) &= \mathbb{P}\left(\langle \theta, e_1\rangle \le \left(\frac{r^2}{2} - \epsilon\right)\|x\|^{-1}\right)\\
&= 1 - \mathbb{P}\left(\langle \theta, e_1\rangle > \left(\frac{r}{2} - \frac{\epsilon}{r}\right)\right)\\
&\ge 1 - \exp\left(-\left(\frac{r}{2} - \frac{\epsilon}{r}\right)^2 \frac{d}{2}\right)\\
&\ge 1 - \exp\left(-\left(4\sqrt{\frac{\log(d)}{d}} - \frac{\sqrt{d}\log(d)}{d\sqrt{\log(d)}}\right)^2 \frac{d}{2}\right)\\
&= 1 - \exp\left(-\frac{9\log(d)}{2}\right)\\
&\ge 1 - \exp\left(-\log(d^4)\right)\\
&\ge 1 - \frac{1}{d^4},
\end{aligned}$$

where the first inequality follows from Theorem 25, and the second inequality follows from the fact that $r \ge \tau \ge \sqrt{2\epsilon}$. □

Lemma 28. *For all $x \in \mathbb{B}_1$ and $\theta \in \mathbb{S}_1$, we have*

$$f_\theta(x) \ge \langle \theta, x\rangle - 1 + \sqrt{1 + 2\epsilon}\,\|\theta - x\|\,.$$

Proof. The proof follows by setting $r = 0$ in Equation (8.1). □

Lemma 29. *For $d \geq 2^8$ and $x \in \mathbb{B}_1$, we have $\Delta_x \geq 2\frac{\log(d)}{d}$.*

Proof. Let $r = \|x\|$. We prove this result by considering two cases.

CASE 1 If $r \geq \tau$, then

$$\mathbb{E}\left[f_\theta(x)\right] \geq \mathbb{P}\left(f_\theta(x) = f(x)\right) f(x) \overset{\text{(a)}}{\geq} \left(1 - \frac{1}{d^4}\right)\left(\frac{1}{2}r^2 + \epsilon\right) \geq \frac{1}{2}\left(\frac{32\log(d)}{d} + \frac{8\log(d)}{d}\right) \geq \frac{20\log(d)}{d},$$

where (a) follows from Lemma 27.

CASE 2 If $r < \tau$, using the lower bound on $f_\theta(x)$ from Lemma 28,

$$\begin{aligned}
\mathbb{E}\left[f_\theta(x)\right] &\geq \mathbb{E}\left[\langle\theta, x\rangle - 1 + \sqrt{1+2\epsilon}\,\|\theta - x\|\right] \\
&\overset{\text{(a)}}{=} \sqrt{1+2\epsilon}\mathbb{E}\left[\sqrt{1 + r^2 - 2\langle\theta, x\rangle}\right] - 1 \\
&\overset{\text{(b)}}{\geq} \left(1 + \frac{2\epsilon - 4\epsilon^2}{2}\right)\mathbb{E}\left[1 + \frac{r^2 - 2\langle\theta, x\rangle - (r^2 - 2\langle\theta, x\rangle)^2}{2}\right] - 1 \\
&\overset{\text{(c)}}{=} \left(1 + \epsilon - 2\epsilon^2\right)\left(1 + \frac{r^2 - r^4}{2} - \mathbb{E}\left[2\langle\theta, x\rangle^2\right]\right) - 1 \\
&\overset{\text{(d)}}{=} \left(1 + \epsilon - 2\epsilon^2\right)\left(1 + \frac{r^2 - r^4}{2} - \frac{2r^2}{16}\right) - 1 \\
&= \left(1 + \epsilon - 2\epsilon^2\right)\left(1 + \frac{3r^2}{8} - \frac{r^4}{2}\right) - 1,
\end{aligned}$$

where (a) follows from $\mathbb{E}\left[\langle\theta, x\rangle\right] = 0$, (b) follows from the inequality $\sqrt{1+a} \geq 1 + \frac{a - a^2}{2}$ for $a \geq -1$, (c) follows from $\mathbb{E}\left[\langle\theta, x\rangle\right] = 0$, (d) follows from $\mathbb{E}[\langle\theta, x\rangle^2] = \frac{r^2}{d}$ and $d > 16$. Note that $(1 + 3r^2/8 - r^4/2) \geq 1$ for $0 \leq r \leq \sqrt{3/4}$, and is decreasing in r for $r \in [\sqrt{3/4}, 1]$. Therefore,

$$1 + \frac{3r^2}{8} - \frac{r^4}{2} \geq \min(1, 1 + \frac{3\tau^2}{8} - \frac{\tau^4}{2}) = 1 + \min(0, 3\epsilon - 4\epsilon^2).$$

This let us further lower bound $\mathbb{E}[f_\theta(x)]$ as

$$\mathbb{E}\left[f_\theta(x)\right] \geq \left(1 + \epsilon - 2\epsilon^2\right)\left(1 + \min(0, 3\epsilon - 4\epsilon^2)\right) - 1.$$

Now, since we have $\epsilon \leq 1/3$, we have $3\epsilon - 4\epsilon^2 \geq 0$, which gives

$$\mathbb{E}\left[f_\theta(x)\right] \geq 1 + \epsilon - 2\epsilon^2 - 1 = \epsilon - 2\epsilon^2 \geq \epsilon - \frac{2}{3}\epsilon \geq \frac{8\log(d)}{3d}\,.$$

□

Lemma 30. *For all $x \in \mathbb{B}_1$ and $d > 256$, $\mathcal{I}_x \leq 2^{40}\frac{\log(d)^4}{d^4}$.*

Proof. Let $x \in \mathbb{B}_1$ and $r = \|x\|$. We prove this result by considering two cases.

CASE 1 If $r \geq \tau$, then we have

$$\begin{aligned}
\mathbb{E}\left[\left(f_\theta(x) - \mathbb{E}\left[f_\theta(x)\right]\right)^2\right] &\leq \mathbb{E}\left[\left(f_\theta(x) - f(x)\right)^2\right] \qquad (9.1)\\
&\leq \mathbb{P}\left(f_\theta(x) = f(x)\right)\left(f(x) - f(x)\right)^2 + \mathbb{P}\left(f_\theta(x) \neq f(x)\right)\left(\frac{1}{2}r^2 + \epsilon\right)^2\\
&\leq \frac{1}{d^4}\left(\frac{1}{2}r^2 + \epsilon\right)^2 \leq \frac{1}{d^4}\left(\frac{1}{2} + \frac{1}{3}\right)^2 \leq \frac{25}{36d^4}\,,
\end{aligned}$$

where the first inequality holds since the mean minimizes the squared deviation, and the second inequality holds since $0 \leq f(x) - f_\theta(x) \leq \frac{1}{2}r^2 + \epsilon$.

CASE 2 Now suppose that $r < \tau$. We have

$$\begin{aligned}
0 &\overset{\text{(a)}}{\leq} f(x) - f_\theta(x)\\
&\overset{\text{(b)}}{\leq} f(x) - \langle \theta, x\rangle + 1 - \sqrt{1+2\epsilon}\,\|\theta - x\|\\
&= f(x) - \langle \theta, x\rangle + 1 - \sqrt{1+2\epsilon}\sqrt{1 + r^2 - 2\langle \theta, x\rangle}\\
&\overset{\text{(c)}}{\leq} \epsilon + \frac{r^2}{2} - \langle \theta, x\rangle + 1 - \left(1 + \epsilon - 2\epsilon^2\right)\left(1 + \frac{r^2}{2} - \langle \theta, x\rangle - \frac{(r^2 - 2\langle \theta, x\rangle)^2}{2}\right)\\
&= \frac{(r^2 - 2\langle \theta, x\rangle)^2}{2} + 2\epsilon^2 - \left(\epsilon - 2\epsilon^2\right)\left(\frac{r^2}{2} - \langle \theta, x\rangle - \frac{\left(r^2 - \langle \theta, x\rangle\right)^2}{2}\right)\\
&= \frac{r^4}{2} + 2\langle \theta, x\rangle^2 - 2\langle \theta, x\rangle r^2 + 2\epsilon^2 - \left(\epsilon - 2\epsilon^2\right)\left(\frac{r^2}{2} - \langle \theta, x\rangle - \frac{r^4}{2} - 2\langle \theta, x\rangle^2 + 2\langle \theta, x\rangle r^2\right)\\
&\overset{\text{(d)}}{\leq} 30\left(r^2 + |\langle \theta, x\rangle| + \epsilon\right)^2,
\end{aligned}$$

where `(a)` follows from $f(x) \geq f_\theta(x)$, the `(b)` follows from Lemma 28, `(c)` follows from $\sqrt{1+x} \geq 1 + \frac{x}{2} - \frac{x^2}{2}$ for all $x \geq -1$, and `(d)` follows from the fact that $r, \epsilon, |\langle \theta, x\rangle| \leq 1$, and the expression in the previous line a polynomial of these terms with degree at least 2 and sum of coefficients at most 30.

Next, starting from the right-hand side of Eq. (9.1), we have

$$\begin{aligned}
\mathbb{E}\left[(f(x) - f_\theta(x))^2\right] &\leq \mathbb{E}\left[30^2\left(r^2 + |\langle\theta, x\rangle| + \epsilon\right)^4\right] \\
&\overset{\texttt{(a)}}{\leq} 30^2\mathbb{E}\left[27\left(r^8 + \langle\theta, x\rangle^4 + \epsilon^4\right)\right] \\
&\leq 30^3\left(r^8 + \mathbb{E}\left[\langle\theta, x\rangle^4\right] + \epsilon^4\right) \\
&\overset{\texttt{(b)}}{\leq} 30^3\left(r^8 + \frac{3r^4}{d^2} + \epsilon^4\right) \\
&\overset{\texttt{(c)}}{\leq} 30^3\left(8^4\epsilon^4 + \frac{3\cdot 8^2\epsilon^2}{d^2} + \epsilon^4\right) \\
&\leq 30^3\left(8^4\epsilon^4 + 3\epsilon^4 + \epsilon^4\right) \\
&\leq 2^{28}\epsilon^4 = 2^{40}\frac{\log(d)^4}{d^4}\,,
\end{aligned}$$

where `(a)` follows from $(a+b+c)^4 \leq 27(a^4+b^4+c^4)$, `(b)` follows from the fact that $\mathbb{E}[\langle\theta, x\rangle^4] = \frac{3r^4}{d(d+2)} < \frac{3r^4}{d^2}$, and `(c)` follows from $r < \tau = \sqrt{8\epsilon}$. □

Proof of Theorem 26. Let ξ be the law of f_θ when θ has law σ. Then for any policy π, and $X \sim \pi$, from Lemma 29 we have

$$\Delta(\pi, \xi) = \mathbb{E}\left[\Delta_X\right] \geq \frac{\log(d)}{2d}\,,$$

and from Lemma 30 we have

$$\mathcal{I}(\pi, \xi) = \mathbb{E}[\mathcal{I}_X] \leq \frac{2^{40}\log(d)^4}{d^4}\,,$$

which together imply

$$\frac{\Delta(\pi, \xi)}{\sqrt{\mathcal{I}(\pi, \xi)}} \geq \frac{d}{2^{19}\log(d)}\,.$$

□

Chapter 10

TS for Known Link Function

In this chapter we provide a tighter d upper bound for the information ratio of TS in d dimensions, when the link function ℓ is known and the action set is the unit ball.

Lemma 31. $(0, d) \in \textsc{ir}(\mathscr{F}_{\texttt{blrm}})$ *whenever* $\mathcal{K} = \mathbb{B}_1^d$ *and the link function* ℓ *is known to the learner.*

We let $\alpha = 0$ in this chapter, which leads to the alternate definition of the $\textsc{ir}$ set,

$$\begin{aligned}
\textsc{ir}(\mathscr{F}) &= \left\{(\alpha, \beta) \in \mathbb{R}_+^2 : \sup_{\xi \in \mathscr{P}(\mathscr{F})} \left[\Delta(\pi_{\text{TS}}^{\xi}, \xi) - \alpha - \sqrt{\beta \mathcal{I}(\pi_{\text{TS}}^{\xi}, \xi)}\right] \leq 0\right\} \\
&\overset{\texttt{(a)}}{=} \left\{(0, \beta) \in \mathbb{R}_+^2 : \Delta(\pi_{\text{TS}}^{\xi}, \xi) - \sqrt{\beta \mathcal{I}(\pi_{\text{TS}}^{\xi}, \xi)} \leq 0, \forall \xi \in \mathscr{P}(\mathscr{F})\right\} \\
&\overset{\texttt{(b)}}{=} \left\{(0, \beta) \in \mathbb{R}_+^2 : \frac{\Delta(\pi_{\text{TS}}^{\xi}, \xi)^2}{\mathcal{I}(\pi_{\text{TS}}^{\xi}, \xi)} \leq \beta, \forall \xi \in \mathscr{P}(\mathscr{F})\right\},
\end{aligned}$$

where `(a)` follows from setting $\alpha = 0$ and replacing $\sup$ with a universal quantifier, and `(b)` follows from algebra. Therefore, to prove that $(0, \beta) \in \textsc{ir}(\mathscr{F})$, it suffices to show that

$$\Psi(\xi) \triangleq \frac{\Delta(\pi_{\text{TS}}^{\xi}, \xi)^2}{\mathcal{I}(\pi_{\text{TS}}^{\xi}, \xi)} \leq \beta\,, \qquad \forall \xi \in \mathscr{P}(\mathscr{F}) \tag{10.1}$$

The proof is followed from different techniques compared to the proof of Theorem 18, which is mainly reducing the analysis to that of the linear setting, i.e. ℓ is the identity function. Hence, we start by presenting the analysis of the linear setting which is by [RV16], and is included for completeness.

10.1 Linear TS

Theorem 32. *For $\ell(x) = x$, $\psi(\xi) \leq d$ for all $\xi \in \mathscr{P}(\mathscr{F}_{\texttt{blrm}})$.*

Proof. First, we have

$$\bar{f}(\cdot) = \mathbb{E}[f(\cdot)] = \mathbb{E}[\langle \cdot, \theta \rangle] = \langle \cdot, \mathbb{E}[\theta] \rangle \,,$$

from which we define $\bar{\theta} = \mathbb{E}[\theta]$. Moreover, for $X \sim \pi^{\xi}_{\text{TS}}$ let

$$V = \mathbb{E}[XX^{\top}].$$

For the remainder of the proof we suppose that V is invertible, which holds as long X is not supported on any subspace of $\mathbb{R}^d$. If X was supported on such a subspace, the proof can be adapted to a basis of that subspace.

Let $\theta \sim \xi$, and $X = \arg\min_{x \in \mathcal{C}} \langle x, \theta \rangle$, and X' be an i.i.d copy of X, which is equivalent to $X' \sim \pi^{\xi}_{\text{TS}}$. Also let $\theta' \sim \xi$, be an i.i.d. copy of θ. We have

$$\begin{aligned} \Delta(\pi^{\xi}_{\text{TS}}, \xi) &= \mathbb{E}[\ell(X'^{\top}\theta) - \ell(X^{\top}\theta)] \\ &\overset{\texttt{(a)}}{=} \mathbb{E}[X'^{\top}\theta - X^{\top}\theta] \\ &\overset{\texttt{(b)}}{=} \mathbb{E}[X^{\top}\theta' - X^{\top}\theta] \\ &= \mathbb{E}[X^{\top}(\bar{\theta} - \theta)] \\ &\overset{\texttt{(c)}}{\leq} \mathbb{E}[\|X\|_{V^{-1}} \|\bar{\theta} - \theta\|_V] \\ &\overset{\texttt{(d)}}{\leq} \sqrt{\mathbb{E}[\|X\|^2_{V^{-1}}]\mathbb{E}[\|\bar{\theta} - \theta\|^2_V]} \,, \end{aligned}$$

where `(a)` is by ℓ being identity, `(b)` is by X' and θ' being i.i.d. copies of X and θ, and `(c)` and `(d)` are by Cauchy-Schwarz. Further, we have

$$\begin{aligned} \mathbb{E}\left[\|X\|^2_{V^{-1}}\right] &= \mathbb{E}\left[X^{\top}\mathbb{E}[XX^{\top}]^{-1}X\right] \\ &= \mathbb{E}\left[\text{trace}(X^{\top}\mathbb{E}[XX^{\top}]^{-1}X)\right] \\ &= \mathbb{E}\left[\text{trace}(XX^{\top}\mathbb{E}[XX^{\top}]^{-1})\right] \\ &= \text{trace}(\mathbb{E}[XX^{\top}]\mathbb{E}[XX^{\top}]^{-1}) \qquad \text{(linearity of trace and expectation)} \\ &= d. \end{aligned}$$

Moreover, we have

$$\begin{aligned}
\mathcal{I}(\pi^{\xi}_{\text{TS}}, \xi) &= E[(f(X) - \bar{f}(X))^2] \\
&= \mathbb{E}[(\ell(X^\top \theta') - \mathbb{E}[\ell(X^\top \theta')|X])^2] \\
&= \mathbb{E}[(X^\top(\theta' - \bar{\theta}))^2] \\
&= \mathbb{E}[(\theta' - \bar{\theta})^\top X X^\top (\theta' - \bar{\theta})] \\
&\overset{\texttt{(a)}}{=} \mathbb{E}[(\theta' - \bar{\theta})^\top V (\theta' - \bar{\theta})] \\
&= \mathbb{E}[\|\theta' - \bar{\theta}\|_V^2]\,,
\end{aligned}$$

where `(a)` follows from tower rule. Putting these together, we have

$$\begin{aligned}
\Psi(\xi) &= \frac{\Delta(\pi^{\xi}_{\text{TS}}, \xi)^2}{\mathcal{I}(\pi^{\xi}_{\text{TS}}, \xi)} \\
&\leq \frac{\mathbb{E}[\|X\|^2_{V^{-1}}]\mathbb{E}[\|\bar{\theta} - \theta\|_V^2]}{\mathbb{E}[\|\bar{\theta} - \theta\|_V^2]} \\
&= \mathbb{E}[\|X\|^2_{V^{-1}}] = d\,,
\end{aligned}$$

and the proof terminates. □

10.2 An Inequality for Convex Functions

To prove Lemma 36, we first provide two lemmas about convex functions, the first of which is only used to prove the second one.

