Statistical Reinforcement Learning and Decision Making: Course Notes

Dylan J. Foster and Alexander Rakhlin

Taught Fall 2022 @ MIT

Last Updated: April 2023

Caution: This is a live draft, and all parts will be updated regularly. Please send us an email if you find a mistake or a typo.

Contents

1. INTRODUCTION

1.1 Decision Making

This is a course about learning to make decisions in an interactive, data-driven fashion. When we say interactive decision making, we are thinking of problems such as:

- Medical treatment: based on a patient's medical history and vital signs, we need to decide what treatment will lead to the most positive outcome.
- Controlling a robot: based on sensor signals, we need to decide what signals to send to a robot's actuators in order to navigate to a goal.

For both problems, we (the *learner*/*agent*) are interacting with an *unknown environment*. In the robotics example, we do not necessarily a-priori know how the signals we send to our robot's actuators change its configuration, or what the landscape it's trying to navigate looks like. However, because we are able to *actively* control the agent, we can *learn* to model the environment on the fly as we make decisions and collect data, which will reduce uncertainty and allow us to make better decisions in the future. The crux of the interactive decision making problem is to make decisions in a way that balances (i) exploring the environment to reduce our uncertainty and (ii) maximizing our overall performance (e.g., reaching a goal state as fast as possible).

Figure 1 depicts an idealized interactive decision making setting, which we will return to throughout this course. Here, at each round t, the agent (doctor) observes the medical history and vital signs of a patient, summarized in a *context* x^t, makes a treatment *decision* π^t, and then observes the outcomes of the treatment in the form of a *reward* r^t, and an auxiliary *observation* o^t about, say, illness progression. With time, we hope that the doctor will learn a good mapping from contexts to decisions. How can we develop an automated system that can achieve this goal?

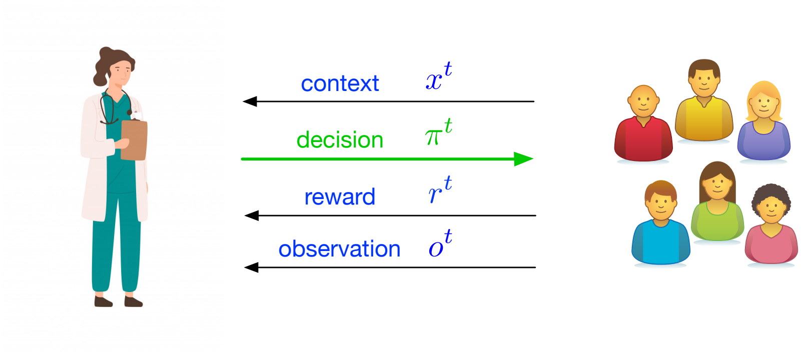


Figure 1: A general decision making problem.

The decision making framework in Figure 1 generalizes many interactive decision making problems the reader might already be familiar with, including multi-armed bandits, contextual bandits, and reinforcement learning. We will cover the foundations of algorithm design and analysis for all of these settings from a unified perspective, with an emphasis on *sample efficiency* (i.e., how to learn a good decision making policy using as few rounds of interaction as possible).

1.2 A Spectrum of Decision Making Problems

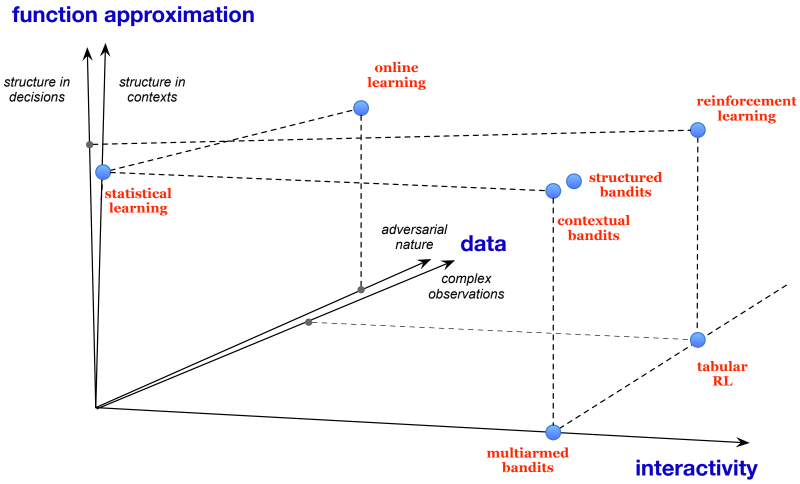


Figure 2: Landscape of decision making problems.

To design algorithms for general interactive decision making problems such as Figure 1, there are many complementary challenges we must overcome. These challenges correspond to different assumptions we can place on the underlying environment and decision making protocol, and give rise to what we describe as a *spectrum* of decision making problems, which is illustrated in Figure 2. There are three core challenges we will focus on throughout the course, which are given by the axes of Figure 2.

- **Interactivity.** Does the learning agent observe data passively, or do the decisions they make actively influence what data we collect? In the setting of Figure 1, the doctor observes the effects of the prescribed treatments, but not the counterfactuals (the effects of the treatments not given). Hence, doctor's decisions influence the data they can collect, which in turn may significantly alter the ability to estimate the effects of different treatments. On the other hand, in classical machine learning, a dataset is typically given to the learner upfront, with no control over how it is collected.

- **Function approximation and generalization.** In supervised statistical learning and estimation, one typically employs *function approximation* (e.g., models such as neural networks, kernels, or forests) to generalize across the space of covariates. For decision making, we can employ function approximation in a similar fashion, either to generalize across a space of contexts, or to generalize across the space of *decisions*. In the setting of Figure 1, the context x^t summarizing the medical history and vital signs might be a highly structured object. Likewise, the treatment π^t might be a high-dimensional vector with interacting components, or a complex multi-stage treatment strategy.

- **Data.** Is the data (e.g., rewards or observations) observed by our learning algorithm produced by a fixed data-generating process, or does it evolve arbitrarily, and even adversarially in response to our actions? If there is fixed data-generating process, do we wish to directly model it, or should we instead aim to be agnostic? Do we observe only the labels of images, as in supervised learning, or a full trajectory of states/actions/rewards for a policy employed by the robot?

As shown in Figure 2, many basic decision making and learning frameworks (contextual bandits, structured bandits, statistical learning, online learning) can be thought of as idealized problems that each capture one or more of the possible challenges, while richer settings such as reinforcement learning encompass all of them.

Figure 2 can be viewed as a roadmap for the course. We start with a brief introduction to Statistical Learning (Section 1.4) and Online Learning (Section 1.6); the concepts and results stated here will serve as a backbone for the rest of the course. We will then study, in order, the problems of Multi-Armed Bandits (Section 2), Contextual Bandits (Section 3), Structured Bandits (Section 4), Tabular Reinforcement Learning (Section 5), General Decision Making (Section 6), and Reinforcement Learning with General Function Approximation (Section 7). Each of these topics will add a layer of complexity, and our aim is to develop a unified approach to all the aforementioned problems, both in terms of statistical complexity (the number of interactions required to achieve the goal), and in terms of algorithm design.

1.3 Minimax Perspective

For much of the course, we take a *minimax* point of view. Abstractly, let $\mathcal{M}$ be a set of possible models (or, choices for the environment) that can be encountered by the learner/decision maker. The set $\mathcal{M}$ can be thought of as representing the prior knowledge of the learner about the underlying environment. Let Alg denote a learning algorithm, and $\mathsf{Perf}_T(\mathsf{Alg}, M)$ be some notion of performance of algorithm Alg on model $M \in \mathcal{M}$ after T rounds of interaction (or—in passive learning—after observing T datapoints). We would like to develop algorithms that perform well, no matter what the model $M \in \mathcal{M}$ is, in the sense that Alg

approximately solves the minimax problem

$$\min_{\mathsf{Alg}} \max_{M\in\mathcal{M}} \mathsf{Perf}_T(\mathsf{Alg}, M). \tag{1.1}$$

Understanding the *statistical complexity* (or, difficulty) of a given problem amounts to establishing matching (or nearly matching) upper bounds $\bar{\phi}_T(\mathcal{M})$ and lower bounds $\underline{\phi}_T(\mathcal{M})$ on the minimax value in (1.1). While developing such upper and lower bounds for specific model classes $\mathcal{M}$ of interest might be a simple task, the grand aim of this course is to develop a more fundamental, unified understanding of what makes *any* model class $\mathcal{M}$ easy verus hard, and to give sharp results for all (or nearly all) $\mathcal{M}$.

On the algorithmic side, we would like to better understand the scope of optimal algorithms that solve (1.1). While the minimax problem is itself an optimization problem, the space of all algorithms is typically prohibitively large. One of the key insights to be leveraged in this course is that for general decision making problems, we can restrict ourselves to algorithms that interleave a type of supervised learning called *online estimation* (this will be described in Sections 1.4 and 1.6), with a principled choice of *exploration strategy* that balances greedily maximizing performance (exploitation) with information acquisition (exploration). As we show, such algorithms achieve or nearly achieve optimality in (1.1) for a surprisingly wide range of decision making problems.

1.4 Statistical Learning: Brief Recap

We begin with a short refresher on the statistical learning problem. Statistical learning is a purely passive problem in which the learner does not directly interact with the environment, but captures the challenge of *generalization and function approximation* the context of Figure 2.

In the statistical learning problem, we receive examples $(x^1, y^1), \ldots, (x^T, y^T) \in \mathcal{X} \times \mathcal{Y}$, i.i.d. from a (unknown) distribution $M^\star$. Here $x^t \in \mathcal{X}$ are *features* (sometimes called contexts or covariates), and $\mathcal{X}$ is the feature space. $y^t \in \mathcal{Y}$ are called *outcomes*, and $\mathcal{Y}$ is the outcome space. Given $(x^1, y^1), \ldots, (x^T, y^T)$, the goal is to produce a model (or, estimator) $\widehat{f} : \mathcal{X} \to \mathcal{Y}'$ that will do a good job *predicting* outcomes from features for future examples (x, y) drawn from $M^\star$.[1]

To measure prediction performance, we take as given a *loss function* $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$. Standard examples include:

- Regression, where common losses include the *square loss* $\ell(a, b) = (a - b)^2$ when $\mathcal{Y} = \mathcal{Y}' = \mathbb{R}$.
- Classification, where $\mathcal{Y} = \mathcal{Y}' = \{0, 1\}$ and we consider the indicator (or 0-1) loss $\ell(a, b) = \mathbb{I}\{a \neq b\}$.
- Conditional density estimation with the *logarithmic loss* (log loss). Here $\mathcal{Y}' = \Delta(\mathcal{Y})$, the set of distributions on $\mathcal{Y}$, and for $p \in \mathcal{Y}'$,

$$\ell_{\log}(p, y) = -\log p(y). \tag{1.2}$$

For a function $f : \mathcal{X} \to \mathcal{Y}'$, we measure the prediction performance via the *population* (or, "test") loss:

$$L(f) := \mathbb{E}_{(x,y)\sim M^\star}[\ell(f(x), y)]. \tag{1.3}$$

[1] Note that we allow the outcome space $\mathcal{Y}$ to be different from the prediction space $\mathcal{Y}'$.

Letting $\mathcal{H}^T := \{(x^t, y^t)\}_{t=1}^T$ denote the dataset, a (deterministic) *algorithm* is a map that takes the dataset as input and returns a function/predictor:

$$\widehat{f}(\cdot; \mathcal{H}^T) : \mathcal{X} \to \mathcal{Y}'. \tag{1.4}$$

The goal in designing algorithms is to ensure that $\mathbb{E}\big[L(\widehat{f})\big]$ is minimized, where $\mathbb{E}[\cdot]$ denotes expectation with respect to the draw of the dataset $\mathcal{H}^T$. Without any assumptions, it is not possible to learn a good predictor unless the number of examples T scales with $|\mathcal{X}|$ (this is sometimes called the *no-free-lunch theorem*). The basic idea behind statistical learning is to work with a restricted class of functions

$$\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$$

in order to facilitate generalization. The class $\mathcal{F}$ can be thought of as (implicitly) encoding prior knowledge about the structure of the data. For example, in computer vision, if the features x^t correspond to images and the outcomes y^t are labels (e.g., "cat" or "dog"), one might expect that choosing $\mathcal{F}$ to be a class of convolutional neural networks will work well, since this encodes spatial structure.

Remark 1 (Conditional density estimation): For the problem of conditional density estimation, we shall overload the notation and interchangeably write $f(x)$ and $f(\cdot|x)$ for the conditional distribution. In this setting, the learner is required to compute a distribution for each x rather than form a point estimate (see Figure 3). For an outcome y, the loss is the negative log of the conditional density for the outcome.

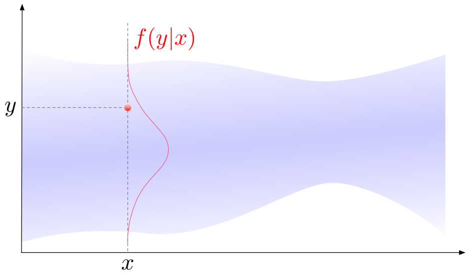


Figure 3: Conditional density estimation.

Empirical risk minimization and excess risk. The most basic and well-studied algorithmic principle for statistical learning is *Empirical Risk Minimization* (ERM). Define the empirical loss for the dataset $\mathcal{H}^T$ as

$$\widehat{L}(f) = \frac{1}{T}\sum_{i=1}^{T} \ell(y^i, f(x^i)). \tag{1.5}$$

Then, the empirical risk minimizer with respect to the class $\mathcal{F}$ is given by

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{L}(f). \tag{1.6}$$

To measure the performance of ERM and other algorithms that attempt to learn with $\mathcal{F}$, we consider *excess loss* (or, regret)

$$\mathcal{E}(f) = L(f) - \min_{f' \in \mathcal{F}} L(f'). \tag{1.7}$$

Intuitively, the quantity $\min_{f' \in \mathcal{F}} L(f')$ in (1.7) captures the best prediction performance any function in $\mathcal{F}$ can achieve, even with knowledge of the *true distribution*. If an algorithm $\widehat{f}$ has low excess risk, this means that we are predicting future outcomes nearly as well as any algorithm based on samples can hope to perform. ERM an other algorithms can ensure that $\mathcal{E}(\widehat{f})$ is small in expectation or with high probability over draw of the dataset $\mathcal{H}^T$.

Connection to estimation. An appealing feature of the formulation in (1.7) is that it does not presuppose any relationship between the class $\mathcal{F}$ and the data distribution; in other words, it is agnostic. However, if $\mathcal{F}$ does happen to be good at modeling the data distribution, the excess loss has an additional interpretation based on estimation.

Definition 1: For prediction with square loss, we say that the problem is *well-specified* (or, realizable) if the regression function $f^\star(x) := \mathbb{E}[Y|X = x]$ is in $\mathcal{F}$.

The regression function $f^\star$ can also be seen as a minimizer of $L(f)$ over measurable functions f, for the same reason that $\mathbb{E}(Z - a)^2$ is minimized at $a = \mathbb{E}[Z]$.

Lemma 1: For the square loss, if the problem is *well-specified*, then for all $f : \mathcal{X} \to \mathcal{Y}$,

$$\mathcal{E}(f) = \mathbb{E}\big[(f(X) - f^\star(X))^2\big] \tag{1.8}$$

Proof of Lemma 1. Adding and subtracting $f^\star$ in the first term, we have

$$\mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^\star(X) - Y)^2 = \mathbb{E}(f(X) - f^\star(X))^2 + 2\,\mathbb{E}[(f^\star(X) - Y)(f(X) - f^\star(X))].$$

□

Inspecting (1.8), we see that achieving low excess loss implies that we are estimating the true regression function $f^\star$. In this case, the aim of prediction and estimation coincide.

Guarantees for ERM. We give bounds on the excess loss of ERM for perhaps the simplest special case, in which $\mathcal{F}$ is finite.

Proposition 1: For any finite class $\mathcal{F}$, empirical risk minimization satisfies

$$\mathbb{E}\big[\mathcal{E}(\widehat{f})\big] \lesssim \mathsf{comp}(\mathcal{F}, T), \tag{1.9}$$

where

1. For any bounded loss (including classification), $\mathsf{comp}(\mathcal{F}, T) = \sqrt{\frac{\log|\mathcal{F}|}{T}}$.

2. For square loss regression, if the problem is well-specified, $\mathsf{comp}(\mathcal{F}, T) = \frac{\log|\mathcal{F}|}{T}$.

In addition, there exists a (different) algorithm that achieves $\mathsf{comp}(\mathcal{F},T) = \frac{\log|\mathcal{F}|}{T}$ for both square loss regression and conditional density estimation, even when the problem is not well-specified.

The rate $\mathsf{comp}(\mathcal{F},T) = \sqrt{\frac{\log|\mathcal{F}|}{T}}$ above is sometimes referred to as a *slow rate*, and is optimal for generic losses. The rate $\mathsf{comp}(\mathcal{F},T) = \frac{\log|\mathcal{F}|}{T}$ is referred to as a *fast rate*, and takes advantage of additional structure (curvature, or strong convexity) of the square loss. Critically, both bounds scale only with the cardinality of $\mathcal{F}$, and do not depend on the size of the feature space $\mathcal{X}$, which could be infinite. This reflects the fact that working with a restricted function class is allowing us to generalize across the feature space $\mathcal{X}$. In this context the cardinality $\log|\mathcal{F}|$ should be thought of a notion of *capacity*, or *expressiveness* for $\mathcal{F}$. Intuitively, choosing a larger, more expressive class will require a larger amount of data, but will make the excess loss bound in (1.7) more meaningful, since the benchmark will be stronger.

Remark 2 (From finite to infinite classes): Throughout these lecture notes, we restrict our attention to finite classes whenever possible in order to simplify presentation. If one wishes to move beyond finite classes, a well-developed literature within statistical learning provides various notions of complexity for $\mathcal{F}$ that lead to bounds on $\mathsf{comp}(\mathcal{F},T)$ for ERM and other algorithms. These include the Vapnik-Chervonenkis (VC) dimension for classification, Rademacher complexity, and covering numbers. Standard references include Bousquet et al. [16], Boucheron et al. [15], Anthony and Bartlett [8], Shalev-Shwartz and Ben-David [66].

1.5 Refresher: Random Variables and Averages

To prove Proposition 1 and similar generalization bounds, the main tools we will use are *concentration inequalities* (or, tail bounds) for random variables.

Definition 2: A random variable Z is sub-Gaussian with variance factor (or variance proxy) σ^2 if

$$\forall \eta \in \mathbb{R}, \qquad \mathbb{E}\, e^{\eta(Z-\mathbb{E}[Z])} \leq e^{\sigma^2\eta^2/2}.$$

Note that if $Z \sim \mathcal{N}(0,\sigma^2)$ is *Gaussian* with variance σ^2, then it is sub-Gaussian with variance proxy σ^2. In this sense, sub-Gaussian random variables generalize the tail behavior of Gaussians. A standard application of Chernoff method yields the following result.

Lemma 2: If $Z_1,\ldots,Z_T$ are i.i.d. random variables with variance proxy σ^2, then

$$\mathbb{P}\left(\frac{1}{T}\sum_{i=1}^{T} Z_i - \mathbb{E}[Z] \geq u\right) \leq \exp\left\{-\frac{Tu^2}{2\sigma^2}\right\} \qquad (1.10)$$

Applying this result with Z and $-Z$ and taking a union bound yields the following two-sided guarantee.

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{i=1}^{T} Z_i - \mathbb{E}[Z]\right| \geq u\right) \leq 2\exp\left\{-\frac{Tu^2}{2\sigma^2}\right\}. \tag{1.11}$$

This implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\frac{1}{T}\sum_{i=1}^{T} Z_i - \mathbb{E}[Z]\right| \leq \sqrt{\frac{2\sigma^2\log(2/\delta)}{T}}; \tag{1.12}$$

to see this, set the right-hand side of (1.11) to δ and solve for u.

Remark 3 (Union bound): The factor 2 under the logarithm in (1.12) is the result of applying union bound to (1.10). Throughout the course, we will frequently apply the union bound to multiple—say N—high probability events involving sub-Gaussian random variables. In this case, the union bound will read as $\log(N/\delta)$. The mild logarithmic dependence is due to the sub-Gaussian tail behavior of the averages.

The following result shows that any bounded random variable is sub-Gaussian.

Lemma 3 (Hoeffding's inequality): Any random variable Z taking values in $[a,b]$ is sub-Gaussian with variance proxy $(b-a)^2/4$, i.e.

$$\forall \eta \in \mathbb{R}, \qquad \ln \mathbb{E}\exp\{-\eta(Z-\mathbb{E}[Z])\} \leq \frac{\eta^2(b-a)^2}{8}. \tag{1.13}$$

As a consequence, for i.i.d. random variables $Z_1,\ldots,Z_T$ taking values in $[a,b]$ almost surely, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{i=1}^{T} Z_i - \mathbb{E}[Z] \leq (b-a)\sqrt{\frac{\log(1/\delta)}{2T}} \tag{1.14}$$

Using Hoeffding's inequality, we can prove now prove Part 1 (the slow rate) from Proposition 1.

Lemma 4 (Proposition 1, Part 1): Let $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be finite, and assume $\ell \circ f \in [0,1]$ almost surely. Then with probability at least $1-\delta$, ERM satisfies

$$L(\widehat{f}) - \min_{f\in\mathcal{F}} L(f) \leq 2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2T}}.$$

Proof of Lemma 4. For any $f \in \mathcal{F}$, we can write

$$L(\widehat{f}) - L(f) = \left[L(\widehat{f}) - \widehat{L}(\widehat{f})\right] + \left[\widehat{L}(\widehat{f}) - \widehat{L}(f)\right] + \left[\widehat{L}(f) - L(f)\right].$$

Observe that for all $f : \mathcal{X} \to \mathcal{Y}$, we have

$$\left|L(f) - \widehat{L}(f)\right| = \left|\mathbb{E}\,\ell(f(X), Y) - \frac{1}{T}\sum_{i=1}^{T} \ell(f(X_i), Y_i)\right|.$$

By union bound and Lemma 3, with probability at least $1 - |\mathcal{F}|\delta$,

$$\forall f \in \mathcal{F}, \qquad \left|\mathbb{E}\,\ell(f(X), Y) - \frac{1}{T}\sum_{i=1}^{T} \ell(f(X_i), Y_i)\right| \leq \sqrt{\frac{\log(2/\delta)}{2T}} \tag{1.15}$$

□

To deduce the in-expectation bound of Proposition 1 from the high-probability tail bound of Lemma 4, a standard technique of "integrating out the tail" is employed. More precisely, for a nonnegative random variable U, it holds that $\mathbb{E}[U] \leq \tau + \int_{\tau}^{\infty} \mathbb{P}\left(U \geq z\right) dz$ for all $\tau > 0$; choosing $\tau \propto T^{-1/2}$ concludes the proof.

To prove the Part 2 (the fast rate) from Proposition 1, we need a more refined concentration inequality (Bernstein's inequality), which gives tighter guarantees for random variables with small variance.

Lemma 5 (Bernstein's inequality): Let $Z_1, \ldots, Z_T, Z$ be i.i.d. with variance $\mathbb{V}(Z_i) = \sigma^2$, and range $|Z - \mathbb{E}\,Z| \leq B$ almost surely. Then with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{i=1}^{T} Z_i - \mathbb{E}\,Z \leq \sigma\sqrt{\frac{2\log(1/\delta)}{T}} + \frac{B\log(1/\delta)}{3T}. \tag{1.16}$$

The proof for Part 2 is given as an exercise in Section 1.7. We refer the reader to Appendix A.1 for further background on tail bounds.

1.6 Online Learning/Prediction

We now move on to the problem of *online learning*, or sequential prediction. The online learning problem generalizes statistical learning on two fronts:

- Rather than receiving a batch dataset of T examples all at once, we receive the examples (x^t, y^t) one by one, and must predict y^t from x^t only using the examples we have already observed.
- Instead of assuming that examples are drawn from a fixed distribution, we allow examples to be generated in an arbitrary, potentially adversarial fashion.

```
Online Learning Protocol
for $t = 1, \ldots, T$ do
    Compute predictor $\widehat{f}^t : \mathcal{X} \to \mathcal{Y}$
    Observe $(x^t, y^t) \in \mathcal{X} \times \mathcal{Y}$
```

In more detail, at each timestep t, given the examples

$$\mathcal{H}^{t-1} = \{(x^1, y^1), \ldots, (x^{t-1}, y^{t-1})\} \tag{1.17}$$

observed so far, the algorithm produces a predictor

$$\widehat{f}^t = \widehat{f}^t(\cdot \mid \mathcal{H}^{t-1}),$$

which aims to predict the outcome y^t from the features x^t. The algorithm's goal is to minimize the cumulative loss over T rounds, given by

$$\sum_{t=1}^{T} \ell(\widehat{f}^t(x^t), y^t)$$

for a known loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$; the cumulative loss can be thought of as a sum of "out-of-sample" prediction errors. Since we will not be placing assumptions on the data-generating process, it is not possible to make meaningful statements about the cumulative loss itself. However, we can aim to ensure that this cumulative loss is not much worse than the best empirical explanation of the data by functions in a given class $\mathcal{F}$. That is, we measure the algorithm's performance via *regret* to $\mathcal{F}$:

$$\mathbf{Reg} = \sum_{t=1}^{T} \ell(\widehat{f}^t(x^t), y^t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x^t), y^t). \tag{1.18}$$

Our aim is to design prediction algorithms that keep regret small for *any* sequence of data. As in statistical learning, the class $\mathcal{F}$ should be thought of as capturing our prior knowledge about the problem, and might be a linear model or neural network. At first glance, keeping the regret small for arbitrary sequences might seem like an impossible task, as it stands in stark contrast with statistical learning, where data is generated i.i.d. from a fixed distribution. Nonetheless, we will that algorithms with guarantees similar to those for statistical learning are available.

Let us remark that it is often useful to apply online learning methods in settings where data is not fully adversarial, but evolves according to processes too difficult to directly model. For example, in the chapters that follow, we will apply online methods as a subroutine with more sophisticated algorithms for decision making. Here, the choice of past decisions, while in our purview, does not look like i.i.d. or simple time-series data.

Remark 4 (Proper learning, improper learning, and randomization): The online learning protocol does not require that $\widehat{f}^t$ lies in $\mathcal{F}$ ($\widehat{f}^t \in \mathcal{F}$). A method that chooses functions from $\mathcal{F}$ will be called *proper*, and the one that selects predictors outside of $\mathcal{F}$ will be called *improper*. It will also be useful to allow for *randomized* predictions of the form

$$\widehat{f}^t \sim q^t(\cdot|\mathcal{H}^{t-1}),$$

where q^t is a distribution on functions, typically on elements of $\mathcal{F}$. For randomized predictions, we slightly abuse notation and write regret as

$$\mathbf{Reg} = \sum_{i=1}^{T} \mathbb{E}_{\widehat{f}^t \sim q^t}\left[\ell(\widehat{f}^t(x^t), y^t)\right] - \min_{f \in \mathcal{F}} \sum_{i=1}^{T} \ell(f(x^t), y^t). \tag{1.19}$$

The algorithms we introduce in the sequel below ensure small regret even if data are adversarially and adaptively chosen. More precisely, for deterministic algorithms, (x^t, y^t) may be chosen based on $\widehat{f}^t$ and all the past data, while for randomized algorithms, Nature can only base this choice on q^t.

In the context of Figure 2, online learning generalizes statistical learning by considering arbitrary sequences of data, but still allows for general-purpose function approximation and generalization via the class $\mathcal{F}$. While the setting involves making predictions in an online fashion, we do not think of this as an *interactive* decision making problem, because the predictions made by the learning agent do not directly influence what data the agent gets to observe.

1.6.1 Connection to Statistical Learning

Online learning can be thought of as a generalization of statistical learning, and in fact, algorithms for online learning immediately yield algorithms for statistical learning via a technique called *online-to-batch conversion*. This result, which is formalized by the following proposition, rests on two observations: the cumulative loss of the algorithm looks like a sum of out-of-sample errors, and the minimum empirical fit to realized data (over $\mathcal{F}$) is, on average, a harder (that is, smaller) benchmark than the minimum expected loss in $\mathcal{F}$.

Proposition 2: Suppose the examples $(x^1, y^1), \dots, (x^T, y^T)$ are drawn i.i.d. from a distribution $M^\star$, and suppose the loss function $a \mapsto \ell(a, b)$ is convex in the first argument for all b. Then for any online learning algorithm, if we define

$$\widehat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} \widehat{f}^t(x),$$

we have

$$\mathbb{E}\big[\mathcal{E}(\widehat{f})\big] \leq \frac{1}{T} \cdot \mathbb{E}[\mathbf{Reg}].$$

Proof of Proposition 2. Let $(x, y) \sim M^\star$ be a fresh sample which is independent of the history $\mathcal{H}^T$. First, by Jensen's inequality,

$$\mathbb{E}\Big[L(\widehat{f})\Big] = \mathbb{E}\left[\mathbb{E}_{(x,y)}\, \ell\left(\frac{1}{T}\sum_{t=1}^{T} \widehat{f}^t(x), y\right)\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(x,y)}\, \ell\left(\widehat{f}^t(x), y\right)\right] \tag{1.20}$$

which is equal to

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(x^t,y^t)}\, \ell\left(\widehat{f}^t(x^t), y^t\right)\right] \tag{1.21}$$

since $\widehat{f}^t$ is a function of $\mathcal{H}^{t-1}$ and (x, y) and (x^t, y^t) are i.i.d. Second,

$$\min_{f\in\mathcal{F}} L(f) = \min_{f\in\mathcal{F}} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell(f(x^t), y^t)\right] \geq \mathbb{E}\left[\min_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f(x^t), y^t)\right] \tag{1.22}$$

□

In light of Proposition 2, one can interpret regret as generalizing the notion of excess risk from i.i.d. data to arbitrary sequences.

Similar to statistical learning, the regret for online learning has an additional interpretation in terms of *estimation* if the outcomes for the problem are well-specified.