Lemma 33. *Assume that $g : \mathbb{R} \to \mathbb{R}$ is a convex and non-decreasing function such that $g(0) = 0$. Then for any random variable X supported on $\mathbb{R}_+$ it holds that*

$$\frac{\mathbb{E}[g(X)^2]}{\mathbb{E}[g(X)]^2} \geq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2}.$$

Proof. Define the function $h(x) = \frac{\mathbb{E}[g(x)]}{\mathbb{E}[x]} g(x)$, and observe that

$$\frac{\mathbb{E}[g(X)^2]}{\mathbb{E}[g(X)]^2} = \frac{\mathbb{E}[h(X)^2]}{\mathbb{E}[h(X)]^2},$$

therefore it suffices to show the inequality for the function h. Note that $\mathbb{E}[h(X)] = \mathbb{E}[X]$ and the desired inequality is equivalent to

$$\mathbb{E}[h(X)^2] \geq \mathbb{E}[X^2].$$

Also, since h is convex there exists a point x_0 such that $h(x) \geq x_0$ for all $x \geq x_0$ and $h(x) \leq x_0$ for all $x \leq x_0$. We have

$$\begin{aligned}\mathbb{E}\left[(h(X)-X)\mathbb{I}_{\{h(X)\geq X\}}\right] &= \mathbb{E}\left[(h(X)-X)(1-\mathbb{I}_{\{h(X)\leq X\}})\right]\\ &= \mathbb{E}\left[-(h(X)-X)\mathbb{I}_{\{h(X)\leq X\}}\right]\\ &= \mathbb{E}\left[(X-h(X))\mathbb{I}_{\{h(X)\leq X\}}\right].\end{aligned}$$

Further, by the properties of x_0 we get

$$\begin{aligned}\mathbb{E}\left[\left(h(X)^2-X^2\right)\mathbb{I}_{\{h(X)\geq X\}}\right] &= \mathbb{E}\left[\left(h(X)-X\right)\left(h(X)+X\right)\mathbb{I}_{\{h(X)\geq X\}}\right]\\ &\geq \mathbb{E}\left[2x_0\left(h(X)-X\right)\mathbb{I}_{\{h(X)\geq X\}}\right]\\ &= \mathbb{E}\left[2x_0\left(X-h(X)\right)\mathbb{I}_{\{h(X)\leq X\}}\right]\\ &\geq \mathbb{E}\left[\left(X+h(X)\right)\left(X-h(X)\right)\mathbb{I}_{\{h(X)\leq X\}}\right]\\ &= \mathbb{E}\left[\left(X^2-h(X)^2\right)\mathbb{I}_{\{h(X)\leq X\}}\right],\end{aligned}$$

which by rearranging gives

$$\begin{aligned}0 &\geq -\mathbb{E}\left[\left(h(X)^2-X^2\right)\mathbb{I}_{\{h(X)\leq X\}}\right]+\mathbb{E}\left[\left(X^2-h(X)^2\right)\mathbb{I}_{\{h(X)\geq X\}}\right]\\ &= \mathbb{E}[X^2]-\mathbb{E}[h(X)^2]\end{aligned}$$

proving the desired inequality. □

Lemma 34. *Assume that $f:\mathbb{R}\to\mathbb{R}$ is a convex and non-decreasing function. Further, let X be a random variable supported on $\mathbb{R}$ and $x_1 = \operatorname{ess\,inf} X$. Then the following inequality holds*

$$\frac{\mathbb{E}[f(X)]-f(x_1)}{\sqrt{\mathbb{E}\left[\left(f(X)-\mathbb{E}[f(X)]\right)^2\right]}} \leq \frac{\mathbb{E}[X]-x_1}{\sqrt{\mathbb{E}\left[\left(X-\mathbb{E}[X]\right)^2\right]}}$$

Proof. First let $X' = X - x_1$, and observe that

$$\frac{\mathbb{E}[X]-x_1}{\sqrt{\mathbb{E}\left[\left(X-\mathbb{E}[X]\right)^2\right]}} = \frac{\mathbb{E}[X']}{\sqrt{\mathbb{E}\left[\left(X'-\mathbb{E}[X']\right)^2\right]}}.$$

Next define $g(x') = f(x' + x_1) - f(x_1)$ so that $g(X') = f(X) - f(x_1)$, and observe that

$$\frac{\mathbb{E}[f(X)] - f(x_1)}{\sqrt{\mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right]}} = \frac{\mathbb{E}[g(X')]}{\sqrt{\mathbb{E}\left[(g(X') - \mathbb{E}[g(X')])^2\right]}}.$$

Therefore, the desired inequality is proved once we show

$$\frac{\mathbb{E}[g(X')]}{\sqrt{\mathbb{E}\left[(g(X') - \mathbb{E}[g(X')])^2\right]}} \geq \frac{\mathbb{E}[X']}{\sqrt{\mathbb{E}\left[(X' - \mathbb{E}[X'])^2\right]}}.$$

Note that g is convex and non-decreasing. Also, we know that $\operatorname{ess\,inf} X' = \operatorname{ess\,inf} X - x_1 = 0$, and $g(0) = f(x_1) - f(x_1) = 0$, which implies that both $\mathbb{E}[X']$, $\mathbb{E}[g(X')]$ are positive, Therefore, we may further simplify the inequality by squaring and rearranging as

$$\begin{aligned} \frac{\mathbb{E}[g(X')]^2}{\mathbb{E}\left[(g(X') - \mathbb{E}[g(X')])^2\right]} &\geq \frac{\mathbb{E}[X']^2}{\mathbb{E}\left[(X' - \mathbb{E}[X'])^2\right]} \\ \Leftrightarrow \frac{\mathbb{E}[g(X')^2] - \mathbb{E}[g(X')]^2}{\mathbb{E}[g(X')]^2} &\leq \frac{\mathbb{E}[X'^2] - \mathbb{E}[X']^2}{\mathbb{E}[X']^2} \\ \Leftrightarrow \frac{\mathbb{E}[g(X')^2]}{\mathbb{E}[g(X')]^2} &\geq \frac{\mathbb{E}[X'^2]}{\mathbb{E}[X']^2}, \end{aligned} \tag{Alternative Inequality}$$

which holds by Lemma 33 and the fact that X' is supported on $\mathbb{R}_+$. □

Lemma 35. $(0, d) \in \mathrm{IR}(\mathscr{F}_{\mathtt{blrm}})$ *whenever* $\mathcal{K} = \mathbb{B}_1^d$ *and the link function* ℓ *is known to the learner.*

Proof. Let $\xi_{\mathrm{TS}} \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$ be the joint law of the parameter θ and the greedy action X, i.e.

$$(\theta, X) \sim \xi_{\mathrm{TS}}, \qquad \text{where } X = \operatorname*{arg\,min}_{x \in \mathcal{K}} \ell(\langle x, \theta \rangle).$$

It follows that

$$\Delta(\pi^{\xi}_{\mathrm{TS}}, \xi) = \mathbb{E}\left[\ell(X^\top \theta') - \ell(X^\top \theta)\right] \,, \qquad \mathcal{I}(\pi^{\xi}_{\mathrm{TS}}, \xi) = \mathbb{E}\left[(\ell(X^\top \theta') - \mathbb{E}[\ell(X^\top \theta')|X])^2\right] \,.$$

Define

$$\phi(X) = \frac{\mathbb{E}[\ell(X^\top \theta')|X] - \ell(X^\top \theta)}{X^\top \theta - \mathbb{E}[X^\top \theta' | X]}.$$

Let $(\tilde{\theta}, \tilde{X}) \sim \xi_{\mathrm{TS}}$ be such that

$$\tilde{\theta}^\top \tilde{X} = \operatorname{ess\,inf}_{(X,\theta)\sim\xi_{\mathrm{TS}}} X^\top \theta,$$

and write

$$\begin{aligned}
\frac{\mathbb{E}[\ell(\tilde{X}^\top \theta')|\tilde{X}] - \ell(\tilde{X}^\top \tilde{\theta})}{\sqrt{\mathbb{E}\left[\left(\ell(\tilde{X}^\top \theta) - \mathbb{E}[\ell(\tilde{X}^\top \theta')|\tilde{X}]\right)^2 |\tilde{X}\right]}} &\overset{\texttt{(a)}}{\leq} \frac{\mathbb{E}[\tilde{X}^\top \theta' | \tilde{X}] - \tilde{X}^\top \tilde{\theta}}{\sqrt{\mathbb{E}\left[\left(\tilde{X}^\top \theta' - \mathbb{E}[\tilde{X}^\top \theta' | \tilde{X}]\right)^2 |\tilde{X}\right]}} \\
&\overset{\texttt{(b)}}{=} \frac{\phi(\tilde{X})\tilde{X}^\top \theta - \mathbb{E}[\phi(\tilde{X})\tilde{X}^\top \theta' | \tilde{X}]}{\sqrt{\mathbb{E}\left[\left(\phi(\tilde{X})\tilde{X}^\top \theta' - \mathbb{E}[\phi(\tilde{X})\tilde{X}^\top \theta' | \tilde{X}]\right)^2 |\tilde{X}\right]}} \\
&\overset{\texttt{(c)}}{=} \frac{\ell(\tilde{X}^\top \tilde{\theta}) - \mathbb{E}[\ell(\tilde{X}^\top \theta')|\tilde{X}]}{\sqrt{\mathbb{E}\left[\left(\phi(\tilde{X})\tilde{X}^\top \theta' - \mathbb{E}[\phi(\tilde{X})\tilde{X}^\top \theta' | \tilde{X}]\right)^2 |\tilde{X}\right]}},
\end{aligned}$$

where `(a)` follows from

$$\ell(\tilde{X}^\top \theta') \overset{X' \ \texttt{greedy}}{\geq} \ell(X'^\top \theta') \overset{\tilde{X} \ \texttt{definition}}{\geq} \ell(\tilde{X}^\top \tilde{\theta})\,,$$

and Lemma 34, `(b)` follows from multiplying both numerator and denominator by $\phi(X)$, and

(c) follows from the definition of $\phi(X)$. Thus, we have

$$\mathbb{E}\left[\left(\ell(\tilde{X}^\top\theta') - \mathbb{E}[\ell(\tilde{X}^\top\theta')|\tilde{X}]\right)^2 |\tilde{X}\right] \geq \mathbb{E}\left[\left(\phi(\tilde{X})\tilde{X}^\top\theta' - \mathbb{E}[\phi(\tilde{X})\tilde{X}^\top\theta'|\tilde{X}]\right)^2 |\tilde{X}\right]. \tag{10.2}$$

Therefore, we have

$$\begin{aligned}\frac{\mathbb{E}[\ell(\tilde{X}^\top\theta')|\tilde{X}] - \ell(\tilde{X}^\top\tilde{\theta})}{\sqrt{\mathbb{E}\left[\left(\ell(\tilde{X}^\top\theta') - \mathbb{E}[\ell(\tilde{X}^\top\theta')|\tilde{X}]\right)^2 |\tilde{X}\right]}} &= \frac{\mathbb{E}[\phi(\tilde{X})\tilde{X}^\top\theta'|\tilde{X}] - \phi(\tilde{X})\tilde{X}^\top\tilde{\theta}}{\sqrt{\mathbb{E}\left[\left(\ell(\tilde{X}^\top\theta') - \mathbb{E}[\ell(\tilde{X}^\top\theta')|\tilde{X}]\right)^2 |\tilde{X}\right]}} \\ &\leq \frac{\mathbb{E}[\phi(\tilde{X})\tilde{X}^\top\theta'|\tilde{X}] - \phi(\tilde{X})\tilde{X}^\top\tilde{\theta}}{\sqrt{\mathbb{E}\left[\left(\phi(\tilde{X})\tilde{X}^\top\theta' - \mathbb{E}[\phi(\tilde{X})\tilde{X}^\top\theta'|\tilde{X}]\right)^2 |\tilde{X}\right]}} \\ &= \frac{\mathbb{E}[\tilde{X}^\top\theta'|\tilde{X}] - \tilde{X}^\top\tilde{\theta}}{\sqrt{\mathbb{E}\left[\left(\tilde{X}^\top\theta' - \mathbb{E}[\tilde{X}^\top\theta'|\tilde{X}]\right)^2 |\tilde{X}\right]}}\,. \end{aligned} \tag{10.3}$$

where the first equality follows from the definition of $\phi(\tilde{X})$ and the second inequality follows from 10.5.

Now, observe that

$$\begin{aligned}\psi(\xi) &= \frac{\Delta(\pi^\xi_{\text{TS}},\xi)^2}{\mathcal{I}(\pi^\xi_{\text{TS}},\xi)} \\ &= \frac{\mathbb{E}\left[\ell(X^\top\theta') - \ell(X^\top\theta)\right]^2}{\mathbb{E}\left[(\ell(X^\top\theta') - \mathbb{E}[\ell(X^\top\theta')|X])^2\right]} \\ &= \frac{\left(\mathbb{P}(X=\tilde{X})\mathbb{E}\left[\ell(\tilde{X}^\top\theta') - \ell(\tilde{X}^\top\tilde{\theta})|X=\tilde{X}\right] + \mathbb{P}(X\neq\tilde{X})\mathbb{E}\left[\ell(X^\top\theta') - \ell(X^\top\theta)|X\neq\tilde{X}\right]\right)^2}{\mathbb{P}(X=\tilde{X})\mathbb{E}\left[\left(\ell(\tilde{X}^\top\theta') - \mathbb{E}[\ell(\tilde{X}^\top\theta')|\tilde{X}]\right)^2 |X=\tilde{X}\right] + \mathbb{P}(X\neq\tilde{X})\mathbb{E}\Big[(\ell(X^\top\theta') - \mathbb{E}[\ell(X^\top\theta')|X])^2 |X}\end{aligned}$$

Further, define $\tilde{X} = \phi(X)X$ and rewrite the inequality as

$$\frac{\mathbb{E}\left[\ell(X^\top\theta) - \mathbb{E}[\ell(X^\top\theta')|X]\right]}{\sqrt{\mathbb{E}\left[(\ell(X^\top\theta') - \mathbb{E}[\ell(X^\top\theta')|X])^2\right]}} \leq \frac{\mathbb{E}\left[\tilde{X}^\top\theta - \mathbb{E}[\tilde{X}^\top\theta'|X]\right]}{\sqrt{\mathbb{E}\left[\left(\tilde{X}^\top\theta' - \mathbb{E}[\tilde{X}^\top\theta'|X]\right)^2\right]}} \leq \sqrt{d}\,,$$

where the $\sqrt{d}$ upper bound follows from Theorem 32. □

10.3 Information Ratio of TS for Known Ridge

We are now in position to prove Lemma 36.

Lemma 36. $(0, d) \in \mathrm{IR}(\mathscr{F}_{\mathtt{blrm}})$ *whenever* $\mathcal{K} = \mathbb{B}_1^d$ *and the link function* ℓ *is known to the learner.*

Proof. Similar to the proof of Theorem 32, let $X, X' \sim_{\text{i.i.d.}} \pi_{\text{TS}}^{\xi}$ and $\theta, \theta' \sim_{\text{i.i.d.}} \xi$, with $X = \arg\min_{x \in \mathcal{K}} \ell(\langle x, \theta \rangle)$.

Also note that since $\mathcal{K} = \mathbb{B}_1^d$, we have both

$$X = \arg\min_{x \in \mathcal{K}} \ell(\langle \theta, x \rangle), \quad \text{and} \quad \theta = \arg\min_{y \in \Theta} \ell(\langle y, X \rangle)\,. \tag{10.4}$$

Define

$$\phi(X) = \frac{\ell(X^\top \theta) - \mathbb{E}[\ell(X^\top \theta')|X]}{X^\top \theta - \mathbb{E}[X^\top \theta'|X]},$$

and write

$$\begin{aligned}
\frac{\ell(X^\top \theta) - \mathbb{E}[\ell(X^\top \theta')|X]}{\sqrt{\mathbb{E}\left[\left(\ell(X^\top \theta) - \mathbb{E}[\ell(X^\top \theta')|X]\right)^2 |X\right]}} &\overset{\texttt{(a)}}{\leq} \frac{X^\top \theta - \mathbb{E}[X^\top \theta'|X]}{\sqrt{\mathbb{E}\left[\left(X^\top \theta' - \mathbb{E}[X^\top \theta'|X]\right)^2 |X\right]}} \\
&\overset{\texttt{(b)}}{=} \frac{\phi(X) X^\top \theta - \mathbb{E}[\phi(X) X^\top \theta'|X]}{\sqrt{\mathbb{E}\left[\left(\phi(X) X^\top \theta' - \mathbb{E}[\phi(X) X^\top \theta'|X]\right)^2 |X\right]}} \\
&\overset{\texttt{(c)}}{=} \frac{\ell(X^\top \theta) - \mathbb{E}[\ell(X^\top \theta')|X]}{\sqrt{\mathbb{E}\left[\left(\phi(X) X^\top \theta' - \mathbb{E}[\phi(X) X^\top \theta'|X]\right)^2 |X\right]}},
\end{aligned}$$

where (a) follows from Lemma 34, (b) follows from multiplying both numerator and denominator by $\phi(X)$, and (c) follows from the definition of $\phi(X)$. Thus, we have

$$\mathbb{E}\left[\left(\ell(X^\top \theta') - \mathbb{E}[\ell(X^\top \theta')|X]\right)^2 |X\right] \geq \mathbb{E}\left[\left(\phi(X) X^\top \theta' - \mathbb{E}[\phi(X) X^\top \theta'|X]\right)^2 |X\right]. \tag{10.5}$$

Using this inequality, we can write

$$\begin{aligned}
\frac{\mathbb{E}\left[\ell(X^\top\theta) - \mathbb{E}[\ell(X^\top\theta')|X]\right]}{\sqrt{\mathbb{E}\left[\left(\ell(X^\top\theta') - \mathbb{E}[\ell(X^\top\theta')|X]\right)^2\right]}} &= \frac{\mathbb{E}\left[\phi(X)X^\top\theta - \mathbb{E}[\phi(X)X^\top\theta'|X]\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}\left[\left(\ell(X^\top\theta') - \mathbb{E}[\ell(X^\top\theta')|X]\right)^2|X\right]\right]}} \\
&\leq \frac{\mathbb{E}\left[\phi(X)X^\top\theta - \mathbb{E}[\phi(X)X^\top\theta'|X]\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}\left[\left(\phi(X)X^\top\theta' - \mathbb{E}[\phi(X)X^\top\theta'|X]\right)^2|X\right]\right]}}
\end{aligned}$$

where the first equality follows from the definition of $\phi(X)$ and the second inequality follows from 10.5. Further, define $\tilde{X} = \phi(X)X$ and rewrite the inequality as

$$\frac{\mathbb{E}\left[\ell(X^\top\theta) - \mathbb{E}[\ell(X^\top\theta')|X]\right]}{\sqrt{\mathbb{E}\left[\left(\ell(X^\top\theta') - \mathbb{E}[\ell(X^\top\theta')|X]\right)^2\right]}} \leq \frac{\mathbb{E}\left[\tilde{X}^\top\theta - \mathbb{E}[\tilde{X}^\top\theta'|X]\right]}{\sqrt{\mathbb{E}\left[\left(\tilde{X}^\top\theta' - \mathbb{E}[\tilde{X}^\top\theta'|X]\right)^2\right]}} \leq \sqrt{d}\,,$$

where the $\sqrt{d}$ upper bound follows from Theorem 32. □

Chapter 11

Empirical Performance of TS

In this chapter we study the empirical performance of TS for 1-dimensional $\texttt{BCO}$ problems. The main challenge for bringing TS to life for $\texttt{BCO}$ is designing a prior ξ on $\mathscr{F}_{\texttt{bl}}$ and a posterior ξ_t which can be sampled from efficiently. We design such a modelling but only for the 1-dimensional case. We don't provide any theoretical guarantees for our sampling scheme, but instead we show that it works well in practice.