Lemma 6: Suppose that the features $x^1, \dots, x^T$ are generated in an arbitrary fashion, but that for all t,

$$\mathbb{E}[y^t \mid x^t = x] = f^\star(x)$$

for some $f^\star \in \mathcal{F}$. Then for the problem of prediction with square loss,

$$\mathbb{E}[\mathbf{Reg}] \geq \mathbb{E}\left[\sum_{t=1}^{T} (\widehat{f}^t(x^t) - f^\star(x^t))^2\right].$$

Notably, this result holds even if the features $x^1, \dots, x^T$ are generated adversarially, with no prior knowledge of the sequence. This is a significant departure from classical estimation results in statistics, where estimation of an unknown function is typically done over a fixed, known sequence ("design") $x^1, \dots, x^T$, or with respect to an i.i.d. dataset.

1.6.2 The Exponential Weights Algorithm

The main online learning algorithm is the *Exponential Weights* algorithm, which is applicable to finite classes $\mathcal{F}$. At each time t, the algorithm computes a distribution $q^t \in \Delta(\mathcal{F})$ via

$$q^t(f) \propto \exp\left\{-\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i)\right\}, \tag{1.23}$$

where $\eta > 0$ is a learning rate. Based on q^t, the algorithm forms the prediction $\widehat{f}^t$. We give two variants here of the method here.

```
Exponential Weights (averaged)
for $t = 1, \dots, T$ do
    Compute $q^t$ in (1.23).
    Let $\widehat{f}^t = \mathbb{E}_{f \sim q^t}[f]$.
    Observe $(x^t, y^t)$, incur $\ell(\widehat{f}^t(x^t), y^t)$.
```

```
Exponential Weights (randomized)
for $t = 1, \dots, T$ do
    Compute $q^t$ in (1.23).
    Sample $\widehat{f}^t \sim q^t$.
    Observe $(x^t, y^t)$, incur $\ell(\widehat{f}^t(x^t), y^t)$.
```

The only difference between these variants lies in whether we compute the prediction $\widehat{f}^t$ from q^t via

$$\widehat{f}^t = \mathbb{E}_{f \sim q^t}[f], \qquad \text{or} \qquad \widehat{f}^t \sim q^t. \tag{1.24}$$

The latter can be applied to any bounded loss functions, while the former leads to faster rates for specific losses such as the square loss and log loss, but is only applicable when $\mathcal{Y}'$ is convex. Note that the averaged version is inherently improper, while the second is proper, yet randomized. From the point of view of regret, the key difference between these two versions is the placement of "$\mathbb{E}_{f \sim q^t}$": For the averaged version it is inside the loss function, and for the randomized version it is outside (see (1.19)). The averaged version can therefore take advantage of the structure of the loss function, such as strong convexity, leading to faster rates. The following result shows that Exponential Weights leads to regret bounds for online learning, with rates that parallel those in Proposition 1.

Proposition 3: For any finite class $\mathcal{F}$, the Exponential Weights algorithm (with appropriate choice of η) satisfies

$$\frac{1}{T}\mathbf{Reg} \lesssim \mathsf{comp}(\mathcal{F}, T) \tag{1.25}$$

for any sequence, where:

1. For arbitrary bounded losses (including classification), $\mathsf{comp}(\mathcal{F}, T) = \sqrt{\frac{\log|\mathcal{F}|}{T}}$. This is achieved by the randomized variant.

2. For regression with the square loss and conditional density estimation with the log loss, $\mathsf{comp}(\mathcal{F}, T) = \frac{\log|\mathcal{F}|}{T}$. This is achieved by the averaged variant.

We now turn to the proof of Proposition 3. Since we are not placing any assumptions on the data generating process, we cannot hope to control the algorithm's loss at any particular time t, but only cumulatively. It is then natural to employ amortized analysis with a potential function.

In more detail, the proof of Proposition 3 relies on several steps, common to standard analyses of online learning: (i) define a potential function, (ii) relate the increase in potential at each time step, to the loss of the algorithm, (iii) relate cumulative loss of any expert $f \in \mathcal{F}$ to the final potential. For the Exponential Weights Algorithm, the proof relies on the following potential for time t, parameterized by $\eta > 0$:

$$\Phi_\eta^t = -\log \sum_{f\in\mathcal{F}} \exp\left\{-\eta \sum_{i=1}^{t} \ell(f(x^i), y^i)\right\}. \tag{1.26}$$

The choice of this potential is rather opaque, and a full explanation of its origin is beyond the scope of the course, but we mention in passing that there are principled ways of coming up with potentials in general online learning problems.

Proof of Proposition 3. We first prove the second statement, focusing on conditional density with the logarithmic loss; for the square loss, see Remark 6 below.
Proof for Part 2: Log loss. Recall that for each x, $f(x)$ is a distribution over $\mathcal{Y}$, and $\ell_{\log}(f(x), y) = -\log f(y|x)$ where we abuse the notation and write $f(x)$ and $f(\cdot|x)$ interchangeably. With $\eta = 1$, the averaged variant of exponential weights satisfies

$$\widehat{f}^t(y^t|x^t) = \sum_{f\in\mathcal{F}} q^t(f) f(y^t|x^t) = \sum_{f\in\mathcal{F}} f(y^t|x^t) \frac{\exp\left\{-\sum_{i=1}^{t-1} \ell_{\log}(f(x^i), y^i)\right\}}{\sum_{f\in\mathcal{F}} \exp\left\{-\sum_{i=1}^{t-1} \ell_{\log}(f(x^i), y^i)\right\}}, \tag{1.27}$$

and thus

$$\ell_{\log}(\widehat{f}(x^t), y^t) = -\log \widehat{f}^t(y^t|x^t) = \Phi_1^t - \Phi_1^{t-1}. \tag{1.28}$$

Hence, by telescoping

$$\sum_{t=1}^{T} \ell_{\log}(\widehat{f}(x^t), y^t) = \Phi_1^T - \Phi_1^0.$$

Finally, observe that $\Phi_1^0 = -\log|\mathcal{F}|$ and, since $-\log$ is monotonically decreasing, we have

$$\Phi_1^T \leq -\log\exp\left\{-\sum_{i=1}^{T}\ell_{\log}(f^\star(x^i), y^i)\right\} = \sum_{i=1}^{T}\ell_{\log}(f^\star(x^i), y^i), \tag{1.29}$$

for any $f^\star \in \mathcal{F}$. This establishes the result for conditional density estimation with the log loss. As already discussed, the above proof follows the strategy: the loss on each round related to change in potential (1.28), and the cumulative loss of any expert is related to the final potential (1.29). We now aim to replicate these steps for arbitrary bounded losses.

Proof for Part 1: Generic loss. To prove this result, we build on the log loss result above. First, observe that without loss of generality, we may assume that $\ell \circ f \in [0, 1]$ for all $f \in \mathcal{F}$ and (x, y), as we can always re-scale the problem. The randomized variant of exponential weights (1.24) satisfies

$$\mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] = \sum_{f\in\mathcal{F}} \ell(f(x^t), y^t)\frac{\exp\left\{-\eta\sum_{i=1}^{t-1}\ell(f(x^i), y^i)\right\}}{\sum_{f\in\mathcal{F}}\exp\left\{-\eta\sum_{i=1}^{t-1}\ell(f(x^i), y^i)\right\}}. \tag{1.30}$$

Hoeffding's inequality (1.13) implies that

$$\eta\,\mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] \leq -\log\sum_{f\in\mathcal{F}}\frac{\exp\{-\eta\ell(f(x^t), y^t)\}\exp\left\{-\eta\sum_{i=1}^{t-1}\ell(f(x^i), y^i)\right\}}{\sum_{f\in\mathcal{F}}\exp\left\{-\eta\sum_{i=1}^{t-1}\ell(f(x^i), y^i)\right\}} + \frac{\eta^2}{8}. \tag{1.31}$$

Note that the right-hand side of this inequality is simply

$$\Phi_\eta^t - \Phi_\eta^{t-1} + \frac{\eta^2}{8},$$

establishing the analogue of (1.28). Summing over t, this gives

$$\eta\sum_{t=1}^{T}\mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] \leq \Phi_\eta^T - \Phi_\eta^0 + \frac{T\eta^2}{8}. \tag{1.32}$$

As in the first part, for any $f^\star \in \mathcal{F}$, we can upper bound

$$\Phi_\eta^T \leq \eta\sum_{t=1}^{T}\ell(f^\star(x^t), y^t),$$

while $\Phi_\eta^0 = -\log|\mathcal{F}|$. Hence, we have that for any $f^\star \in \mathcal{F}$,

$$\sum_{t=1}^{T}\mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] - \ell(f^\star(x^t), y^t) \leq \frac{T\eta}{8} + \frac{\log|\mathcal{F}|}{\eta}.$$

With $\eta = \sqrt{\frac{8\log|\mathcal{F}|}{T}}$, we conclude that

$$\sum_{t=1}^{T} \mathbb{E}_{\widehat{f}^t\sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] - \ell(f^\star(x^t), y^t) \leq \sqrt{\frac{T\log|\mathcal{F}|}{2}}. \tag{1.33}$$

□

Observe that Hoeffding's inequality was all that was needed for Lemma 4. Curiously enough, it was also the only nontrivial step in the proof of Proposition 3. In fact, the connection between probabilistic inequalities and online learning regret inequalities (that hold for arbitrary sequences) runs much deeper.

Remark 5 (Beyond finite classes): As in statistical learning, there are (sequential) complexity measures for $\mathcal{F}$ that can be used to generalize the regret bounds in Proposition 3 to infinite classes. In general, the optimal regret for a class $\mathcal{F}$ will reflect the statistical capacity of the class [60].

Remark 6 (Mixable losses): We did not provide a proof of Proposition 3 for square loss. It is tempting to reduce square loss regression to density estimation by taking the conditional density to be a Gaussian distribution. Indeed, the log loss of a distribution with density proportional to $\exp\{-(\widehat{f}^t(x^t)-y^t)^2\}$ is, up to constants, the desired square loss. However, the mixture in (1.27) does not immediately lead to a prediction strategy for the square loss, as the expectation appears in the wrong location. This issue is fixed by a notion known as *mixability.*

We say that a loss ℓ is *mixable* with parameter η if there exists a constant $c > 0$ such that the following holds: for any x and a distribution $q \in \Delta(\mathcal{F})$, there exists a prediction $\widehat{f}(x) \in \mathcal{Y}'$ such that for all $y \in \mathcal{Y}$,

$$\ell(\widehat{f}(x), y) \leq -\frac{c}{\eta}\log\left(\sum_{f\in\mathcal{F}} q(f)\exp\{-\eta\ell(f(x), y)\}\right). \tag{1.34}$$

If loss is mixable, then given the exponential weights distribution q^t, the best prediction $\widehat{y}^t = \widehat{f}^t(x^t)$ can be written (by bringing the right-hand side of (1.34) to the left side) as an optimization problem

$$\arg\min_{\widehat{y}^t\in\mathcal{Y}'}\max_{y^t\in\mathcal{Y}}\left[\ell(\widehat{y}^t, y^t) + \frac{c}{\eta}\log\left(\sum_{f\in\mathcal{F}} q^t(f)\exp\{-\eta\ell(f(x^t), y^t)\}\right)\right] \tag{1.35}$$

which is equivalent to

$$\arg\min_{\widehat{y}^t\in\mathcal{Y}'}\max_{y^t\in\mathcal{Y}}\left[\ell(\widehat{y}^t, y^t) + \frac{c}{\eta}\log\left(\sum_{f\in\mathcal{F}} \exp\{-\eta\sum_{i=1}^{t}\ell(f(x^i), y^i)\}\right)\right] \tag{1.36}$$

once we remove the normalization factor. With this choice, mixability allows one to replicate the proof of Proposition 3 for the logarithmic loss, with the only difference being that (1.27) (after applying $-\log$ to both sides) becomes an inequality. It can be verified that square loss is mixable with parameter $\eta = 2$ and $c = 1$ when $\mathcal{Y} = \mathcal{Y}' = [0, 1]$, leading to the desired fast rate for square loss in Proposition 3. The idea of translating the English statement "there exists a strategy such that for any outcome..." into a min-max inequality will come up again in the course.

Remark 7 (Online linear optimization): For the slow rate in Proposition 3, the nature of the loss and the dependence on the function f is immaterial for the proof. The guarantee can be stated in a more abstract form that depends only on the vector of losses for functions in $\mathcal{F}$ as follows. Let $|\mathcal{F}| = N$. For timestep t, define $\boldsymbol{\ell}^t_f = \ell(f(x^t), y^t)$ and $\boldsymbol{\ell}^t = (\boldsymbol{\ell}^t_{f_1}, \ldots, \boldsymbol{\ell}^t_{f_N}) \in \mathbb{R}^N$ for $\mathcal{F} = \{f_1, \ldots, f_N\}$. For a randomized strategy $q^t \in \Delta([N])$, expected loss of the learner can be written as

$$\mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] = \langle q^t, \boldsymbol{\ell}^t \rangle ,$$

and the expected regret can be written as

$$\mathbf{Reg} = \sum_{t=1}^{T} \langle q^t, \boldsymbol{\ell}^t \rangle - \min_{j \in \{1, \ldots, N\}} \sum_{t=1}^{T} \langle \boldsymbol{e}_j, \boldsymbol{\ell}^t \rangle \tag{1.37}$$

where $\boldsymbol{e}_j \in \mathbb{R}^N$ is the standard basis vector with 1 in jth position. In its most general form, the exponential weights algorithm gives bounds on the regret in (1.37) for any sequence of vectors $\boldsymbol{\ell}^1, \ldots, \boldsymbol{\ell}^T$, and the update takes the form

$$q^t(k) \propto \exp\left\{ -\eta \sum_{i=1}^{t-1} \boldsymbol{\ell}^t(k) \right\} .$$

This formulation can be viewed as a special case of a problem known as *online linear optimization*, and the exponential weights method can be viewed as an instance of an algorithm known as *mirror descent*.

1.7 Exercises

Exercise 1 (Proposition 1, Part 2.): Consider the setting of Proposition 1, where $(x^1, y^1), \ldots, (x^T, y^T)$ are i.i.d., $\mathcal{F} = \{f : \mathcal{X} \to [0, 1]\}$ is finite, the true regression function satisfies $f^\star \in \mathcal{F}$, and $Y_i \in [0, 1]$ almost surely. Prove that empirical risk minimizer $\widehat{f}$ with respect to square loss satisfies the following bound on excess risk. With probability at least $1 - \delta$,

$$\mathcal{E}(\widehat{f}) \lesssim \frac{\log(|\mathcal{F}|/\delta)}{T}. \tag{1.38}$$

Follow these steps:

1. For a fixed function $f \in \mathcal{F}$, consider the random variable

$$Z_i(f) = (f(x^i) - y^i)^2 - (f^\star(x^i) - y^i)^2$$

for $i = 1, \dots, T$. Show that

$$\mathbb{E}[Z_i(f)] = \mathbb{E}(f(x^i) - f^\star(x^i))^2 = \mathcal{E}(f).$$

2. Show that for any fixed $f \in \mathcal{F}$, the variance $\mathbb{V}(Z_i(f))$ is bounded as

$$\mathbb{V}(Z_i(f)) \le 4\,\mathbb{E}(f(x^i) - f^\star(x^i))^2.$$

3. Apply Bernstein's inequality (Lemma 5) to show that with for any $f \in \mathcal{F}$, with probability at least $1 - \delta$,

$$\mathcal{E}(f) \le 2(\widehat{L}(f) - \widehat{L}(f^\star)) + \frac{C \log(1/\delta)}{T}, \tag{1.39}$$

for an absolute constant C, where $\widehat{L}(f) = \frac{1}{T}\sum_{t=1}^{T}(f(x^t) - y^t)^2$.

4. Extend this probabilistic inequality to simultaneously hold for all $f \in \mathcal{F}$ by taking the union bound over $f \in \mathcal{F}$. Conclude as a consequence that the bound holds for $\widehat{f}$, the empirical minimizer, implying (1.38).

Exercise 2 (ERM in Online Learning): Consider the problem of Online Supervised Learning with indicator loss $\ell(f(x), y) = \mathbb{I}\{f(x) \neq y\}$, $\mathcal{Y} = \mathcal{Y}' = \{0, 1\}$, and a finite class $\mathcal{F}$.

1. Exhibit a class $\mathcal{F}$ for which ERM cannot ensure sublinear growth of regret for all sequences, i.e. there exists a sequence $(x^1, y^1), \dots, (x^T, y^T)$ such that

$$\sum_{t=1}^{T} \ell(\widehat{f}^t(x_t), y_t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) = \Omega(T),$$

where $\widehat{f}^t$ is the empirical minimizer for the indicator loss on $(x^1, y^1), \dots, (x^1, y^1)$. Note: The construction must have $|\mathcal{F}| \le C$, where C is an absolute constant that does not depend on T.

2. Show that if data are i.i.d., then in expectation over the data, ERM attains a sublinear bound $O(\sqrt{T \log|\mathcal{F}|})$ on regret for any finite class $\mathcal{F}$.

Exercise 3 (Low Noise): 1. For a nonnegative random variable X, prove that for any $\eta \ge 0$,

$$\ln \mathbb{E} \exp\{-\eta(X - \mathbb{E}[X])\} \le \frac{\eta^2}{2}\,\mathbb{E}[X^2]. \tag{1.40}$$

Hint: use the fact that $\ln x \le x - 1$ and $\exp(-x) \le 1 - x + x^2/2$ for $x \ge 0$.

2. Consider the setting of Proposition 3, Part 1 (Generic Loss). Prove that the randomized variant of the Exponential Weights Algorithm satisfies, for any $f^\star \in \mathcal{F}$,

$$\sum_{t=1}^{T} \mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)] - \ell(f^\star(x^t), y^t) \le \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}_{\widehat{f}^t \sim q^t}[\ell(\widehat{f}^t(x^t), y^t)^2] + \frac{\log|\mathcal{F}|}{\eta}. \tag{1.41}$$

for any sequence of data. Hint: replace Hoeffding's Lemma by (1.40).

> 3. Suppose $\ell(f(x), y) \in [0,1]$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $f \in \mathcal{F}$. Suppose that there is a "perfect expert $f^\star \in \mathcal{F}$ such that $\ell(f^\star(x_t), y_t) = 0$ for all $t \in [T]$. Conclude that the above algorithm, with an appropriate choice of η, enjoys a bound of $O(\log|\mathcal{F}|)$ on the cumulative loss of the algorithm (equivalently, the fast rate $\frac{\log|\mathcal{F}|}{T}$ for the average regret). This setting is called "zero-noise" or "realizable."
>
> 4. Consider the binary classification problem with indicator loss, and suppose $\mathcal{F}$ contains a perfect expert, as above. The *Halving Algorithm* maintains a version space $\mathcal{F}_t = \{f \in \mathcal{F} : f(x_s) = y_s, s < t\}$ and, given x_t, follows the majority vote of remaining experts in $\mathcal{F}_t$. Show that this algorithm incurs cumulative loss at most $O(\log|\mathcal{F}|)$. Hence, the Exponential Weights Algorithm can be viewed as an extension of the Halving algorithm to settings where the optimal loss is non-zero.

2. MULTI-ARMED BANDITS

This chapter introduces the *multi-armed bandit* problem, which is the simplest interactive decision making framework we will consider in this course.

```
Multi-Armed Bandit Protocol
for $t = 1, \ldots, T$ do
    Select decision $\pi^t \in \Pi := \{1, \ldots, A\}$
    Observe reward $r^t$
```

The protocol (see above) proceeds in T rounds. At each round $t \in [T]$, the learning agent selects a discrete *decision*[2] $\pi^t \in \Pi = \{1, \ldots, A\}$ using the data

$$\mathcal{H}^{t-1} = \{(\pi^1, r^1), \ldots, (\pi^{t-1}, r^{t-1})\}$$

collected so far; we refer to Π as the *decision space* or action space, with $A \in \mathbb{N}$ denoting the size of the space. Based on the decision π^t, the learner receives a reward r^t, and their goal is to maximize the cumulative reward across all T rounds. As an example, one might consider an application in which the learner is a doctor (or personalized medical assistant) who aims to select a treatment (the decision) in order to make a patient feel better (maximize reward); see Figure 4.

The multi-armed bandit problem can be studied in a stochastic framework, in which rewards are generated from a fixed (conditional) distribution, or an non-stochastic/adversarial framework in the vein of online learning (Section 1.6). We will focus on the stochastic framework, and make the following assumption.

Assumption 1 (Stochastic Rewards): Rewards are generated independently via

$$r^t \sim M^\star(\cdot \mid \pi^t), \tag{2.1}$$

where $M^\star(\cdot \mid \cdot)$ is the underlying *model* (conditional distribution).

We define

$$f^\star(\pi) := \mathbb{E}\left[r \mid \pi\right] \tag{2.2}$$

[2] In the literature on bandits, decisions are often referred to as *actions*. We will use these terms interchangeably throughout this section.

as the mean reward function under $r \sim M^\star(\cdot \mid \pi)$. We measure the learner's performance via regret to the action $\pi^\star := \arg\max_{\pi \in \Pi} f^\star(\pi)$ with highest reward:

$$\mathbf{Reg} := \sum_{t=1}^{T} f^\star(\pi^\star) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} f^\star(\pi^t). \tag{2.3}$$

Regret is a natural notion of performance for the multi-armed bandit problem because it is *cumulative*: it measures not just how well the learner can identify an action with good reward, but how well it can maximize reward as it goes. This notion is well-suited to settings like the personalized medicine example in Figure 4, where regret captures the *overall* quality of treatments, not just the quality of the final treatment. As in the online learning framework, we would like to develop algorithms that enjoy sublinear regret, i.e.

$$\frac{\mathbb{E}[\mathbf{Reg}]}{T} \to 0 \quad \text{as} \quad T \to \infty.$$

The most important feature of the multi-armed bandit problem, and what makes the problem fundamentally *interactive*, is that the learner only receives a reward signal for the single decision $\pi^t \in \Pi$ they select at each round. That is, the observed reward r^t gives a noisy estimate for $f^\star(\pi^t)$, but reveals no information about the rewards for other decisions $\pi \neq \pi^t$. For example in Figure 4, if the doctor prescribes a particular treatment to the patient, they can observe whether the patient responds favorably, but they do not directly observe whether other possible treatments might have led to an even better outcome. This issue is often referred to as *partial feedback* or *bandit feedback*. Partial feedback introduces an element of *active data collection*, as it means that the information contained in the dataset $\mathcal{H}^t$ depends on the decisions made by the learner, which we will see necessitates *exploring* different actions. This should be contrasted with statistical learning (where the dataset is generated independently from the learner) and online learning (where losses may be chosen by nature in response to the learner's behavior, but where the outcome y^t— and hence the full loss function $\ell(\cdot, y^t)$—is always revealed).

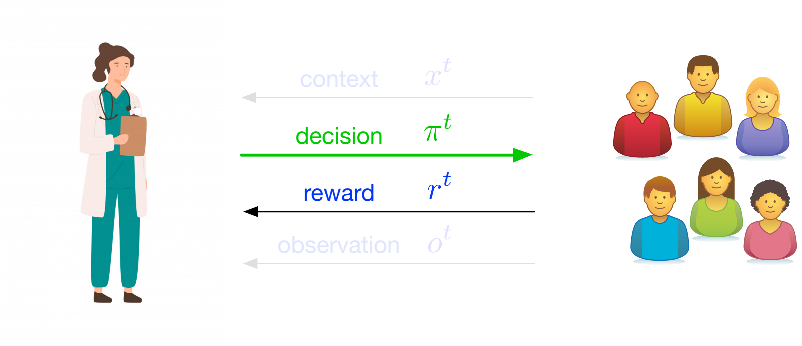


Figure 4: An illustration of the multi-armed bandit problem. A doctor (the learner) aims to select a treatment (the decision) to improve a patient's vital signs (the reward).

In the context of Figure 2, the multi-armed bandit problem constitutes our first step along the "interactivity" axis, but does not incorporate any structure in the decision space (and does not involve features/contexts/covariates). In particular, information about one action does not reveal information about any other actions, so there is no hope of using function approximation to generalize across actions.[3] As a result, the algorithms we will cover in this section will have regret that scales with $\Omega(|\Pi|) = \Omega(A)$. This shortcoming is addressed by the *structured bandit* framework we will introduce in Section 4, which allows for the use of function approximation to model structure in the decision space.[4]

Remark 8 (Other notions of regret): It is also reasonable to consider empirical regret, defined as

$$\max_{\pi \in \Pi} \sum_{t=1}^{T} r^t(\pi) - \sum_{t=1}^{T} r^t(\pi^t), \tag{2.4}$$

where, for $\pi \neq \pi^t$, $r^t(\pi)$ denotes the *counterfactual reward* the learner would have received if they had played π at round t. Using Hoeffding's inequality, one can show that this is equivalent to the definition in (2.3) up to $O(\sqrt{T})$ factors.

2.1 The Need for Exploration

In statistical learning, we saw that the empirical risk minimization algorithm, which greedily chooses the function that best fits the data, leads to interesting bounds on excess risk. For multi-armed bandits, since we assume the data generating process is stochastic, a natural first attempt at designing an algorithm is to apply the greedy principle here in the same fashion. Concretely, at time t, we can compute an empirical estimate for the reward function $f^\star$ via

$$\widehat{f}^t(\pi) = \frac{1}{n^t(\pi)} \sum_{s<t} r^s \mathbb{I}\left\{\pi^s = \pi\right\}, \tag{2.5}$$

where $n^t(\pi)$ is the number of times π has been selected up to time t.[5] Then, we can choose the greedy action

$$\pi^t = \arg\max_{\pi \in \Pi} \widehat{f}^t(\pi).$$

Unfortunately, due to the interactive nature of the bandit problem, this strategy can fail, leading to linear regret ($\mathbf{Reg} = \Omega(T)$). Consider the following problem with $\Pi = \{1, 2\}$ ($A = 2$).

- Decision 1 has reward $\frac{1}{2}$ almost surely.
- Decision 2 has reward $\mathrm{Ber}(3/4)$.

Suppose we initialize by playing each decision a single time to ensure that $n^t(\pi) > 0$, then follow the greedy strategy. One can see that with probability $1/4$, the greedy algorithm will get stuck on action 1, leading to regret $\Omega(T)$.

The issue in this example is that the greedy algorithm immediately gives up on the optimal action and never revisits it. To address this, we will consider algorithms that

[3] Another way to say this is that we take $\mathcal{F} = \mathbb{R}^A$, so that $f^\star \in \mathcal{F}$.

[4] Throughout the lecture notes, we will exclusively use the term "multi-armed bandit" to refer to bandit problems with finite action spaces, and use the term "structured bandit" for problems with large action spaces.

[5] If $n^t(\pi) = 0$, we will set $\widehat{f}^t(\pi) = 0$.

deliberately *explore* less visited actions to ensure that their estimated rewards are not misleading.

2.2 The ε-Greedy Algorithm

The greedy algorithm for bandits can fail because it can insufficiently explore good decisions that initially seem bad, leading it to get stuck playing suboptimal decisions. In light of this failure, a reasonable solution is to manually force the algorithm to explore, so as to ensure that this situation never occurs. This leads us to what is known as the ε-*Greedy* algorithm.

Let $\varepsilon \in [0,1]$ be the *exploration parameter*. At each time $t \in [T]$, the ε-Greedy algorithm computes the estimated reward function $\widehat{f}^t$ as in (2.5). With probability $1-\varepsilon$, the algorithm chooses the greedy decision

$$\widehat{\pi}^t = \arg\max_{\pi} \widehat{f}^t(\pi), \tag{2.6}$$

and with probability ε it samples a uniform random action $\pi^t \sim \mathrm{unif}(\{1,\ldots,A\})$. As the name suggests, ε-Greedy usually plays the greedy action (*exploiting* what it has already learned), but the uniform sampling ensures that the algorithm will also *explore* unseen actions. We can think of the parameter ε as modulating the tradeoff between exploiting and exploring.

Proposition 4: Assume that $f^\star(\pi) \in [0,1]$ and r^t is 1-sub-Gaussian. Then for any T, by choosing ε appropriately, the ε-Greedy algorithm ensures that with probability at least $1-\delta$,

$$\mathbb{E}[\mathbf{Reg}] \lesssim A^{1/3}T^{2/3} \cdot \log^{1/3}(AT/\delta).$$

This regret bound has $\frac{\mathbb{E}[\mathbf{Reg}]}{T} \to 0$ with $T \to \infty$ as desired, though we will see in the sequel that more sophisticated strategies can attain improved regret bounds that scale with $\sqrt{AT}$.[6]

Proof of Proposition 4. Recall that $\widehat{\pi}^t := \arg\max_\pi \widehat{f}^t(\pi)$ denotes the greedy action at round t, and that p^t denotes the distribution over π^t. We can decompose the regret into two terms, representing the contribution from choosing the greedy action and the contribution from exploring uniformly:

$$\begin{aligned}
\mathbf{Reg} &= \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t}[f^\star(\pi^\star) - f^\star(\pi^t)] \\
&= (1-\varepsilon)\sum_{t=1}^T f^\star(\pi^\star) - f^\star(\widehat{\pi}^t) + \varepsilon \sum_{t=1}^T \mathbb{E}_{\pi^t \sim \mathrm{unif}([A])}[f^\star(\pi^\star) - f^\star(\widehat{\pi}^t)] \\
&\leq \sum_{t=1}^T f^\star(\pi^\star) - f^\star(\widehat{\pi}^t) + \varepsilon T.
\end{aligned}$$

In the last inequality, we have simply written off the contribution from exploring uniformly by using that $f^\star(\pi) \in [0,1]$. It remains to bound the regret we incur from playing the

[6] Note that $\sqrt{AT} \leq A^{1/3}T^{2/3}$ whenever $A \leq T$, and when $A \geq T$ both guarantees are vacuous.

greedy action. Here, we bound the per-step regret in terms of estimation error using a similar decomposition to Lemma 4 (note that we are now working with rewards rather than losses):

$$f^\star(\pi^\star) - f^\star(\widehat{\pi}^t) = [f^\star(\pi^\star) - \widehat{f}^t(\pi^\star)] + \underbrace{[\widehat{f}^t(\pi^\star) - \widehat{f}^t(\widehat{\pi}^t)]}_{\leq 0} + [\widehat{f}^t(\widehat{\pi}^t) - f^\star(\widehat{\pi}^t)] \tag{2.7}$$

$$\leq 2 \max_{\pi \in \{\pi^\star, \widehat{\pi}^t\}} |f^\star(\pi) - \widehat{f}^t(\pi)| \leq 2 \max_{\pi} |f^\star(\pi) - \widehat{f}^t(\pi)|. \tag{2.8}$$

Note that this regret decomposition can also be applied to the pure greedy algorithm, which we have already shown can fail. The reason why ε-Greedy succeeds, which we use in the argument that follows, is that because we explore, the "effective" number of times that each arm will be pulled prior to round t is of the order $\varepsilon t/A$, which will ensure that the sample mean converges to $f^\star$. In particular, we will show that the event

$$\mathcal{E}_t = \left\{ \max_{\pi} |f^\star(\pi) - \widehat{f}^t(\pi)| \lesssim \sqrt{\frac{A \log(AT/\delta)}{\varepsilon t}} \right\} \tag{2.9}$$

occurs for all t with probability at least $1 - \delta$.