11.1 A Probability Distribution on $\mathscr{F}_{\texttt{bl}}$

We consider the case where $\mathcal{K} = [-1, 1]$ and $d = 1$. We define a prior ξ on $\mathscr{F}_{\texttt{bl}}$ as follows. Let

$$\phi_y(x) = |x - y| \,,$$

which is a convex function. We use ϕ_ys as building blocks for making convex functions. Let $\epsilon > 0$ be a small constant, which specify the level of discretization of convex functions. Now let

$$\Phi = \{\phi_{\epsilon y} : \epsilon y \in \mathcal{K}, y \in \mathbb{Z}\} \,. \tag{11.1}$$

Let $d = \|\Phi\|$, and $\mathbb{R}_+$ as the set of non-negative reals. Then for any $\theta \in \mathbb{R}_+^d$,

$$f_\theta(x) = \sum_{i=1}^{d} \theta_i \phi_{y_i}(x) \,,$$

where y_i is the i-th element of Φ. Since $\theta \in \mathbb{R}_+^d$, the function f_θ is convex and non-negative. This construction allows us to define a probability distribution over the space of convex functions, by defining a probability measure on $\mathbb{R}_+^d$.

We suppose that $f_\star$, the true loss function, has the form

$$f_\star(x) = f_{\theta_\star}(x) = \sum_{i=1}^{d} \theta_{\star,i}\phi_{y_i}(x)\,, \tag{11.2}$$

and we assume that $\theta_\star$ is sampled from a $\mathbb{R}_+-$truncated multivariate Gaussian distribution with mean μ and covariance matrix Σ. We denote the law of $\theta_\star$ by ξ_θ, and the law of $f_\star$ by ξ. Further, we assume that

$$Y_t = f_{\theta_\star}(X_t) + \sigma\varepsilon_t\,, \tag{11.3}$$

where ε_t is a standard Gaussian noise, and σ is a positive constant. The posterior ξ_t is defined as the law of $\theta_\star$ given the observations $(X_s, Y_s)_{s=1}^t$.

11.1.1 Hit-and-Run Samplers

We use Hit-and-run (HAR) methods to sample from the posterior ξ_t. HAR methods generate proposals by first choosing a random direction and then moving along that line according to the target density restricted to the line. Originally developed for uniform sampling from convex bodies [e.g. Smi84, Lov99], they extend naturally to general log-concave (and even non log-concave) densities by sampling from the one-dimensional conditional along each line.

Let $\pi(\theta) \propto \tilde{\pi}(\theta)$ be the target on a convex subset $\mathcal{X} \subseteq \mathbb{R}^d$ (here $\mathcal{X} = \mathbb{R}_+^d$). Given the current state $\theta^{(m)}$, HAR proceeds by

(a) Draw a direction v from a spherically symmetric distribution (commonly $v \sim \mathcal{N}(0, I_d)$ then normalise $v \leftarrow v/\|v\|$).

(b) Determine the feasible interval

$$I(\theta^{(m)}, v) = \{t \in \mathbb{R} : \theta^{(m)} + tv \in \mathcal{X}\}.$$

In the positive orthant this is simply

$$t_{\min} = \max_{i:v_i<0} \frac{-\theta_i^{(m)}}{v_i}, \qquad t_{\max} = \min_{i:v_i>0} \frac{-\theta_i^{(m)}}{v_i}, \qquad I = [t_{\min}, t_{\max}] \,.$$

(c) Sample t from the one-dimensional density proportional to $\tilde{\pi}(\theta^{(m)} + tv)$ on I.

(d) Set $\theta^{(m+1)} = \theta^{(m)} + tv$.

The only non-trivial step is the one-dimensional draw in Step 3; different variants of HAR use different inner samplers. In our setting, we can sample from the truncated Gaussian posterior restricted to I.

11.2 Bayesian Regret

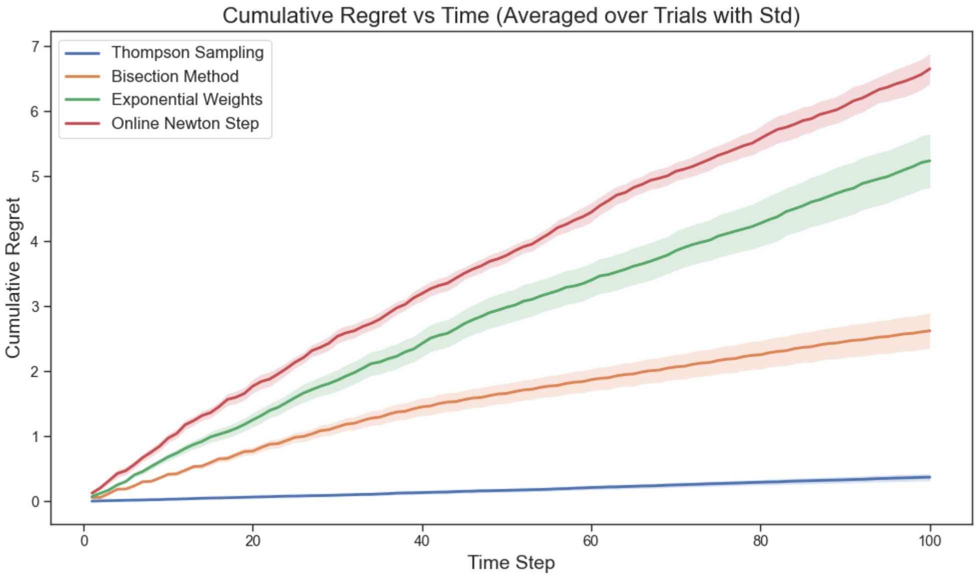


Figure 11.1: Bayesian regret for 1-dimensional bandit convex optimization. The plot shows the cumulative regret over $n = 100$ rounds, averaged over 10 independent runs. The loss functions are sampled from a multivariate Gaussian prior with $d = 100$ basis functions, and observations are corrupted with Gaussian noise $\mathcal{N}(0, 0.1^2)$.

First we study Bayesian regret. We first sample $\theta_\star$ from a multivariate Gaussian distribution that is known to the learner. We set the observation noise $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. The number of basis functions is set to $d = 100$. We run the experiment for $n = 100$ rounds, and report the average performance over 10 runs in Figure 11.1. The performance is compared to the Bisection Method [ADX10], the Exponential Weights with kernel estimation [BLE17], and Online Newton Step [FvdHLM24a]. The superiority of TS can be explained by the fact that it is able to use the knowledge of prior, and also the fact that most of the other algorithms are designed for the adversarial setting primarily.

11.3 Regret Against Specific Losses

One might wonder how well TS will perform if the loss function is not sampled from the prior ξ, and is instead fixed to a specific function $f_\star$. We study three loss functions:

(a) Absolute loss: $f_\star(x) = m\,|x - c|$, for some $c \in [-1, 1]$ and $m > 0$.

(b) Quadratic loss: $f_\star(x) = m(x - c)^2$, for some $c \in [-1, 1]$ and $m > 0$.

(c) Linear loss: $f_\star(x) = mx + b$, such that $\inf_{x \in [-1,1]} f_\star(x) \geq 0$.

We show the behavior of TS alongside other algorithms in Fig 11.2. The dashed lines are functions f_t that are sampled from ξ_t for $t \in [100]$, with newer samples been drawn with higher color density. The circles are the actions taken by different algorithms (the color specify the algorithm) with bigger circles showing more recent actions.

The regret of different algorithms against these losses is pictured in Fig 11.3. It can be seen that TS remains highly competitive even when the losses are chosen in an arbitrary way.

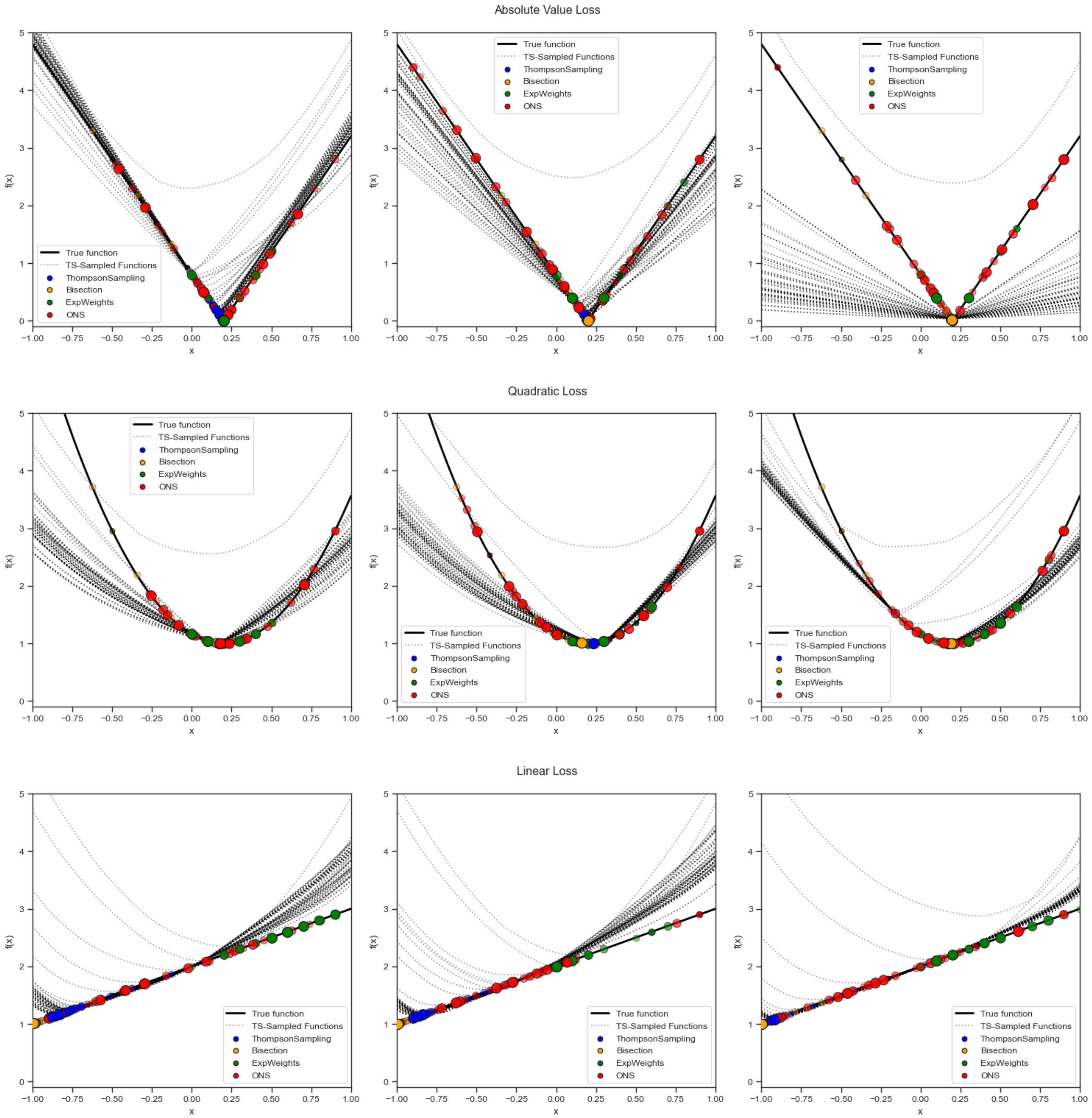
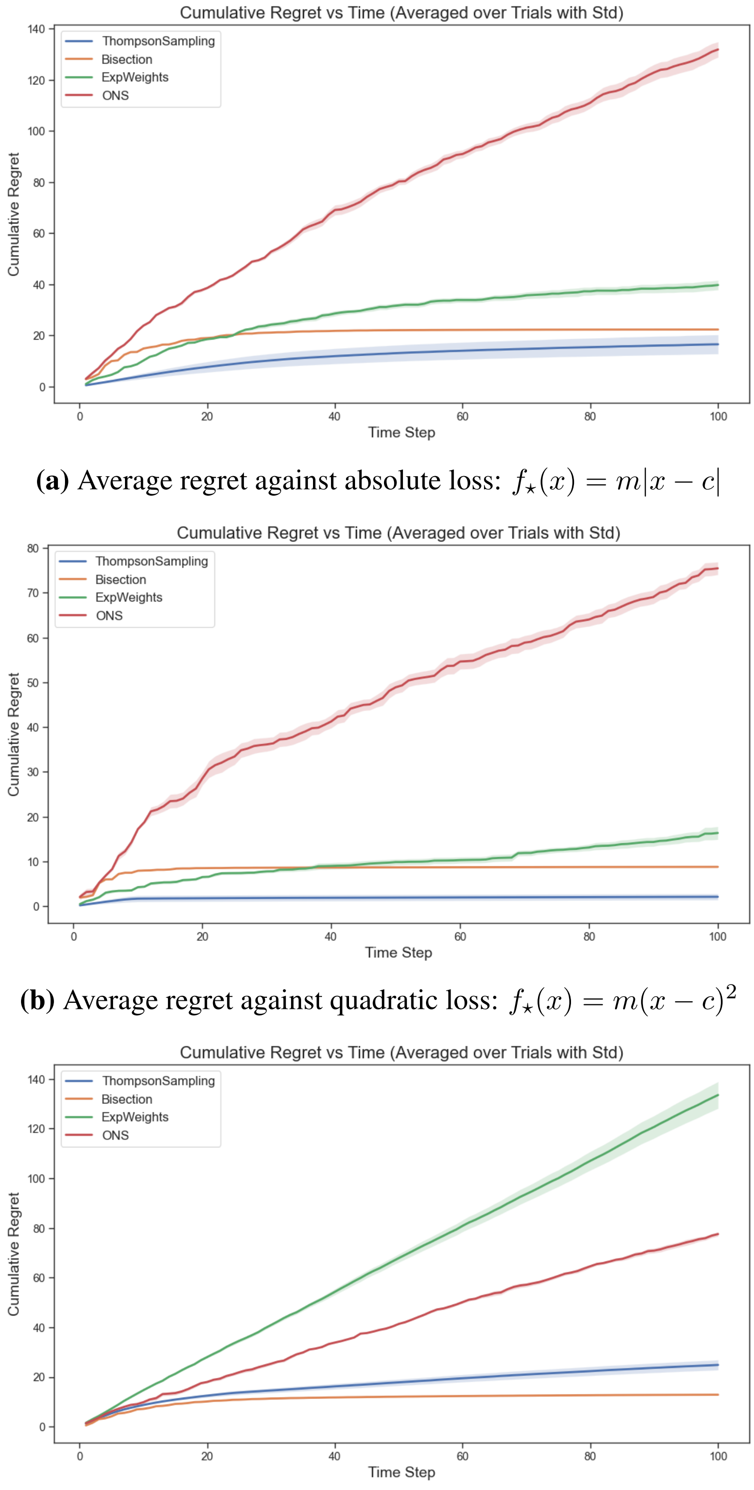


Figure 11.2: Performance of BCO algorithms against specific loss functions. Top: Absolute loss $f_\star(x) = m|x-c|$. Middle: Quadratic loss $f_\star(x) = m(x-c)^2$. Bottom: Linear loss $f_\star(x) = mx+b$. Solid line shows $f_\star$, dashed lines show TS sampled functions, and circles show actions taken by different algorithms.



(a) Average regret against absolute loss: $f_\star(x) = m|x - c|$

(b) Average regret against quadratic loss: $f_\star(x) = m(x - c)^2$

(c) Average regret against linear loss: $f_\star(x) = mx + b$

Figure 11.3: Average regret of TS and other algorithms against specific loss functions.

Chapter 12

Thompson Sampling for Adversarial Problems

In the Bayesian adversarial setting the prior is a probability measure on $\mathscr{F}^n$ and a whole sequence of loss functions is sampled in secret by the environment. The natural generalisations of TS in this setting are the following:

(a) Sample $(f_s)_{s=1}^n$ from the posterior and play $X_t = \arg\min_{x \in K} f_t(x)$.

(b) Sample $(f_s)_{s=1}^n$ from the posterior and play $X_t = \arg\min_{x \in K} \sum_{s=1}^t f_s(x)$.

The version in (a) suffers linear regret as the following example shows. Let $d = 1$ and $K = [-1, 1]$ and $f(x) = \epsilon + \max(\epsilon x, (1-\epsilon)x)$ and $g(x) = f(-x)$. Note that $f \in \mathscr{F}_{\mathtt{bl}}$ and is piecewise linear and minimised at -1 with $f(-1) = 0$ and $f(0) = \epsilon$ and $f(1) = 1$. The function g is the mirror image of f. Now let ν be the uniform distribution on $\{f, g\}$ and $\xi = \nu^n$ be the product measure. TS as defined in (a) plays uniformly on $\{1, -1\}$ and an elementary calculation shows that the regret is $\Omega(n)$.