To prove that (2.9) holds, we first use Hoeffding's inequality for adaptive stopping times (Lemma 35), which gives that for any fixed π, with probability at least $1 - \delta$ over the draw of rewards,

$$|f^\star(\pi) - \widehat{f}^t(\pi)| \leq \sqrt{\frac{2 \log(2T/\delta)}{n^t(\pi)}}. \tag{2.10}$$

From here, taking a union bound over all $t \in [T]$ and $\pi \in \Pi$ ensures that

$$|f^\star(\pi) - \widehat{f}^t(\pi)| \leq \sqrt{\frac{2 \log(2AT^2/\delta)}{n^t(\pi)}} \tag{2.11}$$

for all π and t simultaneously. It remains to show that the number of pulls $n^t(\pi)$ is sufficiently large.

Let $e^t \in \{0, 1\}$ be a random variable whose value indicates whether the algorithm explored uniformly at step t, and let $m^t(\pi) = |\{i < t : \pi^i = \pi, e^i = 1\}|$, which has $n^t(\pi) \geq m^t(\pi)$. Let $Z^t = \mathbb{I}\{\pi^t = \pi, e^t = 1\}$. Observe that we can write

$$m^t(\pi) = \sum_{i<t} Z^i.$$

In addition, $Z^t \sim \mathrm{Ber}(\varepsilon/A)$, so we have $\mathbb{E}[m^t(\pi)] = \varepsilon(t-1)/A$. Using Bernstein's inequality (Lemma 5) with $Z^1, \ldots, Z^{t-1}$, we have that for any fixed π and all $u > 0$, with probability at least $1 - 2e^{-u}$,

$$\left| m^t(\pi) - \frac{\varepsilon(t-1)}{A} \right| \leq \sqrt{2\mathbb{V}[Z](t-1)u} + \frac{u}{3} \leq \sqrt{\frac{2\varepsilon(t-1)u}{A}} + \frac{u}{3} \leq \frac{\varepsilon(t-1)}{2A} + \frac{4u}{3},$$

where we have used that $\mathbb{V}[Z] = \varepsilon/A \cdot (1 - \varepsilon/A) \leq \varepsilon/A$, and then applied the arithmetic mean-geometric mean (AM-GM) inequality, which states that $\sqrt{xy} \leq \frac{x}{2} + \frac{y}{2}$ for $x, y \geq 0$. Rearranging, this gives

$$m^t(\pi) \geq \frac{\varepsilon(t-1)}{2A} - \frac{4u}{3}. \tag{2.12}$$

Setting $u = \log(2AT/\delta)$ and taking a union bound, we are guaranteed that with probability at least $1 - \delta$, for all $\pi \in \Pi$ and $t \in [T]$

$$m^t(\pi) \geq \frac{\varepsilon(t-1)}{2A} - \frac{4\log(2AT/\delta)}{3}. \tag{2.13}$$

As long as $\varepsilon t \gtrsim A\log(AT/\delta)$ (we can write off the rounds where this does not hold), this yields

$$n^t(\pi) \geq m^t(\pi) \gtrsim \frac{\varepsilon t}{A}.$$

Taking a union bound and combining with (2.11), this implies that with probability at least $1 - \delta$, for all t,

$$\max_{\pi} |f^\star(\pi) - \widehat{f}^t(\pi)| \lesssim \sqrt{\frac{A\log(AT/\delta)}{\varepsilon t}}.$$

which leads to the overall regret bound

$$\begin{aligned}\mathbf{Reg} \leq \sum_{t=1}^{T} \max_{\pi} |f^\star(\pi) - \widehat{f}^t(\pi)| + \varepsilon T \lesssim \sum_{t=1}^{T} \sqrt{\frac{A\log(AT/\delta)}{\varepsilon t}} + \varepsilon T \\ \leq \sqrt{\frac{AT\log(AT/\delta)}{\varepsilon}} + \varepsilon T. \end{aligned}\tag{2.14}$$

To balance the terms on the right-hand side, we set

$$\varepsilon \propto \left(\frac{A\log(AT/\delta)}{T}\right)^{1/3},$$

which gives the final result. □

This proof shows that the ε-Greedy strategy allows the learner to acquire information uniformly for all actions, but we pay for this in terms of regret (specifically, through the εT factor in the final regret bound (2.14)). This issue here is that the ε-Greedy strategy continually explores all actions, even though we might expect to rule out actions with very low reward after a relatively small amount of exploration. To address this shortcoming, we will consider more adaptive strategies.

Remark 9 (Explore-then-commit): A relative of ε-Greedy is the explore-then-commit (ETC) algorithm, which uniformly explores actions for the first N rounds, then estimates rewards based on the data collected and commits to the greedy action for the remaining $T - N$ rounds. This strategy can be shown to attain $\mathbf{Reg} \lesssim A^{1/3}T^{2/3}$ for an appropriate choice of N, matching ε-Greedy.

2.3 The Upper Confidence Bound (UCB) Algorithm

The next algorithm we will study for bandits is the Upper Confidence Bound (UCB) algorithm. The UCB algorithm attains a regret bound of the order $\widetilde{O}(\sqrt{AT})$, which improves upon the regret bound for ε-Greedy, and is optimal (in a worst-case sense) up to logarithmic factors. In addition to optimality, the algorithm offers several secondary benefits, including adaptivity to favorable structure in the underlying reward function.

The UCB algorithm is based on the notion of *optimism in the face of uncertainty*, which is a general principle we will revisit throughout this text in increasingly rich settings. The idea behind the principle is that at each time t, we should adopt the most optimistic perspective of the world possible given the data collected so far, and then choose the decision π^t based on this perspective.

To apply the idea of optimism to the multi-armed bandit problem, suppose that for each step t, we can construct "confidence intervals"

$$\underline{f}^t, \bar{f}^t : \Pi \to \mathbb{R}, \tag{2.15}$$

with the following property: with probability at least $1 - \delta$,

$$\forall t \in [T], \pi \in \Pi, \quad f^*(\pi) \in [\underline{f}^t(\pi), \bar{f}^t(\pi)]. \tag{2.16}$$

We refer to $\underline{f}^t$ as a *lower confidence bound* and $\bar{f}^t$ as a *upper confidence bound*, since we are guaranteed that with high probability, they lower (resp. upper) bound $f^\star$. Given confidence intervals, the UCB algorithm simply chooses π^t as the "optimistic" action that maximizes the upper confidence bound:

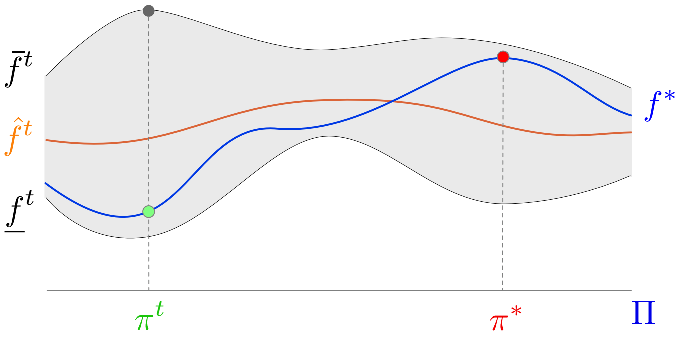


Figure 5: Illustration of the UCB algorithm. Selecting the action π^t optimistically ensures that the suboptimality never greater exceeds the confidence width.

$$\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi).$$

The following lemma shows that the instantaneous regret for this strategy is bounded by the width of the confidence interval; see Figure 5 for an illustration.

Lemma 7: Fix t, and suppose that $f^\star(\pi) \in [\underline{f}^t(\pi), \bar{f}^t(\pi)]$ for all π. Then the optimistic action

$$\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$$

has

$$f^\star(\pi^\star) - f^\star(\pi^t) \leq \bar{f}^t(\pi^t) - f^\star(\pi^t) \leq \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t). \tag{2.17}$$

Proof of Lemma 7. The result follows immediate from the observation that for any $t \in [T]$ and any $\pi^\star \in \Pi$, we have

$$f^\star(\pi^\star) \leq \bar{f}^t(\pi^\star) \leq \bar{f}^t(\pi^t) \quad \text{and} \quad -f^\star(\pi^t) \leq -\underline{f}^t(\pi^t).$$

□

Lemma 7 implies that as long as we can build confidence intervals for which the width $\bar{f}^t(\pi^t) - \underline{f}^t(\pi^t)$ shrinks, the regret for the UCB strategy will be small. To construct such intervals, here we appeal to Hoeffding's inequality for adaptive stopping times (Lemma 35).[7] As long as $r^t \in [0,1]$, a union bound gives that with probability at least $1-\delta$, for all $t \in [T]$ and $\pi \in \Pi$,

$$|\widehat{f}^t(\pi) - f^\star(\pi)| \leq \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}}, \tag{2.18}$$

where we recall that $\widehat{f}^t$ is the sample mean and $n^t(\pi) := \sum_{i<t} \mathbb{I}\{\pi^i = \pi\}$. This suggests that by choosing

$$\bar{f}^t(\pi) = \widehat{f}^t(\pi) + \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}}, \quad \text{and} \quad \underline{f}^t(\pi) = \widehat{f}^t(\pi) - \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}}, \tag{2.19}$$

we obtain a valid confidence interval. With this choice—along with Lemma 7—we are in a favorable position, because for a given round t, one of two things must happen:

- The optimistic action has high reward, so the instantaneous regret is small.
- The instantaneous regret is large, which by Lemma 7 implies that confidence width is large as well (and $n^t(\pi^t)$ is small). This can only happen a small number of times, since $n^t(\pi^t)$ will increase as a result, causing the width to shrink.

Using this idea, we can prove the following regret bound.

Proposition 5: Using the confidence bounds in (2.19), the UCB algorithm ensures that with probability at least $1-\delta$,

$$\mathbf{Reg} \lesssim \sqrt{AT\log(AT/\delta)}.$$

This result is optimal up to the $\log(AT)$ factor, which can be removed by using the same algorithm with a slightly more sophisticated confidence interval construction [9]. Note that compared to the statistical learning and online learning setting, where we were able to attain regret bounds that scaled logarithmically with the size of the benchmark class, here the optimal regret scales *linearly* with $|\Pi| = A$. This is the price we pay for partial/bandit feedback, and reflects that fact that we must explore all actions to learn.

Proof of Proposition 5. Let us condition on the event in (2.18). Whenever this occurs, we have that $f^\star(\pi) \in \left[\underline{f}^t(\pi), \bar{f}^t(\pi)\right]$ for all $t \in [T]$ and $\pi \in \Pi$, so the confidence intervals are valid. As a result, Lemma 7 bounds regret in terms of the confidence width:

$$\sum_{t=1}^T f^\star(\pi^\star) - f^\star(\pi^t) \leq \sum_{t=1}^T \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t) = \sum_{t=1}^T 2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}} \wedge 1; \tag{2.20}$$

here, the "$\wedge 1$" term appears because we can write off the regret for early rounds where $n^t(\pi^t) = 0$ as 1.

[7] While asymptotic confidence intervals in classical statistics arise from limit theorems, we are interested in valid *non-asymptotic* intervals, and thus appeal to concentration inequalities.

To bound the right-hand side, we use a potential argument. The basic idea is that at every round, $n^t(\pi)$ must increase for some action π, and since there are only A actions, this means that $1/\sqrt{n^t(\pi^t)}$ can only be large for a small number of rounds. This can be thought of as a quantitative instance of the pigeonhole principle.

Lemma 8 (Confidence width potential lemma): We have

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \lesssim \sqrt{AT}.$$

Proof of Lemma 8. We begin by writing.

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 = \sum_{\pi}\sum_{t=1}^{T} \frac{\mathbb{I}\{\pi^t = \pi\}}{\sqrt{n^t(\pi)}} \wedge 1 = \sum_{\pi}\sum_{t=1}^{n^{T+1}(\pi)} \frac{1}{\sqrt{t-1}} \wedge 1. \tag{2.21}$$

For any $n \in \mathbb{N}$, we have $\sum_{t=1}^{n} \frac{1}{\sqrt{t-1}} \wedge 1 \leq 1 + 2\sqrt{n}$, which allows us to bound by

$$A + 2\sum_{\pi} \sqrt{n^T(\pi)}.$$

The factor of A above is a lower-order term (recall that we have $A \leq \sqrt{AT}$ whenever $A \leq T$, and if $A > T$ the regret bound we are proving is vacuous). To bound the second term, using Jensen's inequality, we have

$$\sum_{\pi} \sqrt{n^T(\pi)} \leq A\sqrt{\sum_{\pi} \frac{n^T(\pi)}{A}} = A\sqrt{T/A} = \sqrt{AT}.$$

□

The main regret bound now follows from Lemma 8 and (2.20).

□

To summarize, the key steps in the proof of Proposition 5 were to:

1. Use the optimistic property and validity of the confidence bounds to bound regret by the sum of confidence widths.
2. Use a potential argument to show that the sum of confidence widths is small.

We will revisit and generalize both ideas in subsequent chapters for more sophisticated settings, including contextual bandits, structured bandits, and reinforcement learning.

Remark 10 (Instance-dependent regret for UCB): The $\widetilde{O}(\sqrt{AT})$ regret bound attained by UCB holds uniformly for all models, and is (nearly) minimax-optimal, in the sense that for any algorithm, there exists a model $M^\star$ for which the regret must scale as $\Omega(\sqrt{AT})$. Minimax optimality is a useful notion of performance, but may be overly pessimistic. As an alternative, it is possible to show that the UCB attains what is known as an *instance-dependent* regret bound, which adapts to the underlying reward function, and can be smaller for "nice" problem instances.

Let $\Delta(\pi) := f^\star(\pi^\star) - f^\star(\pi)$ be the *suboptimality gap* for decision π. Then, when $f^\star(\pi) \in [0,1]$, UCB can be shown to achieve

$$\mathbf{Reg} \lesssim \sum_{\pi : \Delta(\pi) > 0} \frac{\log(AT/\delta)}{\Delta(\pi)}.$$

If we keep the underlying model fixed and take $T \to \infty$, this regret bound scales only *logarithmically* in T, which improves upon the $\sqrt{T}$-scaling of the minimax regret bound.

2.4 Bayesian Bandits and the Posterior Sampling Algorithm*

Up to this point, we have been designing and analyzing algorithms from a *frequentist* viewpoint, in which we aim to minimize regret for a *worst-case* choice of the underlying model $M^\star$. An alterative is to adopt a *Bayesian* viewpoint, and assume that the underlying model is drawn from a known *prior* $\mu \in \Delta(\mathcal{M})$.[8] In this case, rather than worst-case performance, we will be concerned with average regret under the prior, defined via

$$\mathbf{Reg}_{\mathsf{Bayes}}(\mu) := \mathbb{E}_{M^\star \sim \mu}\, \mathbb{E}^{M^\star}[\mathbf{Reg}],$$

where $\mathbb{E}^{M^\star}[\cdot]$ denotes the algorithm's expected regret when $M^\star$ is the underlying reward distribution.

Working in the Bayesian setting opens up additional avenues for designing algorithms, because we can take advantage of our knowledge of the prior to compute quantities of interest that are not available in the frequentist setting, such as posterior distribution over $\pi^\star$ after observing the dataset $\mathcal{H}^{t-1}$. The most basic and well-known strategy here is *posterior sampling* (also known as Thompson sampling or probability matching) [70].

```
Posterior Sampling
for $t = 1, \ldots, T$ do
    Set $p^t(\pi) = \mathbb{P}(\pi^\star = \pi \mid \mathcal{H}^{t-1})$, where $\mathcal{H}^{t-1} = (\pi^1, r^1), \ldots, (\pi^{t-1}, r^{t-1})$.
    Sample $\pi^t \sim p^t$ and observe $r^t$.
```

The basic idea is as follows. At each time t, we can use our knowledge of the prior to compute the distribution $\mathbb{P}(\pi^\star = \cdot \mid \mathcal{H}^{t-1})$, which represents the posterior distribution over $\pi^\star$ given all of the data we have collected from rounds $1, \ldots, t-1$. The posterior sampling algorithm simply samples the learner's action π^t from this distribution, thereby "matching" the posterior distribution of $\pi^\star$.

Proposition 6: For any prior μ, the posterior sampling algorithm ensures that

$$\mathbf{Reg}_{\mathsf{Bayes}}(\mu) \leq \sqrt{AT\log(A)}. \tag{2.22}$$

In what follows, we prove a simplified version of Proposition 6; the full proof is given in Section 2.6.

Proof of Proposition 6 (simplified version). We will make the following simplified assumptions:

[8] It is important that μ is known, otherwise this is no different from the frequentist setting.

- We restrict to reward distributions where $M^\star(\cdot \mid \pi) = \mathcal{N}(f^\star(\pi), 1)$. That is, $f^\star$ is the only part of the reward distribution that is unknown.

- $f^\star$ belongs to a known class $\mathcal{F}$, and rather than proving the regret bound in Proposition 6, we will prove a bound of the form

  $$\mathbf{Reg}_{\mathsf{Bayes}}(\mu) \lesssim \sqrt{AT\log|\mathcal{F}|},$$

  which replaces the $\log A$ factor in the proposition with $\log|\mathcal{F}|$.

Since the mean reward function $f^\star$ is the only part of the reward distribution $M^\star$ that is unknown, we can simplify by considering an equivalent formulation where the prior has the form $\mu \in \Delta(\mathcal{F})$. That is, we have a prior over $f^\star$ rather than $M^\star$.

Before proceeding, let us introduce some notation. The process through which we sample $f^\star \sim \mu$ and the run the bandit algorithm induces a joint law over $(f^\star, \mathcal{H}^T)$, which we call $\mathbb{P}$. Throughout the proof, we use $\mathbb{E}[\cdot]$ to denote the expectation under this law. We also define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}^t]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid \mathcal{H}^t]$.

We begin by using the law of total expectation to express the expected regret as

$$\mathbf{Reg}_{\mathsf{Bayes}}(\mu) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{t-1}[f^\star(\pi_{f^\star}) - f^\star(\pi^t)]\right].$$

Above, we have written $\pi^\star = \pi_{f^\star}$ to make explicit the fact that this is a random variable whose value is a function of $f^\star$.

We first simplify the expected regret for each step t. Let $\mu^t(f) := \mathbb{P}(f^\star = f \mid \mathcal{H}^{t-1})$ be the posterior distribution at timestep t. The learner's decision π^t is *conditionally independent* of $f^\star$ given $\mathcal{H}^{t-1}$, so we can write

$$\mathbb{E}_{t-1}[f^\star(\pi_{f^\star}) - f^\star(\pi^t)] = \mathbb{E}_{f^\star\sim\mu^t, \pi^t\sim p^t}[f^\star(\pi_{f^\star}) - f^\star(\pi^t)].$$

If we define $\bar{f}^t(\pi) = \mathbb{E}_{f^\star\sim\mu^t}[f^\star(\pi)]$ as the expected reward function under the posterior, we can further write this as

$$\mathbb{E}_{f^\star\sim\mu^t, \pi^t\sim p^t}\left[f^\star(\pi_{f^\star}) - \bar{f}^t(\pi^t)\right].$$

By the design of the posterior sampling algorithm, $\pi^t \sim p^t$ is identical in distribution to $\pi_{f^\star}$ under $f^\star \sim \mu^t$, so this is equal to

$$\mathbb{E}_{f^\star\sim\mu^t}\left[f^\star(\pi_{f^\star}) - \bar{f}^t(\pi_{f^\star})\right].$$

This quantity captures—on average—how far a given realization of $f^\star$ deviates from the posterior mean $\bar{f}^t$, for a specific decision $\pi_{f^\star}$ which is coupled to $f^\star$. The expression above might appear to be unrelated to the learner's decision distribution, but the next lemma shows that it is possible to relate this quantity back to the learner's decision distribution using a notion of *information gain* (or, estimation error).

Lemma 9 (Decoupling): We have

$$\mathbb{E}_{f^\star\sim\mu^t}\left[f^\star(\pi_{f^\star}) - \bar{f}^t(\pi_{f^\star})\right] \le \sqrt{A \cdot \mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\left[(f^\star(\pi^t) - \bar{f}^t(\pi^t))^2\right]}. \tag{2.23}$$

Proof of Lemma 9. We will show a more general result. Namely, for any $\nu \in \Delta(\mathcal{F})$ and $\bar{f} : \Pi \to \mathbb{R}$, if we define $p(\pi) = \mathbb{P}_{f\sim\nu}(\pi_f = \pi)$, then

$$\mathbb{E}_{f\sim\nu}\big[f(\pi_f) - \bar{f}(\pi_f)\big] \leq \sqrt{A \cdot \mathbb{E}_{f\sim\nu}\,\mathbb{E}_{\pi\sim p}\big[(f(\pi) - \bar{f}(\pi))^2\big]}. \tag{2.24}$$

This can be thought of as a "decoupling" lemma. On the left-hand side, the random variables f and π_f are coupled, but on the right-hand side, π is drawn from the *marginal distribution* over π_f, independent of the draw of f itself.

To prove the result, we use Cauchy-Schwarz as follows:

$$\begin{aligned}\mathbb{E}_{f\sim\nu}\big[f(\pi_f) - \bar{f}(\pi_f)\big] &= \mathbb{E}_{f\sim\nu}\left[\frac{p^{1/2}(\pi_f)}{p^{1/2}(\pi_f)}\big(f(\pi_f) - \bar{f}(\pi_f)\big)\right] \\ &\leq \left(\mathbb{E}_{f\sim\nu}\left[\frac{1}{p(\pi_f)}\right]\right)^{1/2} \cdot \left(\mathbb{E}_{f\sim\nu}\Big[p(\pi_f)\big(f(\pi_f) - \bar{f}(\pi_f)\big)^2\Big]\right)^{1/2}.\end{aligned}$$

For the first term, we have

$$\mathbb{E}_{f\sim\nu}\left[\frac{1}{p(\pi_f)}\right] = \sum_f \frac{\nu(f)}{p(\pi_f)} = \sum_\pi \sum_{f:\pi_f=\pi} \frac{\nu(f)}{p(\pi)} = \sum_\pi \frac{p(\pi)}{p(\pi)} = A.$$

For the second term, we have

$$\mathbb{E}_{f\sim\nu}\Big[p(\pi_f)\big(f(\pi_f) - \bar{f}(\pi_f)\big)^2\Big] \leq \mathbb{E}_{f\sim\nu}\left[\sum_\pi p(\pi)\big(f(\pi) - \bar{f}(\pi)\big)^2\right] = \mathbb{E}_{f\sim\nu}\,\mathbb{E}_{\pi\sim p}\big[(f(\pi) - \bar{f}(\pi))^2\big].$$

Putting these bounds together yields (2.24). □

Using Lemma 9, we have that

$$\begin{aligned}\mathbb{E}[\mathbf{Reg}] &\leq \mathbb{E}\left[\sum_{t=1}^T \sqrt{A \cdot \mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[(f^\star(\pi^t) - \bar{f}^t(\pi^t))^2\big]}\right] \\ &\leq \sqrt{AT \cdot \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[(f^\star(\pi^t) - \bar{f}^t(\pi^t))^2\big]\right]}.\end{aligned}$$

To finish up we will show that $\sum_{t=1}^T \mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[(f^\star(\pi^t) - \bar{f}^t(\pi^t))^2\big] \leq \log|\mathcal{F}|$. To do this, we need some additional information-theoretic tools.

- For a random variable X with distribution $\mathbb{P}$, $\mathsf{Ent}(X) \equiv \mathsf{Ent}(\mathbb{P}) := \sum_x p(x)\log(1/p(x))$.
- For random variables X and Y, $\mathsf{Ent}(X \mid Y = y) := \mathsf{Ent}(\mathbb{P}_{X|Y=y})$ and $\mathsf{Ent}(X \mid Y) := \mathbb{E}_{y\sim p_Y}[\mathsf{Ent}(X \mid Y = y)]$.
- For distributions $\mathbb{P}$ and $\mathbb{Q}$, $D_{\mathsf{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) = \sum_x p(x)\log(p(x)/q(x))$.

To keep notation as clear as possible going forward, let us use boldface script ($\boldsymbol{\pi}^t$, $\boldsymbol{\pi}^\star$, $\boldsymbol{f}^\star$, $\boldsymbol{\mathcal{H}}^t$) to refer to the abstract random variables under consideration, and use non-boldface script (π^t, $\pi^\star$, $f^\star$, $\mathcal{H}^t$) to refer to their realizations. Our aim will be to use the conditional entropy $\mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^t)$ as a potential function, and show that for each t,

$$\frac{1}{2}\,\mathbb{E}\big[\mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[(f^\star(\pi^t)-\bar{f}^t(\pi^t))^2\big]\big] = \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^{t-1}) - \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^t). \tag{2.25}$$

From here the result will follow, because

$$\begin{aligned}\frac{1}{2}\,\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[(f^\star(\pi^t)-\bar{f}^t(\pi^t))^2\big]\right] &= \sum_{t=1}^T \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^{t-1}) - \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^t)\\ &= \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^0) - \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^T)\\ &\le \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^0)\\ &\le \log|\mathcal{F}|,\end{aligned}$$

where the last inequality follows because the entropy of a random variable X over a set $\mathcal{X}$ is always bounded by $\log|\mathcal{X}|$.

We proceed to prove (2.25). To begin, we use Lemma 40, which implies that

$$\frac{1}{2}(f^\star(\pi^t)-\bar{f}^t(\pi^t))^2 \le D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{r}^t\mid f^\star,\pi^t,\mathcal{H}^{t-1}} \,\|\, \mathbb{P}_{\boldsymbol{r}^t\mid\pi^t,\mathcal{H}^{t-1}}\big).$$

and

$$\frac{1}{2}\,\mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[(f^\star(\pi^t)-\bar{f}^t(\pi^t))^2\big] = \mathbb{E}_{f^\star\sim\mu^t}\,\mathbb{E}_{\pi^t\sim p^t}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{r}^t\mid f^\star,\pi^t,\mathcal{H}^{t-1}} \,\|\, \mathbb{P}_{\boldsymbol{r}^t\mid\pi^t,\mathcal{H}^{t-1}}\big)\big]$$

Since KL divergence satisfies $\mathbb{E}_{x\sim\mathbb{P}_X}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{Y\mid X=x}\,\|\,\mathbb{P}_Y\big)\big] = \mathbb{E}_{y\sim\mathbb{P}_Y}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{X\mid Y=y}\,\|\,\mathbb{P}_X\big)\big]$, this is equal to

$$\mathbb{E}_{t-1}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{f}^\star\mid\pi^t,r^t,\mathcal{H}^{t-1}} \,\|\, \mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^{t-1}}\big)\big] = \mathbb{E}_{t-1}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^{t-1}}\big)\big]. \tag{2.26}$$

Taking the expectation over $\mathcal{H}^{t-1}$, we can write this as

$$\mathbb{E}\big[\mathbb{E}_{t-1}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^{t-1}}\big)\big]\big] = \mathbb{E}_{\mathcal{H}^{t-1}}\,\mathbb{E}_{\mathcal{H}^t\mid\mathcal{H}^{t-1}}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^{t-1}}\big)\big].$$

A simple exercise shows that for random variables X, Y, Z,

$$\mathbb{E}_{(x,y)\sim\mathbb{P}_{X,Y}}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{Z\mid X=x,Y=y}\,\|\,\mathbb{P}_{Z\mid X=x}\big)\big] = \mathsf{Ent}(Z \mid X) - \mathsf{Ent}(Z \mid X, Y).$$

Applying this result above (and using that $\mathcal{H}^{t-1} \subset \mathcal{H}^t$) gives

$$\mathbb{E}_{\mathcal{H}^{t-1}}\,\mathbb{E}_{\mathcal{H}^t\mid\mathcal{H}^{t-1}}\big[D_{\mathsf{KL}}\big(\mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{f}^\star\mid\mathcal{H}^{t-1}}\big)\big] = \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^{t-1}) - \mathsf{Ent}(\boldsymbol{f}^\star \mid \boldsymbol{\mathcal{H}}^t)$$

as desired.

□

The analysis above critically makes use of the fact that we are concerned with Bayesian regret, and have access to the *true prior.* One might hope that by choosing a sufficiently uninformative prior, this approach might continue to work in the frequentist setting. In fact, this indeed the case for bandits, though a different analysis is required [6, 7]. However, one can show (Sections 4 and 6) that the Bayesian analysis we have given here extends to significantly richer decision making settings, while the frequentist counterpart is limited to simple variants of the multi-armed bandit.

Remark 11 (Equivalence of min-max frequentist regret and max-min Bayesian regret): Using the minimax theorem, it is possible to show that under appropriate technical conditions

$$\min_{\mathsf{Alg}} \max_{M^\star} \mathbb{E}^{M^\star}[\mathbf{Reg}] = \max_{\mu \in \Delta(\mathcal{M})} \min_{\mathsf{Alg}} \mathbb{E}_{M^\star \sim \mu} \mathbb{E}^{M^\star}[\mathbf{Reg}].$$

That is, if we take the worst-case value of the Bayesian regret over all possible choices of prior, this coincides with the minimax value of the frequentist regret.

2.5 Adversarial Bandits and the Exp3 Algorithm*

We conclude this section with a brief introduction to the multi-armed bandit problem with non-stochastic/adversarial rewards, which dispenses with Assumption 1. In the context of Figure 2, the non-stochastic nature of rewards adds a new "adversarial data" dimension to the problem. As one might expect, the solution we will present for non-stochastic bandits will leverage the the online learning tools introduced in Section 1.6.

To simplify the presentation, suppose that the collection of rewards

$$\{r^t(\pi) \in [0,1] : \pi \in [A], t \in [T]\}$$

for each action and time step is arbitrary and fixed ahead of the interaction by an oblivious adversary. Since we do not posit a stochastic model for rewards, we define regret as in (2.4).