We do not know if the version of TS defined in (b) has $\tilde{O}(\sqrt{n})$ Bayesian regret. However, the following example shows that in general the adversarial version of the information ratio is not bounded. Because the loss function changes from round to round, the action X_t may not minimise f_t. This must be reflected in the definition of the information ratio. Let ξ be a probability measure on $\mathscr{F} \times K$ and π be a probability measure on K and let

$$\Delta(\pi, \xi) = \mathbb{E}[f(X) - f(X_\star)] \qquad \text{and} \qquad \mathcal{I}(\pi, \xi) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)|X])^2]\,,$$

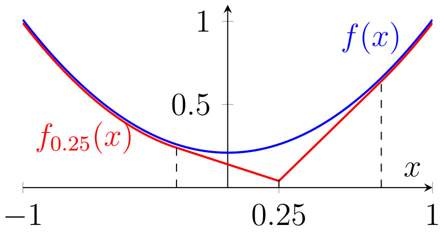


Figure 12.1: The function $f(x) = 0.2 + 0.8x^2$ and the function $f_{0.25}$ with $\epsilon = 0.2$. The function $f_{0.25}$ is the largest convex function that is smaller than f and has $f_{0.25}(0.25) = f(0.25) - 0.2$.

where $(f, X_\star, X)$ has law $\xi \otimes \pi$. Thompson sampling as in Item (b) is the policy π with the same law as $X_\star$. The claim is that in general it does not hold that

$$\Delta(\pi, \xi) \leq \alpha + \sqrt{\beta \mathcal{I}(\pi, \xi)}\,,$$

unless α is unreasonably large. Let $d = 1, \epsilon \in (0, 2^{-7}), K = [-1, 1]$, and $f(x) = \epsilon + (1 - \epsilon)x^2$. Given $\theta \in [-1, 1]$ let $f_\theta(x)$ be defined as

$$f_\theta(x) = \begin{cases} (1-\epsilon)(\theta^2 + 2(x-\theta)(\theta + \sqrt{\frac{\epsilon}{1-\epsilon}})) & \theta \leq x \leq \theta + \sqrt{\frac{\epsilon}{1-\epsilon}} \\ (1-\epsilon)(\theta^2 + 2(x-\theta)(\theta - \sqrt{\frac{\epsilon}{1-\epsilon}})) & \theta - \sqrt{\frac{\epsilon}{1-\epsilon}} \leq x < \theta \\ f(x) & \text{otherwise,} \end{cases}$$

which is convex and smaller than f for all $x \in K$. Essentially, f_θ should be thought of as the largest convex function that is smaller than f and has $f_\theta(\theta) = f(\theta) - \epsilon$ (see Fig. 12.1). Moreover, an elementary calculation shows that $\max_{x \in K} |f(x) - f_\theta(x)| = \epsilon$ for all $\theta \in [-1, 1]$. Let ξ be the law of (f_θ, θ) when θ is sampled uniformly from $[-1, 1]$ and π be uniform on $[-1, 1]$ which is the TS policy as defined in (b). Then, by letting θ' be an i.i.d. copy of θ we have

$$\Delta(\pi, \xi) = \mathbb{E}\left[f_\theta(\theta') - f_\theta(\theta)\right] = \mathbb{E}\left[f_\theta(\theta') - f(\theta)\right] + \epsilon = \mathbb{E}\left[f_{\theta'}(\theta) - f(\theta)\right] + \epsilon$$

where the second equality follows from the definition of f_θ and the third equality follows from

$f_\theta(\theta) = f(\theta) - \epsilon$. Next, we have

$$\begin{aligned}\mathbb{E}\left[f_{\theta'}(\theta) - f(\theta)\right] &= \mathbb{E}\left[\mathbf{1}_{\left\{|\theta-\theta'|\le\sqrt{\frac{\epsilon}{1-\epsilon}}\right\}}\left(f_{\theta'}(\theta) - f(\theta)\right) + \mathbf{1}_{\left\{|\theta-\theta'|\ge\sqrt{\frac{\epsilon}{1-\epsilon}}\right\}}\left(f_{\theta'}(\theta) - f(\theta)\right)\right] \\ &\overset{\texttt{(a)}}{=} \mathbb{E}\left[\mathbf{1}_{\left\{|\theta-\theta'|\le\sqrt{\frac{\epsilon}{1-\epsilon}}\right\}}\left(f_{\theta'}(\theta) - f(\theta)\right)\right] \\ &\overset{\texttt{(b)}}{\ge} -\mathbb{P}\left(|\theta-\theta'| \le \sqrt{\frac{\epsilon}{1-\epsilon}}\right)\epsilon \\ &\overset{\texttt{(c)}}{\ge} -2\sqrt{\frac{\epsilon}{1-\epsilon}}\epsilon\,,\end{aligned}$$

where `(a)` follows from the fact that $f_{\theta'}(\theta) = f(\theta)$ if $|\theta - \theta'| \ge \sqrt{\frac{\epsilon}{1-\epsilon}}$; `(b)` follows from the fact that $f_{\theta'} \le f(\theta)$ and the fact that $\max_{x\in K}|f(x) - f_\theta(x)| = \epsilon$; and `(c)` follows from the fact that θ and θ' are i.i.d. on $[-1, 1]$. Therefore,

$$\Delta(\pi, \xi) \ge \epsilon\left(1 - 2\sqrt{\frac{\epsilon}{1-\epsilon}}\right)\,.$$

Next, we turn our attention to $\mathcal{I}(\pi, \xi)$, which can be upper bounded as

$$\begin{aligned}\mathcal{I}(\pi, \xi) &= \mathbb{E}\left[\left(f_{\theta'}(\theta) - \mathbb{E}\left[f_{\theta'}(\theta)|\theta\right]\right)^2\right] \\ &\overset{\texttt{(a)}}{\le} \mathbb{E}\left[\left(f_{\theta'}(\theta) - f(\theta)\right)^2\right] \\ &\overset{\texttt{(b)}}{\le} \mathbb{P}\left(|\theta-\theta'| \le \sqrt{\frac{\epsilon}{1-\epsilon}}\right)\epsilon^2 \\ &\overset{\texttt{(c)}}{\le} 2\epsilon^2\sqrt{\frac{\epsilon}{1-\epsilon}}\,,\end{aligned}$$

where `(a)` follows from the fact that the mean minimizes the squared deviation; `(b)` follows from the fact that $f_{\theta'}(\theta) = f(\theta)$ if $|\theta - \theta'| \ge \sqrt{\frac{\epsilon}{1-\epsilon}}$; and `(c)` follows from the fact that θ and θ' are i.i.d. on $[-1, 1]$. Therefore, by putting the two inequalities together we have

$$\frac{\Delta(\pi,\xi)}{\sqrt{\mathcal{I}(\pi,\xi)}} \ge \frac{\epsilon\left(1 - 2\sqrt{\frac{\epsilon}{1-\epsilon}}\right)}{\sqrt{2\epsilon^2\sqrt{\frac{\epsilon}{1-\epsilon}}}} = \frac{1 - \sqrt{\frac{4\epsilon}{1-\epsilon}}}{\sqrt[4]{\frac{4\epsilon}{1-\epsilon}}} = \sqrt[4]{\frac{1-\epsilon}{4\epsilon}} - \sqrt[4]{\frac{4\epsilon}{1-\epsilon}}\,,$$

which can be further lower bounded by

$$\frac{\Delta(\pi,\xi)}{\sqrt{\mathcal{I}(\pi,\xi)}} \ge \sqrt[4]{\frac{1-\epsilon}{4\epsilon}} - \sqrt[4]{\frac{4\epsilon}{1-\epsilon}} \ge \sqrt[4]{\frac{1}{4\epsilon} - \frac{1}{4}} - \sqrt[4]{8\epsilon} \ge \sqrt[4]{\frac{1}{8\epsilon}} - \sqrt[4]{8\epsilon} \ge \frac{1}{4}\epsilon^{-\frac{1}{4}}\,,$$

where the all inequalities follow from $\epsilon \in (0, 2^{-7})$. Therefore, the information ratio is unbounded as $\epsilon \to 0$.

Chapter 13

Discussion

In this thesis, we explored the performance of Thompson sampling (TS) for Bayesian bandit convex optimization (BCO), revealing a nuanced picture of its capabilities. We established its near-optimal performance in one-dimensional settings and for the class of monotone ridge functions, demonstrating its efficacy in structured scenarios. In stark contrast, we also showed that TS can fail dramatically in general high-dimensional problems and that standard analytical tools face fundamental limitations. This chapter synthesizes these findings, discusses their broader implications for algorithm design in online optimization, and explores the structural properties that determine the success or failure of TS.

13.1 Adversarial setup

In the Bayesian adverarial setting a sequence of loss functions $f_1, \dots, f_n$ are sampled from a joint distribution on $\mathscr{F}^n$. The learner plays X_t and observes $Y_t = f_t(X_t)$ and the Bayesian regret is $\mathrm{BReg}(\mathscr{A}, \xi) = \mathbb{E}[\sup_{x \in K} \sum_{t=1}^n (f_t(X_t) - f_t(x))]$. One can envisage two possible definitions of Thompson sampling in this setting. One samples g_t from the marginal of the posterior and plays $X_t = x_{g_t}$. The second samples $g_1, \dots, g_n$ from the posterior and plays X_t as the minimiser of $\sum_{t=1}^n g_t$. The former has linear regret, while [BDKP15] notes that the latter has an unbounded information ratio. The situation was discussed in more details in Chapter 12.

13.2 Tightness of bounds

At present we are uncertain whether or not the monotonicity assumption is needed in the ridge setting. Our best guess is that it is not. One may also wonder if the bound on the information ratio in Theorem 18 can be improved. We are cautiously believe that when the loss has the form $f(x) = \ell(\langle x, \theta \rangle)$ for *known* convex link function $\ell : \mathbb{R} \to \mathbb{R}$, then the information ratio is at most d. This would mean that convex generalised linear bandits are no harder than linear bandits.

13.3 TS vs IDS

Theorem 22 shows that TS can have more-or-less linear regret in high-dimensional problems. On the other hand, [BE18] and [Lat20] show that IDS has a well-controlled information ratio, but is much harder to compute. An obvious question is whether some simple adaptation of Thompson sampling has a well-controlled information ratio.

13.4 Applications

Many problems are reasonably modelled as 1-dimensional convex bandits, with the classical example being dynamic pricing where K is a set of prices and convexity is a reasonable assumption based on the response of demand to price. The monotone ridge function class is a natural model for resource allocation problems where a single resource (e.g., money) is allocated to d locations. The success of some global task increases as more resources are allocated, but with diminishing returns. Problems like this can reasonably be modelled by convex monotone ridge functions with $K = \{x \geq \mathbf{0} : \|x\|_1 \leq 1\}$.

13.5 Lipschitz assumption

Our bounds depend logarithmically on the Lipschitz constant associated with the class of loss functions. There is a standard trick to relax this assumption based on the observation that bounded convex functions must be Lipschitz on a suitably defined interior of the constraint set K. Concretely, suppose that K is a convex body and $f : K \to [0, 1]$ is convex and $\mathbb{B}_r \subset K$ and $K_\epsilon = (1 - \epsilon)K$. Then $\min_{x \in K_\epsilon} f(x) \leq \inf_{x \in K} f(x) + \epsilon$ and f is $1/(r\epsilon)$-Lipschitz on K_ϵ [Lat24, Chapter 3]. Hence, you can run TS on K_ϵ with $\epsilon = 1/n$ and the Lipschitz constant is at most n/r. Moreover, if K is in (approximate) isotropic or John's position, then $\mathbb{B}_1 \subset K \subset \mathbb{B}_{2d}$ by [KLS95] and John's theorem, respectively.

13.6 Frequentist regret

An ambitious goal would be to prove a bound on the frequentist regret of TS for some well-chosen prior. This is already quite a difficult problem in multi-armed [KKM12, AG12] and linear bandits [AG13] and is out of reach of the techniques developed here. On the other hand, the Bayesian algorithm has the advantage of being able to specify a prior that makes use of background knowledge and the theoretical guarantees for TS provide a degree of comfort.

13.7 Choice of prior

The choice of the prior depends on the application. A variety of authors constructed priors supported on non-parametric classes of 1-dimensional convex functions using a variety of methods [RLS93, CCH$^+$07, SWD11]. In many cases you may know the loss belongs to a simple parametric class, in which case the prior and posterior computations may simplify dramatically.

Part II

Anytime Estimation to Decision Algorithm

Chapter 14

Decision Making With Structured Observations

Regret minimization is a widely studied objective in bandits and reinforcement learning theory [LS20a] that has inspired practical algorithms, for example, in noisy zero-order optimization[e.g., SKKS10a] and deep reinforcement learning [e.g., OBPVR16]. Cumulative regret measures the online performance of the algorithm by the total loss suffered due to choosing suboptimal decisions. Regret is unavoidable to a certain extent as the learner collects information to reduce uncertainty about the environment. In other words, a learner will inevitably face the exploration-exploitation trade-off where it must balance collecting rewards and collecting information. Finding the right balance is the central challenge of sequential decision-making under uncertainty.In this part of the thesis, we study the problem of regret minimization in sequential decision-making with structured observations, as defined in the following section.

14.1 Problem Setup

More formally, denote by Π a decision space and $\mathcal{O}$ an observation space. Let $\mathcal{H}$ be a class of models, where $f = (r_f, M_f) \in \mathcal{H}$ associated with a reward function $r_f : \Pi \to \mathbb{R}$ and observation map $M_f : \Pi \to \mathcal{P}(\mathcal{O})$, where $\mathcal{P}(\mathcal{O})$ is the set of all probability distributions over $\mathcal{O}$.[1] The learner's objective is to collect as much reward as possible in n steps when facing a model $f^* \in \mathcal{H}$.

[1] To simplify the presentation, we ignore tedious measure-theoretic details in this part. The reader could either fill out the missing details, or just assume that all sets, unless otherwise stated, are discrete.

The learner's prior information is $\mathcal{H}$ and the associated reward and observation maps, but does not know the true instance $f^* \in \mathcal{H}$. The learner constructs a stochastic sequence $\pi_1, \ldots, \pi_n$ of decisions taking values in Π and adapted to the history of observations $y_t \sim M_{f^*}(\pi_t)$. The policy of the learner is the sequence of probability kernels $\mu_{1:n} = (\mu_t)_{t=1}^n$ that are used to take decisions. The expected regret of a policy $\mu_{1:n}$ and model f^* after $n \in [N]$ steps is

$$R_n(\mu_{1:n}, f^*) = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^n r_{f^*}(\pi) - r_{f^*}(\pi_t)\right]$$

The literature studies regret minimization for various objectives, including worst-case and instance-dependent frequentist regret [LS20a], Bayesian regret [RV16] and robust variants [GRM+20, KBJK20]. For the frequentist analysis, all prior knowledge is encoded in the model class $\mathcal{H}$. The worst-case regret of policy $\mu_{1:n}$ on $\mathcal{H}$ is $\sup_{f \in \mathcal{H}} R_n(\mu_{1:n}, f)$, and therefore the optimal minimax regret $\inf_\mu \sup_{f \in \mathcal{H}} R_n(\mu_{1:n}, f)$ only depends on $\mathcal{H}$ and the horizon n. The Bayesian, in addition, assumes access to a prior $\nu \in \mathcal{P}(\mathcal{H})$, which leads to the Bayesian regret $\mathbb{E}_{f \sim \nu}[R_n(\mu_{1:n}, f)]$. Interestingly, the worst-case frequentist regret and Bayesian regret are dual in the following sense [LS19]:[2]

$$\inf_{\mu_{1:n}} \sup_{f \in \mathcal{H}} R_n(\mu_{1:n}, f) = \sup_{\nu \in \mathcal{P}(\mathcal{H})} \inf_{\mu_{1:n}} \mathbb{E}_{f \sim \nu}[R_n(\mu_{1:n}, f)] \tag{14.1}$$

Unfortunately, directly solving for the minimax policy (or the worst-case prior) is intractable, except in superficially simple problems. This is because the optimization is over the exponentially large space of adaptive policies. However, the relationship in Eq. (14.1) has been directly exploited in prior works, for example, to derive non-constructive upper bounds on the worst-case regret via a Bayesian analysis [BDKP15]. Moreover, it can be seen as inspiration underlying "optimization-based" algorithms for regret minimization: The crucial step is to carefully relax the saddle point problem in a way that preserves the statistical complexity, but can be analyzed and computed more easily. This idea manifests in several closely related algorithms, including information-directed sampling [RV14, KK18], ExpByOpt [LS20b, LG21], and most recently, the Estimation-To-Decisions (E2D) framework [FKQR21, FGH23]. These algorithms have in common that they optimize the information trade-off directly, which in structured settings leads to large improvements compared to standard optimistic exploration approaches and Thompson sampling. Yet, the precise relation among the different approaches are not yet fully understood. On

[2] The result by [LS19] was only shown for finite action, reward and observation spaces, but can likely be extended to the infinite case under suitable continuity assumptions.

the other hand, algorithms that directly optimize the information trade-off can be computationally more demanding and, consequently, are often not the first choice of practitioners. This is partly due to the literature primarily focusing on statistical aspects, leaving computational and practical considerations underexplored.

14.1.1 Related Work

There is a broad literature on regret minimization in bandits [LS20a] and reinforcement learning [JAZBJ18, AOM17, ZGS21, DKL$^+$21, ZLKB20]. Arguably the most popular approaches are based on optimism, leading to the widely analysed upper confidence bound (UCB) algorithms [LS20a], and Thompson sampling (TS) [Tho33b, RV16].

A long line of work approaches regret minimization as a saddle point problem. [DSK20] showed that in the structured bandit setting, an algorithm based on solving a saddle point equation achieves asymptotically optimal regret bounds, while explicitly controlling the finite-order terms. [LS20b] propose an algorithm based on exponential weights in the partial monitoring setting [Rus99] that finds a distribution for exploration by solving a saddle-point problem. The saddle-point problem balances the trade-off between the exponential weights distribution and an information or stability term. The same approach was further refined by [LG21]. In stochastic linear bandits, [KLVS21] demonstrated that information-directed sampling can be understood as a primal-dual method solving the asymptotic lower bound, which leads to an algorithm that is both worst-case and asymptotically optimal. The saddle-point approach has been further explored in the PAC setting [e.g., DSK20, DMSV20].

The work in this part or the thesis is closely related to recent work by [FKQR21, FGH23]. They consider *decision making with structured observations* (DMSO), which generalizes the bandit and RL setting. They introduce a complexity measure, the *offset decision-estimation coefficient* (offset DEC), defined as a min-max game between a learner and an environment, and provide lower bounds in terms of the offset DEC. Further, they provide an algorithm, *Estimation-to-Decisions* (E2D) with corresponding worst-case upper bounds in terms of the offset DEC. Notably, the lower and upper bound nearly match and recover many known results in bandits and RL. The offset DEC is More recently, [FGH23] refined the previous bounds by introducing the *constrained* DEC and a corresponding algorithm E2D$^+$. Although achieving better bounds, this algorithm and the analysis is significantly more involved than for the E2D algorithm.

There are various other results related to the DEC and the E2D algorithm. [FGQ$^+$22] show that the E2D achieves improved bounds in model-free RL when combined with optimistic estimation (as introduced by [Zha22]). [CMB22] introduced two new complexity measures based on the DEC that are necessary and sufficient for reward-free learning and PAC learning. They also introduced new algorithms based on the E2D algorithm for the above two settings and various other improvements. [FRSS22] have shown that the DEC is necessary and sufficient to obtain low regret for *adversarial* decision-making. An asymptotically instance-optimal algorithm for DMSO has been proposed by [DM22], extending a similar approach for the linear bandit setting [LS17].

The decision-estimation coefficient is also related to the information ratio [RV14] and the decoupling coefficient [Zha22]. The information ratio has been studied under both the Bayesian [RV14] and the frequentist regret [KK18, KLK20, KLVS21, KLK23] in various settings including bandits, reinforcement learning, and partial monitoring. The decoupling coefficient was studied for the Thompson sampling algorithm in contextual bandits [Zha22], and RL [DMZZ21, AZ22].

14.2 The Problem Setup

Recall that Π is a compact decision space and $\mathcal{O}$ is an observation space. The model class $\mathcal{H}$ is a set of tuples $f = (r_f, M_f)$ containing a reward function $r_f : \Pi \to \mathbb{R}$ and an observation distribution $M_f : \Pi \to \mathscr{P}(\mathcal{O})$. Further bounded- and finiteness assumptions will be introduced in the context of the relevant statements. We make the assumption that both $\mathcal{H}$ and Π are finite, however, our results extend to continuous action and hypothesis spaces under appropriate technical conditions using standard arguments. In particular, the sample complexity bounds we provide scale with $\log(|\mathcal{H}|)$ in many cases, facilitating covering arguments for the continuous setting. Computationally, we will assume that $\mathcal{H}$ can be computationally enumerated (or rely on subsampling). While this is a strong assumption, it is less clear under what conditions one can achieve better results; and even the case where $\mathcal{H}$ is finite, computational aspects of E2D have not been explored in the literature. We define the gap function

$$\Delta(\pi, g) = r_g(\pi_g^*) - r_g(\pi)\,,$$

where $\pi_g^* = \arg\max_{\pi\in\Pi} r_g(\pi)$ is an optimal decision for model g, chosen arbitrarily if not unique. A randomized policy is a sequence of kernels $\mu_{1:n} = (\mu_t)_{t=1}^n$ from histories $h_{t-1} = (\pi_1, y_1, \dots, \pi_{t-1}, y_{t-1}) \in (\Pi \times \mathcal{O})^{t-1}$ to sampling distributions $\mathscr{P}(\Pi)$. The filtration generated by the history h_t is $\mathcal{F}_t$. The learner's decisions $\pi_1, \dots, \pi_n$ are sampled from the policy $\pi_t \sim \mu_t$ and observations $y_t \sim M_{f^*}(\pi_t)$

are generated by an unknown true model $f^* \in \mathcal{H}$. The expected regret under model f^* is formally defined as follows:

$$R_n(\mu_{1:n}, f^*) = \mathbb{E}[\sum_{t=1}^{n} \mathbb{E}_{\pi_t \sim \mu_t(h_t)}[\Delta(\pi_t, f^*)]]$$

For now, we do not make any assumption about the reward being observed. This provides additional flexibility to model a wide range of scenarios, including for example, duelling and ranking feedback [YJ09, RKJ08, CMPL15, LKLS18, KK21] (e.g. used in reinforcement learning with human feedback, RLHF) or dynamic pricing[dB15]. The setting is more widely known as partial monitoring [Rus99]. The special case where the reward is part of the observation distribution is called *decision-making with structured observations* [DMSO, FKQR21]. Earlier work studies the closely related *structured bandit* setting [CMP17].

A variety of examples across bandit models and reinforcement learning are discussed in [CMP17, FKQR21, FGH23, KLK23]. For the purpose of this part, we focus on simple cases for which we can provide tractable implementations. Besides the finite setting where $\mathcal{M}$ can be enumerated, these are the following linearly parametrized feedback models.

Example 37 (Linear Bandits, [AL99])**.** The model class is identified with a subset of $\mathbb{R}^d$ and features $\phi_\pi \in \mathbb{R}^d$ for each $\pi \in \Pi$. The reward function is $r_f(\pi) = \langle \phi(\pi), f \rangle$ and the observation distribution is $M_f(\pi) = \mathcal{N}(\langle \phi_\pi, f \rangle, 1)$.

The linear bandit setting can be generalized by separating reward and feedback maps, which leads to the *linear partial monitoring* framework [LAK$^+$14, KLK20]. Here we restrict our attention to the special case of *linear bandits with side-observations* [c.f. KLK23], which, for example, generalizes the classical semi-bandit setting [MS11]

Example 38 (Linear Bandits with Side-Observations)**.** As in the linear bandit setting, we have $\mathcal{H} \subset \mathbb{R}^d$, and features $\phi_\pi \in \mathbb{R}^d$ that define the reward functions $r_f(\pi) = \langle \phi_\pi, f \rangle$. Observation matrices $M_\pi \in \mathbb{R}^{m_\pi \times d}$ for each $\pi \in \Pi$ define m_π-dimensional observation distributions $M_f(\pi) = \mathcal{N}(M_\pi f, \sigma^2 \mathbf{1}_{m_\pi})$. In addition, we assume that $\phi_\pi \phi_\pi^\top \preceq M_\pi^\top M_\pi$, which is automatically satisfied if $\phi_\pi^\top$ is included in the rows of M_π, i.e. when the reward is part of the observations.

Chapter 15

Regret Minimization via Saddle-Point Optimization

The goal of the learner is to choose decisions $\pi \in \Pi$ that achieve a small gap $\Delta(\pi, f^*)$ under the true model $f^* \in \mathcal{H}$. Since the true model is unknown, the learner has to collect data that provides statistical evidence to reject models $g \neq f^*$ for which the regret $\Delta(\pi, g)$ is large. To quantify the information-regret trade-off, we use a divergence $D(\cdot\|\cdot)$ defined for distributions in $\mathscr{P}(\mathcal{O})$. For a reference model f, the information (or divergence) function is defined by:

$$I_f(\pi, g) = \mathrm{KL}(M_g(\pi)\|M_f(\pi))\,,$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the KL divergence. Intuitively, $I_f(\pi, g)$ is the rate at which the learner collects statistical information to reject $g \in \mathcal{H}$ when choosing $\pi \in \Pi$ and data is generated under the reference model f. Note that $I_f(\pi, f) = 0$ for all $f \in \mathcal{H}$ and $\pi \in \Pi$. As we will see shortly, the regret-information trade-off can be written precisely as a combination of the gap function, Δ, and the information function, I_f. We remark in passing that other choices such as the Hellinger distance are also possible, and the KL divergence is mostly for concreteness and practical reasons.

To simplify the notation and emphasize the bilinear nature of the saddle point problem that we study, we will view $\Delta, I_f \in \mathbb{R}_+^{\Pi\times\mathcal{H}}$ as $|\Pi| \times |\mathcal{H}|$ matrices (by fixing a canonical ordering on Π and $\mathcal{H}$). For vectors $\mu \in \mathbb{R}^\Pi$ and $\nu \in \mathbb{R}^{\mathcal{H}}$, we will frequently write bilinear forms $\mu\Delta\nu$ and $\mu I_f \nu$. This also means that by convention, μ will always denote a row vector, while ν will always denote a column vector. The standard basis for $\mathbb{R}^\Pi$ and $\mathbb{R}^{\mathcal{H}}$ is $(e_\pi)_{\pi\in\Pi}$ and $(e_g)_{g\in\mathcal{H}}$.

15.0.1 The Decision-Estimation Coefficient

To motivate our approach, we recall the *decision-estimation coefficient* (DEC) introduced by [FKQR21, FGH23], before introducing the main quantity of interest, the *average-constrained DEC*. First, the *offset decision-estimation coefficient* (without localization) [FKQR21] is

$$\operatorname{dec}^o_\lambda(f) = \min_{\mu \in \mathscr{P}(\Pi)} \max_{g \in \mathcal{H}} \mu \Delta e_g - \lambda \mu I_f e_g$$

The tuning parameter $\lambda > 0$ controls the weight of the information matrix relative to the gaps: Viewing the above as a two-player zero-sum game, we see that increasing λ forces the max-player to avoid models that differ significantly from f under the min-player's sampling distribution. The advantage of this formulation is that the information term $\mu I_f e_g$ can be telescoped in the analysis, which directly leads to regret bounds in terms of the estimation error (introduced below in Eq. (16.1)). The disadvantage of the λ-parametrization is that the trade-off parameter is chosen by optimizing the final regret upper bound. This is inconvenient because the optimal choice requires knowledge of the horizon and a bound on $\max_{f \in \mathcal{H}} \operatorname{dec}^o_\lambda(f)$. Moreover, any choice informed by the upper bound may be conservative, leading to sub-optimal performance.

The *constrained decision-estimation coefficient* [FGH23] is

$$\operatorname{dec}^c_\epsilon(f) = \min_{\mu \in \mathscr{P}(\Pi)} \max_{g \in \mathcal{H}} \mu \Delta e_g \qquad \text{s.t.} \qquad \mu I_f e_g \leq \epsilon^2 \tag{15.1}$$

In this formulation, the max player is restricted to choose models g that differ from f at most by ϵ^2 in terms of the observed divergence under the min-player's sampling distribution. Note that because $e_\pi I_f e_f = 0$ for all $e_\pi \in \Pi$, there always exists a feasible solution. For horizon n, the radius can be set to $\epsilon^2 \approx \frac{\beta_{\mathcal{H}}}{n}$, where $\beta_{\mathcal{H}}$ is a model estimation complexity parameter, thereby essentially eliminating the trade-off parameter from the algorithm. However, because of the hard constraint, strong duality of the Lagrangian saddle point problem (for fixed μ) fails, and consequently, telescoping the information gain in the analysis is no longer easily possible (or at least, with the existing analysis). To achieve sample complexity $\operatorname{dec}^c_\epsilon(f)$, [FGH23] propose a sophisticated scheme that combines phased exploration with a refinement procedure (E2D$^+$).

As the main quantity of interest in the current work, we now introduce the *average-constrained decision-estimation coefficient*, defined as follows:

$$\operatorname{dec}^{ac}_\epsilon(f) = \min_{\mu \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \mu \Delta \nu \qquad \text{s.t.} \qquad \mu I_f \nu \leq \epsilon^2 \tag{15.2}$$

Similar to the dec^c_ϵ, the parameterization of the $\mathrm{dec}^{ac}_\epsilon$ is via the confidence radius ϵ^2, making the choice of the hyperparameter straightforward in many cases. By convexifying the domain $\mathscr{P}(\mathcal{H})$ of the max-player, we recover strong duality of the Lagrangian (for fixed μ). Thereby, the formulation inherits the ease of choosing the ϵ-parameter from the dec^c_ϵ, while, at the same time, admitting a telescoping argument in the analysis and a much simpler algorithm.

Specifically, Sion's theorem implies three equivalent Lagrangian representations for Eq. (15.2):

$$\begin{aligned} \mathrm{dec}^{ac}_\epsilon(f) &= \min_{\mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathscr{P}(\mathcal{H})} \min_{\lambda\geq 0} \mu\Delta\nu - \lambda(\mu I_f \nu - \epsilon^2) &(15.3)\\ &= \min_{\lambda\geq 0, \mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathscr{P}(\mathcal{H})} \mu\Delta\nu - \lambda(\mu I_f \nu - \epsilon^2) &(15.4)\\ &= \min_{\lambda\geq 0} \max_{\nu\in\mathscr{P}(\mathcal{H})} \min_{\mu\in\mathscr{P}(\Pi)} \mu\Delta\nu - \lambda(\mu I_f \nu - \epsilon^2)\,. &(15.5) \end{aligned}$$

When fixing the outer problem, strong duality holds for the inner saddle-point problem in each line, however, the joint program in Eq. (15.4) is not convex-concave. An immediate consequence of relaxing the domain of the max player and Eq. (15.4) is that

$$\mathrm{dec}^c_\epsilon(f) \leq \mathrm{dec}^{ac}_\epsilon(f) = \min_{\lambda\geq 0}\{\mathrm{dec}^o_\lambda(f) + \lambda\epsilon^2\} \tag{15.6}$$

The $\mathrm{dec}^{ac}_\epsilon$ can therefore be understood as setting the λ parameter of the dec^o_λ optimally for the given confidence radius ϵ^2. On the other hand, the cost paid for relaxing the program is that there exist model classes $\mathcal{H}$ where the inequality in Eq. (15.6) is strict, and $\mathrm{dec}^{ac}_\epsilon$ does not lead to a tight characterization of the regret [FGH23, Proposition 4.4]. The remedy is that under a stronger regularity condition and localization, the two notions are essentially equivalent [FGH23, Proposition 4.8].

Chapter 16

Anytime Estimation-To-Decisions (Anytime-E2D)

Estimations-To-Decisions (E2D) is an algorithmic framework that directly leverages the decision-estimation coefficient for choosing a decision in each round. The key idea is to compute a sampling distribution $\mu_t \in \mathscr{P}(\Pi)$ attaining the minimal DEC for an estimate $\hat{f}_t$ of the underlying model, and then define the policy to sample $\pi_t \sim \mu_t$. The E2D approach, using the $\operatorname{dec}_\epsilon^{ac}$ formulation, is summarized in Algorithm 3. To compute the estimate $\hat{f}_t$, the E2D algorithm takes an abstract estimation oracle EST as input, that, given the collected data, returns $\hat{f}_t \in \mathcal{M}$. The final guarantee depends on the *estimation error* (or estimation regret), defined as the sum over divergences of the observation distributions under the estimate $\hat{f}_t$ and the true model f^*:

$$\mathrm{Est}_n = \mathbb{E}\left[\sum_{t=1}^{n} \mu_t I_{\hat{f}_t} e_{f^*}\right] \tag{16.1}$$

Intuitively, the estimation error is well-behaved if $\hat{f}_t \approx f^*$, since $\mu_t I_{f^*} e_{f^*} = 0$. Equation (16.1) is closely related to the *total information gain* used in the literature on information-directed sampling [RV14] and kernel bandits [SKKS10b].

To bound the estimation error, [FKQR21] rely on *online density estimation* (also, *online regression* or *online aggregation*) [CBL06, Chapter 9]. For finite $\mathcal{M}$, the default approach is the *exponential weights algorithm* (EWA), which we provide for reference in Chapter 19. When using this algorithm, the estimation error always satisfies $\mathrm{Est}_n \leq \log(|\mathcal{H}|)$, see [CBL06, Proposition 3.1]. While these bounds extend to continuous model classes via standard covering arguments, the resulting algorithm is often not tractable without additional assumptions. For linear feedback models (Examples 37 and 38), one can rely on the more familiar ridge regression estimator, which, we show, achieves bounded estimation regret $\mathrm{Est}_n \leq \mathcal{O}(d\log(n))$. For further discussion, see Section 19.1.

```
args: Hypothesis class $\mathcal{M}$, estimation oracle EST, sequence $\epsilon_t \geq 0$
$\mathcal{D}_0 = \emptyset$
for $t = 1$ to $\infty$:
    estimate $\hat{f}_t = \mathrm{EST}(\mathcal{D}_{t-1})$
    compute gap and information matrices: $\Delta$, $I_{\hat{f}_t} \in \mathbb{R}^{\Pi\times\mathcal{H}}$
    $\mu_t = \arg\min_{\mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathscr{P}(\mathcal{H})}\{\mu\Delta\nu : \mu I_{\hat{f}_t}\nu \leq \epsilon_t^2\}$
    sample $\pi_t \sim \mu_t$ and observe $y_t \sim M_{f^*}(\pi_t)$
    update $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(\pi_t, y_t)\}$
```

Algorithm 3: ANYTIME-E2D

With this in mind, we state our main result.

Theorem 39. *Let $\lambda_t \geq 0$ be any sequence adapted to the filtration $\mathcal{F}_t$. Then the regret of* ANYTIME-E2D *(Algorithm 3) with input sequence ϵ_t satisfies for all $n \geq 1$:*

$$R_n \leq \operatorname{ess\,sup}_{t\in[n]} \left\{ \frac{\mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(\hat{f}_t)}{\epsilon_t^2} \right\} \left(\sum_{t=1}^{n} \epsilon_t^2 + \mathrm{Est}_n \right)$$

where we define

$$\mathrm{dec}^{ac}_{\epsilon,\lambda}(f) = \min_{\mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathscr{P}(\mathcal{M})} \mu\Delta\nu - \lambda(\mu I_f \nu - \epsilon^2) = \mathrm{dec}^o_\lambda + \lambda\epsilon^2 \,.$$

Proof of Theorem 39. Let μ_t^* and ν_t^* be a saddle-point solution to the offset dec,

$$\mathrm{dec}^o_{\lambda_t}(\hat{f}_t) = \min_{\mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathscr{P}(\mathcal{M})} \mu\Delta\nu - \lambda_t \mu I_{\hat{f}_t}\nu$$

Note that $\mu_t^*\Delta\nu_t^* - \lambda_t\mu_t^* I_f\nu_t^* \geq \mu_t^*\Delta e_f - \lambda_t\mu_t^* I_f e_f \geq 0$, which implies that $\lambda_t\epsilon_t^2 \leq \mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}$. Next,

$$\begin{aligned}
R_n = \mathbb{E}\left[\sum_{t=1}^n \mu_t\Delta e_{f^*}\right] &= \sum_{t=1}^n \mathbb{E}\left[\mu_t\Delta e_{f^*} - \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2)\right] \\
&\leq \sum_{t=1}^n \mathbb{E}[\max_{g\in\mathcal{H}} \mu_t\Delta_{\hat{f}_t} e_g - \lambda_t(\mu_t I_{\hat{f}_t} e_g - \epsilon_t^2) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2)] \\
&\leq \sum_{t=1}^n \mathbb{E}[\min_{\mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathscr{P}(\mathcal{H})} \mu\Delta\nu - \lambda_t(\mu I_{\hat{f}_t}\nu - \epsilon_t^2) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2)]
\end{aligned}$$

So far, we only introduced the saddle point problem by maximizing over f^*. The last inequality is by our choice of λ_t and μ_t, and noting that $\nu \in \mathscr{P}(\mathcal{H})$ can always be realized as a Dirac. Continuing,

$$\begin{aligned}
R_n &\leq \sum_{t=1}^n \mathbb{E}\left[\mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(\hat{f}_t) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2)\right] \\
&\overset{(i)}{\leq} \sum_{t=1}^n \mathbb{E}[\mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(\hat{f}_t) + \frac{1}{\epsilon_t^2}\mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(\hat{f}_t)\mu_t I_{\hat{f}_t} e_{f^*}] \\
&\overset{(ii)}{\leq} \operatorname{ess\,sup}_{t\in[n]} \max_{f\in\mathcal{H}} \left\{\frac{1}{\epsilon_t^2}\mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(f)\right\} \sum_{t=1}^n \left(\epsilon_t^2 + \mathbb{E}[\mu_t I_{\hat{f}_t} e_{f^*}]\right)
\end{aligned}$$

We first drop the negative term in (i) and use the beforehand stated fact that $\lambda_t\epsilon_t^2 \leq \mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(\hat{f}_t)$. The last step, (ii), is taking the maximum out of the sum. □

As an immediate corollary, we obtain a regret bound for Algorithm 3 where the sampling distribution μ_t is chosen to optimize $\mathrm{dec}^{ac}_{\epsilon_t}$ for any sequence ϵ_t.

Corollary 40. *The regret of* ANYTIME-E2D *(Algorithm 3) with input $\epsilon_t \geq 0$ satisfies for all $n \geq 1$:*

$$R_n \leq \max_{t\in[n], f\in\mathcal{H}} \left\{\frac{\mathrm{dec}^{ac}_{\epsilon_t}(f)}{\epsilon_t^2}\right\}\left(\sum_{t=1}^n \epsilon_t^2 + \mathrm{Est}_n\right)$$

Importantly, the regret of Algorithm 3 is directly controlled by the worst-case DEC, $\max_{f\in\mathcal{H}} \mathrm{dec}^{ac}_{\epsilon}(f)$, and the estimation error Est_n. It remains to set ϵ_t^2 (respectively λ_t) appropriately. For a fixed horizon n, we let $\epsilon_t^2 = \frac{\mathrm{Est}_n}{n}$. With the reasonable assumption that $\max_{f\in\mathcal{H}}\{\epsilon^{-2}\mathrm{dec}^{ac}_{\epsilon}(f)\}$ is non-

Setting	dec^o_γ	$\mathrm{dec}^{ac}_\epsilon$
Multi-Armed Bandits	$\lvert\Pi\rvert/\gamma$	$2\epsilon\sqrt{\lvert\Pi\rvert}$
Linear Bandits	$d/4\gamma$	$\epsilon\sqrt{d}$
Lipschitz Bandits	$2\gamma^{-\frac{1}{d+1}}$	$2^{\frac{d+1}{d+2}}\epsilon^{\frac{2}{d+2}}$
Convex Bandits	$\tilde{O}(d^4/\gamma)$	$\tilde{O}(\epsilon d^2)$

Table 16.1: Comparison of dec^o_γ and $\mathrm{dec}^{ac}_\epsilon$ for different settings. Bounds between dec^o_γ and $\mathrm{dec}^{ac}_\epsilon$ can be converted using Eq. (15.6). Refined bounds for linear bandits with side-observations are in Lemma 47.

decreasing in ϵ, Corollary 40 reads

$$R_n \le 2n \max_{f\in\mathcal{H}} \left\{ \mathrm{dec}^{ac}_{\sqrt{\mathrm{Est}_n/n}}(f) \right\} . \tag{16.2}$$

This almost matches the lower bound $R_n \ge \Omega(n\mathrm{dec}^c_{1/\sqrt{n}}(\mathcal{F}))$[1] [FGH23, Theorem 2.2], up to the estimation error and the beforehand mentioned gap between dec^c_ϵ and $\mathrm{dec}^{ac}_\epsilon$.

$$\frac{1}{\epsilon^2}\mathrm{dec}^{ac}_\epsilon(f) = \min_{\lambda\ge 0}\{\mathrm{dec}^o_\lambda(f)\epsilon^{-2} + \lambda\} \tag{16.3}$$

To get an anytime algorithm with essentially the same scaling as in Eq. (16.2), we set $\epsilon_t^2 = \log(|\mathcal{M}|)/t$ for finite model classes, and $\epsilon_t^2 = \frac{\beta_\mathcal{H}}{t}$ if $\mathrm{Est}_t \le \beta_\mathcal{H}\log(t)$ for $\beta_\mathcal{H} > 0$. For linear bandits, $\mathrm{dec}^{ac}_\epsilon \le \epsilon\sqrt{d}$ (see Chapter 17), and $\mathrm{Est}_n \le d\log(n)$. Choosing $\epsilon_t^2 = d/t$ recovers the optimal regret bound $R_n \le \tilde{\mathcal{O}}(d\sqrt{n})$ [LS20a]. Alternatively, one can also choose λ_t by minimizing an upper bound on $\max_{t\in[n],f\in\mathcal{H}}\left\{\mathrm{dec}^{ac}_{\epsilon_t,\lambda_t}(f)/\epsilon_t^2\right\}$. For example, in linear bandits, $\mathrm{dec}^{ac}_{\epsilon_t,\lambda} \le \frac{d}{4\lambda} + \lambda\epsilon_t^2$ (see Table 16.1); hence, for $\epsilon_t^2 = d/t$, we can set $\lambda_t = t/4$. Further discussion and refined upper bound for linear feedback models are in Chapter 17.