The algorithm we present will build upon the exponential weights algorithm studied in the context of online supervised learning in Section 1.6. To make the connection as clear as possible, we make a temporary switch from rewards to losses, mapping r^t to $1 - r^t$, a transformation that does not change the problem itself.

Recall that p^t denotes the randomization distribution for the learner at round t. As discussed in Remark 7, we can write expected regret as

$$\mathbf{Reg} = \sum_{t=1}^{T} \langle p^t, \boldsymbol{\ell}^t \rangle - \min_{\pi \in [A]} \sum_{t=1}^{T} \langle \boldsymbol{e}_\pi, \boldsymbol{\ell}^t \rangle \tag{2.27}$$

where $\boldsymbol{\ell}^t \in [0,1]^A$ is the vector of losses for each of the actions at time t.

Since only the loss (equivalently, reward) of the chosen action $\pi^t \sim p^t$ is observed, we cannot directly appeal to the exponential weights algorithm, which requires knowledge of the full vector $\boldsymbol{\ell}^t$. To address this, we build an *unbiased estimate* of the vector $\boldsymbol{\ell}^t$ from a single real-valued observation $\boldsymbol{\ell}^t(\pi^t)$. At first, this might appear impossible, but it is straightforward to show that

$$\widetilde{\boldsymbol{\ell}}^t(\pi) = \frac{\boldsymbol{\ell}^t(\pi)}{p^t(\pi)} \times \mathbb{I}\{\pi^t = \pi\} \tag{2.28}$$

is an unbiased estimate for all $\pi \in [A]$, or in vector notation

$$\mathbb{E}_{\pi^t \sim p^t}\left[\widetilde{\boldsymbol{\ell}}^t\right] = \boldsymbol{\ell}^t. \tag{2.29}$$

If we apply the exponential weights algorithm with the loss vectors $\widetilde{\boldsymbol{\ell}}^t$, it can be shown to attain regret

$$\mathbb{E}[\mathbf{Reg}] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p^t, \boldsymbol{\ell}^t \rangle\right] - \min_{\pi} \sum_{t=1}^{T} \langle e_\pi, \boldsymbol{\ell}^t \rangle \tag{2.30}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle p^t, \widetilde{\boldsymbol{\ell}}^t \right\rangle\right] - \min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \left\langle e_\pi, \widetilde{\boldsymbol{\ell}}^t \right\rangle\right] \lesssim \sqrt{AT \log A}. \tag{2.31}$$

This algorithm is known as *Exp3* ("Exponential Weights for Exploration and Exploitation"). A full proof of this result is left as an exercise in Section 2.7.

2.6 Deferred Proofs

Proof of Proposition 6 (full version) . Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}^t]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid \mathcal{H}^t]$. We begin by using the law of total expectation to express the expected regret as

$$\mathbf{Reg}_{\mathsf{Bayes}}(\mu) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{t-1}[f^\star(\pi^\star) - f^\star(\pi^t)]\right].$$

Here and throughout the proof, $\mathbb{E}[\cdot]$ will denote the joint expectation over both $M^\star \sim \mu$ and over the sequence $\mathcal{H}^T = (\pi^1, r^1), \ldots, (\pi^T, r^T)$ that the algorithm generates by interacting with $M^\star$.

We first simplify the (conditional) expected regret for each step t. Let $\bar{f}^t(\pi) := \mathbb{E}_{t-1}[f^\star(\pi)]$ denote the *posterior mean reward* function at time t, which should be thought of as the expected value of $f^\star$ given everything we have learned so far. Next, let $\bar{f}^t_{\pi'}(\pi) = \mathbb{E}_{t-1}[f^\star(\pi) \mid \pi^\star = \pi']$, which is the expected reward given everything we have learned so far, assuming that $\pi^\star = \pi'$. We proceed to write the expression

$$\mathbb{E}_{t-1}[f^\star(\pi^\star) - f^\star(\pi^t)]$$

in terms of these quantities. For the learner's reward, we observe that $f^\star$ is conditionally independent of π^t given $\mathcal{H}^{t-1}$, we have

$$\mathbb{E}_{t-1}[f^\star(\pi^t)] = \mathbb{E}_{\pi \sim p^t}[\bar{f}^t(\pi)].$$

For the reward of the optimal action, we begin by writing

$$\begin{aligned}
\mathbb{E}_{t-1}[f^\star(\pi^\star)] &= \sum_{\pi \in \Pi} \mathbb{P}_{t-1}(\pi^\star = \pi)\, \mathbb{E}_{t-1}[f^\star(\pi) \mid \pi^\star = \pi] \\
&= \sum_{\pi \in \Pi} \mathbb{P}_{t-1}(\pi^\star = \pi) \bar{f}^t_\pi(\pi) \\
&= \mathbb{E}_{\pi \sim p^t}\left[\bar{f}^t_\pi(\pi)\right],
\end{aligned}$$

where we have used that p^t was chosen to match the posterior distribution over $\pi^\star$. This establishes that

$$\mathbb{E}_{t-1}[f^\star(\pi^\star) - f^\star(\pi^t)] = \mathbb{E}_{\pi \sim p^t}\left[\bar{f}^t_\pi(\pi) - \bar{f}^t(\pi)\right].$$

We now require a decoupling-type lemma, which can be proven through the same reasoning as Lemma 9.

Lemma 10:

$$\mathbb{E}_{\pi\sim p^t}\left[\bar{f}^t_{\pi}(\pi) - \bar{f}^t(\pi)\right] \leq \sqrt{A \cdot \mathbb{E}_{\pi,\pi^\star\sim p^t}\left[(\bar{f}^t_{\pi^\star}(\pi) - \bar{f}^t(\pi))^2\right]}. \tag{2.32}$$

To keep notation as clear as possible going forward, let us use boldface script ($\boldsymbol{\pi}^t$, $\boldsymbol{\pi}^\star$, $\boldsymbol{f}^\star$, $\boldsymbol{\mathcal{H}}^t$) to refer to the abstract random variables under consideration, and use non-boldface script (π^t, $\pi^\star$, $f^\star$, $\mathcal{H}^t$) to refer to their realizations. As in the simplified proof, we will show that the right-hand side in (2.32) is related to a notion of *information gain* (that is, information about $\pi^\star$ acquired at step t). Using Pinsker's inequality, we have

$$\mathbb{E}_{\pi^t,\pi^\star\sim p^t}\left[(\bar{f}^t_{\pi^\star}(\pi^t) - \bar{f}^t(\pi^t))^2\right] \leq \mathbb{E}_{t-1}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{r}^t|\pi^\star,\pi^t,\mathcal{H}^{t-1}} \,\|\, \mathbb{P}_{\boldsymbol{r}^t|\pi^t,\mathcal{H}^{t-1}}\right)\right].$$

Since KL divergence satisfies $\mathbb{E}_X\left[D_{\mathsf{KL}}\left(\mathbb{P}_{Y|X} \,\|\, \mathbb{P}_Y\right)\right] = \mathbb{E}_Y\left[D_{\mathsf{KL}}\left(\mathbb{P}_{X|Y} \,\|\, \mathbb{P}_X\right)\right]$, this is equal to

$$\mathbb{E}_{t-1}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{\pi}^\star|\pi^t,r^t,\mathcal{H}^{t-1}} \,\|\, \mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^{t-1}}\right)\right] = \mathbb{E}_{t-1}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^{t-1}}\right)\right]. \tag{2.33}$$

This is quantifying how much information about $\pi^\star$ we gain by playing π^t and observing r^t at step t, relative to what we knew at step $t-1$. Applying Lemma 10 and (2.33), we have

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_{\pi\sim p^t}\left[\bar{f}^t_{\pi}(\pi) - \bar{f}^t(\pi)\right] &\leq \sum_{t=1}^T \sqrt{A \cdot \mathbb{E}_{\pi,\pi^\star\sim p^t}\left[(\bar{f}^t_{\pi^\star}(\pi) - \bar{f}^t(\pi))^2\right]} \\
&\leq \sum_{t=1}^T \sqrt{A \cdot \mathbb{E}_{t-1}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^{t-1}}\right)\right]} \\
&\leq \sqrt{AT \cdot \sum_{t=1}^T \mathbb{E}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^{t-1}}\right)\right]}.
\end{aligned}$$

We can write

$$\mathbb{E}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^{t-1}}\right)\right] = \mathsf{Ent}(\boldsymbol{\pi}^\star \mid \boldsymbol{\mathcal{H}}^{t-1}) - \mathsf{Ent}(\boldsymbol{\pi}^\star \mid \boldsymbol{\mathcal{H}}^t),$$

so telescoping gives

$$\sum_{t=1}^T \mathbb{E}\left[D_{\mathsf{KL}}\left(\mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^t} \,\|\, \mathbb{P}_{\boldsymbol{\pi}^\star|\mathcal{H}^{t-1}}\right)\right] = \mathsf{Ent}(\boldsymbol{\pi}^\star \mid \boldsymbol{\mathcal{H}}^0) - \mathsf{Ent}(\boldsymbol{\pi}^\star \mid \boldsymbol{\mathcal{H}}^T) \leq \log(A).$$

□

2.7 Exercises

Exercise 4 (Adversarial Bandits): In this exercise, we will prove a regret bound for *adversarial bandits* (Section 2.5), where the sequence of rewards (losses) is non-stochastic. To make a direct connection to the Exponential Weights Algorithm, we switch from rewards to losses, mapping r^t to $1 - r^t$, a transformation that does not change the problem itself. To simplify the presentation, suppose that a collection of losses

$$\{\boldsymbol{\ell}^t(\pi) \in [0,1] : \pi \in [A], t \in [T]\}$$

for each action π and time step t is arbitrary and chosen before round $t = 1$; this is referred to as an *oblivious adversary*. We denote by $\boldsymbol{\ell}^t = (\boldsymbol{\ell}^t(1), \ldots, \boldsymbol{\ell}^t(A))$ the vector of losses at time t.

The protocol for the problem of adversarial multi-armed bandits (with losses) is as follows:

```
Multi-Armed Bandit Protocol
for $t = 1, \ldots, T$ do
    Select decision $\pi^t \in \Pi := \{1, \ldots, A\}$ by sampling $\pi^t \sim p^t$
    Observe loss $\boldsymbol{\ell}^t(\pi^t)$
```

Let p^t be the randomization distribution of the decision-maker on round t. Expected regret can be written as

$$\mathbb{E}[\mathbf{Reg}] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p^t, \boldsymbol{\ell}^t \rangle\right] - \min_{\pi \in [A]} \sum_{t=1}^{T} \langle \boldsymbol{e}_\pi, \boldsymbol{\ell}^t \rangle . \tag{2.34}$$

Since only the loss of the chosen action $\pi^t \sim p^t$ is observed, we cannot directly appeal to the Exponential Weights Algorithm. The solution is to build an unbiased estimate of the vector $\boldsymbol{\ell}^t$ from the single real-valued observation $\boldsymbol{\ell}^t(\pi^t)$.

1. Prove that the vector $\widetilde{\boldsymbol{\ell}}^t(\cdot \mid \pi^t)$ defined by

$$\widetilde{\boldsymbol{\ell}}^t(\pi \mid \pi^t) = \frac{\boldsymbol{\ell}^t(\pi)}{p^t(\pi)} \times \mathbb{I}\left\{\pi^t = \pi\right\} \tag{2.35}$$

is an *unbiased estimate* for $\boldsymbol{\ell}^t(\pi)$ for all $\pi \in [A]$. In vector notation, this means

$$\mathbb{E}_{\pi^t \sim p^t}[\widetilde{\boldsymbol{\ell}}^t(\cdot \mid \pi^t)] = \boldsymbol{\ell}^t.$$

Conclude that

$$\mathbb{E}[\mathbf{Reg}] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left\langle p^t, \widetilde{\boldsymbol{\ell}}^t \right\rangle\right] - \min_{\pi \in [A]} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left\langle e_\pi, \widetilde{\boldsymbol{\ell}}^t \right\rangle\right] \tag{2.36}$$

Above, we use the shorthand $\widetilde{\boldsymbol{\ell}}^t = \widetilde{\boldsymbol{\ell}}(\cdot \mid \pi^t)$.

2. Show that given π',

$$\mathbb{E}_{\pi \sim p^t}\left[\widetilde{\boldsymbol{\ell}}^t(\pi \mid \pi')^2\right] = \frac{\boldsymbol{\ell}^t(\pi')^2}{p^t(\pi')}, \quad \text{so that} \quad \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}_{\pi \sim p^t}\left[\widetilde{\boldsymbol{\ell}}^t(\pi \mid \pi^t)^2\right] \leq A. \tag{2.37}$$

3. Define

$$p^t(\pi) \propto \exp\left\{-\eta \sum_{s=1}^{t-1} \left\langle e_\pi, \widetilde{\boldsymbol{\ell}}^s(\cdot \mid \pi^s) \right\rangle\right\},$$

which corresponds to the exponential weights algorithm on the estimated losses $\widetilde{\boldsymbol{\ell}}^s$. Apply (1.41) to the estimated losses to show that for any $\pi \in [A]$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left\langle p^t, \widetilde{\boldsymbol{\ell}}^t \right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left\langle e_\pi, \widetilde{\boldsymbol{\ell}}^t \right\rangle\right] \lesssim \sqrt{AT \log A}$$

Hence, the price of bandit feedback in the adversarial model, as compared to full-information online learning, is only $\sqrt{A}$.

3. CONTEXTUAL BANDITS

In the last section, we studied the multi-armed bandit problem, which arguably the simplest framework for interactive decision making. This simplicity comes at a cost: few real-world problems can be modeled as a multi-armed bandit problem directly. For example, for the problem of selecting medical treatments, the multi-armed bandit formulation presuppose that one treatment rule (action/decision) is good for all patients, which is clearly unreasonable. To address this, we augment the problem formulation by allowing the decision-maker to select the action π^t after observing a *context* x^t; this is called the *contextual bandit* problem. The context x^t, which may also be thought of as a feature vector or collection of covariates (e.g., a patient's medical history, or the profile of a user arriving at a website), can be used by the learner to better maximize rewards by tailoring decisions to the specific patient or user under consideration.

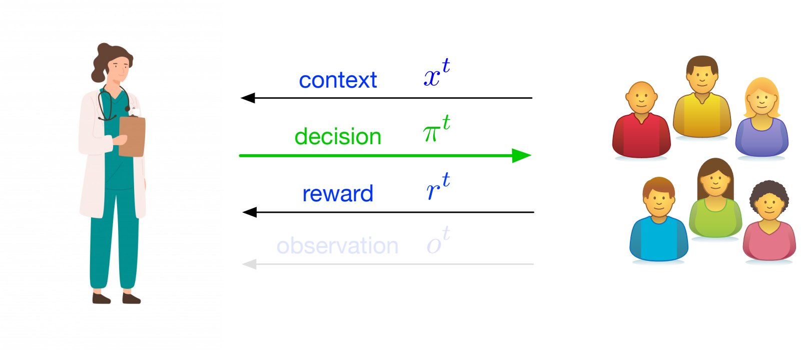


Figure 6: An illustration of the contextual multi-armed bandit problem. A doctor (the learner) aims to select a treatment based on the context (medical history, symptoms).

```
Contextual Bandit Protocol
for $t = 1, \dots, T$ do
    Observe context $x^t \in \mathcal{X}$.
    Select decision $\pi^t \in \Pi = \{1, \dots, A\}$.
    Observe reward $r^t \in \mathbb{R}$.
```

As with multi-armed bandits, contextual bandits can be studied in a stochastic framework or in an adversarial framework. In this course, we will allow the contexts $x^1, \dots, x^T$ to be generated in an arbitrary, potentially adversarially fashion, but assume that rewards are generated from a fixed conditional distribution.

Assumption 2 (Stochastic Rewards): Rewards are generated independently via

$$r^t \sim M^\star(\cdot \mid x^t, \pi^t), \tag{3.1}$$

where $M^\star(\cdot \mid \cdot, \cdot)$ is the underlying *model* (or conditional distribution).

This generalizes the stochastic multi-armed bandit framework in Section 2. We define

$$f^\star(x, \pi) := \mathbb{E}\left[r \mid x, \pi\right] \tag{3.2}$$

as the mean reward function under $r \sim M^\star(\cdot \mid x, \pi)$, and define $\pi^\star(x) := \arg\max_{\pi \in \Pi} f^\star(x, \pi)$ as the optimal *policy*, which maps each context x to the optimal action for the context. We measure performance via regret relative to $\pi^\star$:

$$\mathbf{Reg} := \sum_{t=1}^{T} f^\star(x^t, \pi^\star(x^t)) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^\star(x^t, \pi^t)], \tag{3.3}$$

where $p^t \in \Delta(\Pi)$ is the learner's action distribution at step t (conditioned on the $\mathcal{H}^{t-1}$ and x^t). This provides a (potentially) much stronger notion of performance than what we considered for the multi-armed bandit: Rather than competing with the reward of the single best action, we are competing with the reward of the best sequence of decisions tailored to the context sequence we observe.

Remark 12 (Contextual bandits versus reinforcement learning): To readers already familiar with reinforcement learning, the contextual bandit setting may appear quite similar at first glance, with the term "context" replacing "state". The key difference is that in reinforcement learning, we aim to control the evolution of $x^1, \ldots, x^T$ (which is why they are referred to as state), whereas in contextual bandits, we take the sequence as a given, and only aim to maximize our rewards conditioned on the sequence.

Function approximation and desiderata. If $\mathcal{X}$, the set of possible contexts, is finite, one might imagine running a separate MAB algorithm for each context. In this case, the regret bound would scale with $|\mathcal{X}|$,[9] an undesirable property which reflects the fact that this approach does not allow for generalization across contexts. Instead, we would like to share information between different contexts. After all, a doctor prescribing treatments might never observe exactly the same medical history and symptoms twice, but they might see similar patients or recognize underlying patterns. In the spirit of statistical learning (Section 1) this means assuming access to a class $\mathcal{F}$ that can model the mean reward function, and aiming for regret bounds that scale with $\log|\mathcal{F}|$ (reflecting the statistical capacity of $\mathcal{F}$), with *no dependence* on the cardinality of $\mathcal{X}$. To facilitate this, we will assume a well-specified/realizable model.

Assumption 3: The decision-maker has access to a class $\mathcal{F} \subset \{f : \mathcal{X} \times \Pi \to \mathbb{R}\}$ such that $f^\star \in \mathcal{F}$.

Using the class $\mathcal{F}$, we would like to develop algorithms that can model the underlying reward function for better decision making performance. With this goal in mind, it is reasonable to try leveraging the algorithms and respective guarantees we have already seen for statistical and online supervised learning. At this point, however, the decision-making problem—with its exploration-exploitation dilemma—appears to be quite distinct from these supervised learning frameworks. Indeed, naively applying supervised learning methods, which do not account for the interactive nature of the problem, can lead to failure, as we saw with the

[9] For example, one can show that running an independent instance of UCB for each context leads to regret $\widetilde{O}(\sqrt{AT \cdot |\mathcal{X}|})$.

greedy algorithm in Section 2. In spite of these apparent difficulties, in the next few lectures, we will show that it is possible to leverage supervised learning methods to develop provable decision making methods, thereby bridging the two methodologies.

3.1 Optimism: Generic Template

What algorithmic principles should we employ to solve the contextual bandit problem? One approach is to adapt solutions from the multi-armed bandit setting. There, we saw that the principle of *optimism* (in particular, the UCB algorithm) led to (nearly) optimal rates for bandits, so a natural question is whether optimism will continue to succeed in the presence of contexts. The answer to this last question is: *it depends.* We will first describe a some positive results under assumptions on $\mathcal{F}$, then provide a negative example, and finally turn to an entirely different algorithmic principle.

Optimism via confidence sets. Let us describe a general approach (or, template) for applying the principle of optimism to contextual bandits. Suppose that at each time, we have a way to construct a *confidence set*

$$\mathcal{F}^t \subseteq \mathcal{F}$$

based on the data observed so far, with the property that $f^\star \in \mathcal{F}^t$. Given such a confidence set we can define upper and lower *confidence functions* $\underline{f}^t, \bar{f}^t : \mathcal{X} \times \Pi \to \mathbb{R}$ via

$$\underline{f}^t(x,\pi) = \min_{f \in \mathcal{F}^t} f(x,\pi), \quad \bar{f}^t(x,\pi) = \max_{f \in \mathcal{F}^t} f(x,\pi). \tag{3.4}$$

These functions generalize the upper and lower confidence bounds we constructed in Section 2. Since $f^\star \in \mathcal{F}^t$, they have the property that

$$\underline{f}^t(x,\pi) \le f^\star(x,\pi) \le \bar{f}^t(x,\pi)$$

for all $x \in \mathcal{X}, \pi \in \Pi$. As such, if we consider a contextual analogue of the UCB algorithm, given by

$$\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(x^t,\pi), \tag{3.5}$$

then as in Lemma 7, the optimistic action satisfies

$$f^\star(x^t,\pi^\star) - f^\star(x^t,\pi^t) \le \bar{f}^t(x^t,\pi^t) - \underline{f}^t(x^t,\pi^t).$$

That is, the suboptimality is bounded by the width of the confidence interval at (x^t,π^t), and the total regret is bounded as

$$\mathbf{Reg} \le \sum_{t=1}^{T} \bar{f}^t(x^t,\pi^t) - \underline{f}^t(x^t,\pi^t). \tag{3.6}$$

To make this approach concrete and derive sublinear bounds on the regret, we need a way to construct the confidence set $\mathcal{F}^t$, ideally so that the width in (3.6) shrinks as fast as possible.

Constructing confidence sets with least squares. We construct confidence sets by appealing to a supervised learning method, empirical risk minimization with the square loss (or, least squares). Assume that $f(x,a) \in [0,1]$ for all $f \in \mathcal{F}$, and that $r^t \in [0,1]$ almost surely. Let

$$\widehat{f}^t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} (f(x^i, \pi^i) - r^i)^2 \tag{3.7}$$

be the empirical risk minimizer at round t, and with $\beta := 8\log(|\mathcal{F}|/\delta)$ define $\mathcal{F}^1 = \mathcal{F}$ and

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} (f(x^i, \pi^i) - r^i)^2 \leq \sum_{i=1}^{t-1} (\widehat{f}^t(x^i, \pi^i) - r^i)^2 + \beta \right\} \tag{3.8}$$

for $t > 1$. That is, our confidence set $\mathcal{F}^t$ is the collection of all functions that have empirical squared error close to that of $\widehat{f}^t$. The idea behind this construction is to set β "just large enough", to ensure that we do not accidentally exclude $f^\star$, with the precise value for β informed by the concentration inequalities we explored in Section 1. The only catch here is that we need to use variants of these inequalities that handle dependent data, since x^t and π^t are not i.i.d. in (3.7). The following result shows that $\mathcal{F}^t$ is indeed valid and, moreover, that all functions $f \in \mathcal{F}^t$ have low estimation error on the history.

Lemma 11: Let $\pi^1, \ldots, \pi^T$ be chosen by an arbitrary (and possibly randomized) decision-making algorithm. With probability at least $1-\delta$, $f^\star \in \mathcal{F}^t$ for all $t \in [T]$. Moreover, with probability at least $1-\delta$, for all $\tau \leq T$, all $f \in \mathcal{F}^\tau$ satisfy:

$$\sum_{t=1}^{\tau-1} \mathbb{E}_{\pi^t \sim p^t} \left[(f(x^t, \pi^t) - f^\star(x^t, \pi^t))^2 \right] \leq 4\beta. \tag{3.9}$$

This result establishes validity of the confidence bounds used within the UCB algorithm, but this is not yet enough to show that the algorithm attains low regret. Indeed, to bound the regret, we need to control the confidence widths in (3.6), but there is a mismatch: for step t, the regret bound in (3.6) considers the width at (x^t, π^t), but (3.9) only ensures closeness of functions in $\mathcal{F}^t$ under $(x^1, \pi^1), \ldots, (x^{t-1}, \pi^{t-1})$. We will show in the sequel that for linear models, it is possible to control this mismatch, but that this is not possible in general.

Proof of Lemma 11. For $f \in \mathcal{F}$, define

$$U^t(f) = (f(x^t, \pi^t) - r^t)^2 - (f^\star(x^t, \pi^t) - r^t)^2. \tag{3.10}$$

It is straightforward to check that[10]

$$\mathbb{E}_{t-1}\, U^t(f) = \mathbb{E}_{t-1}(f(x^t, \pi^t) - f^\star(x^t, \pi^t))^2, \tag{3.11}$$

where $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}^{t-1}, x^t]$. Then $Z^t(f) = \mathbb{E}_{t-1}\, U^t(f) - U^t(f)$ is a martingale difference sequence and $\sum_{t=1}^{\tau} Z^t(f)$ is a martingale. Since increments $Z^t(f)$ are bounded as $|Z^t(f)| \leq 1$

[10] We leave $\mathbb{E}_{t-1}$ on the right-hand side to include the case of randomized decisions $\pi^t \sim p^t$.

(this holds whenever $f \in [0,1], r^t \in [0,1]$), according to Lemma 36 with $\eta = \frac{1}{8}$, with probability at least $1-\delta$, for all $\tau \leq T$,

$$\sum_{t=1}^{\tau} Z^t(f) \leq \frac{1}{8}\sum_{t=1}^{\tau} \mathbb{E}_{t-1}\big[Z^t(f)^2\big] + 8\log(\delta^{-1}). \tag{3.12}$$

To control the right-hand side, we again use that $f, r^t \in [0,1]$ to bound

$$\begin{aligned} \mathbb{E}_{t-1}\big[Z^t(f)^2\big] &\leq \mathbb{E}_{t-1}\Big[\big((f(x^t,\pi^t)-r^t)^2 - (f^\star(x^t,\pi^t)-r^t)^2\big)^2\Big] && (3.13)\\ &\leq 4\,\mathbb{E}_{t-1}\big[(f(x^t,\pi^t)-f^\star(x^t,\pi^t))^2\big] = 4\,\mathbb{E}_{t-1}\,U^t(f) && (3.14)\end{aligned}$$

Then, after rearranging, (3.12) becomes

$$\frac{1}{2}\sum_{t=1}^{\tau}\mathbb{E}_{t-1}\,U^t(f) \leq \sum_{t=1}^{\tau} U^t(f) + 8\log(\delta^{-1}). \tag{3.15}$$

Since the left-hand side is nonnegative, we conclude that with probability at least $1-\delta$,

$$\sum_{t=1}^{\tau}(f^\star(x^t,\pi^t)-r^t)^2 \leq \sum_{t=1}^{\tau}(f(x^t,\pi^t)-r^t)^2 + 8\log(\delta^{-1}). \tag{3.16}$$

Taking a union bound over $f \in \mathcal{F}$, gives that probability at least $1-\delta$,

$$\forall f \in \mathcal{F},\ \forall \tau \in [T], \qquad \sum_{t=1}^{\tau}(f^\star(x^t,\pi^t)-r^t)^2 \leq \sum_{t=1}^{\tau}(f(x^t,\pi^t)-r^t)^2 + 8\log(|\mathcal{F}|/\delta), \tag{3.17}$$

and in particular

$$\forall \tau \in [T+1], \qquad \sum_{t=1}^{\tau-1}(f^\star(x^t,\pi^t)-r^t)^2 \leq \sum_{t=1}^{\tau-1}(\widehat{f}^\tau(x^t,\pi^t)-r^t)^2 + 8\log(|\mathcal{F}|/\delta); \tag{3.18}$$

that is, we have$f^\star \in \mathcal{F}^\tau$ for all $\tau \in \{1,\ldots,T+1\}$, proving the first claim. For the second part of the claim, observe that any $f \in \mathcal{F}^\tau$ must satisfy

$$\sum_{t=1}^{\tau-1} U^t(f) \leq \beta$$

since the empirical risk of $f^\star$ is never better than the empirical risk of the minimizer $\widehat{f}^t$. Thus from (3.15), with probability at least $1-\delta$, for all $\tau \leq T$,

$$\sum_{t=1}^{\tau-1}\mathbb{E}_{t-1}\,U^t(f) \leq 2\beta + 16\log(\delta^{-1}). \tag{3.19}$$

The second claim follows by taking union bound over $f \in \mathcal{F}^\tau \subseteq \mathcal{F}$, and by (3.11). □

3.2 Optimism for Linear Models: The LinUCB Algorithm

We now instantiate the general template for optimistic algorithms developed in the previous section for the special case where $\mathcal{F}$ is a class of *linear functions*.

Linear models. We fix a *feature map* $\phi : \mathcal{X} \times \Pi \to \mathsf{B}_2^d(1)$, where $\mathsf{B}_2^d(1)$ is the unit-norm Euclidean ball in $\mathbb{R}^d$. The feature map is assumed to be known to the learning agent. For example, in the case of medical treatments, ϕ transforms the medical history and symptoms x for the patient, along with a possible treatment π, to a representation $\phi(x, \pi) \in \mathsf{B}_2^d(1)$. We take $\mathcal{F}$ to be the set of linear functions given by

$$\mathcal{F} = \{(x, \pi) \mapsto \langle \theta, \phi(x, \pi) \rangle \mid \theta \in \Theta\}, \tag{3.20}$$

where $\Theta \subseteq \mathsf{B}_2^d(1)$ is the parameter set. As before, we assume $f^\star \in \mathcal{F}$; we let θ^* denote the corresponding parameter vector, so that $f^\star(x, \pi) = \langle \theta^\star, \phi(x, \pi) \rangle$.

To apply the technical results in the previous section, we assume for simplicity that $|\Theta| = |\mathcal{F}|$ is finite. To extend the results we will give to potentially non-finite sets, one can work with an ε-discretization or ε-net, which is of size at most $O(\varepsilon^{-d})$ using standard arguments. Taking $\varepsilon \sim 1/T$ ensures only a constant loss in cumulative regret relative to the continuous set of parameters, while $\log |\mathcal{F}| \lesssim d \log T$.