[1] Here, $\mathrm{dec}^c_\epsilon(\mathcal{F}) = \max_{f\in\mathrm{co}(\mathcal{H})}\min_{\mu\in\mathscr{P}(\Pi)}\max_{g\in\mathcal{H}\cup\{f\}}\{\mu\Delta\nu : \mu I_f e_g \le \epsilon^2\}$.

Chapter 17

Certifying Upper Bounds

As shown by Corollary 40, the regret of Algorithm 3 scales directly with the $\operatorname{dec}_{\epsilon}^{ac}$. For analysis purposes, it is however useful to compute upper bounds on the $\operatorname{dec}_{\epsilon}^{ac}$ to verify the scaling w.r.t. parameters of interest. Via the equivalence Eq. (15.6), bounds on the $\operatorname{dec}_{\lambda}^{o}$ directly translate to the $\operatorname{dec}_{\epsilon}^{ac}$ (see Table 16.1). For a detailed discussion of upper bounds in various models, we refer to [FKQR21]. Below, we highlight three connections that are directly facilitated by the $\operatorname{dec}_{\epsilon}^{ac}$. In particular, we show a hierarchy of *decoupling* arguments ([RV14]) that lead to increasingly weaker bounds on the dec^{ac}.

To this end, we first introduce a variant of the $\operatorname{dec}_{\epsilon}^{ac}$ where the gap function depends on f:

$$\operatorname{dec}_{\epsilon}^{ac,f}(f) = \min_{\mu \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \mu \Delta_f \nu \qquad \text{s.t.} \qquad \mu I_f \nu \leq \epsilon^2 \,, \tag{17.1}$$

where $\Delta_f(\pi, g) = r_g(\pi_g^*) - r_f(\pi)$. This choice additively decouples the reference model f and the alternative model g, making Δ_f the sum of two rank-one matrices. More explicitly, we denote $\delta_f(g) = r_g(\pi_g^*) - r_f(\pi_f^*)$, so that we get

$$\Delta_f(\pi, g) = \delta_f(g) + \Delta_f(\pi, f) \tag{17.2}$$

For distributions $\nu \in \mathscr{P}(\mathcal{H})$ and $\mu \in \mathscr{P}(\Pi)$, we further get $\mu \Delta_f \nu = \delta_f \nu + \mu \Delta_f e_f$.

The following assumption implies that the observations for a decision π are at least as informative as observing the rewards.

Assumption 41 (Reward Data Processing)**.** The rewards and information matrices are related via the following data-processing inequality that holds for any $\mu \in \mathscr{P}(\Pi)$:

$$|\mathbb{E}_{\pi\sim\mu}[r_f(\pi) - r_g(\pi)]| \leq \sqrt{\mathbb{E}_{\pi\sim\mu}[D(M_f(\pi)\|M_g(\pi))]}$$

The next lemma shows that under Assumption 41, $\operatorname{dec}_\epsilon^{ac}(f)$ and $\operatorname{dec}_\epsilon^{ac,f}(f)$ are essentially equivalent, at least for the typical worst-case bounds where $\max_{f\in\mathcal{H}} \operatorname{dec}_\epsilon^{ac}(f) \geq \Omega(\epsilon)$.

Lemma 42. *If Assumption 41 holds, then*

$$\operatorname{dec}_\epsilon^{ac,f}(f) - \epsilon \leq \operatorname{dec}_\epsilon^{ac}(f) \leq \operatorname{dec}_\epsilon^{ac,f}(f) + \epsilon$$

Proof of Lemma 42. Note that

$$\begin{aligned}
\operatorname{dec}_\epsilon^{ac}(f) &= \min_{\mu\in\mathcal{P}(\Pi)} \max_{\nu\in\mathcal{P}(\mathcal{H})} \mu\Delta\nu \qquad \text{s.t.} \qquad \mu I_f \nu \leq \epsilon^2 \\
&= \min_{\mu\in\mathcal{P}(\Pi)} \max_{\nu\in\mathcal{P}(\mathcal{H})} \mu\Delta_f\nu + \sum_{g\in\mathcal{H}}\sum_{\pi\in\Pi} \mu_\pi(r_f(\pi) - r_g(\pi))\nu_g \qquad \text{s.t.} \qquad \mu I_f \nu \leq \epsilon^2 \\
&\leq \min_{\mu\in\mathcal{P}(\Pi)} \max_{\nu\in\mathcal{P}(\mathcal{H})} \mu\Delta_f\nu + \sqrt{\mu I_f \nu} \qquad \text{s.t.} \qquad \mu I_f \nu \leq \epsilon^2 \\
&\leq \epsilon + \min_{\mu\in\mathcal{P}(\Pi)} \max_{\nu\in\mathcal{P}(\mathcal{H})} \mu\Delta_f\nu \qquad \text{s.t.} \qquad \mu I_f \nu \leq \epsilon^2 \\
&\leq \epsilon + \operatorname{dec}_\epsilon^{ac,f}(f),
\end{aligned}$$

where the first inequality follows from

$$\left(\sum_g \mu(r_f - r_g)\nu_g\right)^2 \leq \sum_g \big(\mu(r_f - r_g)\big)^2 \nu_g \leq \mu I_f \nu\,.$$

Also, by lower bounding the sum in the second equality by $-\sqrt{\mu I_f \nu}$ we get left inequality. □

We remark that Algorithm 3 where the sampling distribution is computed for $\operatorname{dec}_\epsilon^{ac,f}(\hat{f}_t)$ and Δ_f achieves a bound analogous to Theorem 39, as long as Assumption 41 holds.

Lemma 43. *If Assumption 41 holds, then the regret of* ANYTIME-E2D *(Algorithm 3) with Δ replaced with Δ_f is bounded as follows:*

$$R_n \leq \max_{t\in[n], f\in\mathcal{H}} \left\{\frac{\operatorname{dec}_{\epsilon_t}^{ac,f}(f)}{\epsilon_t^2}\right\} \left(\sum_{t=1}^n \epsilon_t^2 + \operatorname{Est}_n\right) + \sqrt{n\operatorname{Est}_n}$$

Proof. The proof follows along the lines of the proof of Theorem 39. The main difference is that when introducing Δ_f, we get a term that captures the reward estimation error:

$$\begin{aligned}
R_n &= \mathbb{E}[\sum_{t=1}^{n} \mu_t \Delta_{f^*} e_{f^*}] \\
&= \sum_{t=1}^{n} \mathbb{E}[\mu_t \Delta_{\hat{f}_t} e_{f^*} - \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2) + \mu_t(r_{\hat{f}_t} - r_{f^*})] \\
&\leq \sum_{t=1}^{n} \mathbb{E}[\max_{g \in \mathcal{H}} \mu_t \Delta_{\hat{f}_t} e_g - \lambda_t(\mu_t I_{\hat{f}_t} e_g - \epsilon_t^2) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2) + \mu_t(r_{\hat{f}_t} - r_{f^*})] \\
&= \sum_{t=1}^{n} \mathbb{E}[\min_{\lambda \geq 0} \min_{\mu \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \mu \Delta_{\hat{f}_t} \nu - \lambda(\mu I_{\hat{f}_t} \nu - \epsilon_t^2) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2)] \\
&\quad + \sum_{t=1}^{n} \mathbb{E}[\mu_t(r_{\hat{f}_t} - r_{f^*})]
\end{aligned}$$

So far, we only introduced the saddle point problem by maximizing over f^*. The last equality is by our choice of λ_t and μ_t. Continuing,

$$\begin{aligned}
R_n &\leq \sum_{t=1}^{n} \mathbb{E}[\text{dec}_{\epsilon_t}^{ac}(\hat{f}_t) + \lambda_t(\mu_t I_{\hat{f}_t} e_{f^*} - \epsilon_t^2)] + \sum_{t=1}^{n} \mathbb{E}[\mu_t(r_{\hat{f}_t} - r_{f^*})] && (17.3) \\
&\leq \max_{t \in [n]} \max_{f \in \mathcal{H}} \left\{ \frac{1}{\epsilon_t^2} \text{dec}_{\epsilon_t}^{ac}(f) \right\} \sum_{t=1}^{n} \left(\epsilon_t^2 + \mathbb{E}[\mu_t I_{\hat{f}_t} e_{f^*}] \right) + \sqrt{n \text{Est}_n} && (17.4)
\end{aligned}$$

For the last inequality, we used Cauchy-Schwarz and Assumption 41 to bound the error term,

$$\sum_{t=1}^{n} \mathbb{E}[\mu_t(r_{\hat{f}_t} - r_{f^*})] \leq \sqrt{n \sum_{t=1}^{n} \mathbb{E}[(\mu_t(r_{\hat{f}_t} - r_{f^*}))^2]} \leq \sqrt{n \sum_{t=1}^{n} \mathbb{E}[\mu_t I_{\hat{f}_t} e_f]} = \sqrt{n \text{Est}_n}\,.$$

□

17.1 Upper Bounds via Decoupling

First, we introduce the *information ratio*,

$$\Psi_f(\mu, \nu) = \frac{(\mu \Delta_f \nu)^2}{\mu I_f \nu}$$

The definition is closely related to the Bayesian information ratio [RV16], where ν takes the role of a prior over $\mathcal{H}$. The Thompson sampling distribution is $\mu_\nu^{\text{TS}} = \sum_{h\in\mathcal{H}} \nu_h e_{\pi_h^*}$. The decoupling coefficient, $\text{dc}(f)$, [Zha22, Definition 1] is defined as the smallest number $K \geq 0$, such that for all distributions $\nu \in \mathscr{P}(\mathcal{H})$,

$$\mu_\nu^{\text{TS}}\Delta_f\nu \leq \inf_{\eta\geq 0}\left\{\eta \sum_{g,h\in\mathcal{H}} \nu_g\nu_h e_{\pi_h^*}(r_g - r_f)^2 + \frac{K}{4\eta}\right\} = \sqrt{K\sum_{g,h\in\mathcal{H}}\nu_g\nu_h e_{\pi_h^*}(r_g-r_f)^2} \tag{17.5}$$

In other words, the decoupling coefficient is equal to the information ratio for the Thompson sampling distribution and the worst-case prior, $\text{dc}(f) = \max_{\nu\in\mathscr{P}(\nu)} \Psi_f(\mu_\nu^{\text{TS}})$.

The next lemma provides upper bounds on the $\text{dec}_\epsilon^{ac}(f)$ in terms of the information ratio, which is further upper-bounded by the decoupling coefficient.

Lemma 44. *With $\Psi(f) = \max_{\nu\in\mathcal{H}} \min_{\mu\in\mathscr{P}(\Pi)} \Psi_f(\mu,\nu)$ and Assumption 41 satisfied, we have*

$$\text{dec}_\epsilon^{ac,f}(f) \leq \epsilon\sqrt{\Psi(f)} \leq \epsilon\sqrt{\text{dc}(f)}\,.$$

By [Zha22, Lemma 2], this further implies $\text{dec}_\epsilon^{ac,f} \leq \epsilon\sqrt{d}$.

Proof of Lemma 44. For the first inequality, using the definition of $\text{dec}_\epsilon^{ac}(f,\Delta_f)$ and the AM-GM inequality:

$$\begin{aligned}\text{dec}_\epsilon^{ac,f}(f) &= \min_{\lambda\geq 0}\max_{\nu\in\mathscr{P}(\mathcal{H})}\min_{\mu\in\mathscr{P}(\Pi)} \mu\Delta_f\nu - \lambda\mu I_f\nu + \lambda\epsilon^2 \\ &\leq \min_{\lambda> 0}\max_{\nu\in\mathscr{P}(\mathcal{H})}\min_{\mu\in\mathscr{P}(\Pi)} \frac{(\mu\Delta_f\nu)^2}{4\lambda\mu I_f\nu} + \lambda\epsilon^2 && (17.6)\\ &= \min_{\lambda>0}\frac{\Psi(f)}{4\lambda} + \lambda\epsilon^2 = \epsilon\sqrt{\Psi(f)}\,. && (17.7)\end{aligned}$$

Further, by Eq. (17.5) and Assumption 41 we have $\mu_\nu^{\text{TS}}\Delta_f\nu \leq \sqrt{K\sum_{g,h\in\mathcal{H}}\nu_g\nu_h e_{\pi_h^*}(r_g-r_f)^2} \leq \sqrt{K\mu_f^{\text{TS}} I_f\nu}$, which gives $\Psi(f)\leq K$. Plugging this into Eq. (17.7) gives the second inequality. □

The generalized information ratio [LG21] for $\mu\in\mathscr{P}(\Pi)$, $\nu\in\mathscr{P}(\mathcal{H})$, and $\alpha > 1$ is defined as

$$\Psi_{\alpha,f}(\mu,\nu) = \frac{(\mu\Delta_f\nu)^\alpha}{\mu I_f\nu} \tag{17.8}$$

For $\alpha = 2$, we get the standard information ratio introduced by [RV16] with ν as a prior over the model class $\mathcal{M}$. Define $\Psi_\alpha(f) = \max_{\nu\in\mathcal{H}} \min_{\mu\in\mathscr{P}(\Pi)} \Psi_{\alpha,f}(\mu,\nu)$. To upper bound dec_ϵ^{ac}, we have the following lemma.

Lemma 45. *For the reference model f, the ac-dec can be upper bounded as*

$$\text{dec}_\epsilon^{ac}(f) \le \min_{\lambda>0} \left\{\lambda^{\frac{1}{1-\alpha}} \alpha^{\frac{\alpha}{1-\alpha}} (\alpha-1)\Psi_\alpha(f)^{\frac{1}{\alpha-1}} + \lambda\epsilon^2\right\} \tag{17.9}$$

for $\alpha > 1$.

Proof. First, note that by Jensen inequality and concavity of $\ln$, we have

$$\begin{aligned} p_1\ln(x_1) + p_2\ln(x_2) &\le \ln\left(p_1x_1 + p_2x_2\right) \\ \Rightarrow x_1^{p_1} + x_2^{p_2} &\le p_1x_1 + p_2x_2\,, \end{aligned}$$

where $p_1 + p_2 = 1$ and $x_1, x_2 > 0$. This implies that for $\alpha > 1$,

$$\alpha \cdot x_1^{\frac{1}{\alpha}} \cdot x_2^{\frac{\alpha-1}{\alpha}} - x_1 \le (\alpha-1)x_2$$

Writing $x_1 = \lambda\mu I_f\nu$ and $x_2 = \alpha^{\frac{\alpha}{1-\alpha}}\left(\frac{(\mu\Delta\nu)^\alpha}{\lambda\mu I_f\nu}\right)^{\frac{1}{\alpha-1}}$ and using the previous inequality with the dec_ϵ^{ac} definition gives the result. □

17.2 PAC to Regret

Another useful way to upper bound the $\text{dec}_\epsilon^{ac,f}$ is via an analogous definition for the PAC setting [c.f. Eq. (10), FGH23]:

$$\text{pac-dec}_\epsilon^{ac,f}(f) = \min_{\mu\in\mathscr{P}(\Pi)} \max_{\nu\in\mathcal{H}} \delta_f\nu \qquad \text{s.t.} \qquad \mu I_f\nu \le \epsilon^2 \tag{17.10}$$

Lemma 46. *Under Assumption 41,*

$$\text{dec}_\epsilon^{ac,f}(f) \le \min_{p\in[0,1]} \left\{\text{pac-dec}_{\epsilon p^{-1/2}}^{ac,f}(f) + p\Delta_{\max}\right\}$$

Proof. The Lagrangian for Eq. (17.10) is

$$\text{pac-dec}_\epsilon^{ac,f}(f) = \min_{\lambda \geq 0} \min_{\mu \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{M})} \delta_f \nu - \lambda(\mu I_f \nu - \epsilon^2).$$

Reparametrize any $\mu \in \mathscr{P}(\Pi)$ as $\bar{\mu}(p) = (1-p)e_{\pi_f^*} + p\mu_2$. We bound $\text{dec}_\epsilon^{ac,f}$ by a function of $\text{pac-dec}_\epsilon^{ac,f}$. Starting from Eq. (15.4), we have

$$\begin{aligned}
\text{dec}_\epsilon^{ac,f}(f) &= \min_{\lambda \geq 0} \min_{\mu \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \mu \Delta_f \nu - \lambda(\mu I_f \nu - \epsilon^2) \\
&= \min_{\lambda \geq 0} \min_{\bar{\mu} \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \bar{\mu} \Delta_f \nu - \lambda(\bar{\mu} I_f \nu - \epsilon^2) \\
&= \min_{\lambda \geq 0} \min_{0 \leq p \leq 1} \min_{\mu_2 \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \delta_f \nu + p\mu_2 \Delta_f e_f - \lambda \bar{\mu} I_f \nu - \lambda \epsilon^2 \\
&\leq \min_{\lambda \geq 0} \min_{0 \leq p \leq 1} \min_{\mu_2 \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \delta_f \nu + p\mu_2 \Delta_f e_f - \lambda p \mu_2 I_f \nu - \lambda \epsilon^2 \\
&\leq \min_{0 \leq p \leq 1} \min_{\lambda' \geq 0} \min_{\mu_2 \in \mathscr{P}(\Pi)} \max_{\nu \in \mathscr{P}(\mathcal{H})} \delta_f \nu - \lambda'(\mu_2 I_f \nu - \frac{\epsilon^2}{p}) + p\Delta_{\max} \\
&\leq \min_{0 \leq p \leq 1} \text{pac-dec}_{\frac{\epsilon}{\sqrt{p}}}^{ac,f}(f) + p\Delta_{\max}\,.
\end{aligned}$$

□

17.3 Application to Linear Feedback Models

To illustrate the techniques introduced, we compute a regret bound for Algorithm 3 for linear bandits with side-observations (Examples 37 and 38).