The LinUCB algorithm. The following figure displays an algorithm we refer to as LinUCB, which adapts the generic template for optimistic algorithms to the case where $\mathcal{F}$ is linear in the sense of (3.20).

```
LinUCB
Input: $R > 0$
for $t = 1, \ldots, T$ do
    Compute the least squares solution $\widehat{\theta}^t$ (over $\theta \in \Theta$) given by
        $\widehat{\theta}^t = \arg\min_{\theta \in \Theta} \sum_{i<t} (\langle \theta, \phi(x^i, \pi^i) \rangle - r^i)^2.$
    Define
        $\widetilde{\Sigma}^t = \sum_{i=1}^{t-1} \phi(x^i, \pi^i) \phi(x^i, \pi^i)^\intercal + I.$
    Given $x^t$, select action
        $\pi^t \in \arg\max_{\pi} \max_{\theta : \|\theta - \widehat{\theta}^t\|^2_{\widetilde{\Sigma}^t} \le R} \langle \theta, \phi(x^t, \pi) \rangle .$
    Observe reward $r^t$
```

The following result shows that LinUCB enjoys a regret bound that scales with the complexity $\log|\mathcal{F}|$ of the model class and the feature dimension d.

Proposition 7: Let $\Theta \subseteq \mathsf{B}_2^d(1)$ and fix $\phi : \mathcal{X} \times \Pi \to \mathsf{B}_2^d(1)$. For a finite set $\mathcal{F}$ of linear functions (3.20), taking $\beta = 8 \log(|\mathcal{F}|/\delta)$, LinUCB with $R = 16\beta + 4$ satisfies, with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{\beta d T \log(1 + T/d)} \lesssim \sqrt{dT \log(|\mathcal{F}|/\delta) \log(1 + T/d)}$$

for any sequence of contexts $x^1, \ldots, x^T$. More generally, for infinite $\mathcal{F}$, we may take

$\beta = O(d \log(T))$ and

$$\mathbf{Reg} \lesssim d\sqrt{T} \log(T).$$

Notably, this regret bound has no explicit dependence on the context space size $|\mathcal{X}|$. Interestingly, the bound is also independent of the number of actions $|\Pi|$, which is replaced by the dimension d; this reflects that the linear structure of $\mathcal{F}$ allows the learner to generalize not just across actions, but across decisions. We will expand upon the idea of generalizing across actions in Section 4.

Proof of Proposition 7. The confidence set (3.8) in the generic optimistic algorithm template is equivalent to

$$\mathcal{F}^t = \left\{ \theta \in \Theta : \sum_{i=1}^{t-1} (\langle \theta, \phi(x^i, \pi^i) \rangle - r^i)^2 \le \sum_{i=1}^{t-1} (\langle \widehat{\theta}^t, \phi(x^i, \pi^i) \rangle - r^i)^2 + \beta \right\}, \tag{3.21}$$

where $\widehat{\theta}^t$ is the least squares solution computed in LinUCB. According to Lemma 11, with probability at least $1 - \delta$, for all $t \in [T]$, all $\theta \in \mathcal{F}^t$ satisfy

$$\sum_{i=1}^{t-1} (\langle \theta - \theta^*, \phi(x^i, \pi^i) \rangle)^2 \le 4\beta, \tag{3.22}$$

which means that $\mathcal{F}^t$ is a subset of[11]

$$\Theta' = \left\{ \theta \in \Theta : \|\theta - \theta^*\|_{\Sigma^t}^2 \le 4\beta \right\}, \quad \text{where} \quad \Sigma^t = \sum_{i=1}^{t-1} \phi(x^i, \pi^i) \phi(x^i, \pi^i)^\mathsf{T}. \tag{3.23}$$

Since $\widehat{\theta}^t \in \mathcal{F}^t$, we have that for any $\theta \in \Theta'$, by triangle inequality, $\left\|\theta - \widehat{\theta}^t\right\|_{\Sigma^t}^2 \le 16\beta$. Furthermore, since $\widehat{\theta}^t \in \Theta \subseteq \mathsf{B}_2^d(1)$, $\left\|\theta - \widehat{\theta}^t\right\|_2 \le 2$. Combining the two constraints into one, we find that Θ' is a subset of

$$\Theta'' = \left\{ \theta \in \mathbb{R}^d : \left\|\theta - \widehat{\theta}^t\right\|_{\widetilde{\Sigma}^t}^2 \le 16\beta + 4 \right\}, \quad \text{where} \quad \widetilde{\Sigma}^t = \sum_{i=1}^{t-1} \phi(x^i, \pi^i) \phi(x^i, \pi^i)^\mathsf{T} + I. \tag{3.24}$$

The definition of $\bar{f}^t$ in (3.4) and the inclusion $\Theta' \subseteq \Theta''$ implies that

$$\bar{f}^t(x, \pi) \le \max_{\theta : \left\|\theta - \widehat{\theta}^t\right\|_{\widetilde{\Sigma}^t} \le \sqrt{16\beta + 4}} \langle \theta, \phi(x, \pi) \rangle = \left\langle \widehat{\theta}^t, \phi(x, \pi) \right\rangle + \sqrt{16\beta + 4} \, \|\phi(x, \pi)\|_{(\widetilde{\Sigma}^t)^{-1}}, \tag{3.25}$$

and similarly $\underline{f}^t(x, \pi) \ge \left\langle \widehat{\theta}^t, \phi(x, \pi) \right\rangle - \sqrt{16\beta + 4} \|\phi(x, \pi)\|_{(\widetilde{\Sigma}^t)^{-1}}$. We conclude that regret of the UCB algorithm, in view of Lemma 7, is

$$\mathbf{Reg} \le 2\sqrt{\beta} \sum_{t=1}^{T} \|\phi(x^t, \pi^t)\|_{(\widetilde{\Sigma}^t)^{-1}} \le \sqrt{\beta T \sum_{t=1}^{T} \|\phi(x^t, \pi^t)\|_{(\widetilde{\Sigma}^t)^{-1}}^2}. \tag{3.26}$$

[11] For a PSD matrix $\Sigma \succeq 0$, we define $\|x\|_\Sigma = \sqrt{\langle x, \Sigma x \rangle}$.

The above upper bound has the same flavor as the one in Lemma 8: as we obtain more and more information in some direction v, the matrix $\widetilde{\Sigma}^t$ has a larger and larger component in that direction, and for that direction v, the term $\|v\|^2_{(\widetilde{\Sigma}^t)^{-1}}$ becomes smaller and smaller. To conclude, we apply a potential argument, Lemma 12 below, to bound

$$\sum_{t=1}^{T} \|\phi(x^t, \pi^t)\|^2_{(\widetilde{\Sigma}^t)^{-1}} \lesssim d \log(1 + T/d).$$

□

The following result is referred to as the elliptic potential lemma, and it can be thought of as a generalization of Lemma 8.

Lemma 12 (Elliptic potential lemma): Let $a_1, \ldots, a_T \in \mathbb{R}^d$ satisfy $\|a_t\| \leq 1$ for all $t \in [T]$, and let $V_t = I + \sum_{s \leq t} a_s a_s^\intercal$. Then

$$\sum_{t=1}^{T} \|a_t\|^2_{V_{t-1}^{-1}} \leq 2d \log(1 + T/d). \tag{3.27}$$

Proof Lemma 12 (sketch). First, the determinant of V_t evolves as

$$\det(V_t) = \det(V_{t-1})\left(1 + \|a_t\|^2_{V_{t-1}^{-1}}\right).$$

Second, using the identity $u \wedge 1 \leq 2\ln(1 + u)$ for $u \geq 0$, the left-hand side of (7.38) is at most $2\sum_{t=1}^{T} \log\left(1 + \|a_t\|^2_{V_{t-1}^{-1}}\right)$. The proof concludes by upper bounding the determinant of V_n via the AM-GM inequality. We leave the details as an exercise; see also Lattimore and Szepesvári [51]. □

3.3 Moving Beyond Linear Classes: Challenges

We now present an example of a class $\mathcal{F}$ for which optimistic methods necessarily incur regret that scales linearly with either the cardinality of $\mathcal{F}$ or with cardinality of $\mathcal{X}$, meaning that we do not achieve the desired $\log|\mathcal{F}|$ scaling of regret that one might expect in (offline or online) supervised learning.

Example 3.1 (Failure of optimism for contextual bandits)**.** Let $A = 2$, and let $N \in \mathbb{N}$ be given. Let π_g and π_b be two actions available in each context, so that $\mathcal{A} = \{\pi_g, \pi_b\}$, and $|\mathcal{A}| = 2$. Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be a set of distinct contexts, and define a class $\mathcal{F} = \{f^\star, f_1, \ldots, f_N\}$ of cardinality $N + 1$ as follows. Fix $0 < \varepsilon < 1$. Let $f^\star(x, \pi_g) = 1 - \varepsilon$ and $f^\star(x, \pi_b) = 0$ for any $x \in \mathcal{X}$. For each $i \in [N]$, $f_i(x_j, \pi_g) = 1 - \varepsilon$ and $f_i(x_j, \pi_b) = 0$ for $j \neq i$, while $f_i(x_i, \pi_g) = 0$ and $f_i(x_i, \pi_b) = 1$.

Now, consider a (well-specified) problem instances in which rewards are deterministic and given by

$$r^t = f^\star(x^t, \pi^t),$$

which we note is a constant function with respect to the context. Since $f^\star$ is the true model, π_g is always the best action, bringing a reward of $1 - \varepsilon$ per round. Any time π_b is chosen, the decision-maker incurs instantaneous regret $1 - \varepsilon$. We will now argue that if we apply the generic optimistic algorithm from Section 3.1, it will choose π_b every time a new context is encountered, leading to $\Omega(N)$ regret.

Let S^t be the set of distinct contexts encountered before round t. Clearly, the exact minimizers of empirical square loss (see (3.7)) are $f^\star$, and all f_i where i is such that $x_i \notin S^t$. Hence, for any choice of $\beta \geq 0$, the confidence set in (3.8) contains all f_i for which $x_i \notin S^t$. This implies that for each $t \in [T]$ where $x^t = x_i \notin S^t$, action π_b has a higher upper confidence bound than π_g, since

$$\bar{f}^t(x^t, \pi_b) = f_i(x_i, \pi_b) = 1 > \bar{f}^t(x^t, \pi_g) = f^\star(x^t, \pi_g) = 1 - \varepsilon.$$

Hence, the cumulative regret grows by $1 - \varepsilon$ every time a new context is presented, and thus scales as $\Omega(N(1-\varepsilon))$ if the contexts are presented in order. That is, since $N = |\mathcal{X}| = |\mathcal{F}| - 1$, the confidence-based algorithm fails to achieve logarithmic dependence on $\mathcal{F}$ (note that we may take $\varepsilon = 1/2$ for concreteness).

Let us remark that this failure continues even if contexts are stochastic. If the contexts are chosen via the uniform distribution on $\mathcal{X}$, then for $T \geq N$, at least a constant proportion of the domain will be presented, which still leads to a lower bound of

$$\mathbb{E}[\mathbf{Reg}] = \Omega(N) = \Omega(\min\{|\mathcal{F}|, |\mathcal{X}|\}).$$

◁

What is behind the failure of optimism in this example? The structure of $\mathcal{F}$ forces optimistic methods to *over-explore*, as the algorithm puts too much hope into trying the arm π_b for each new context. As a result, the confidence widths in (3.6) do not shrink quickly enough. Below, we will see that there are alternative methods which *do* enjoy logarithmic dependence on the size of $\mathcal{F}$, with the best of these methods achieving regret $O(\sqrt{AT \log|\mathcal{F}|})$.

We mention in passing that even though optimism does not succeed in general, it is useful to understand in what cases it works. We saw that the structure of linear classes in $\mathbb{R}^d$ only allowed for d "different" directions, while in the example above, the optimistic algorithm gets tricked by each new context, and is not able to shrink the confidence band quickly enough over the domain. In a few lectures (Section 4), we will introduce the eluder dimension, a structural property of the class $\mathcal{F}$ which is sufficient for optimistic methods to experience low regret, generalizing the linear setting.

3.4 The ε-Greedy Algorithm for Contextual Bandits

Given that the principle of optimism only leads to low regret for classes $\mathcal{F}$ with special structure, we are left wondering whether there are more general algorithmic principles for decision making that can succeed for *any* class $\mathcal{F}$. In this section and the following one, we will present two such principles. Both approaches will still make use of supervised learning with the class $\mathcal{F}$, but will build upon *online* supervised learning as opposed to offline/statistical learning. To make the use of supervised learning as modular as possible, we will abstract this away using the notion of an *online regression oracle.*

Definition 3 (Online Regression Oracle): At each time $t \in [T]$, an *online regression oracle* returns, given

$$(x^1, \pi^1, r^1), \ldots, (x^{t-1}, \pi^{t-1}, r^{t-1})$$

with $\mathbb{E}[r^i | x^i, \pi^i] = f^\star(x^i, \pi^i)$ and $\pi^i \sim p^i$, a function $\widehat{f}^t : \mathcal{X} \times \Pi \to \mathbb{R}$ such that

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} (\widehat{f}^t(x^t, \pi^t) - f^\star(x^t, \pi^t))^2 \leq \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$$

with probability at least $1 - \delta$. For the results that follow, $p^i = p^i(\cdot | x^i, \mathcal{H}^{i-1})$ will represent the randomization distribution of a decision-maker.

For example, for finite classes, the (averaged) exponential weights method introduced in Section 1.6 is an online regression oracle with $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta) = \log(|\mathcal{F}|/\delta)$. More generally, in view of Lemma 6, any online learning algorithm that attains low square loss regret or the problem of predicting of r^t based on (x^t, π^t) can leads to a valid online regression oracle.

Note that we make use of *online learning* oracles for the results that follow because we aim to derive regret bounds that hold for arbitrary, potentially adversarial sequences $x^1, \ldots, x^T$. If we instead assume that contexts are i.i.d., it is reasonable to make use of algorithms for *offline estimation*, or statistical learning with $\mathcal{F}$. See Section 3.5.1 for further discussion.

The first general-purpose contextual bandit algorithm we will study, illustrated below, is a contextual counterpart to the ε-Greedy method introduced in Section 2.

```
$\varepsilon$-Greedy for Contextual Bandits
Input: $\varepsilon \in (0, 1)$.
for $t = 1, \ldots, T$ do
    Obtain $\widehat{f}^t$ from online regression oracle for $(x^1, \pi^1, r^1), \ldots, (x^{t-1}, \pi^{t-1}, r^{t-1})$.
    Observe $x^t$.
    With prob. $\varepsilon$, select $\pi^t \sim \mathrm{unif}([A])$, and with prob. $1 - \varepsilon$, choose the greedy action
        $\widehat{\pi}^t = \arg\max_{\pi \in [A]} \widehat{f}^t(x^t, \pi)$.
    Observe reward $r^t$.
```

At each step t, the algorithm uses an online regression oracle to compute a reward estimator $\widehat{f}^t(x, a)$ based on the data $\mathcal{H}^{t-1}$ collected so far. Given this estimator, the algorithm uses the same sampling strategy as in the non-contextual case: with probability $1 - \varepsilon$, the algorithm chooses the greedy decision

$$\widehat{\pi}^t = \arg\max_{\pi} \widehat{f}^t(x^t, \pi), \tag{3.28}$$

and with probability ε it samples a uniform random action $\pi^t \sim \mathrm{unif}(\{1, \ldots, A\})$. The following theorem shows that whenever the online estimation oracle has low estimation error $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$, this method achieves low regret.

Proposition 8: Assume $f^\star \in \mathcal{F}$ and $f^\star(x,a) \in [0,1]$. Suppose the decision-maker has access to an online regression oracle (Definition 3) with a guarantee $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F},T,\delta)$. Then by choosing ε appropriately, the ε-Greedy algorithm ensures that with probability at least $1-\delta$,

$$\mathbf{Reg} \lesssim A^{1/3}T^{2/3} \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F},T,\delta)^{1/3}$$

for any sequence $x^1,\ldots,x^T$. As a special case, when $\mathcal{F}$ is finite, if we use the (averaged) exponential weights algorithm as an online regression oracle, the ε-Greedy algorithm has

$$\mathbf{Reg} \lesssim A^{1/3}T^{2/3} \cdot \log^{1/3}(|\mathcal{F}|/\delta).$$

Notably, this result scales with $\log|\mathcal{F}|$ for *any finite class*, analogous to regret bounds for offline/online supervised learning. The $T^{2/3}$-dependence in the regret bound is suboptimal (as seen for the special case of non-contextual bandits), which we will address using more deliberate exploration methods in the sequel.

Proof of Proposition 8. Recall that p^t denotes the randomization strategy on round t, computed after observing x^t. Following the same steps as the proof of Proposition 4, we can bound regret by

$$\mathbf{Reg} = \sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t}[f^\star(x^t,\pi^\star(x^t)) - f^\star(x^t,\pi^t)] \leq \sum_{t=1}^T f^\star(x^t,\pi^\star(x^t)) - f^\star(x^t,\widehat{\pi}^t) + \varepsilon T,$$

where the εT term represents the bias incurred by exploring uniformly.

Fix t and abbreviate $\pi^\star(x^t) = \pi^\star$. We have

$$\begin{aligned}
&f^\star(x^t,\pi^\star) - f^\star(x^t,\widehat{\pi}^t)\\
&= [f^\star(x^t,\pi^\star) - \widehat{f}^t(x^t,\pi^\star)] + [\widehat{f}^t(x^t,\pi^\star) - \widehat{f}^t(x^t,\widehat{\pi}^t)] + [\widehat{f}^t(x^t,\widehat{\pi}^t) - f^\star(x^t,\widehat{\pi}^t)]\\
&\leq \sum_{\pi\in\{\widehat{\pi}^t,\pi^\star\}} |f^\star(x^t,\pi) - \widehat{f}^t(x^t,\pi)|\\
&= \sum_{\pi\in\{\widehat{\pi}^t,\pi^\star\}} \frac{1}{\sqrt{p^t(\pi)}}\sqrt{p^t(\pi)}|f^\star(x^t,\pi) - \widehat{f}^t(x^t,\pi)|.
\end{aligned}$$

By the Cauchy-Schwarz inequality, the last expression is at most

$$\left\{\sum_{\pi\in\{\widehat{\pi}^t,\pi^\star\}} \frac{1}{p^t(\pi)}\right\}^{1/2}\left\{\sum_{\pi\in\{\widehat{\pi}^t,\pi^\star\}} p^t(\pi)\Big(f^\star(x^t,\pi) - \widehat{f}^t(x^t,\pi)\Big)^2\right\}^{1/2} \tag{3.29}$$

$$\leq \sqrt{\frac{2A}{\varepsilon}}\left\{\mathbb{E}_{\pi^t\sim p^t}\Big(f^\star(x^t,\pi^t) - \widehat{f}^t(x^t,\pi^t)\Big)^2\right\}^{1/2}. \tag{3.30}$$

Summing across t, this gives

$$\sum_{t=1}^{T} f^{\star}(x^t, \pi^{\star}(x^t)) - f^{\star}(x^t, \widehat{\pi}^t) \leq \sqrt{\frac{2A}{\varepsilon}} \sum_{t=1}^{T} \left\{ \mathbb{E}_{\pi^t \sim p^t} \left(f^{\star}(x^t, \pi^t) - \widehat{f}^t(x^t, \pi^t) \right)^2 \right\}^{1/2} \tag{3.31}$$

$$\leq \sqrt{\frac{2AT}{\varepsilon}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left(f^{\star}(x^t, \pi^t) - \widehat{f}^t(x^t, \pi^t) \right)^2 \right\}^{1/2}. \tag{3.32}$$

Now observe that the online regression oracle guarantees that with probability $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left(f^{\star}(x^t, \pi^t) - \widehat{f}^t(x^t, \pi^t) \right)^2 \leq \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta).$$

Whenever this occurs, we have

$$\mathbf{Reg} \lesssim \sqrt{\frac{AT\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)}{\varepsilon}} + 2\varepsilon.$$

Choosing ε to balance

$$\varepsilon T \asymp \sqrt{\frac{AT\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)}{\varepsilon}}$$

leads to the claimed result. □

Remark 13:

3.5 Inverse Gap Weighting: An Optimal Algorithm for General Model Classes

To conclude this section, we present a general, oracle-based algorithm for contextual bandits which achieves

$$\mathbf{Reg} \lesssim \sqrt{AT \log|\mathcal{F}|}$$

for any finite class $\mathcal{F}$. As with ε-Greedy, this approach has no dependence on the cardinality $|\mathcal{X}|$ of the context space, reflecting the ability to generalize across contexts. The dependence on T improves upon ε-Greedy, and is optimal.

To motivate the approach, recall that conceptually, the key step of the proof of Proposition 8 involved relating the instantaneous regret

$$\mathbb{E}_{\pi^t \sim p^t}[f^{\star}(x^t, \pi^{\star}(x^t)) - f^{\star}(x^t, \pi^t)] \tag{3.33}$$

of the decision maker at time t to the instantaneous estimation error

$$\mathbb{E}_{\pi^t \sim p^t} \left(f^{\star}(x^t, \pi^t) - \widehat{f}^t(x^t, \pi^t) \right)^2 \tag{3.34}$$

between $\widehat{f}^t$ and $f^{\star}$ under the randomization distribution p^t. The ε-Greedy exploration distribution gives a way to relate these quantities, but the algorithm's regret is suboptimal because the randomization distribution puts mass at least ε/A on every action, even those that are clearly suboptimal and should be discarded. One can ask whether there exists a better randomization strategy that still admits an upper bound on (3.33) in terms of (3.34). Proposition 9 below establishes exactly that. At first glance, this distribution might appear to be somewhat arbitrary or "magical", but we will show in subsequent chapters that it arises as a special case of more general—and in some sense, universal—principle for designing decision making algorithms, which extends well beyond contextual bandits.

Definition 4 (Inverse Gap Weighting [2, 32]): Given a vector $\widehat{f} = (\widehat{f}(1), \ldots, \widehat{f}(A)) \in \mathbb{R}^A$, the Inverse Gap Weighting distribution $p = \mathsf{IGW}_\gamma(\widehat{f}(1), \ldots, \widehat{f}(A))$ with parameter $\gamma \geq 0$ is defined as

$$p(\pi) = \frac{1}{\lambda + 2\gamma(\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi))}, \tag{3.35}$$

where $\widehat{\pi} = \arg\max_\pi \widehat{f}(\pi)$ is the greedy action, and where $\lambda \in [1, A]$ is chosen such that $\sum_\pi p(\pi) = 1$.

Above, the normalizing constant $\lambda \in [1, A]$ is always guaranteed to exist, because we have $\frac{1}{\lambda} \leq \sum_\pi p(\pi) \leq \frac{A}{\lambda}$, and because $\lambda \mapsto \sum_\pi p(\pi)$ is continuous over $[1, A]$.

Let us give some intuition behind the distribution in (3.35). We can interpret the parameter γ as trading off exploration and exploitation. Indeed, $\gamma \to 0$ gives a uniform distribution, while $\gamma \to \infty$ amplifies the gap between the greedy action $\widehat{\pi}$ and any action with $\widehat{f}(\pi) < \widehat{f}(\widehat{\pi})$, resulting in a distribution supported only on actions that achieve the largest estimated value $\widehat{f}(\widehat{\pi})$.

The following fundamental technical result shows that playing the Inverse Gap Weighting distribution always suffices to link the instantaneous regret in (3.33) in to the instantaneous estimation error in (3.34).

Proposition 9: Consider a finite decision space $\Pi = \{1, \ldots, A\}$. For any vector $\widehat{f} \in \mathbb{R}^A$ and $\gamma > 0$, define $p = \mathsf{IGW}_\gamma(\widehat{f}(1), \ldots, \widehat{f}(A))$. This strategy guarantees that for all $f^\star \in \mathbb{R}^A$,

$$\mathbb{E}_{\pi\sim p}[f^\star(\pi^\star) - f^\star(\pi)] \leq \frac{A}{\gamma} + \gamma \cdot \mathbb{E}_{\pi\sim p}\Big[(\widehat{f}(\pi) - f^\star(\pi))^2\Big]. \tag{3.36}$$

Proof of Proposition 9. We break the "regret" term on the left-hand side of (3.36) into three terms:

$$\mathbb{E}_{\pi\sim p}\Big[f^\star(\pi^\star) - f^\star(\pi)\Big] = \underbrace{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)\Big]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi) - f^\star(\pi)\Big]}_{\text{(II) est error on policy}} + \underbrace{f^\star(\pi^\star) - \widehat{f}(\widehat{\pi})}_{\text{(III) est error at opt}}.$$

The first term asks "how much would we lose by exploring, if $\widehat{f}$ were the true reward function?", and is equal to

$$\sum_\pi \frac{\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)}{\lambda + 2\gamma\left(\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)\right)} \leq \frac{A-1}{2\gamma},$$

while the second term is at most

$$\sqrt{\mathbb{E}_{\pi\sim p}(\widehat{f}(\pi) - f^\star(\pi))^2} \leq \frac{1}{2\gamma} + \frac{\gamma}{2}\,\mathbb{E}_{\pi\sim p}(\widehat{f}(\pi) - f^\star(\pi))^2.$$

The third term can be further written as

$$\begin{aligned} f^\star(\pi^\star) - \widehat{f}(\pi^\star) - (\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star)) &\leq \frac{\gamma}{2}p(\pi^\star)\left(f^\star(\pi^\star) - \widehat{f}(\pi^\star)\right)^2 + \frac{1}{2\gamma p(\pi^\star)} - (\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star)) \\ &\leq \frac{\gamma}{2}\,\mathbb{E}_{\pi\sim p}(f^\star(\pi) - \widehat{f}(\pi))^2 + \left[\frac{1}{2\gamma p(\pi^\star)} - (\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star))\right]. \end{aligned}$$

The term in brackets above is equal to

$$\frac{\lambda + 2\gamma(\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star))}{2\gamma} - (\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star)) = \frac{\lambda}{2\gamma} \leq \frac{A}{2\gamma}.$$

□

The simple result we just proved is remarkable. The special IGW strategy guarantees a relation between regret and estimation error for *any* estimator $\widehat{f}$ and any $f^\star$, irrespective of the problem structure or the class $\mathcal{F}$. Proposition 9 will be at the core of the development for the rest of the course, and will be greatly generalized to general decision making problems and reinforcement learning.

Below, we present a contextual bandit algorithm which makes use of the Inverse Gap Weighting distribution.

```
SquareCB
Input: Exploration parameter $\gamma > 0$.
for $t = 1, \ldots, T$ do
    Obtain $\widehat{f}^t$ from online regression oracle with $(x^1, \pi^1, r^1), \ldots, (x^{t-1}, \pi^{t-1}, r^{t-1})$.
    Observe $x^t$.
    Compute $p^t = \mathsf{IGW}_\gamma\left(\widehat{f}^t(x^t, 1), \ldots, \widehat{f}^t(x^t, A)\right)$.
    Select action $\pi^t \sim p^t$.
    Observe reward $r^t$.
```

At each step t, the algorithm uses an online regression oracle to compute a reward estimator $\widehat{f}^t(x, a)$ based on the data $\mathcal{H}^{t-1}$ collected so far. Given this estimator, the algorithm uses Inverse Gap Weighting to compute $p^t = \mathsf{IGW}_\gamma(\widehat{f}^t(x^t, \cdot))$ as an exploratory distribution, then samples $\pi^t \sim p^t$.

The following result, which is a near-immediate consequence of Proposition 9, gives a regret bound for this algorithm.

Proposition 10: Given a class $\mathcal{F}$ with $f^\star \in \mathcal{F}$, assume the decision-maker has access to an online regression oracle (Definition 3) with estimation error $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$. Then SquareCB with $\gamma = \sqrt{TA/\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)}$ attains a regret bound of

$$\mathbf{Reg} \lesssim \sqrt{AT\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)}$$

with probability at least $1 - \delta$ for any sequence $x^1, \ldots, x^T$. As a special case, when $\mathcal{F}$ is finite, the averaged exponential weights algorithm achieves $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$, leading to

$$\mathbf{Reg} \lesssim \sqrt{AT \log(|\mathcal{F}|/\delta)}.$$

Proof of Proposition 10. We begin with regret, then add and subtract the squared estimation error as follows:

$$\begin{aligned}\mathbf{Reg} &= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^\star(x^t, \pi^\star) - f^\star(x^t, \pi^t)] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\Big[f^\star(x^t, \pi^\star) - f^\star(x^t, \pi^t) - \gamma \cdot (f^\star(x^t, \pi^t) - \widehat{f}^t(x^t, \pi^t))^2\Big] + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta).\end{aligned}$$

By appealing to Proposition 9 with $\widehat{f}(x^t, \cdot)$ and $f^\star(x^t, \cdot)$, for each step t, we have

$$\mathbb{E}_{\pi^t \sim p^t}\Big[f^\star(x^t, \pi^\star) - f^\star(x^t, \pi^t) - \gamma \cdot (f^\star(x^t, \pi^t) - \widehat{f}^t(x^t, \pi^t))^2\Big] \leq \frac{A}{\gamma},$$

and thus

$$\mathbf{Reg} \leq \frac{TA}{\gamma} + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta).$$

Choosing γ to balance these terms yields the result. □

If the online regression oracle is minimax optimal (that is, $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$ is the "best possible" for $\mathcal{F}$) then SquareCB is also minimax optimal for $\mathcal{F}$. Thus, IGW not only provides a connection between online supervised learning and decision making, but it does so in an optimal fashion. Establishing minimax optimality is beyond the scope of this course: it requires understanding of minimax optimality of online regression with arbitrary $\mathcal{F}$, as well as lower bound on regret of contextual bandits with arbitrary sequences of contexts. We refer to Foster and Rakhlin [32] for details.

3.5.1 Extending to Offline Regression

When $x^1, \ldots, x^T$ are i.i.d., it is natural to ask whether an online regression method that works for arbitrary sequences is necessary, or whether one can work with a weaker oracle tuned to i.i.d. data. For SquareCB, it turns out that any oracle for *offline regression* (defined below) is sufficient.

Definition 5 (Offline Regression Oracle): Given

$$(x^1, \pi^1, r^1), \ldots, (x^{t-1}, \pi^{t-1}, r^{t-1})$$

where $x^1, \ldots, x^{t-1}$ are i.i.d., $\pi^i \sim p(x^i)$ for fixed $p : \mathcal{X} \to \Delta(\Pi)$ and $\mathbb{E}[r^i | x^i, \pi^i] = f^\star(x^i, \pi^i)$, an *offline regression oracle* returns a function $\widehat{f} : \mathcal{X} \times \Pi \to \mathbb{R}$ such that

$$\mathbb{E}_{x, \pi \sim p(x)}(\widehat{f}(x, \pi) - f^\star(x, \pi))^2 \leq t^{-1} \mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, t, \delta)$$

with probability at least $1 - \delta$.

Note that the normalization t^{-1} above is introduced to keep the scaling consistent with our conventions for offline estimation.

Below, we state a variant of SquareCB which is adapted to offline oracles. Compared to the SquareCB for online oracles, the main change is that we update the estimation oracle and exploratory distribution on an epoched schedule as opposed to updating at every round. In addition, the parameter γ for the Inverse Gap Weighting distribution changes as a function of the epoch.

```
SquareCB with offline oracles
Input: Exploration parameters $\gamma_1, \gamma_2, \ldots$ and epoch sizes $\tau_1, \tau_2, \ldots$
for $m = 1, 2, \ldots$ do
    Obtain $\widehat{f}^m$ from offline regression oracle with
        $(x^{\tau_{m-2}+1}, \pi^{\tau_{m-2}+1}, r^{\tau_{m-2}+1}), \ldots, (x^{\tau_{m-1}}, \pi^{\tau_{m-1}}, r^{\tau_{m-1}})$.
    for $t = \tau_{m-1} + 1, \ldots, \tau_m$ do
        Observe $x^t$.
        Compute $p^t = \mathsf{IGW}_{\gamma_m}\left(\widehat{f}^m(x^t, 1), \ldots, \widehat{f}^m(x^t, A)\right)$.
        Select action $\pi^t \sim p^t$.
        Observe reward $r^t$.
```

While this algorithm is quite intuitive, proving a regret bound for it is quite non-trivial—much more so than the online oracle variant. They key challenge is that, while the contexts $x^1, \ldots, x^T$ are i.i.d., the decisions $\pi^1, \ldots, \pi^T$ evolve in a time-dependent fashion, which makes it unclear to invoke the guarantee in Definition 5. Nonetheless, the following remarkable result shows that this algorithm attains a regret bound similar to that of Proposition 10.

Proposition 11 (Simchi-Levi and Xu [68]): Let $\tau_m = 2^m$ and $\gamma_m = \sqrt{AT/\mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, \tau_{m-1}, \delta)}$ for $m = 1, 2, \ldots$. Then with probability at least $1-\delta$, regret of $\mathsf{SquareCB}$ with an offline oracle is at most

$$\mathbf{Reg} \lesssim \sum_{m=1}^{\lceil \log T \rceil} \sqrt{A \cdot \tau_m \cdot \mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, \tau_m, \delta/m^2)}.$$

Under mild assumptions, above bound scales as

$$\mathbf{Reg} \lesssim \sqrt{A \cdot T \cdot \mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, \tau_m, \delta/\log T)}.$$

For a finite class $\mathcal{F}$, we recall from Section 1 that empirical risk with the square loss (least squares) achieves $\mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$, which gives

$$\mathbf{Reg} \lesssim \sqrt{AT \log(|\mathcal{F}|/\delta)}.$$

3.6 Exercises

Exercise 5 (Unstructured Contextual Bandits): Consider a contextual bandit problem with a finite set $\mathcal{X}$ of possible contexts, and a finite set of actions $\mathcal{A}$. Show that running UCB independently for each context yields a regret bound of the order $\widetilde{O}(\sqrt{|X||\mathcal{A}|T})$ in expectation, ignoring logarithmic factors. In the setting where $\mathcal{F} = \mathcal{X} \times \mathcal{A} \to [0,1]$ is unstructured, and consists of all possible functions, this is essentially optimal.

Exercise 6 (ε-Greedy with Offline Oracles): In Proposition 8, we analyzed the ε-Greedy contextual bandit algorithm assuming access to an online regression oracle. Because we appeal to online learning, this algorithm was able to handle adversarial contexts $x^1, \ldots, x^T$. In the present problem, we will modify the ε-greedy algorithm and proof to show that if contexts are stochastic (that is $x^t \sim \mathcal{D}\ \forall t$, where $\mathcal{D}$ is a fixed distribution), ε-greedy works even if we use an *offline oracle* (Definition 5).

We consider the following variant of ε-greedy. The algorithm proceeds in epochs $m = 0, 1, \ldots$ of doubling size

$$\{2\}, \{3,4\}, \{5\ldots 8\}, \ldots, \underbrace{\{2^m+1, 2^{m+1}\}}_{\text{epoch } m}, \ldots, \{T/2+1, T\};$$

we assume without loss of generality that T is a power of 2, and that an arbitrary decision is made on round $t = 1$. At the end of each epoch $m-1$, the offline oracle is invoked with the data from the epoch, producing an estimated model $\widehat{f}^m$. This model is used for the greedy step in the next epoch m. In other words, for any round $t \in [2^m+1, 2^{m+1}]$ of epoch m, the algorithm observes $x^t \sim \mathcal{D}$, chooses an action $\pi^t \sim \mathrm{unif}[A]$ with probability ε and chooses the greedy action

$$\pi^t = \operatorname*{arg\,max}_{\pi \in [A]} \widehat{f}^m(x^t, \pi)$$

with probability $1 - \varepsilon$. Subsequently, the reward r^t is observed.

1. Prove that for any $T \in \mathbb{N}$ and $\delta > 0$, by setting ε appropriately, this method ensures that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim A^{1/3}T^{1/3}\left(\sum_{m=1}^{\log_2 T} 2^{m/2}\mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, 2^{m-1}, \delta/m^2)^{1/2}\right)^{2/3}$$

2. Recall that for a finite class, ERM achieves $\mathbf{Est}^{\mathsf{off}}_{\mathsf{Sq}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$. Show that with this choice, the above upper bound matches that in Proposition 8, up to logarithmic in T factors.

Exercise 7 (Model Misspecification in Contextual Bandits): In Proposition 10, we showed that for contextual bandits with a general class $\mathcal{F}$, SquareCB attains regret

$$\mathbf{Reg} \lesssim \sqrt{AT \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)}. \tag{3.37}$$

To do so, we assumed that $f^\star \in \mathcal{F}$, where $f^\star(x,a) := \mathbb{E}_{r \sim M^\star(\cdot|x,a)}[r]$; that is, we have a well-specified model. In practice, it may be unreasonable to assume that we have $f^\star \in \mathcal{F}$. Instead, a weaker assumption is that there exists some function $\bar{f} \in \mathcal{F}$ such that

$$\max_{x \in \mathcal{X}, a \in \mathcal{A}} |\bar{f}(x,a) - f^\star(x,a)| \le \varepsilon$$

for some $\varepsilon > 0$; that is, the model is *ε-misspecified*. In this problem, we will generalize the regret bound for SquareCB to handle misspecification. Recall that in the lecture notes, we assumed (Definition 3) that the regression oracle satisfies

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} (\widehat{f}^t(x^t, \pi^t) - f^\star(x^t, \pi^t))^2 \leq \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta).$$

In the misspecified setting, this is too much to ask for. Instead, we will assume that the oracle satisfies the following guarantee for *every sequence*:

$$\sum_{t=1}^{T} (\widehat{f}^t(x^t, \pi^t) - r^t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(x^t, \pi^t) - r^t)^2 \leq \mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T).$$

Whenever $f^\star \in \mathcal{F}$, we have $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta) \lesssim \mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T) + \log(1/\delta)$ with probability at least $1 - \delta$. However, it is possible to keep $\mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T)$ small even when $f^\star \notin \mathcal{F}$. For example, the averaged exponential weights algorithm satisfies this guarantee with $\mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim \log|\mathcal{F}|$, regardless of whether $f^\star \in \mathcal{F}$.

We will show that for every $\delta > 0$, with an appropriate choice of γ, SquareCB (that is, the algorithm that chooses $p^t = \mathsf{IGW}_\gamma(\widehat{f}^t(x^t, \cdot))$) ensures that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{AT \cdot (\mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T) + \log(1/\delta))} + \varepsilon \cdot A^{1/2} T.$$

Assume that all functions in $\mathcal{F}$ and rewards take values in $[0, 1]$.

1. Show that for any sequence of estimators $\widehat{f}^1, \ldots, \widehat{f}^t$, by choosing $p^t = \mathsf{IGW}_\gamma(\widehat{f}^t(x^t, \cdot))$, we have that

$$\mathbf{Reg} = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^\star(x^t, \pi^\star(x^t)) - f^\star(x^t, \pi^t)] \lesssim \frac{AT}{\gamma} + \gamma \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2\right] + \varepsilon T.$$

If we had $f^\star = \bar{f}$, this would follow from Proposition 9, but the difference is that in general ($\bar{f} \neq f^\star$), the expression above measures estimation error with respect to the best-in-class model $\bar{f}$ rather than the true model $f^\star$ (at the cost of an extra εT factor).

2. Show that the following inequality holds for every sequence

$$\sum_{t=1}^{T} (\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 \leq \mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T) + 2 \sum_{t=1}^{T} (r^t - \bar{f}(x^t, \pi^t))(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t)).$$

3. Using Freedman's inequality (Lemma 36), show that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2\right] \leq 2 \sum_{t=1}^{T} (\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 + O(\log(1/\delta)).$$

4. Using Freedman's inequality once more, show that with probability at least $1 - \delta$,

$$2 \sum_{t=1}^{T} (r^t - \bar{f}(x^t, \pi^t))(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t)) \leq \frac{1}{4} \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2\right] + O(\varepsilon^2 T + \log(1/\delta)).$$

Conclude that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2\right] \lesssim \mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T) + \varepsilon^2 T + \log(1/\delta).$$

> 5. Combining the previous results, show that for any $\delta > 0$, by choosing $\gamma > 0$ appropriately, we have that with probability at least $1 - \delta$,
>
> $$\mathbf{Reg} \lesssim \sqrt{AT \cdot (\mathbf{Reg}_{\mathsf{Sq}}(\mathcal{F}, T) + \log(1/\delta))} + \varepsilon \cdot A^{1/2} T.$$

4. STRUCTURED BANDITS

Up to this point, we have focused our attention on bandit problems (with or without contexts) in which the decision space Π is a small, finite set. This section introduces the *structured bandit* problem, which generalizes the basic (non-contextual) multi-armed bandit problem by allowing for large, potentially infinite or continuous decision spaces. The protocol for the setting is as follows.

```
Structured Bandit Protocol
for $t = 1, \ldots, T$ do
    Select decision $\pi^t \in \Pi$.          //$\Pi$ is large and potentially continuous.
    Observe reward $r^t \in \mathbb{R}$.
```

This protocol is exactly the same as for multi-armed bandits (Section 2), except that we have removed the restriction that $\Pi = \{1, \ldots, A\}$, and now allow it to be arbitrary. This added generality is natural in many applications:

- In medicine, the treatment may be a continuous variable, such as a dosage. The treatment could even by a high-dimensional vector (such as dosages for many different medications). See Figure 7.
- In pricing applications, a seller might aim to select a continuous price or vector or prices in order to maximize their returns.
- In routing applications, the decision space may be finite, but combinatorially large. For example, the decision might be a path or flow in a graph.

Both contextual bandits and structured bandits generalize the basic multi-armed bandit problem, by incorporating function approximation and generalization, but in different ways:

- The contextual bandit formulation in Section 3 assumes structure in the context space. The aim here was to *generalize across contexts*, but we restricted the decision space to be finite (unstructured).
- In structured bandits, we will focus our attention on the case of *no contexts*, but will assume the decision space is structured, and aim to *generalize across decisions*.

Clearly, both ideas above can be combined, and we will touch on this in Section 4.5.

Assumptions and regret. To build intuition as to what it means to generalize across decisions, and to give a sense for what sort of guarantees we might hope to prove, let us first give the formal setup for the structured bandit problem. As in preceding sections, we will assume that rewards are stochastic, and generated from a fixed model.

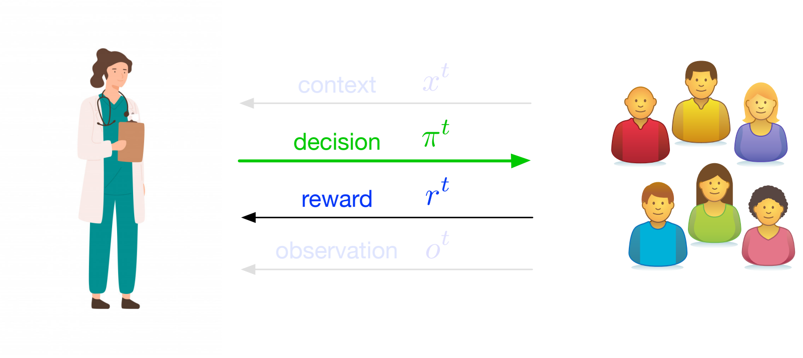


Figure 7: An illustration of the structured bandit problem. A doctor aims to select a continuous, high-dimensional treatment.

Assumption 4 (Stochastic Rewards): Rewards are generated independently via

$$r^t \sim M^\star(\cdot \mid \pi^t), \tag{4.1}$$

where $M^\star(\cdot \mid \cdot)$ is the underlying *model*.

We define

$$f^\star(\pi) := \mathbb{E}\left[r \mid \pi\right] \tag{4.2}$$

as the mean reward function under $r \sim M^\star(\cdot \mid \pi)$, and measure regret via

$$\mathbf{Reg} := \sum_{t=1}^{T} f^\star(\pi^\star) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^\star(\pi^t)]. \tag{4.3}$$

Here, $\pi^\star := \arg\max_{\pi \in \Pi} f^\star(\pi)$ as usual. We will define the history as $\mathcal{H}^t = (\pi^1, r^1), \ldots, (r^t, \pi^t)$.

Function approximation. A first attempt to tackle the structured bandit problem might be to apply algorithms for the multi-armed bandit setting, such as UCB. This would give regret $\widetilde{O}(\sqrt{|\Pi|T})$, which could be vacuous if Π is large relative to T. However, with no further assumptions on the underlying reward function $f^\star$, this is unavoidable. To allow for better regret, we will make assumptions on the structure of $f^\star$ that will allow us to share information across decisions, and to generalize to decisions that we may not have played. This is well-suited for the applications described above, where Π is a continuous set (e.g., $\Pi \subseteq \mathbb{R}^d$), but we expect $f^\star$ to be continuous, or perhaps even linear with respect some well-designed set of features. To make this idea precise, we follow the same approach as in statistical learning and contextual bandits, and assume access to a *well-specified* function class $\mathcal{F}$ that aims to capture our prior knowledge about $f^\star$.

Assumption 5: The decision-maker has access to a class $\mathcal{F} \subset \{f : \Pi \to \mathbb{R}\}$ such that $f^\star \in \mathcal{F}$.

Given such a class, a reasonable goal—particularly in light of the development in Section 1 and Section 3—would be to achieve guarantees that scale with the complexity of supervised learning or estimation with $\mathcal{F}$, e.g. $\log|\mathcal{F}|$ for finite classes; this is what we were able to achieve for contextual bandits, after all. Unfortunately, this is too good to be true, as the following example shows.

Example 4.1 (Necessity of structural assumptions)**.** Let $\Pi = [A]$, and let $\mathcal{F} = \{f_i\}_{i\in[A]}$, where

$$f_i(\pi) := \frac{1}{2} + \frac{1}{2}\mathbb{I}\{\pi = i\}.$$

It is clear that one needs $\mathbf{Reg} \gtrsim A$ for this setting, yet $\log|\mathcal{F}| = \log(A)$, so a regret bound of the form $\mathbf{Reg} \lesssim \sqrt{T\log|\mathcal{F}|}$ is not possible if A is large relative to T. ◁

What this example highlights is that generalizing across decisions is fundamentally different (and, in some sense, more challenging) than generalizing across contexts. In light of this, we will aim for guarantees that scale with $\log|\mathcal{F}|$, but additionally scale with an appropriate notion of *complexity of exploration* for the decision space Π. Such a notion of complexity should reflect how much information is shared across decisions, which depends on the interplay between Π and $\mathcal{F}$.

4.1 Building Intuition: Optimism for Structured Bandits

Our goal is to obtain regret bounds for structured bandits that reflect the intrinsic difficulty of exploring the decision space Π, which should reflect the structure of the function class $\mathcal{F}$ under consideration. To build intuition as to what such guarantees will look like, and how they can be obtained, we first investigate the behavior of the optimism principle and the UCB algorithm when applied to structured bandits. We will see that:

1. UCB attains guarantees that scale with $\log|\mathcal{F}|$, and additionally scale with a notion of complexity called the *eluder dimension*, which is small for simple problems such as bandits with linear rewards.
2. In general, UCB is not optimal, and can have regret that is exponentially large compared to the optimal rate.

4.1.1 UCB for Structured Bandits

We can adapt the UCB algorithm from multi-armed bandits to structured bandit by appealing to least squares and confidence sets, similar to the approach we took for contextual bandits. Assume $\mathcal{F} = \{f : \Pi \to [0,1]\}$ and $r^t \in [0,1]$ almost surely. Let

$$\widehat{f}^t = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{t-1}(f(\pi^i) - r^i)^2 \tag{4.4}$$

be the empirical minimizer on round t, and with $\beta := 8\log(|\mathcal{F}|/\delta)$, define confidence sets $\mathcal{F}^1 = \mathcal{F}$ and

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1}(f(\pi^i) - r^i)^2 \le \sum_{i=1}^{t-1}(\widehat{f}^t(\pi^i) - r^i)^2 + \beta \right\}. \tag{4.5}$$

Defining $\bar{f}^t(\pi) := \max_{f\in\mathcal{F}^t} f(\pi)$ as the upper confidence bound, the generalized UCB algorithm is given by

$$\pi^t = \arg\max_{\pi\in\Pi} \bar{f}^t(\pi). \tag{4.6}$$

When does the confidence width shrink? UsingProposition 7, one can see the generalized UCB algorithm ensures that $f^\star \in \mathcal{F}^t$ for all t with high probability. Whenever this happens, regret is bounded by the upper confidence width:

$$\mathbf{Reg} \leq \sum_{t=1}^{T} \bar{f}^t(\pi^t) - f^\star(\pi^t). \tag{4.7}$$

This bound holds for all structured bandit problems, with no assumption on the structure of Π and $\mathcal{F}$. Hence, to derive a regret bound, the only question we need to answer is *when will the confidence widths shrink?*

For the unstructured multi-armed bandit, we need to shrink the width for every arm separately, and the best bound on (4.7) we can hope for is $O(\sqrt{|\Pi|T})$. One might hope that if Π and $\mathcal{F}$ have nice structure, we can do better. In fact, we have already seen one such case: For linear models, where

$$\mathcal{F} = \left\{\pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta \subset \mathsf{B}_2^d(1)\right\}, \tag{4.8}$$

Proposition 7 shows that we can bound (4.7) by $\sqrt{dT \log|\mathcal{F}|}$. Here, the number of decisions $|\Pi|$ is replaced by the dimension d, which reflects the fact that there are only d truly unique directions to explore before we can start *extrapolating* to new actions. Is there a more general version of this phenomenon when we move beyond linear models?

4.1.2 The Eluder Dimension

The eluder dimension [63] is a complexity measure that aims to capture the extent to which the function class $\mathcal{F}$ facilitates extrapolation (i.e., generalization to unseen decisions), and gives a generic way of bounding the confidence width in (4.7). It is defined for a class $\mathcal{F}$ as follows.

Definition 6 (Eluder Dimension): Let $\mathcal{F} \subset (\Pi \to \mathbb{R})$ and $f^\star : \Pi \to \mathbb{R}$ be given, and define $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon)$ as the length of the longest sequence of decisions $\pi^1, \ldots, \pi^d \in \Pi$ such that for all $t \in [d]$, there exists $f^t \in \mathcal{F}$ such that

$$|f^t(\pi^t) - f^\star(\pi^t)| > \varepsilon, \quad \text{and} \quad \sum_{i<t} (f^t(\pi^i) - f^\star(\pi^i))^2 \leq \varepsilon^2. \tag{4.9}$$

The eluder dimension is defined as $\mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon') \vee 1$. We abbreviate $\mathsf{Edim}(\mathcal{F}, \varepsilon) = \max_{f^\star \in \mathcal{F}} \mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon)$.

The intuition behind the eluder dimension is simple: It asks, for a worst-case sequence of decisions, how many times we can be “surprised” by a new decision π^t if we can estimate the underlying model $f^\star$ well on all of the preceding points. In particular, if we form confidence sets as in (4.5) with $\beta = \varepsilon^2$, then the number of times the upper confidence width in (4.7) can be larger than ε is at most $\mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon)$. We consider the definition $\mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon') \vee 1$ instead of directly working with $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon)$ to ensure monotonicity with respect to ε, which will be useful in the proofs that follow.

The following result gives a regret bound for UCB for generic structured bandit problems. The regret bound has no dependence on the size of the decision space, and scales only with $\mathsf{Edim}(\mathcal{F}, \varepsilon)$ and $\log|\mathcal{F}|$.

Proposition 12: For a finite set of functions $\mathcal{F} \subset (\Pi \to [0,1])$, using $\beta = 8\log(|\mathcal{F}|/\delta)$, the generalized UCB algorithm guarantees that with probability at least $1-\delta$,

$$\mathbf{Reg} \lesssim \min_{\varepsilon>0}\left\{\sqrt{\mathsf{Edim}(\mathcal{F},\varepsilon)\cdot T\log(|\mathcal{F}|/\delta)} + \varepsilon T\right\} \lesssim \sqrt{\mathsf{Edim}(\mathcal{F}, T^{-1/2})\cdot T\log(|\mathcal{F}|/\delta)}. \tag{4.10}$$

For the case of linear models in (4.8), it is possible to use the elliptic potential lemma (Lemma 12) to show that

$$\mathsf{Edim}(\mathcal{F},\varepsilon) \lesssim d\log(\varepsilon^{-1}).$$

For finite classes, this gives $\mathbf{Reg} \lesssim \sqrt{dT\log(|\mathcal{F}|/\delta)\log(T)}$, which recovers the guarantee in Proposition 7. Another well-known example is that of *generalized linear models*. Here, we fix *link function* $\sigma : [-1,+1] \to \mathbb{R}$ and define

$$\mathcal{F} = \left\{\pi \mapsto \sigma\big(\langle\theta,\phi(\pi)\rangle\big) \mid \theta \in \Theta \subset \mathsf{B}_2^d(1)\right\}.$$

This is a more flexible model than linear bandits. A well-known special case is the logistic bandit problem, where $\sigma(z) = 1/(1+e^{-z})$. One can show [63] that for any choice of σ, if there exist $\mu, L > 0$ such that $\mu < \sigma'(z) < L$ for all $z \in [-1,+1]$, then

$$\mathsf{Edim}(\mathcal{F},\varepsilon) \lesssim \frac{L^2}{\mu^2}\cdot d\log(\varepsilon^{-1}). \tag{4.11}$$

This leads to a regret bound that scales with $\frac{L}{\mu}\sqrt{dT\log|\mathcal{F}|}$, generalizing the regret bound for linear bandits.

In general, the eluder dimension can be quite large. Consider the generalized linear model setup above with $\sigma(z) = +\mathsf{relu}(z)$ or $\sigma(z) = -\mathsf{relu}(z)$ (either choice of sign works), where $\mathsf{relu}(z) := \max\{z,0\}$ is the ReLU function; this can be interpreted as a neural network with a single neuron. Here, we can have $\sigma'(z) = 0$, so (4.11) does not apply, and it turns out [52] that

$$\mathsf{Edim}(\mathcal{F},\varepsilon) \gtrsim e^d \tag{4.12}$$

for constant ε. That is, even for a single ReLU neuron, the eluder dimension is already exponential, which is a bit disappointing. Fortunately, we will show in the sequel that the eluder dimension can be overly pessimistic, and it is possible to do better, but this will require changing the algorithm.

Proof of Proposition 12. Define

$$\overline{\mathcal{F}}^t = \left\{f \in \mathcal{F} \mid \sum_{i<t}(f(\pi^i) - f^\star(\pi^i))^2 \le 4\beta\right\}.$$

By Lemma 11, we have that with probability at least $1-\delta$, for all t:

1. $f^\star \in \mathcal{F}^t$.

2. $\mathcal{F}^t \subseteq \overline{\mathcal{F}}^t$.

Let us condition on this event. As in Lemma 7 , since $f^\star \in \mathcal{F}^t$, we can upper bound

$$\mathbf{Reg} \le \sum_{t=1}^{T} \bar{f}^t(\pi^t) - f^\star(\pi^t).$$

Now, define

$$w^t(\pi) = \sup_{f \in \overline{\mathcal{F}}^t} [f(\pi) - f^\star(\pi)],$$

which is a useful upper bound on the upper confidence width at time t. Since $\mathcal{F}^t \subseteq \overline{\mathcal{F}}^t$, we have

$$\mathbf{Reg} \le \sum_{t=1}^{T} w^t(\pi^t).$$

We now appeal to the following technical lemma concerning the eluder dimension.

Lemma 13 (Russo and Van Roy [63], Lemma 3): Fix a function class $\mathcal{F}$, function $f^\star \in \mathcal{F}$, and parameter $\beta > 0$. For any sequence $\pi^1, \ldots, \pi^T$, if we define

$$w^t(\pi) = \sup_{f \in \mathcal{F}} \left\{ f(\pi) - f^\star(\pi) : \sum_{i<t} (f(\pi^i) - f^\star(\pi^i))^2 \le \beta \right\},$$

then for all $\alpha > 0$,

$$\sum_{t=1}^{T} \mathbb{I}\{w^t(\pi^t) > \alpha\} \le \left(\frac{\beta}{\alpha^2} + 1 \right) \cdot \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \alpha).$$

Note that for the special case where $\beta = \alpha^2$, the bound in Lemma 13 immediately follows from the definition of the eluder dimension. The point of this lemma is to show that a similar bound holds for all scales α simultaneously, but with a pre-factor $\frac{\beta}{\alpha^2}$ that grows large when $\alpha^2 \ll \beta$.

To apply this result, fix $\varepsilon > 0$, and bound

$$\sum_{t=1}^{T} w^t(\pi^t) \le \sum_{t=1}^{T} w^t(\pi^t) \mathbb{I}\{w^t(\pi^t) > \varepsilon\} + \varepsilon T. \tag{4.13}$$

Let us order the indices $\{1, \ldots, T\}$ as $\{i_1, \ldots, i_T\}$, so that $w^{i_1}(\pi^{i_1}) \ge w^{i_2}(\pi^{i_2}) \ge \ldots \ge w^{i_T}(\pi^{i_T})$. Consider any index τ for which $w^{i_\tau}(\pi^{i_\tau}) > \varepsilon$. For any $\alpha > \varepsilon$, if we have $w^{i_\tau}(\pi^{i_\tau}) > \alpha$, then Lemma 13 (since $\alpha \le 1 \le \beta$) implies that

$$\tau \le \sum_{t=1}^{T} \mathbb{I}\{w^t(\pi^t) > \alpha\} \le \left(\frac{4\beta}{\alpha^2} + 1 \right) \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \alpha) \le \frac{5\beta}{\alpha^2} \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \alpha). \tag{4.14}$$

Since we have restricted to $\alpha \ge \varepsilon$ and $\alpha \mapsto \mathsf{Edim}_{f^\star}(\mathcal{F}, \alpha)$ is decreasing, rearranging yields

$$w^{i_\tau}(\pi^{i_\tau}) \le \sqrt{\frac{5\beta \mathsf{Edim}(\mathcal{F}, \varepsilon)}{\tau}}.$$

With this, we can bound the main term in (4.13) by

$$\sum_{t=1}^{T} w^t(\pi^t)\mathbb{I}\{w^t(\pi^t) > \varepsilon\} \lesssim \sum_{t=1}^{T} \sqrt{\frac{\beta \mathsf{Edim}(\mathcal{F}, \varepsilon)}{t}} \lesssim \sqrt{\beta \mathsf{Edim}(\mathcal{F}, \varepsilon) T}.$$

Combining this with (4.13) gives $\mathbf{Reg} \lesssim \sqrt{\beta \mathsf{Edim}(\mathcal{F}, \varepsilon) T} + \varepsilon T$. Since $\varepsilon > 0$ was arbitrary, we are free to minimize over it.

□

Proof of Lemma 13. Let us adopt the shorthand $d = \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \alpha)$. We begin with a definition. We say π is α-independent of $\pi^1, \ldots, \pi^t$ if there exists $f \in \mathcal{F}$ such that $|f(\pi) - f^\star(\pi)| > \alpha$ and $\sum_{i=1}^{t}(f(\pi^i) - f^\star(\pi^i))^2 \leq \alpha^2$. We say π is α-dependent on $\pi^1, \ldots, \pi^t$ if for all $f \in \mathcal{F}$ with $\sum_{i=1}^{t}(f(\pi^i) - f^\star(\pi^i))^2 \leq \alpha^2$, $|f(\pi) - f^\star(\pi)| \leq \alpha$.

We first claim that for any t, if $w^t(\pi^t) > \alpha$, then π_t is α-dependent on at most β/α^2 disjoint subsequences of $\pi^1, \ldots, \pi^{t-1}$. Indeed, let f be such that $|f(\pi^t) - f^\star(\pi^t)| > \alpha$. If π^t is α-dependent on a particular subsequence $\pi^{i_1}, \ldots, \pi^{i_k}$ but $w^t(\pi^t) > \alpha$, we must have

$$\sum_{j=1}^{k}(f(\pi^{i_j}) - f^\star(\pi^{i_j}))^2 \geq \alpha^2.$$

If there are M such disjoint sequences, we have

$$M\alpha^2 \leq \sum_{i<t}(f(\pi^i) - f^\star(\pi^i))^2 \leq \beta,$$

so $M \leq \frac{\beta}{\alpha^2}$.

Next, we claim that for τ and any sequence $(\pi^1, \ldots, \pi^\tau)$, there is some j such that π^j is α-dependent on at least $\lfloor \tau/d \rfloor$ disjoint subsequences of $\pi^1, \ldots, \pi^{j-1}$. Let $N = \lfloor \tau/d \rfloor$, and let $B_1, \ldots, B_N$ be subsequences of $\pi^1, \ldots, \pi^\tau$. We initialize with $B_i = (\pi^i)$. If π^{N+1} is α-dependent on $B_i = (\pi^i)$ for all $1 \leq i \leq N$ we are done. Otherwise, choose i such that π^{N+1} is α-independent of B_i, and add it to B_i. Repeat this process until we reach j such that either π^j is α-dependent on all B_i or $j = \tau$. In the first case we are done, while in the second case, we have $\sum_{i=1}^{N}|B_i| \geq \tau \geq dN$. Moreover, $|B_i| \leq d$, since each $\pi^j \in B_i$ is α-independent of its prefix (this follows from the definition of eluder dimension). We conclude that $|B_i| = d$ for all i, so in this case π^τ is α-dependent on all B_i.