Lemma 47. *For linear bandits with side-observations and divergence $I_f(\pi, g) = \|M_\pi(g-f)\|^2$,*

$$\text{pac-dec}_\epsilon^{ac,f}(f) \leq \min_{\mu \in \mathscr{P}(\Pi)} \max_{b \in \Pi} \epsilon \|\phi_b\|_{V(\mu)^{-1}} \leq \epsilon\sqrt{d}$$

where $V(\mu) = \sum_{\pi \in \Pi} \mu_\pi M_\pi M_\pi^\top$. Moreover, denoting $\Omega = \min_{\mu \in \mathscr{P}(\Pi)} \max_{b \in \Pi} \|\phi_b\|_{V(\mu)^{-1}}$,

$$\text{dec}_\epsilon^{ac,f}(f) \leq \min\left(\epsilon\sqrt{\Psi(f)}, 2\epsilon^{2/3}\Omega^{1/3}\Delta_{\max}^{1/3}\right)$$

Proof of Lemma 47. For the first part, note that

$$\begin{aligned}
\text{pac-dec}_{\epsilon}^{ac,f}(f) &= \min_{\mu \in \mathscr{P}(\Pi)} \min_{\lambda \geq 0} \max_{\nu \in \mathscr{P}(\mathcal{H})} \delta_f \nu - \lambda \mu I_f \nu + \lambda \epsilon^2 \\
&= \min_{\mu \in \mathscr{P}(\Pi)} \min_{\lambda \geq 0} \max_{b \in \Pi} \max_{g \in \mathcal{H}} \langle \phi_b, g \rangle - \left\langle \phi_{\pi_f^*}, f \right\rangle - \lambda \|g - f\|_{V(\mu)}^2 + \lambda \epsilon^2 \\
&\overset{(i)}{=} \min_{\mu \in \mathscr{P}(\Pi)} \min_{\lambda \geq 0} \max_{b \in \Pi} \left\langle \phi_b - \phi_{\pi_f^*}, f \right\rangle + \frac{1}{4\lambda} \|\phi_b\|_{V(\mu)^{-1}}^2 + \lambda \epsilon^2 \\
&\overset{(ii)}{\leq} \min_{\mu \in \mathscr{P}(\Pi)} \min_{\lambda \geq 0} \max_{b \in \Pi} \frac{1}{4\lambda} \|\phi_b\|_{V(\mu)^{-1}}^2 + \lambda \epsilon^2 \\
&= \min_{\mu \in \mathscr{P}(\Pi)} \max_{b \in \Pi} \epsilon \|\phi_b\|_{V(\mu)^{-1}} \\
&\overset{(iii)}{\leq} \epsilon \sqrt{d}\,.
\end{aligned}$$

Equation (i) follows by computing the maximizer attaining the quadratic form over $\mathcal{M} = \mathbb{R}^d$. The inequality (ii) is by definition of π_f^* and the last inequality (iii) by the assumption that the reward is observed, respectively, $\phi_\pi \phi_\pi^\top \preceq M_\pi^\top M_\pi$, and the Kiefer–Wolfowitz theorem. The second part of the statement follows by combining Lemmas 44 and 46. □

While in the worst-case for linear bandits, there is no improvement over the standard $\mathcal{O}(d\sqrt{n})$ without further refinement or specification of the upper bounds, in the case of linear side-observations there is an improvement whenever $\Omega \leq \max_{f \in \mathcal{M}} \Psi(f)$. To exemplify the improvement, consider a semi-bandit with a "revealing" action $\hat{\pi}$, e.g. $M_{\hat{\pi}} = \mathbf{1}_d$. Here, the regret bound improves to $R_n \leq \min\{d\sqrt{n}, d^{1/3} n^{2/3}\}$, since then $\text{pac-dec}_{\epsilon}^{ac,f}(f) \leq \epsilon$. The corresponding improvement in the regime $n \leq d^4$ might seem modest, but is relevant in high-dimensional and non-parametric models. Moreover, in (deep) reinforcement learning, high-dimensional models are commonly used and the learner obtains side information in the form of state observations. Therefore, it is plausible that the $n^{2/3}$ rate is dominant even for a moderate horizon. Exploring this effect in reinforcement learning is therefore an important direction for future work.

Notably, this improvement is *not* observed by upper confidence bound algorithms and Thompson sampling, because both approaches discard informative but suboptimal actions early on [c.f. LS17], including the action $\hat{\pi}$ in the example above. E2D for a constant offset parameter $\lambda > 0$, in principle, attains the better rate, but only if one pre-commits to a fixed horizon. Lastly, we note that a similar effect was observed for information-directed sampling in sparse high-dimensional linear bandits [HLW20].

Chapter 18

Computational Aspects

For finite model classes, Algorithm 3 can be readily implemented. Since almost no structure is imposed on the gap and information matrices of size $|\Pi| \times |\mathcal{H}|$, avoiding scaling with $|\Pi| \cdot |\mathcal{H}|$ seems hardly possible without introducing additional assumptions. Even in the finite case, solving Eq. (15.2) is not immediate because the corresponding Lagrangian is not convex-concave. A practical approach is to solve the inner saddle point for Eq. (15.4) as a function of λ. Strong duality holds for the inner problem, and one can obtain a solution efficiently by solving the corresponding linear program using standard solvers. It then remains to optimize over $\lambda \geq 0$. This can be done, for example, via a grid search over the range $[0, \max_{f \in \mathcal{H}} \epsilon^{-2} \mathrm{dec}_\epsilon^{ac}(f)]$.

In the linear setting, the above is not satisfactory because most commonly $\mathcal{M}$ is identified with parameters in $\mathbb{R}^d$. As noted before, ridge regression can be used instead of online aggregation while preserving the optimal scaling of the estimation error (see Section 19.1). The next lemma further shows that the saddle point problem Eq. (15.2) can be rewritten to only scale with the size of the decision set $|\Pi|$.

Lemma 48. *Consider linear bandits with side observations, $\mathcal{H} = \mathbb{R}^d$ and quadratic divergence, $I_f(\pi, g) = \|M_\pi(g - f)\|^2$, and denote $\phi_\mu = \sum_{\pi \in \Pi} \mu_\pi \phi_\pi$ and $V(\mu) = \sum_{\pi \in \Pi} \mu_\pi M_\pi^\top M_\pi$. Then*

$$\mathrm{dec}_\epsilon^{ac,f}(\hat{f}_t) = \min_{\lambda \geq 0} \min_{\mu \in \mathscr{P}(\Pi)} \max_{b \in \Pi} \left\langle \phi_b - \phi_\mu, \hat{f}_t \right\rangle + \frac{1}{4\lambda} \|\phi_b\|^2_{V(\mu)^{-1}} + \lambda \epsilon^2$$

Moreover, the objective is convex in $\mu \in \mathscr{P}(\Pi)$.

Proof of Lemma 48.

$$
\begin{aligned}
\operatorname{dec}_{\epsilon}^{ac,f}(\hat{f}_t) &= \min_{\mu\in\mathscr{P}(\Pi)} \min_{\lambda\geq 0} \max_{b\in\Pi} \max_{g\in\mathcal{H}} \langle \phi_b, g\rangle - \left\langle \phi_\mu, \hat{f}_t \right\rangle - \lambda \|g - \hat{f}_t\|_{V(\mu)}^2 + \lambda\epsilon^2 \\
&= \min_{\lambda\geq 0} \min_{\mu\in\mathscr{P}(\Pi)} \max_{b\in\Pi} \left\langle \phi_b - \phi_\mu, \hat{f}_t \right\rangle + \frac{1}{4\lambda}\|\phi_b\|_{V(\mu)^{-1}}^2 + \lambda\epsilon^2 \,. \qquad (18.1)
\end{aligned}
$$

The first equality is by definition, and the second equality follows from solving the quadratic maximization over $g \in \mathcal{M} = \mathbb{R}^d$. To show that the problem is convex in μ, note that taking inverses of positive semi-definite matrices X, Y is a convex function, i.e. $((1-\eta)X + \eta Y)^{-1} \preceq (1-\eta)X^{-1} + \eta Y^{-1}$. In particular, $V((1-\eta)\mu_1 + \eta\mu_2)^{-1} \preceq (1-\eta)V(\mu_1)^{-1} + \eta V(\mu_2)^{-1}$. With this the claim follows. □

Note that the saddle point expression is analogous to Eq. (15.4), and in fact, one can linearize the inner maximization over $\mathscr{P}(\Pi)$, such that the inner saddle point becomes convex-concave. This leads to expressions equivalent to Eqs. (15.3) and (15.5), albeit the objective is no longer linear in $\mu \in \mathscr{P}(\Pi)$. We use Lemma 48 to employ the same strategy as before: As a function of $\lambda \geq 0$, solve the inner problem of the expression in Lemma 48, for example, as a convex program with $|\Pi|$ variables and $|\Pi|$ constraints (Section 18.1). Then all that remains is to solve a one-dimensional optimization problem over $\lambda \in [0, \max_{f\in\mathcal{H}} \epsilon^{-2}\operatorname{dec}_{\epsilon}^{ac}(f)]$. We demonstrate this approach in Chapter 20 to showcase the performance of E2D on simple examples.

18.1 Convex Program for Fixed λ

Take Eq. (18.1) and fix $\lambda > 0$. Then we have the following saddle-point problem:

$$
\begin{aligned}
&\min_{\mu\in\mathscr{P}(\Pi)} \max_{b\in\Pi} \left\langle \phi_b - \phi_\mu, \hat{f}_t \right\rangle + \frac{1}{4\lambda}\|\phi_b\|_{V(\mu)^{-1}}^2 + \lambda\epsilon^2 \\
&= \lambda\epsilon^2 + \min_{\mu\in\mathscr{P}(\Pi)} \max_{b\in\Pi} \left\langle \phi_b - \phi_\mu, \hat{f}_t \right\rangle + \frac{1}{4\lambda}\|\phi_b\|_{V(\mu)^{-1}}^2
\end{aligned}
$$

Up to the constant additive term, this saddle point problem is equivalent to the following convex program

$$\begin{aligned} \min_{y\in\mathbb{R},\mu\in\mathbb{R}^{\Pi}} y \qquad \text{s.t.} \qquad & y \geq \left\langle \phi_b - \phi_\mu, \hat{f}_t \right\rangle + \frac{1}{4\lambda}\|\phi_b\|^2_{V(\mu)^{-1}} \quad \forall b \in \Pi \\ & \mathbf{1}\mu = 1 \\ & \mu_\pi \geq 0 \quad \forall \pi \end{aligned}$$

Chapter 19

Online Density Estimation

```
args: finite model class $\mathcal{M}$, data $\mathcal{D}_t = \{(y_1, \pi_1), \ldots, (y_t, \pi_t)\}$, $\eta > 0$
define $L(f) = -\sum_{s=1}^{t} \log M_f(y_s \mid \pi_s)$
set $p(f) \propto \exp(-\eta\, L(f))$
if $\mathcal{M}$ is convex:
    return $\sum_{f \in \mathcal{M}} p(f) f$
else:
    sample $f \sim p(\cdot)$ and return $f$
```

Algorithm 4: Exponential Weights Algorithm (EWA) for Density Estimation

For any $f \in \mathcal{H}$ and $\pi \in \Pi$, we denote by $p(\cdot|\pi, f)$ the the density function of the observation distribution $M_f(\pi)$ w.r.t. a reference measure over the observation space $\mathcal{O}$. Consider a finite model class $\mathcal{H}$ and the KL divergence,

$$e_\pi I_f e_g = \mathbb{E}_{y \sim M_g(\pi)}[\log \left(\frac{p(y|\pi, g)}{p(y|\pi, f)} \right)] \tag{19.1}$$

In this case, the estimation error can be written as follows:

$$\begin{aligned} \mathrm{Est}_n = \mathbb{E}[\sum_{t=1}^{n} e_{\pi_t} I_{\hat{f}_t} e_{f^*}] &= \mathbb{E}[\sum_{t=1}^{n} \log(p(y_t|\pi_t, f^*)/p(y_t|\pi_t, \hat{f}_t)] \\ &= \mathbb{E}[\sum_{t=1}^{n} \log \left(\frac{1}{p(y_t|\pi_t, \hat{f}_t)} \right) - \sum_{t=1}^{n} \log \left(\frac{1}{p(y_t|\pi_t, f^*)} \right)] \end{aligned}$$

The last line can be understood as the *estimation regret* of the estimates $\hat{f}_1, \ldots, \hat{f}_n$ under the logarithmic loss. A classical approach to control this term is the *exponential weights algorithm* (EWA) given in Algorithm 4. For the EWA algorithm, we have the following bound.

Lemma 49 (EWA for Online Density Estimation)**.** *For any data stream $\{y_1, \pi_1, \dots, y_n, \pi_n\}$ the predictions $\hat{f}_1, \dots \hat{f}_n$ obtained via Algorithm 4 with $\eta = 1$ satisfy*

$$\mathrm{Est}_n \leq \mathbb{E}[\sum_{t=1}^n \log\left(\frac{1}{p(y_t|\pi_t, \hat{f}_t)}\right) - \inf_{g\in\mathcal{H}} \sum_{t=1}^n \log\left(\frac{1}{p(y_t|\pi_t, g)}\right)] \leq \log(|\mathcal{H}|) \tag{19.2}$$

For a proof, see [CBL06, Proposition 3.1].

19.1 Bounding the Estimation Error of Projected Regularized Least-Squares

In this section, we consider the linear model from Example 38. We denote by $\|\cdot\|$ the Euclidean norm. For simplicity, the observation maps $M_\pi \in \mathbb{R}^{m\times d}$ are assumed to have the same output dimension $m \in \mathbb{N}$. The observation distribution is such that $y_t = M_{\pi_t} f^* + \xi_t$, where $\xi \in \mathbb{R}^m$ is random noise such that $\mathbb{E}_t[\xi] = 0$ and $\mathbb{E}_t[\|\xi\|^2] \leq \sigma^2$. Here, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\pi_1, y_1, \dots, \pi_{t-1}, y_{t-1}, \pi_t]$ is the conditional observation in round t including the decision π_t chosen in round t.

We will use the quadratic divergence[1], $e_\pi I_f e_g = \frac{1}{2}\|M_\pi(g - f)\|^2$ This choice corresponds to the Gaussian KL, but we do not require that the noise distribution is Gaussian is the following. In the linear bandit model, this choice reduces to $e_\pi I_f e_g = \frac{1}{2}\langle \phi_\pi, g - f\rangle^2$.

Let $K \subset \mathbb{R}^d$ be a closed convex set. Our goal is to control the estimation regret for the projected regularized least-squares estimator,

$$\hat{f}_t = \arg\min_{f\in K} \sum_{s=1}^{t-1} \|M_{\pi_s} f - y_s\|^2 + \|f\|_{V_0}^2 = \mathrm{Proj}_{V_t}\left(V_t^{-1} \sum_{s=1}^{t-1} M_{\pi_s}^\top y_s\right) \tag{19.3}$$

where V_0 is a positive definite matrix, $V_t = \sum_{s=1}^{t-1} M_{\pi_s}^\top M_{\pi_s} + V_0$ and $\mathrm{Proj}_{V_t}(\cdot)$ is the orthogonal projection w.r.t. the $\|\cdot\|_{V_t}$ norm. For $K = \mathbb{R}^d$ and $V_0 = \eta\mathbf{1}_d$, this recovers the standard ridge regression. The projection is necessary to bound the magnitude of the squared loss, and the result

[1] We added a factor of $\frac{1}{2}$ for convenience.

will depend on an almost-surely bound on the ‘observed’ diameter,

$$\max_{f,g\in K}\max_{\pi\in\Pi}\|M_\pi(f-g)\|\leq B$$

Recall that our goal is to bound the estimation error,

$$\mathrm{Est}_n=\mathbb{E}[\sum_{t=1}^n e_{\pi_t}I_{\hat{f}_t}e_{f^*}]=\mathbb{E}[\sum_{t=1}^n \tfrac{1}{2}\|M_{\pi_t}(f^*-\hat{f}_t)\|^2] \tag{19.4}$$

We remark that one can get the following naive bound by applying Cauchy-Schwarz:

$$\sum_{t=1}^n\|M_{\pi_t}(f^*-\hat{f}_t)\|^2\leq\sum_{t=1}^n\|M_{\pi_t}\|^2_{V_t^{-1}}\|f^*-\hat{f}_t\|^2_{V_t}\leq\mathcal{O}(d^2\log(n)^2) \tag{19.5}$$

The last inequality follows from the elliptic potential lemma and standard concentration inequalities [LS20a, Lemma 19.4 and Theorem 20.5]. However, this will lead to an additional d-factor in the regret that can be avoided, as we see next.

For $K=\mathbb{R}^d$, one-dimensional observations and noise bounded in the range $[-\bar{B},\bar{B}]$, one can also directly apply [CBL06, Theorem 11.7] to get $\mathrm{Est}_n\leq\mathcal{O}(\bar{B}^2d\log(n))$, thereby improving the naive bound by a factor $d\log(n)$. This result is obtained in a more general setting, where no assumptions, other than boundedness, are placed on the observation sequence $y_1,\dots,y_n$. Here we refine and generalize this result in two directions: First, we allow for the more general feedback model in with multi-dimensional observations (Example 38). Second, we directly exploit the stochastic observation model to obtain a stronger result that does not require the observation noise to be bounded.

Theorem 50. *Consider the linear observation setting with additive noise and quadratic divergence $e_\pi I_f e_g=\frac{1}{2}\|M_\pi(g-f)\|^2$, as described at the beginning of this section. Assume that $\max_{f,g\in\mathcal{H},\pi\in\Pi}\|M_\pi(f-g)\|\leq B$ and $\mathbb{E}[\|\xi_t\|^2]\leq\sigma^2$. Then*

$$\mathrm{Est}_n\leq(\sigma^2+B^2)\mathbb{E}[\log\left(\frac{\det V_n}{\det V_0}\right)]$$

If in addition $\|M_\pi\|\leq L$ and $V_0=\eta\mathbf{1}_d$, then $\mathrm{Est}_n\leq(\sigma^2+B^2)\log\left(1+\frac{nL^2}{\eta d}\right)$.

Remark 51. Note that by the [LS20a, Theorem 19.4], $\log\left(\frac{\det V_n}{\det V_0}\right)$ can further be upper bounded by $d\log\left(\frac{\mathrm{trace}V_0+nL^2}{d\det(V_0)^{1/d}}\right)$, which effectively results the desired bound.

Proof. The proof adapts [GPS13, Theorem 19.8] to multi-dimensional observations and takes advantage of the stochastic loss function by taking the expectation.