Finally, let $(\pi^{t_1}, \ldots, \pi^{t_\tau})$ be the subsequence $\pi^1, \ldots, \pi^T$ consisting of all elements for which $w^{i_i}(\pi^{t_i}) > \alpha$. Each element of the sequence is dependent on at most β/α^2 disjoint subsequences of $(\pi^{t_1}, \ldots, \pi^{t_\tau})$, and by the argument above, one element is dependent on at least $\lfloor \tau/d \rfloor$ disjoint subsequences, so we must have $\lfloor \tau/d \rfloor \leq \beta/\alpha^2$, and which implies that $\tau \leq (\beta/\alpha^2 + 1)d$. □

4.1.3 Suboptimality of Optimism

The following example shows a function class $\mathcal{F}$ for which the regret experienced by UCB is exponentially large compared to the regret obtained by a simple alternative algorithm. This shows that while the algorithm is useful for some special cases, it does not provide a general principle that attains optimal regret for any structured bandit problem.

Example 4.2 (Cheating Code)**.** Let $A \in \mathbb{N}$ be a power of 2 and consider the following function class $\mathcal{F}$.

- The decision space is $\Pi = [A] \cup \mathcal{C}$, where $\mathcal{C} = \{c_1, \ldots, c_{\log_2(A)}\}$ is a set of "cheating" actions.
- For all actions $\pi \in [A]$, $f(\pi) \in [0,1]$ for all $f \in \mathcal{F}$, but we otherwise make no assumption on the reward.
- For each $f \in \mathcal{F}$, rewards for actions in $\mathcal{C}$ take the following form. Let $\pi_f \in [A]$ denote the action in $[A]$ with highest reward. Let $b(f) = (b_1(f), \ldots, b_{\log_2(A)}(f)) \in \{0,1\}^{\log_2(A)}$ be a binary encoding for the index of $\pi_f \in [A]$ (e.g., if $\pi_f = 1$, $b(f) = (0, 0, \ldots, 0)$, if $\pi_f = 2$, $b(f) = (0, 0, \ldots, 0, 1)$, and so on). For each action $c_i \in \mathcal{C}$, we set

$$f(c_i) = -b_i(f).$$

The idea here is that if we ignore the actions $\mathcal{C}$, this looks like a standard multi-armed bandit problem, and the optimal regret is $\Theta(\sqrt{AT})$. However, we can use the actions in $\mathcal{C}$ to "cheat" and get an exponential improvement in sample complexity. The argument is as follows.

Suppose for simplicity that rewards are Gaussian with $r \sim \mathcal{N}(f^\star(\pi), 1)$ under π. For each cheating action $c_i \in \mathcal{C}$, since $f^\star(c_i) = -b_i(f^\star) \in \{0, -1\}$, we can determine whether the value is $b_i(f^\star) = 0$ or $b_i(f^\star) = 1$ with high probability using $\widetilde{O}(1)$ action pulls. If we do this for each $c_i \in \mathcal{C}$, which will incur $\widetilde{O}(\log(A))$ regret (there are $\log(A)$ such actions and each one leads to constant regret), we can infer the binary encoding $b(f^\star) = b_1(f^\star), \ldots, b_{\log_2(A)}(f^\star)$ for the optimal action $\pi_{f^\star}$ with high probability. At this point, we can simply stop exploring, and commit to playing $\pi_{f^\star}$ for the remaining rounds, which will incur no more regret. If one is careful with the details, this gives that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \log^2(A/\delta).$$

In other words, by exploiting the cheating actions, our regret has gone from linear to *logarithmic* in A (we have also improved the dependence on T, which is a secondary bonus).

Now, let us consider the behavior of the generalized UCB algorithm. Unfortunately, since all actions $c_i \in \mathcal{C}$ have $f(c_i) \leq 0$ for all $f \in \mathcal{F}$, we have $\bar{f}^t(c_i) \leq 0$. As a result, the generalized UCB algorithm will only ever pull actions in $[A]$, ignoring the cheating actions and effectively turning this into a vanilla multi-armed bandit problem, which means that

$$\mathbf{Reg} \gtrsim \sqrt{AT}.$$

◁

This example shows that UCB can behave suboptimally in the presence of decisions that reveal useful information but do not necessarily lead to high reward. Since the "cheating" actions are guaranteed to have low reward, UCB avoids them even though they are very informative. We conclude that:

1. Obtaining optimal sample complexity for structured bandits requires algorithms that more deliberately balance the tradeoff between optimizing reward and acquiring information.
2. In general, the optimal strategy for picking decisions can be very different depending on the choice of the class $\mathcal{F}$. This contrasts the contextual bandit setting, where we saw that the Inverse Gap Weighting algorithm attained optimal sample complexity for any choice of class $\mathcal{F}$, and all that needed to change was how to perform estimation.

Remark 14 (Suboptimality of posterior sampling): Recall the Bayesian bandit setting in Section 2.4, where we showed that the posterior sampling algorithm attains regret $\widetilde{O}(\sqrt{AT})$ when $\Pi = \{1, \ldots, A\}$. Posterior sampling is a general-purpose algorithm, and can be applied to directly to arbitrary structured bandit problems (as long as a prior is available). However, similar to UCB, the cheating code construction in Example 4.2 implies that posterior sampling is not optimal in general. Indeed, posterior sampling will never select the cheating arms in $\mathcal{C}$, as these have sub-optimal reward for all models in $\mathcal{F}$. As a result, the Bayesian regret of the algorithm will scale with $\mathbf{Reg} \gtrsim \sqrt{AT}$ for a worst-case prior.

4.2 The Decision-Estimation Coefficient

The discussion in the prequel highlights two challenges in designing algorithms and understanding sample complexity for structured bandits: 1) the optimal regret (in a sense, the complexity of exploration) can depend on the class $\mathcal{F}$ in a subtle, sometimes surprising fashion, and 2) the algorithms required to achieve optimal regret can heavily depend on the choice of $\mathcal{F}$. In light of these challenges, it is natural to ask whether it is possible to have any sort of unified understanding of the optimal regret. We will now show that the answer is *yes*, and this will be achieved by a single, general-purpose principle for algorithm design.

The algorithm we will present in this section reduces the problem of decision making to that of supervised online learning/estimation, in a similar fashion to the SquareCB method for contextual bandits in Section 3. To apply this method, we require the following oracle for supervised estimation.

Definition 7 (Online Regression Oracle): At each time $t \in [T]$, an *online regression oracle* returns, given

$$(\pi^1, r^1), \ldots, (\pi^{t-1}, r^{t-1})$$

with $\mathbb{E}[r^i|\pi^i] = f^\star(\pi^i)$ and $\pi^i \sim p^i$, a function $\widehat{f}^t : \Pi \to \mathbb{R}$ such that

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}(\widehat{f}^t(\pi^t) - f^\star(\pi^t))^2 \le \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$$

with probability at least $1-\delta$. Here, $p^i(\cdot|\mathcal{H}^{i-1})$ is the randomization distribution for the decision-maker.

Recall, following the discussion in Section 3, that the averaged exponential weights algorithm achieves is an online regression oracle with $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, t, \delta) \lesssim \log(|\mathcal{F}|/\delta)$.

The following algorithm, which we call Estimation-to-Decisions or E2D, is a general-purpose meta-algorithm for structured bandits.

```
Estimation-to-Decisions (E2D) for Structured Bandits
Input: Exploration parameter $\gamma > 0$.
for $t = 1, \ldots, T$ do
    Obtain $\widehat{f}^t$ from online regression oracle with $(\pi^1, r^1), \ldots, (\pi^{t-1}, r^{t-1})$.
```

Compute

$$p^t = \arg\min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}^t(\pi))^2\right].$$

Select action $\pi^t \sim p^t$.

At each timestep t, the algorithm calls invokes an online regression oracle to obtain an estimator $\widehat{f}^t$ using the data $\mathcal{H}^{t-1} = (\pi^1, r^1, \ldots, \pi^{t-1}, r^{t-1})$ observed so far. The algorithm then finds a distribution p^t by solving a min-max optimization problem involving the estimator $\widehat{f}^t$ and the class $\mathcal{F}$, then samples the decision π^t from this distribution.

The minimax problem in $\mathsf{E2D}$ is derived from a complexity measure (or, structural parameter) for $\mathcal{F}$ called the *Decision-Estimation Coefficient*, whose value is given by

$$\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) = \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p}\Bigg[\underbrace{f(\pi_f) - f(\pi)}_{\text{regret of decision}} - \gamma\cdot\underbrace{(f(\pi) - \widehat{f}(\pi))^2}_{\text{information gain for obs.}}\Bigg]. \tag{4.15}$$

The Decision-Estimation Coefficient can be thought of as the value of a game in which the learner (represented by the min player) aims to find a distribution over decisions such that for a worst-case problem instance (represented by the max player), the *regret* of their decision is controlled by a notion of *information gain* (or, estimation error) relative to a reference model $\widehat{f}$. Conceptually, $\widehat{f}$ should be thought of as a guess for the true model, and the learner (the min player) aims to—in the face of an unknown environment (the max player)—optimally balance the regret of their decision with the amount information they acquire. With enough information, the learner can confirm or rule out their guess $\widehat{f}$, and scale parameter γ controls how much regret they are willing to incur to do this. In general, the larger the value of $\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f})$, the more difficult it is to explore.

To state a regret bound for $\mathsf{E2D}$, we define

$$\mathsf{dec}_\gamma(\mathcal{F}) = \sup_{\widehat{f}\in\mathrm{co}(\mathcal{F})} \mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}). \tag{4.16}$$

Here, $\mathrm{co}(\mathcal{F})$ denotes the set of all convex combinations of elements in $\mathcal{F}$. The reason we consider the set $\mathrm{co}(\mathcal{F})$ is that in general, online estimation algorithms such as exponential weights will produce improper predictions with $\widehat{f} \in \mathrm{co}(\mathcal{F})$. In fact, it turns out (see **??**) that even if we allow $\widehat{f}$ to be unconstrained above, the maximizer above always lies in $\mathrm{co}(\mathcal{F})$.

The main result for this section shows that the regret for $\mathsf{E2D}$ is controlled by the value of the DEC and the estimation error $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$ for the online regression oracle.

Proposition 13: The $\mathsf{E2D}$ algorithm with exploration parameter $\gamma > 0$ guarantees that with probability at least $1-\delta$,

$$\mathbf{Reg} \le \mathsf{dec}_\gamma(\mathcal{F})\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta). \tag{4.17}$$

We can optimize over the parameter γ in the result above, which yields

$$\begin{aligned}\mathbf{Reg} &\le \inf_{\gamma>0}\left\{\mathsf{dec}_\gamma(\mathcal{F})\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)\right\}\\ &\le 2\cdot\inf_{\gamma>0}\max\left\{\mathsf{dec}_\gamma(\mathcal{F})\cdot T, \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)\right\}.\end{aligned}$$

For finite classes, we can use the exponential weights method to obtain $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F},T,\delta) \lesssim \log(|\mathcal{F}|/\delta)$, and this bound specializes to

$$\mathbf{Reg} \lesssim \inf_{\gamma>0} \max\left\{ \mathsf{dec}_{\gamma}(\mathcal{F}) \cdot T, \gamma \cdot \log(|\mathcal{F}|/\delta) \right\}. \tag{4.18}$$

As desired, this gives a bound on regret that scales only with:

1. the complexity $\log|\mathcal{F}|$ for estimation.

2. the complexity of exploration in the decision space, which is captured by $\mathsf{dec}_{\gamma}(\mathcal{F})$.

Before interpreting the result further, we give the proof, which is a nearly immediate consequence of the definition of the DEC, and bears strong similarity to the proof of the regret bound for SquareCB (Proposition 10), minus contexts.

Proof of Proposition 13. We write

$$\begin{aligned}
\mathbf{Reg} &= \sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t}[f^\star(\pi^\star) - f^\star(\pi^t)] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t}[f^\star(\pi^\star) - f^\star(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t\sim p^t}\left[(f^\star(\pi^t) - \widehat{f}^t(\pi^t))^2\right] + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F},T,\delta).
\end{aligned}$$

For each t, since $f^\star \in \mathcal{F}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^t\sim p^t}[f^\star(\pi^\star) - f^\star(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t\sim p^t}\left[(f^\star(\pi^t) - \widehat{f}^t(\pi^t))^2\right] \\
&\leq \sup_{f\in\mathcal{F}}\left\{ \mathbb{E}_{\pi^t\sim p^t}[f(\pi_f) - f(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t\sim p^t}\left[(f(\pi^t) - \widehat{f}^t(\pi^t))^2\right]\right\} \\
&= \inf_{p\in\Delta(\Pi)} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi^t) - \widehat{f}^t(\pi^t))^2\right] \\
&= \mathsf{dec}_{\gamma}(\mathcal{F}, \widehat{f}^t),
\end{aligned} \tag{4.19}$$

where the first equality above uses that p^t is chosen as the minimizer for $\mathsf{dec}_{\gamma}(\mathcal{F}, \widehat{f}^t)$. Summing across rounds, we conclude that

$$\mathbf{Reg} \leq \sup_{\widehat{f}} \mathsf{dec}_{\gamma}(\mathcal{F}, \widehat{f}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F},T,\delta).$$

□

When designing algorithms for structured bandits, a common challenge is that the connection between decision making (where the learner's decisions influence what feedback is collected) and estimation (where data is collected passively) may not seem apparent a-priori. The power of the Decision-Estimation Coefficient is that it—*by definition*—provides a bridge, which the proof of Proposition 13 highlights. One can select decisions by building an estimate for the model using all of the observations collected so far, then sampling from the distribution p that solves (4.15) with the estimated reward function $\widehat{f}$ plugged in. Boundedness of the DEC implies that at every round, any learner using this strategy either enjoys small regret or acquires information, with their total regret controlled by the cumulative online estimation error.

Example: Multi-Armed Bandit. Of course, the perspective above is only useful if the DEC is indeed bounded, which itself is not immediately apparent. In Section 6, we will show that boundedness of the DEC is not just sufficient, but in fact *necessary* for low regret in a fairly strong quantitative sense. For now, we will build intuition about the DEC through examples. We begin with the multi-armed bandit, where $\Pi = [A]$ and $\mathcal{F} = \mathbb{R}^A$. Our first result shows that $\mathsf{dec}_\gamma(\mathcal{F}) \leq \frac{A}{\gamma}$, and that this is achieved with the Inverse Gap Weighting method introduced in Section 3.

Proposition 14 (IGW minimizes the DEC): For the Multi-Armed Bandit setting, where $\Pi = [A]$ and $\mathcal{F} = \mathbb{R}^A$, the Inverse Gap Weighting distribution $p = \mathsf{IGW}_{4\gamma}(\widehat{f})$ in (3.35) is the *exact* minimizer for $\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f})$, and certifies that $\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) = \frac{A-1}{4\gamma}$.

By rewriting Proposition 9, it is straightforward to deduce that the DEC is bounded by $\frac{A}{\gamma}$, but Proposition 14 shows that IGW is actually the *best possible distribution* for this minimax problem. In this sense, the SquareCB algorithm can be seen as a (contextual) special case of the Estimation-to-Decisions principle. Note that to attain the exact optimal value (instead of a bound that is optimal up to constants), we use $\mathsf{IGW}_{4\gamma}$ as opposed IGW_γ as in Proposition 9; the reason why this choice for γ is optimal is related to the fact that the inequality $xy \leq x^2 + \frac{1}{4}y^2$ is tight in general.

Proof of Proposition 14. We rewrite the minimax problem as

$$\begin{aligned}
&\min_{p\in\Delta([A])} \max_{f\in\mathbb{R}^A} \mathbb{E}_{\pi\sim p}\Big[f(\pi_f) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\Big] \\
&= \min_{p\in\Delta([A])} \max_{f\in\mathbb{R}^A} \max_{\pi^\star\in[A]} \mathbb{E}_{\pi\sim p}\Big[f(\pi^\star) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\Big] \\
&= \min_{p\in\Delta([A])} \max_{\pi^\star\in[A]} \max_{f\in\mathbb{R}^A} \mathbb{E}_{\pi\sim p}\Big[f(\pi^\star) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\Big].
\end{aligned}$$

For any fixed p and $\pi^\star$, first-order conditions for optimality imply that the choice for f that maximizes this expression is

$$f(\pi) = \widehat{f}(\pi) - \frac{1}{2\gamma} + \frac{1}{2\gamma p(\pi^\star)}\mathbb{I}\{\pi = \pi^\star\}.$$

This choice gives

$$\mathbb{E}_{\pi\sim p}[f(\pi^\star) - f(\pi)] = \mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1 - p(\pi^\star)}{2\gamma p(\pi^\star)}$$

and

$$\gamma\,\mathbb{E}_{\pi\sim p}\Big[(f(\pi) - \widehat{f}(\pi))^2\Big] = \frac{1 - p(\pi^\star)}{4\gamma} + \frac{(1 - p(\pi^\star))^2}{4\gamma p(\pi^\star)} = \frac{1}{4\gamma p(\pi^\star)} - \frac{1}{4\gamma}.$$

Plugging in and simplifying, we compute that the original minimax game is equivalent to

$$\min_{p\in\Delta([A])} \max_{\pi^\star\in[A]} \left\{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)}\right\} - \frac{1}{4\gamma}. \tag{4.20}$$

Finishing the proof: Ad-hoc approach. Observe that for any $p \in \Delta(\Pi)$, we have

$$\max_{\pi^\star\in[A]} \left\{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)}\right\} \geq \mathbb{E}_{\pi^\star\sim p}\left[\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)}\right] = \frac{A}{4\gamma},$$

so no p can attain value better than $\frac{A}{4\gamma}$. If we can show that IGW achieves this value, we are done.

Observe that by setting $p = \mathsf{IGW}_{4\gamma}(\widehat{f})$, we have that for all $\pi^\star$,

$$\begin{aligned}\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)} &= \mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{\lambda}{4\gamma} + \widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star) \qquad (4.21)\\ &= \mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)\Big] + \frac{\lambda}{4\gamma}.\end{aligned}$$

Note that the value on the right-hand side is independent of $\pi^\star$. That is, the inverse gap weighting distribution is an equalizing strategy. This means that for this choice of p, we have

$$\begin{aligned}\max_{\pi^\star\in[A]}\bigg\{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)}\bigg\} &= \min_{\pi^\star\in[A]}\bigg\{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)}\bigg\}\\ &= \mathbb{E}_{\pi^\star\sim p}\bigg\{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\Big] + \frac{1}{4\gamma p(\pi^\star)}\bigg\} = \frac{A}{4\gamma}.\end{aligned}$$

Hence, $p = \mathsf{IGW}_{4\gamma}(\widehat{f})$ achieves the optimal value.

Finishing the proof: Principled approach. We begin by relaxing to $p \in \mathbb{R}_+^A$. Define

$$g_{\pi^\star}(p) = \widehat{f}(\pi^\star) + \frac{1}{4\gamma p(\pi^\star)}.$$

Let $\nu \in \mathbb{R}$ be a Lagrange multiplier and $p \in \mathbb{R}_+^A$, and consider the Lagrangian

$$\mathcal{L}(p,\nu) = g_{\pi^\star}(p) - \sum_\pi p(\pi)\widehat{f}(\pi) + \nu\left(\sum_\pi p(\pi) - 1\right).$$

By the KKT conditions, if we wish to show that $p \in \Delta(\Pi)$ is optimal for the objective in (4.20), it suffices to find ν such that[12]

$$\mathbf{0} \in \partial_p \mathcal{L}(p,\nu),$$

where ∂_p denotes the subgradient with respect to p. Recall that for a convex function $h(x) = \max_y g(x,y)$, we have $\partial_x h(x) = \mathrm{co}(\big\{\nabla g(x,y) \mid g(x,y) = \max_{y'} g(x,y')\big\})$. As a result,

$$\partial_p\mathcal{L}(p,\nu) = \nu\mathbf{1} - \widehat{f} + \mathrm{co}(\{\nabla_p g_{\pi^\star}(p) \mid g_{\pi^\star}(p) = \max_{\pi'} g_{\pi'}(p)\}).$$

Now, let $p = \mathsf{IGW}_{4\gamma}(\widehat{f})$. We will argue that $\mathbf{0} \in \partial_p\mathcal{L}(p,\nu)$ for an appropriate choice of ν. By (4.21), we know that $g_\pi(p) = g_{\pi'}(p)$ for all π, π' (p is equalizing), so the expression above simplifies to

$$\partial_p\mathcal{L}(p,\nu) = \nu\mathbf{1} - \widehat{f} + \mathrm{co}(\{\nabla_p g_{\pi^\star}(p)\}_{\pi^\star\in\Pi}). \qquad (4.22)$$

Noting that $\nabla_p g_{\pi^\star}(p) = -\frac{1}{4\gamma p^2(\pi^\star)} e_{\pi^\star}$, we compute

$$\boldsymbol{\delta} := \sum_\pi p(\pi) g_\pi(p) = \left\{-\frac{1}{4\gamma p(\pi)}\right\}_{\pi\in\Pi} = \left\{-\frac{\lambda}{4\gamma} - \widehat{f}(\widehat{\pi}) + \widehat{f}(\pi)\right\}_{\pi\in\Pi},$$

which has $\boldsymbol{\delta} \in \mathrm{co}(\{\nabla_p g_{\pi^\star}(p)\}_{\pi^\star\in\Pi})$. By choosing $\nu = \frac{\lambda}{4\gamma} + \widehat{f}(\widehat{\pi})$, we have

$$\nu\mathbf{1} - \widehat{f} + \boldsymbol{\delta} = \mathbf{0},$$

so (4.22) is satisfied.

□

[12] If $p \in \Delta(\Pi)$, the KKT condition that $\frac{d}{d\nu}\mathcal{L}(p,\nu) = 0$ is already satisfied.

4.3 Decision-Estimation Coefficient: Examples

We now show how to bound the Decision-Estimation Coefficient for a number of examples beyond finite-armed bandits—some familiar and others new—and show how this leads to bounds on regret via E2D.

Approximately solving the DEC. Before proceeding, let us mention that to apply E2D, it is not necessary to exactly solve the minimax problem (4.15). Instead, let us say that a distribution $p = p(\widehat{f}, \gamma)$ *certifies an upper bound* on the DEC if, given $\widehat{f}$ and $\gamma > 0$, it ensures that

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\Big] \le \overline{\mathsf{dec}}_\gamma(\mathcal{F}, \widehat{f})$$

for some known upper bound $\overline{\mathsf{dec}}_\gamma(\mathcal{F}, \widehat{f}) \ge \mathsf{dec}_\gamma(\mathcal{F}, \widehat{f})$. In this case, letting $\overline{\mathsf{dec}}_\gamma(\mathcal{F}) := \sup_{\widehat{f}} \overline{\mathsf{dec}}_\gamma(\mathcal{F}, \widehat{f})$, it is simple to see that if we use this distribution $p^t = p(\widehat{f}^t, \gamma)$ within E2D, we have

$$\mathbf{Reg} \le \overline{\mathsf{dec}}_\gamma(\mathcal{F}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta).$$

4.3.1 Cheating Code

For a first example, we show that the DEC leads to regret bounds that scale with $\log(A)$ for the cheating code example in Example 4.2; that is, unlike UCB and posterior sampling, the DEC correctly adapts to the structured of this problem.

Proposition 15 (DEC for Cheating Code): Consider the cheating code in Example 4.2. For this class $\mathcal{F}$, we have

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{\log_2(A)}{\gamma}.$$

Note that while the strategy p in Proposition 15 certifies a bound on the DEC, it is not necessarily the exact minimizer, and hence the distributions $p^1, \dots, p^T$ played by E2D may be different. Nonetheless, since the regret of E2D is bounded by the DEC, this result (via Proposition 13) implies that its regret is bounded by $\mathbf{Reg} \lesssim \sqrt{\log_2(A) T \log|\mathcal{F}|}$. Using a slightly more refined version of the E2D algorithm [38], one can improve this to match the $\log(T)$ regret bound given in Example 4.2.

Proof of Proposition 15. To simplify exposition, we present a bound on $\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f})$ for this example only for $\widehat{f} \in \mathcal{F}$, not for $\widehat{f} \in \mathrm{co}(\mathcal{F})$. A similar approach (albeit with a slightly different choice for p) leads to the same bound on $\mathsf{dec}_\gamma(\mathcal{F})$. Let $\widehat{f} \in \mathcal{F}$ and $\gamma > 0$ be given, and define

$$p = (1 - \varepsilon)\pi_{\widehat{f}} + \varepsilon \cdot \mathrm{unif}(\mathcal{C}).$$

We will show that if we choose $\varepsilon = 2\frac{\log_2(A)}{\gamma}$, this strategy certifies that

$$\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) \lesssim \frac{\log_2(A)}{\gamma}.$$

Let $f \in \mathcal{F}$ be fixed, and consider the value

$$\mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\Big].$$

We consider two cases. First the first, if $\pi_f = \pi_{\widehat{f}}$, then we can upper bound

$$\mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\Big] \leq \mathbb{E}_{\pi \sim p}[f(\pi_f) - f(\pi)] = \mathbb{E}_{\pi \sim p}\Big[f(\pi_{\widehat{f}}) - f(\pi)\Big] \leq 2\varepsilon,$$

since $f \in [-1, 1]$.

For the second case, suppose that $\pi_f \neq \pi_{\widehat{f}}$. We begin by bounding

$$\mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\Big] \leq 2 - \gamma \cdot \mathbb{E}_{\pi \sim p}\Big[(f(\pi) - \widehat{f}(\pi))^2\Big],$$

using that $f \in [-1, 1]$. To proceed, we want to argue that the negative offset term above is sufficiently large; informally, this means that we are exploring "enough". Observe that since $\pi_f \neq \pi_{\widehat{f}}$, if we let $b_1, \ldots, b_{\log_2(A)}$ and $b'_1, \ldots, b'_{\log_2(A)}$ denote the binary representations for π_f and $\pi_{\widehat{f}}$, there exists i such that $b_i \neq b'_i$. As a result, we have

$$\mathbb{E}_{\pi \sim p}\Big[(f(\pi) - \widehat{f}(\pi))^2\Big] \geq \frac{\varepsilon}{\log_2(A)}(f(c_i) - \widehat{f}(c_i))^2 = \frac{\varepsilon}{\log_2(A)}(b_i - b'_i)^2 = \frac{\varepsilon}{\log_2(A)}.$$

We conclude that in the second case,

$$\mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\Big] \leq 2 - \gamma \frac{\varepsilon}{\log_2(A)}.$$

Putting the cases together, we have

$$\mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\Big] \leq \max\Big\{2\varepsilon, 2 - \gamma \frac{\varepsilon}{\log_2(A)}\Big\}.$$

To balance these terms, we set

$$\varepsilon = 2\frac{\log_2(A)}{\gamma},$$

which leads to the result. □

4.3.2 Linear Bandits

We next consider the problem of linear bandits *linear bandit* [2, 10, 24, 22, 1], which is a special case of the linear contextual bandit problem we saw in Section 3. We let Π be arbitrary, and define $\mathcal{F} = \{\pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta\}$, where $\Theta \subseteq \mathsf{B}_2^d(1)$ is a parameter set and $\phi : \Pi \to \mathsf{B}_2^d(1)$ is a fixed feature map that is known to the learner.

To prove bounds on the DEC for this setting, we make use of a primitive from convex analysis and experimental design known as the *G-optimal design*.

Proposition 16 (G-optimal design [45]): For any compact set $\mathcal{Z} \subseteq \mathbb{R}^d$ with $\dim \operatorname{span}(\mathcal{Z}) = d$, there exists a distribution $p \in \Delta(\mathcal{Z})$, called the *G-optimal design*, which has

$$\sup_{z \in \mathcal{Z}} \langle \Sigma_p^{-1} z, z \rangle \leq d, \tag{4.23}$$

where $\Sigma_p := \mathbb{E}_{z\sim p}\big[zz^\top\big]$.

The G-optimal design ensures coverage in every direction of the decision space, generalizing the notion of uniform exploration for finite action spaces. In this sense, it can be thought of as a "universal" exploratory distribution for linearly structured action spaces. Special cases include:

- When $\mathcal{Z} = \Delta([A])$, we can take $p = \mathrm{unif}(e_1, \ldots, e_A)$ as an optimal design
- When $\mathcal{Z} = \mathsf{B}_2^d(1)$, we can again take $p = \mathrm{unif}(e_1, \ldots, e_A)$ as an optimal design.
- For any positive definite matrix $A \succ 0$, the set $\mathcal{Z} = \big\{z \in \mathbb{R}^d \mid \langle Az, z\rangle \le 1\big\}$ is an ellipsoid. Letting $\lambda_1, \ldots, \lambda_d$ and $v_1, \ldots, v_d$ denote the eigenvalues and eigenvectors for A, respectively, the distribution $p = \mathrm{unif}(\lambda_1^{-1/2}v_1, \ldots, \lambda_d^{-1/2}v_d)$ is an optimal design.

To see how the G-optimal design can be used for exploration, consider the following generalization of the ε-greedy algorithm.

- Let $q \in \Delta(\Pi)$ be the G-optimal design for the set $\{\phi(\pi)\}_{\pi\in\Pi}$.
- At each step t, obtain $\widehat{f}^t$ from a supervised estimation oracle. Play $\widehat{\pi}^t = \pi_{\widehat{f}^t}$ with probability $1-\varepsilon$, and sample $\pi^t \sim q$ otherwise.

It is straightforward to show that this strategy gives $\mathbf{Reg} \lesssim d^{1/3}T^{2/3}\log|\mathcal{F}|$ for linear bandits. The basic idea is to replace (3.29) in the proof of Proposition 8 with the optimal design property (4.23), using that the reward functions under consideration are linear. The intuition is that even though we are no longer guaranteed to explore every single action with some minimum probability, by exploring with the optimal design, we ensure that some fraction of the data we collect covers every possible direction in action space to the greatest extent possible.

The following result shows that by combining optimal design inverse gap weighting, we can obtain a d/γ bound on the DEC, which leads to an improved $\sqrt{dT}$ regret bound.