First, define $l_t(f) = \frac{1}{2}\|M_{\pi_t} f - y_t\|^2$. Then, using that $\mathbb{E}_t[y_t] = M_{\pi_t} f^*$,

$$\begin{aligned}\mathrm{Est}_n = \mathbb{E}[\sum_{t=1}^{n} \tfrac{1}{2}\|M_{\pi_t}(f^* - \hat{f}_t)\|^2] &= \mathbb{E}[\sum_{t=1}^{n} \tfrac{1}{2}\|M_{\pi_t}\hat{f}_t - y_t\|^2 - \tfrac{1}{2}\|M_{\pi_t} f^* - y_t\|^2] \\ &= \mathbb{E}[\sum_{t=1}^{n} l_t(\hat{f}_t) - l_t(f^*)]\end{aligned}$$

Further, by directly generalizing [GPS13, Lemma 19.7], we have that

$$l_t(\hat{f}_{t+1}) - l_t(\hat{f}_t) \leq \nabla l_t(\hat{f}_t) V_t^{-1} \nabla l_t(\hat{f}_t) = (M_{\pi_t}\hat{f}_t - y_t)^\top M_{\pi_t}^\top V_t^{-1} M_{\pi_t}(M_{\pi_t}\hat{f}_t - y_t) \qquad (19.6)$$

We now start upper bounding the estimation error,

$$\begin{aligned}\mathrm{Est}_n &\overset{(i)}{\leq} \|f^*\|^2 + \mathbb{E}[\sum_{t=1}^{n} \big(l_t(w_t) - l_t(w_{t+1})\big)] \\ &\overset{(ii)}{\leq} \|f^*\|^2 + \mathbb{E}[\sum_{t=1}^{n} \big(\xi_t + M_{\pi_t}(f^* - \hat{f}_{t-1})\big) M_{\pi_t} V_t^{-1} M_{\pi_t} \big(\xi_t + M_{\pi_t}(f^* - \hat{f}_{t-1})\big)] \\ &\overset{(iii)}{=} \|f^*\|^2 + \mathbb{E}[\sum_{t=1}^{n} \xi_t M_{\pi_t} V_t^{-1} M_{\pi_t} \xi_t\big)] + \mathbb{E}[\sum_{t=1}^{n} \bar{x}_t M_{\pi_t} V_t^{-1} M_{\pi_t} \bar{x}_t\big)] \\ &\overset{(iv)}{\leq} \|f^*\|^2 + \mathbb{E}[\sum_{t=1}^{n} \lambda_{\max}(M_{\pi_t} V_t^{-1} M_{\pi_t})\|\xi_t\|^2] + \mathbb{E}[\sum_{t=1}^{n} \lambda_{\max}(M_{\pi_t} V_t^{-1} M_{\pi_t})\|\bar{x}_t\|^2\big)] \\ &\overset{(v)}{\leq} \|f^*\|^2 + (\sigma^2 + B^2)\mathbb{E}[\sum_{t=1}^{n} \lambda_{\max}(M_{\pi_t} V_t^{-1} M_{\pi_t})] \qquad (19.7)\end{aligned}$$

The inequality (i) follows from [SS$^+$12, Lemma 2.3]. For (ii) we used Eq. (19.6). For (iii) we used that $\mathbb{E}_t[\xi_t] = 0$. In (iv), we introduce the maximum eigenvalue $\lambda_{\max}(A)$ for $A \in \mathbb{R}^{m \times m}$ and denote $\bar{x}_t = M_{\pi_t}(f^* - f_{t-1})$. Lastly, in (v) we used that $\|\bar{x}_t\|^2 \leq B$ and $\mathbb{E}_t[\|\xi_t\|^2] \leq \sigma^2$.

We conclude the proof with basic linear algebra. Denote by $\lambda_i(A)$ the i-th eigenvalue of a matrix $M \in \mathbb{R}^{m \times m}$. Using the generalized matrix determinant lemma, we get

$$\begin{aligned}\det(V_{t-1}) &= \det(V_t - M_{\pi_t}^\top M_{\pi_t}) \\ &= \det(V_t)\det(I - M_{\pi_t}^\top V_t^{-1} M_{\pi_t}) \\ &= \det(V_t) \prod_{i=1}^m (1 - \lambda_i(M_{\pi_t}^\top V_t^{-1} M_{\pi_t}))\end{aligned}$$

Note that $\lambda_i(M_{\pi_t}^\top V_t^{-1} M_{\pi_t}) \in (0,1]$. Next, using that $\log(1-x) \leq -x$ for all $x < 1$, we get that

$$\log\left(\frac{\det(V_{t-1})}{\det(V_t)}\right) = \sum_{i=1}^m \log(1 - \lambda_i(M_{\pi_t}^\top V_t^{-1} M_{\pi_t})) \leq -\sum_{i=1}^m \lambda_i(M_{\pi_t}^\top V_t^{-1} M_{\pi_t})$$

Rearranging the last display, and bounding the sum by its maximum element, we get

$$\lambda_{\max}(M_{\pi_t}^\top V_t^{-1} M_{\pi_t}) \leq \sum_{i=1}^m \lambda_i(M_{\pi_t}^\top V_t^{-1} M_{\pi_t}) \leq \log\left(\frac{\det(V_t)}{\det(V_{t-1})}\right) \tag{19.8}$$

The proof is concluded by combining Eqs. (19.7) and (19.8). □

Remark 52 (Expected Regret)**.** The beauty of Theorem 39 is that the proof uses *only* in-expectation arguments. This is unlike most previous analysis, that controls the regret via controlling tail-events, and bounds on the expected regret are then derived a-posteriori from high-probability bounds. In the context of linear bandits, Theorem 50 leads to bound on the expected regret that only requires the noise variance to be bounded, whereas most previous work relies on the stronger sub-Gaussian noise assumption [e.g. AYPS11].

Remark 53 (Kernel Bandits / Bayesian Optimization)**.** Using the standard 'kernel-trick', the analysis can further be extended to the non-parametric setting where $\mathcal{H}$ is an infinite-dimensional reproducing kernel Hilbert space (RKHS).

Chapter 20

Experiments

All experiments below were run on a semi-bandit problem with a ”revealing action”, as alluded to in the paragraph below Lemma 47. Specifically, we assume a semi-bandit model where $\mathcal{H} = \mathbb{R}^d$ and the features are $\phi_\pi \in \mathbb{R}^d$. For an instance $f^* \in \mathcal{M}$, the reward function is $r_{f^*} = \langle \phi_\pi, f^* \rangle$ for all $\pi \in \Pi$. There is one revealing (sub-optimal) action $\hat{\pi} \neq \pi^*_{f^*}$. The observation for any action $\pi \neq \hat{\pi}$ is

$$M_{f^*}(\pi) = \mathcal{N}(\langle \phi_\pi, f^* \rangle, 1) \tag{20.1}$$

Define $M_{\hat{\pi}} = [\phi_{\pi_1}, \ldots, \phi_{\pi_{|\Pi|}}]^\top$. Then the observation for action $\hat{\pi}$ is

$$M_{f^*}(\hat{\pi}) = \mathcal{N}(M_{\hat{\pi}} f^*, \mathbf{1}_d) \tag{20.2}$$

Thus, the information for any action $\pi \neq \hat{\pi}$ is

$$I_f(g, \pi) = \frac{\sigma^2}{2} \langle \phi_\pi, g - f \rangle^2 \tag{20.3}$$

while the information for action $\hat{\pi}$ is

$$I_f(g, \hat{\pi}) = \frac{\sigma^2}{2} \| M_{\hat{\pi}}(g - f) \|^2 = \frac{\sigma^2}{2} \sum_\pi \langle \phi_\pi, g - f \rangle^2 \tag{20.4}$$

For this setting $\mathrm{Est}_n \leq \mathcal{O}(d \log(n))$ (see Section 19.1).

20.1 Experimental Setup

Our main objective is to compare our algorithm ANYTIME-E2D to the fixed-horizon E2D algorithm by [FKQR21]. ANYTIME-E2D and E2D were implemented by using the procedure described in Chapter 18. Both ANYTIME-E2D and E2D need to solve the inner convex problem in Lemma 48. To do so we use Frank-Wolfe [FW56, DH78, Jag13] for 100 steps and warm-starting the optimization at the solution from the previous round, μ_{t-1}. For ANYTIME-E2D we further perform a grid search over $\lambda \in [0, \max_{g \in \mathcal{M}} \epsilon^{-2} \text{dec}^{ac,f}(g)]$ (with a discretization of 50 points) to optimize over lambda within each iteration of Frank-Wolfe. For both the E2D and ANYTIME-E2D algorithm we used the version with the gaps Δ replaced with Δ_f, since we noticed that both algorithms performed better with Δ_f. For E2D, the scale hyperparameter λ was set using $\lambda = \sqrt{\frac{n}{4\log(n)}}$ as mentioned in [FKQR21, Section 6.1.1]. While for ANYTIME-E2D we set the hyper-parameter $\epsilon_t^2 = d/t$. Further, we compare to standard bandit algorithms: Upper Confidence Bound (UCB) and Thompson Sampling (TS) [LS20a].

20.2 Experiment 1

In this experiment, we aim to demonstrate the advantage of having an anytime algorithm. Specifically, we tune λ in the E2D algorithm for different horizons $n = 200, 500, 1000, 2000$, but run it for a fixed horizon of $n = 2000$. As such, we expect our algorithm ANYTIME-E2D to perform better than E2D when λ was tuned for the incorrect horizons (i.e. $n = 200, 500, 1000$). The feature dimension is $d = 3$. The number of decisions is $|\Pi| = 10$. We generated the features ϕ_π for each $\pi \in \Pi$ and parameter $f^* \in \mathbb{R}^d$ randomly at the beginning and then kept them fixed throughout the experimentation. 100 independent runs were performed for each algorithm.

The results of the experiment can be seen as the left plot in Fig. 20.1. As expected, our algorithm ANYTIME-E2D performs better than E2D (for $n = 200, 500, 1000$). This indicates that the E2D algorithm is sensitive to different settings of λ, which is problematic when the horizon is not known beforehand. Whereas our ANYTIME-E2D algorithm performs well even when the horizon is not known.

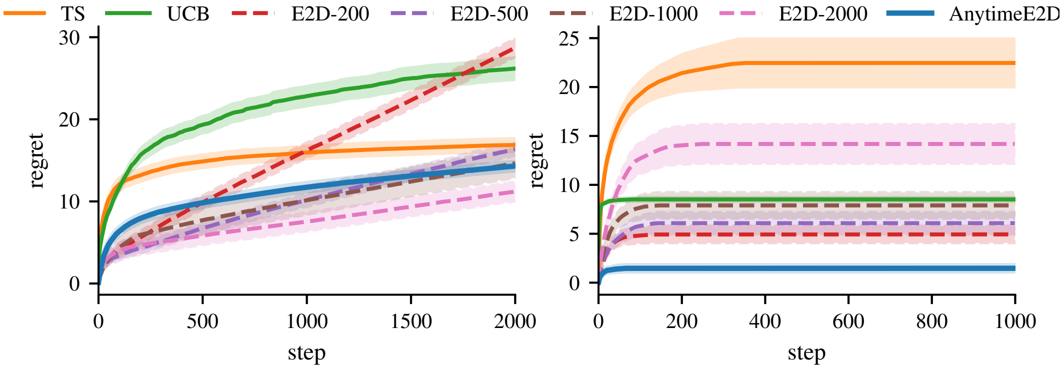


Figure 20.1: Running ANYTIME-E2D, TS, UCB, and E2D optimized for different horizons $n \in \{200, 500, 1000, 2000\}$. Left: The result for horizon $n = 2000$, and the feature space dimension $d = 3$. Right: The result for horizon $n = 1000$, and the feature space dimension $d = 30$.

20.3 Experiment 2

In this experiment, we investigate the case when $n < d^4$. As pointed out below Lemma 46, we expect improvement in this regime as the regret bound of our algorithm is $R_n \leq \min\{d\sqrt{n}, d^{1/3}n^{2/3}\}$, while the default, fixed-horizon E2D algorithm cannot achieve these bounds simultaneously and one has to pick one of $d\sqrt{n}$ or $d^{1/3}n^{2/3}$ beforehand for setting the scale hyperparameter λ. It is standard that the choice of λ is made according to the $d\sqrt{n}$ regret bound for E2D [FKQR21](which is not optimal when $n \ll d^4$), especially, if the horizon is not known beforehand. Thus, we set the horizon to $n = 1000$ and the dimension of the feature space to $d = 30$, which gives us that $n = 1000 \ll 810000 = d^4$. The rest of the setup and parameters are the same as in the previous experiment except for the features ϕ_π and f^* which are again chosen randomly in the beginning and then kept fixed throughout the experiment.

The results of the experiment can be seen as the right plot in Fig. 20.1. As expected, our algorithm ANYTIME-E2D performs better than E2D, UCB, and TS. This indicates that indeed, ANYTIME-E2D is likely setting λ appropriately to achieve the preferred $d^{1/3}n^{2/3}$ regret rate for small horizons. The poor performance of the other algorithms can be justified, since E2D is optimized based on the worse $d\sqrt{n}$ regret rate (for small horizons), while the UCB and TS algorithms are not known to get regret better than $d\sqrt{n}$.

References

[AAGM15] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman. *Asymptotic geometric analysis, Part I*, volume 202. American Mathematical Soc., 2015.

[ADX10] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pages 28–40. Citeseer, 2010.

[AG12] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference on Learning Theory*, 2012.

[AG13] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, Atlanta, GA, USA, 2013. JMLR.org.

[AL99] Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 3–11, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[AL17] M. Abeille and A. Lazaric. Linear Thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 176–184, Fort Lauderdale, FL, USA, 2017. JMLR.org.

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[AZ22] Alekh Agarwal and Tong Zhang. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022.

[BDKP15] S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Conference on Learning Theory*, pages 266–278, Paris, France, 2015. JMLR.org.

[BE18] S. Bubeck and R. Eldan. Exploratory distributions for convex functions. *Mathematical Statistics and Learning*, 1(1):73–100, 2018.

[BLE17] S. Bubeck, Y.T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 72–85, New York, NY, USA, 2017. ACM.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CCH$^+$07] I-S. Chang, L-C. Chien, C. Hsiung, C-C. Wen, and Y-J. Wu. Shape restricted regression with random bernstein polynomials. *Lecture Notes-Monograph Series*, pages 187–202, 2007.

[CMB22] Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022.

[CMP17] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

[CMPL15] R. Combes, S. Magureanu, A. Proutiere, and C. Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 231–244. ACM, 2015.

[dB15] A. V. den Boer. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18, 2015.

[DH78] Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

[DHK08] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Conference on Learning Theory*, pages 355–366, 2008.

[DKL$^+$21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[DM22] Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making. *arXiv preprint arXiv:2206.02326*, 2022.

[DMR19] S. Dong, T. Ma, and B. Van Roy. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160. PMLR, 2019.

[DMSV20] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

[DMZZ21] Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

[DSK20] Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2443–2452. PMLR, 13–18 Jul 2020.

[FCGS10] S. Filippi, O. Cappé, A. Garivier, and Cs. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594. Curran Associates, Inc., 2010.

[FGH23] Dylan J Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023.

[FGQ$^+$22] Dylan J Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*, 2022.

[FKM05] A Flaxman, A Kalai, and HB McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *SODA'05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

[FKQR21] D. J. Foster, S. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[FRSS22] D. J. Foster, A. Rakhlin, A. Sekhari, and K. Sridharan. On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems*, 35:35404–35417, 2022.

[FvdHLM24a] H. Fokkema, D. van der Hoeven, T. Lattimore, and J. Mayo. Improved online Newton method for bandit convex optimisation. 2024.

[FvdHLM24b] H. Fokkema, D. van der Hoeven, T. Lattimore, and J. Mayo. Online Newton method for bandit convex optimisation. In *Conference on Learning Theory*, 2024.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[GPS13] András György, Dávid Pál, and Csaba Szepesvári. Online learning: Algorithms for big data. *Lecture Notes*, 2013.

[GRM$^+$20] Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

[HHK^{+}21] B. Huang, K. Huang, S. Kakade, J. D. Lee, Q. Lei, R. Wang, and J. Yang. Optimal gradient-based algorithms for non-concave bandit optimization. *Advances in Neural Information Processing Systems*, 34:29101–29115, 2021.

[HLW20] Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

[Jag13] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

[JAZBJ18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[KBJK20] Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally Robust Bayesian Optimization. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, August 2020.

[Kha90] L. G. Khachiyan. An inequality for the volume of inscribed ellipsoids. *Discrete & computational geometry*, 5:219–222, 1990.

[KK18] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

[KK21] Johannes Kirschner and Andreas Krause. Bias-robust bayesian optimization via dueling bandits. In *International Conference on Machine Learning*, pages 5595–5605. PMLR, 2021.

[KKM12] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 199–213. Springer Berlin Heidelberg, 2012.

[Kle05] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704. MIT Press, 2005.

[KLK20] Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR, 2020.

[KLK23] Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision making: Algorithms, regret bounds and applications. *Journal of Machine Learning Research*, August 2023.

[KLS95] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13:541–559, 1995.

[KLVS21] Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

[KZS$^+$20] B. Kveton, M. Zaheer, Cs. Szepesvári, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

[LAK$^+$14] Tian Lin, Bruno Abrahao, Robert Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pages 901–909, 2014.

[Lat20] T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3/4):311–334, 2020.

[Lat21] T. Lattimore. Minimax regret for bandit convex optimisation of ridge functions. *arXiv preprint arXiv:2106.00444*, 2021.

[Lat24] T. Lattimore. *Bandit Convex Optimisation*. 2024.

[LG21] T. Lattimore and A. György. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

[LG23] T. Lattimore and A. György. A second-order method for stochastic bandit convex optimisation. *arXiv preprint arXiv:2302.05371*, 2023.

[LH21] T. Lattimore and B. Hao. Bandit phase retrieval. *Advances in Neural Information Processing Systems*, 34:18801–18811, 2021.

[LKLS18] T. Lattimore, B. Kveton, S. Li, and Cs. Szepesvári. Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems*, pages 3949–3958. Curran Associates, Inc., 2018.

[Lov99] László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

[LS17] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

[LS19] T. Lattimore and Cs. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the 32nd Conference on Learning Theory*, pages 2111–2139, Phoenix, USA, 2019. PMLR.

[LS20a] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LS20b] Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pages 2488–2515. PMLR, 2020.

[MS11] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24, 2011.

[Nie92] W. Niemiro. Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, pages 1514–1533, 1992.

[OBPVR16] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

[RKJ08] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791. ACM, 2008.

[RLS93] P. Ramgopal, P. Laud, and A. Smith. Nonparametric bayesian bioassay with prior constraints on the shape of the potency curve. *Biometrika*, 80(3):489–498, 1993.

[Rus99] Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1-2):224–243, 1999.

[RV14] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591. Curran Associates, Inc., 2014.

[RV16] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[RVK$^+$18] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

[SKKS10a] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, page 1015–1022, Madison, WI, USA, 2010. Omnipress.

[SKKS10b] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

[Smi84] Robert L. Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

[SNNJ21] A. Saha, N. Natarajan, P. Netrapalli, and P. Jain. Optimal regret algorithm for pseudo-1d bandit convex optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9255–9264. PMLR, 2021.

[SS$^+$12] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[SWD11] T. Shively, S. Walker, and P. Damien. Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics*, 161(2):166–181, 2011.

[Tho33a] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[Tho33b] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[TKE88] S. Tarasov, L. G. Khachiyan, and I. I. Erlich. The method of inscribed ellipsoids. In *Soviet Mathematics-Doklady*, volume 37, pages 226–230, 1988.

[Tko18] T. Tkocz. Asymptotic convex geometry lecture notes. 2018.

[YJ09] Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

[ZGS21] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[Zha22] Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

[ZL19] J. Zimmert and T. Lattimore. Connections between mirror descent, thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems*, pages 11973–11982. Curran Associates, Inc., 2019.

[ZLKB20] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.