Proposition 17 (DEC for Linear Bandits): Consider the linear bandit setting. Let a linear function $\widehat{f}$ and $\gamma > 0$ be given, consider the following distribution p:

- Define $\bar{\phi}(\pi) = \phi(\pi)/\sqrt{1 + \frac{\gamma}{d}\big(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi)\big)}$, where $\pi_{\widehat{f}} = \arg\max_{\pi\in\Pi} \widehat{f}(\pi)$.
- Let $\bar{q} \in \Delta(\Pi)$ be the G-optimal design for the set $\{\bar{\phi}(\pi)\}_{\pi\in\Pi}$, and define $q = \frac{1}{2}\bar{q} + \frac{1}{2}\mathbb{I}_{\pi_{\widehat{f}}}$.
- For each $\pi \in \Pi$, set
$$p(\pi) = \frac{q(\pi)}{\lambda + \frac{\gamma}{d}(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi))},$$
where $\lambda \in [1/2, 1]$ is chosen such that $\sum_\pi p(\pi) = 1$.[a]

This strategy certifies that
$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{d}{\gamma}.$$

[a] The normalizing constant λ always exists because we have $\frac{1}{2\lambda} \leq \sum_\pi p(\pi) \leq \frac{1}{\lambda}$.

One can show that $\mathsf{dec}_\gamma(\mathcal{F}) \gtrsim \frac{d}{\gamma}$ for this setting as well, so this is the best bound we can hope for. Combining this result with Proposition 13 and using the averaged exponential weights algorithm for estimation as in (4.18) gives $\mathbf{Reg} \lesssim \sqrt{dT \log(|\mathcal{F}|/\delta)}$.

Proof of Proposition 17. Fix $f \in \mathcal{F}$. Let us abbreviate $\eta = \frac{\gamma}{d}$. As in Proposition 9, we break regret into three terms:

$$\mathbb{E}_{\pi\sim p}\Big[f(\pi_f) - f(\pi)\Big] = \underbrace{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi)\Big]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi\sim p}\Big[\widehat{f}(\pi) - f(\pi)\Big]}_{\text{(II) est error on policy}} + \underbrace{f(\pi_f) - \widehat{f}(\pi_{\widehat{f}})}_{\text{(III) est error at opt}} .$$

The first term captures the loss in exploration that we would incur if $\widehat{f}$ we true the reward function, and is equal to:

$$\sum_\pi \frac{q(\pi)(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi))}{\lambda + \eta\left(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi)\right)} \leq \sum_\pi \frac{q(\pi)}{\eta} \leq \frac{1}{\eta},$$

and the second term, as before, is at most

$$\sqrt{\mathbb{E}_{\pi\sim p}(\widehat{f}(\pi) - f(\pi))^2} \leq \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{\pi\sim p}(\widehat{f}(\pi) - f(\pi))^2.$$

The third term can be written as

$$\text{(III)} = f(\pi_f) - \widehat{f}(\pi_f) - (\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi_f)) = \big\langle \theta - \widehat{\theta}, \phi(\pi_f)\big\rangle - (\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi_f)),$$

where $\theta, \widehat{\theta} \in \Theta$ are parameters such that $f(\pi) = \langle \theta, \phi(\pi)\rangle$ and $\widehat{f}(\pi) = \langle \widehat{\theta}, \phi(\pi)\rangle$. Defining $\Sigma_p = \mathbb{E}_{\pi\sim p}\big[\phi(\pi)\phi(\pi)^\top\big]$, we can bound

$$\begin{aligned}\big\langle \theta - \widehat{\theta}, \phi(\pi_f)\big\rangle &= \big\langle \Sigma_p^{1/2}(\theta - \widehat{\theta}), \Sigma_p^{-1/2}\phi(\pi_f)\big\rangle \\ &\leq \|\Sigma_p^{1/2}(\theta - \widehat{\theta})\|_2 \|\Sigma_p^{-1/2}\phi(\pi_f)\|_2 \quad \leq \frac{\gamma}{2}\|\Sigma_p^{1/2}(\theta - \widehat{\theta})\|_2^2 + \frac{1}{2\gamma}\|\Sigma_p^{-1/2}\phi(\pi_f)\|_2^2.\end{aligned}$$

Note that $\|\Sigma_p^{1/2}(\theta - \widehat{\theta})\|_2^2 = \mathbb{E}_{\pi\sim p}[(\widehat{f}(\pi) - f(\pi))^2]$ and $\|\Sigma_p^{-1/2}\phi(\pi_f)\|_2^2 = \langle \phi(\pi_f), \Sigma_p^{-1}\phi(\pi_f)\rangle$, so we have

$$\text{(III)} \leq \frac{\gamma}{2}\mathbb{E}_{\pi\sim p}[(\widehat{f}(\pi) - f(\pi))^2] + \underbrace{\frac{1}{2\gamma}\langle \phi(\pi_f), \Sigma_p^{-1}\phi(\pi_f)\rangle - (\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi_f))}_{\text{(IV)}}.$$

To proceed, observe that

$$\begin{aligned}\Sigma_p &\succeq \frac{1}{2}\sum_\pi \frac{\bar{q}(\pi)}{\lambda + \eta(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi))}\phi(\pi)\phi(\pi)^\top \\ &\succeq \frac{1}{2}\sum_\pi \frac{\bar{q}(\pi)}{1 + \eta(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi))}\phi(\pi)\phi(\pi)^\top \succeq \frac{1}{2}\sum_\pi \bar{q}(\pi)\bar{\phi}(\pi)\bar{\phi}(\pi)^\top =: \frac{1}{2}\overline{\Sigma}_{\bar{q}}\end{aligned}$$

This means that we can bound

$$\begin{aligned}\langle\phi(\pi_f),\Sigma_p^{-1}\phi(\pi_f)\rangle &\le 2\langle\phi(\pi_f),\overline{\Sigma}_{\bar{q}}^{-1}\phi(\pi_f)\rangle\\ &= 2(1+\eta(\widehat{f}(\pi_{\widehat{f}})-\widehat{f}(\pi_f))\langle\bar{\phi}(\pi_f),\overline{\Sigma}_{\bar{q}}^{-1}\bar{\phi}(\pi_f)\rangle\\ &\le 2d(1+\eta(\widehat{f}(\pi_{\widehat{f}})-\widehat{f}(\pi_f)),\end{aligned}$$

where the last line uses that $\bar{q}$ is the G-optimal design for $\{\bar{\phi}(\pi)\}_{\pi\in\Pi}$. We conclude that

$$(\mathrm{IV})\le\frac{2d}{2\gamma}+\frac{2d\eta}{2\gamma}(\widehat{f}(\pi_{\widehat{f}})-\widehat{f}(\pi_f))-(\widehat{f}(\pi_{\widehat{f}})-\widehat{f}(\pi_f))\le\frac{d}{\gamma}.$$

□

Remark 15: In fact, it can be shown [34] that when $\Theta=\mathbb{R}^d$, the *exact* minimizer of the DEC for linear bandits is given by

$$p=\underset{p\in\Delta(\Pi)}{\arg\max}\left\{\mathbb{E}_{\pi\sim p}\big[\widehat{f}(\pi)\big]+\frac{1}{4\gamma}\log\det(\mathbb{E}_{\pi\sim p}[\phi(\pi)\phi(\pi)^\top])\right\}.$$

4.3.3 Nonparametric Bandits

For all of the examples so far, we have shown that

$$\mathsf{dec}_\gamma(\mathcal{F})\lesssim\frac{\mathsf{eff\text{-}dim}(\mathcal{F},\Pi)}{\gamma},$$

where $\mathsf{eff\text{-}dim}(\mathcal{F},\Pi)$ is some quantity that (informally) reflects the amount of exploration required for the class $\mathcal{F}$ under consideration (A for bandits, $\log_2(A)$ for the cheating code, and d for linear bandits). In general though, the Decision-Estimation Coefficient does not always shrink at a γ^{-1} rate, and can have slower decay for problems where the optimal rate is worse than $\sqrt{T}$. We now consider such a setting: a standard *nonparametric* bandit problem called *Lipschitz bandits in metric spaces* [11, 47].

We take Π to be a metric space equipped with metric ρ, and define

$$\mathcal{F}=\{f:\Pi\to[0,1]\mid f\text{ is 1-Lipschitz w.r.t }\rho\}.$$

We give a bound on the Decision-Estimation Coefficient which depends on the *covering number* for the space Π (with respect to the metric ρ). Let us say that $\Pi'\subseteq\Pi$ is an ε-cover with respect to ρ if

$$\forall\pi\in\Pi\quad\exists\pi'\in\Pi'\quad\text{s.t.}\quad\rho(\pi,\pi')\le\varepsilon,$$

and let $\mathcal{N}_\rho(\Pi,\varepsilon)$ denote the size of the smallest such cover.

Proposition 18 (DEC for Lipschitz Bandits): Consider the Lipschitz bandit setting, and suppose that there exists $d>0$ such that $\mathcal{N}_\rho(\Pi,\varepsilon)\le\varepsilon^{-d}$ for all $\varepsilon>0$. Let $\widehat{f}:\Pi\to[0,1]$ and $\gamma\ge1$ be given and consider the following distribution:

1. Let $\Pi'\subseteq\Pi$ witness the covering number $\mathcal{N}_\rho(\Pi,\varepsilon)$ for a parameter $\varepsilon>0$.

2. Let p be the result of applying the inverse gap weighting strategy in (3.35) to $\widehat{f}$, restricted to the (finite) decision space Π'.

By setting $\varepsilon \propto \gamma^{-\frac{1}{d+1}}$, this strategy certifies that

$$\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) \lesssim \gamma^{-\frac{1}{d+1}}.$$

Ignoring dependence on $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$, this result leads to regret bounds that scale as $T^{\frac{d+1}{d+2}}$ (after tuning γ in Proposition 13), which cannot be improved.

Proof of Proposition 18. Let $f \in \mathcal{F}$ be fixed. Let Π' be the ε-cover for Π. Since f is 1-Lipschitz, for all $\pi \in \Pi$ there exists a corresponding covering element $\iota(\pi) \in \Pi'$ such that $\rho(\pi, \iota(\pi)) \leq \varepsilon$, and consequently for any distribution p,

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}[f(\pi_f) - f(\pi)] &\leq \mathbb{E}_{\pi\sim p}[f(\iota(\pi_f)) - f(\pi)] + |f(\pi_f) - f(\iota(\pi_f))| \\
&\leq \mathbb{E}_{\pi\sim p}[f(\iota(\pi_f)) - f(\pi)] + \rho(\pi_f, \iota(\pi_f)) \\
&\leq \mathbb{E}_{\pi\sim p}[f(\iota(\pi_f)) - f(\pi)] + \varepsilon.
\end{aligned}$$

At this point, since $\iota(\pi_f) \in \Pi'$, Proposition 9 ensures that if we choose p using inverse gap weighting over Π', we have

$$\mathbb{E}_{\pi\sim p}[f(\iota(\pi_f)) - f(\pi)] \leq \frac{|\Pi'|}{\gamma} + \gamma \cdot \mathbb{E}_{\pi\sim p}\Big[(f(\pi) - \widehat{f}(\pi))^2\Big].$$

From our assumption on the growth of $\mathcal{N}_\rho(\Pi, \varepsilon)$, $|\Pi'| \leq \varepsilon^{-d}$, so the value is at most

$$\varepsilon + \frac{\varepsilon^{-d}}{\gamma}.$$

We choose $\varepsilon \propto \gamma^{-\frac{1}{d+1}}$ to balance the terms, leading to the result.

□

4.3.4 Further Examples

We state the following additional upper bounds on the DEC without proof.

Example 4.3 (Decision-Estimation Coefficient subsumes Eluder Dimension)**.** Consider any class $\mathcal{F}$ with values in $[0, 1]$. For all $\gamma \geq e$, we have

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \inf_{\varepsilon>0}\left\{\varepsilon + \frac{\mathsf{Edim}(\mathcal{F} - \mathcal{F}, \varepsilon)\log^2(\gamma)}{\gamma}\right\} + \gamma^{-1}. \tag{4.24}$$

◁

As a special case, this example implies that $\mathsf{E2D}$ enjoys a regret bound for generalized linear bandits similar to that of UCB.

Example 4.4 (Bandits with Concave Rewards)**.** The concave (or convex, if one considers losses rather than rewards) bandit problem [46, 31, 4, 17, 19, 49] is a generalization of the linear bandit. We take $\Pi \subseteq \mathsf{B}_2^d(1)$ and define

$$\mathcal{F} = \{f : \Pi \to [0, 1] \mid f \text{ is concave and 1-Lipschitz w.r.t } \ell_2\}.$$

For this setting, whenever $\mathcal{F} \subseteq (\Pi \to [0,1])$, results of Lattimore [49] imply that

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{d^4}{\gamma} \cdot \mathrm{polylog}(d, \gamma) \tag{4.25}$$

for all $\gamma > 0$. ◁

For the function class

$$\mathcal{F} = \left\{ f(\pi) = -\mathsf{relu}(\langle \phi(\pi), \theta \rangle) \mid \theta \in \Theta \subset \mathsf{B}_2^d(1) \right\},$$

(4.25) leads to a $\sqrt{\mathrm{poly}(d)T}$ regret bound for E2D. This highlights a case where the Eluder dimension is overly pessimistic, since we saw that it grows exponentially for this class.

4.4 Relationship to Optimism and Posterior Sampling

We close this section by highlighting some connections between the Decision-Estimation Coefficient and E2D and other techniques we have covered so far: Optimism (UCB) and Posterior Sampling.

4.4.1 Connection to Optimism

The E2D meta-algorithm and the Decision-Estimation Coefficient can be combined with the idea of *confidence sets* that we used in the UCB algorithm. Consider the following variant of E2D.

```
Estimation-to-Decisions (E2D) with Confidence Sets
Input: Exploration parameter $\gamma > 0$, confidence radius $\beta > 0$.
for $t = 1, \ldots, T$ do
    Obtain $\widehat{f}^t$ from online regression oracle with $(\pi^1, r^1), \ldots, (\pi^{t-1}, r^{t-1})$.
    Set
        $\mathcal{F}^t = \left\{ f \in \mathcal{F} \mid \sum_{i<t} \mathbb{E}_{\pi^i \sim p^i}\left[(\widehat{f}^i(\pi^i) - f^\star(\pi^i))^2\right] \le \beta \right\}.$
    Compute
        $p^t = \arg\min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}^t} \mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}^t(\pi))^2\right].$
    Select action $\pi^t \sim p^t$.
```

This strategy is the same as the basic E2D algorithm, except that at each step, we compute a confidence set $\mathcal{F}^t$ and modify the minimax problem so that the max player is restricted to choose $f \in \mathcal{F}^t$.[13] With this change, the distribution p^t can be interpreted as the minimizer for $\mathsf{dec}_\gamma(\mathcal{F}^t, \widehat{f}^t)$.

To analyze this algorithm, we show that as long as $f^\star \in \mathcal{F}^t$ for all t, the same per-step analysis as in Proposition 13 goes through, with $\mathcal{F}$ replaced by $\mathcal{F}^t$. This allows us to prove the following result.

[13] Note that compared to the confidence sets used in UCB, a slight difference is that we compute $\mathcal{F}^t$ using the estimates $\widehat{f}^1, \ldots, \widehat{f}^T$ produced by the online regression oracle (this is sometimes referred to as "online-to-confidence set conversion") as opposed to using ERM; this difference is unimportant, and the later would work as well.

Proposition 19: For any $\delta \in (0,1)$ and $\gamma > 0$, if we set $\beta = \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)$, then E2D with confidence sets ensures that with probability at least $1-\delta$,

$$\mathbf{Reg} \leq \sum_{t=1}^{T} \mathsf{dec}_{\gamma}(\mathcal{F}^t) + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta). \tag{4.26}$$

This bound is never worse than the one in Proposition 13, but it can be smaller if the confidence sets $\mathcal{F}^1, \dots, \mathcal{F}^T$ shrink quickly. For a proof, see Exercise 9.

Remark 16: In fact, the regret bound in (4.26) can be shown to hold for *any* sequence of confidence sets $\mathcal{F}^1, \dots, \mathcal{F}^T$, as long as $f^\star \in \mathcal{F}^t \quad \forall t$ with probability at least $1-\delta$; the specific construction we use within the E2D variant above is chosen only for concreteness.

Relation to confidence width and UCB. It turns out that the usual UCB algorithm, which selects $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$ for $\bar{f}^t(\pi) = \max_{f \in \mathcal{F}^t} f^t(\pi)$, certifies a bound on $\mathsf{dec}_{\gamma}(\mathcal{F}^t)$ which is never worse than usual confidence width we use in the UCB analysis.

Proposition 20: The UCB strategy $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$ certifies that

$$\mathsf{dec}_0(\mathcal{F}^t) \leq \bar{f}^t(\pi^t) - \underline{f}(\pi^t). \tag{4.27}$$

Proof of Proposition 20. By choosing $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$, we have that for any $\widehat{f}$,

$$\begin{aligned}
\mathsf{dec}_0(\mathcal{F}^t, \widehat{f}) &= \inf_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{F}^t} \mathbb{E}_{\pi \sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi)\right] \\
&\leq \sup_{f \in \mathcal{F}^t}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi^t)\right] \\
&\leq \sup_{f \in \mathcal{F}^t}\left[\max_{\pi^\star} \bar{f}^t(\pi^\star) - f(\pi^t)\right] \\
&= \sup_{f \in \mathcal{F}^t}\left[\bar{f}^t(\pi^t) - f(\pi^t)\right] = \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t).
\end{aligned}$$

□

As we saw in the analysis of UCB for multi-armed bandits with $\Pi = \{1, \dots, A\}$ (Section 2.3), the confidence width in (4.27) might be large for a given round t, but by the pigeonhole argument (Lemma 8), when we sum over all rounds we have

$$\sum_{t=1}^{T} \mathsf{dec}_0(\mathcal{F}^t) \leq \sum_{t=1}^{T} \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t) \leq \widetilde{O}(\sqrt{AT}).$$

Hence, even though UCB is not the optimal strategy to minimize the DEC, it can still lead to upper bounds on regret when the confidence width shrinks sufficiently quickly. Of course, as examples like the cheating code show, we should not expect this to happen in general.

Interestingly, the bound on the DEC in Proposition 20 holds for $\gamma = 0$, which only leads to meaningful bounds on regret because $\mathcal{F}^1, \ldots, \mathcal{F}^T$ are shrinking. Indeed, Proposition 14 shows that with $\mathcal{F} = \mathbb{R}^A$, we have

$$\mathsf{dec}_\gamma(\mathcal{F}) \gtrsim \frac{A}{\gamma},$$

so the unrestricted class $\mathcal{F}$ has $\mathsf{dec}_\gamma(\mathcal{F}) \to \infty$ as $\gamma \to 0$. By allowing for $\gamma > 0$, we can prove the following slightly stronger result, which replaces $\underline{f}^t$ by $\widehat{f}^t$.

Proposition 21: For any $\gamma > 0$, the UCB strategy $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$ certifies that

$$\mathsf{dec}_\gamma(\mathcal{F}^t, \widehat{f}^t) \leq \bar{f}^t(\pi^t) - \widehat{f}^t(\pi^t) + \frac{1}{4\gamma}.$$

Proof of Proposition 21. This is a slight generalization of the proof of Proposition 20. By choosing $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$, we have

$$\begin{aligned}
\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}^t) &= \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}_t} \mathbb{E}_{\pi \sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma \cdot (\widehat{f}^t(\pi) - f(\pi))^2\right] \\
&\leq \max_{f \in \mathcal{F}_t}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi^t) - \gamma \cdot (\widehat{f}^t(\pi^t) - f(\pi^t))^2\right] \\
&\leq \max_{f \in \mathcal{F}_t}\left[\bar{f}^t(\pi^t) - f(\pi^t) - \gamma \cdot (\widehat{f}^t(\pi^t) - f(\pi^t))^2\right] \\
&= \underbrace{\max_{f \in \mathcal{F}_t}\left[\widehat{f}^t(\pi^t) - f(\pi^t) - \gamma \cdot (\widehat{f}^t(\pi^t) - f(\pi^t))^2\right]}_{\leq \frac{1}{4\gamma}} + \bar{f}^t(\pi^t) - \widehat{f}^t(\pi^t).
\end{aligned}$$

□

4.4.2 Connection to Posterior Sampling

The Decision-Estimation Coefficient (4.15) is a min-max optimization problem, which we have mentioned can be interpreted as a game in which the learner (the "min" player) aims to find a decision distribution p that optimally trades off regret and information acquisition in the face of an adversary (the "max" player) that selects a worst-case model in $\mathcal{M}$. We can define a natural *dual* (or, max-min) analogue of the DEC via

$$\underline{\mathsf{dec}}_\gamma(\mathcal{F}, \widehat{f}) = \sup_{\mu \in \Delta(\mathcal{F})} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{f \sim \mu} \mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\right]. \tag{4.28}$$

The dual Decision-Estimation Coefficient has the following Bayesian interpretation. The adversary selects a *prior* distribution μ over models in $\mathcal{M}$, and the learner (with knowledge of the prior) finds a decision distribution p that balances the average tradeoff between regret and information acquisition when the underlying model is drawn from μ.

Using the minimax theorem (Lemma 42), one can show that the Decision-Estimation Coefficient and its Bayesian counterpart coincide.

Proposition 22 (Equivalence of primal and dual DEC): Under mild regularity conditions, we have

$$\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) = \underline{\mathsf{dec}}_\gamma(\mathcal{F}, \widehat{f}). \tag{4.29}$$

Thus, any bound on the dual DEC immediately yields a bound on the primal DEC. This perspective is useful because it allows us to bring existing tools for Bayesian bandits and reinforcement learning to bear on the primal Decision-Estimation Coefficient. As an example, we can adapt the posterior sampling/probability matching strategy introduced in Section 2. When applied to the Bayesian DEC—this approach selects p to be the action distribution induced by sampling $f \sim \mu$ and selecting π_f. Using Lemma 9, one can show that this strategy certifies that

$$\underline{\mathsf{dec}}_\gamma(\mathcal{F}) \lesssim \frac{|\Pi|}{\gamma}$$

for the multi-armed bandit. In fact, existing analysis techniques for the Bayesian setting can be viewed as implicitly providing bounds on the dual Decision-Estimation Coefficient [64, 18, 17, 65, 50, 49]. Notably, the dual DEC is always bounded by a Bayesian complexity measure known as the *information ratio*, which is used throughout the literature on Bayesian bandits and reinforcement learning [35].

Beyond the primal and dual Decision-Estimation Coefficient, there are deeper connections between the DEC and Bayesian algorithms, including a Bayesian counterpart to the E2D algorithm itself [35].

4.5 Incorporating Contexts*

The Decision-Estimation Coefficient and E2D algorithm trivially extend to handle *contextual* structured bandits. This approach generalizes the SquareCB method introduced in Section 3 from finite action spaces to general action spaces. Consider the following protocol.

```
Contextual Structured Bandit Protocol
for $t = 1, \dots, T$ do
    Observe context $x^t \in \mathcal{X}$.
    Select decision $\pi^t \in \Pi$.            //$\Pi$ is large and potentially continuous.
    Observe reward $r^t \in \mathbb{R}$.
```

This is the same as the contextual bandit protocol in Section 3, except that we allow Π to be large and potentially continuous. As in that section, we allow the contexts $x^1, \dots, x^T$ to be generated in an arbitrary, potentially adversarial fashion, but assume that

$$r^t \sim M^\star(\cdot \mid x^t, \pi^t),$$

and define $f^\star(x, \pi) = \mathbb{E}_{r \sim M^\star(\cdot|x,\pi)}$. We assume access to a function class $\mathcal{F}$ such that $f^\star \in \mathcal{F}$, and assume access to an estimation oracle for $\mathcal{F}$ that ensures that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}(\widehat{f}^t(x^t, \pi^t) - f^\star(x^t, \pi^t))^2 \le \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta).$$

For $f \in \mathcal{F}$, we define $\pi_f(x) = \arg\max_{\pi \in \Pi} f(x, \pi)$.

To extend the E2D algorithm to this setting, at each time t we solve the minimax problem corresponding to the DEC, but condition on the context x^t.

```
Estimation-to-Decisions (E2D) for Contextual Structured Bandits
Input: Exploration parameter $\gamma > 0$.
for $t = 1, \ldots, T$ do
    Observe $x^t \in \mathcal{X}$.
    Obtain $\widehat{f}^t$ from online regression oracle with $(x^1, \pi^1, r^1), \ldots, (x^{t-1}, \pi^{t-1}, r^{t-1})$.
    Compute
        $p^t = \arg\min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[ f(x^t, \pi_f(x^t)) - f(x^t, \pi) - \gamma \cdot (f(x^t, \pi) - \widehat{f}^t(x^t, \pi))^2 \right].$
    Select action $\pi^t \sim p^t$.
```

For $x \in \mathcal{X}$, define

$$\mathcal{F}(x, \cdot) = \{ f(x, \cdot) \mid f \in \mathcal{F} \}$$

as the projection of $\mathcal{F}$ onto $x \in \mathcal{X}$. The following result shows that whenever the DEC is bounded conditionally—that is, whenever it is bounded for $\mathcal{F}(x, \cdot)$ for all x—this strategy has low regret.

Proposition 23: The E2D algorithm with exploration parameter $\gamma > 0$ guarantees that

$$\mathbf{Reg} \leq \sup_{x \in \mathcal{X}} \mathsf{dec}_{\gamma}(\mathcal{F}(x, \cdot)) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta), \tag{4.30}$$

We omit the proof of this result, which is nearly identical to that of Proposition 13. The basic idea is that for each round, once we condition on the context x^t, the DEC allows us to link regret to estimation error in the same fashion non-contextual setting.

We showed in Proposition 14 that the IGW distribution exactly solves the DEC minimax problem when $\mathcal{F} = \mathbb{R}^A$. Hence, the SquareCB algorithm in Section 3 is precisely the special case of Contextual E2D in which $\mathcal{F} = \mathbb{R}^A$.

Going beyond the finite-action setting, it is simplest to interpret Proposition 23 when $\mathcal{F}(x, \cdot)$ has the same structure for all contexts. One example is *contextual bandits with linearly structured action spaces*. Here, we take

$$\mathcal{F} = \{ f(x, a) = \langle \phi(x, a), g(x) \rangle \mid g \in \mathcal{G} \},$$

where $\phi(x, a) \in \mathbb{R}^d$ is a fixed feature map and $\mathcal{G} \subset (\mathcal{X} \to \mathsf{B}_2^d(1))$ is an arbitrary function class. This setting generalizes the linear contextual bandit problem from Section 3, which corresponds to the case where $\mathcal{G}$ is a set of constant functions. We can apply Proposition 17 to conclude that $\sup_{x \in \mathcal{X}} \mathsf{dec}_{\gamma}(\mathcal{F}(x, \cdot)) \lesssim \frac{d}{\gamma}$, so that Proposition 23 gives $\mathbf{Reg} \lesssim \sqrt{dT \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)}$.

4.6 Additional Properties of the Decision-Estimation Coefficient*

Proposition 24: Using the minimax theorem, one can show that

$$\sup_{\widehat{f}\in\mathrm{co}(\mathcal{F})} \mathsf{dec}_{\gamma}(\mathcal{F},\widehat{f}) \leq \sup_{\widehat{f}:\Pi\to\mathbb{R}} \mathsf{dec}_{\gamma}(\mathcal{F},\widehat{f}) \leq \sup_{\widehat{f}\in\mathrm{co}(\mathcal{F})} \mathsf{dec}_{\gamma/4}(\mathcal{F},\widehat{f}).$$

4.7 Exercises

Exercise 8 (Posterior Sampling for Multi-Armed Bandits): Prove that for the standard multi-armed bandit,

$$\underline{\mathsf{dec}}_{\gamma}(\mathcal{F}) \lesssim \frac{|\Pi|}{\gamma},$$

by using the Posterior Sampling strategy (select p to be the action distribution induced by sampling $f \sim \mu$ and selecting π_f), and applying the decoupling lemma (Lemma 9). Recall that here, $\underline{\mathsf{dec}}_{\gamma}(\mathcal{F})$ is the "maxmin" version of the DEC (4.28).

Exercise 9: Prove Proposition 19.

Exercise 10: Prove Proposition 24.

5. REINFORCEMENT LEARNING: BASICS

We now introduce the framework of *reinforcement learning*, encompasses a rich set of dynamic, stateful decision making problems. Consider the task of repeated medical treatment assignment, depicted in Figure 4. To make the setting more realistic, it is natural to allow the decision-maker to apply *multi-stage strategies* rather simple one-shot decisions such as "prescribe a painkiller." In principle, in the language of structured bandits, nothing is preventing us from having each decision π^t be a complex multi-stage treatment strategy that, at each stage, acts on the patient's dynamic state, which evolves as a function of the treatments at previous stages. As an example, intermediate actions of the type "if patient's blood pressure is above X then do Y" can form a decision tree that defines the complex strategy π^t. Methods from the previous lectures provide guarantees for such a setting, as long as we have a succinct model of expected rewards. What sets RL apart from structured bandits is the *additional information* about the intermediate state transitions and intermediate rewards. This information facilitates *credit assignment*, the mechanism for recognizing which of the actions led to the overall (composite) decision to be good or bad. This extra information can reduce what would otherwise be exponential sample complexity in terms of the number of stages, states, and actions in multi-stage decision making.

This section is structured as follows. We first present the formal reinforcement learning framework and present basic principles including Bellman optimality and dynamic programming, which facilitate efficiently computing optimal decisions when the environment is known. We then consider the case in which the environment is unknown, and give algorithms for perhaps the simplest reinforcement learning setting, *tabular reinforcement learning*, where the state and action spaces are finite. Algorithms for more complex reinforcement learning settings are given in Section 5.

5.1 Finite-Horizon Episodic MDP Formulation

We consider an episodic finite-horizon reinforcement learning framework. With H denoting the *horizon*, a *Markov Decision Process* (MDP) M takes the form

$$M = \left\{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\},$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space,

$$P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$$

is the probability transition kernel at step h,

$$R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$$

is the reward distribution, and $d_1 \in \Delta(\mathcal{S}_1)$ is the initial state distribution. We allow the reward distribution and transition kernel to vary across MDPs, but assume for simplicity that the initial state distribution is fixed and known.

For a fixed MDP M, an *episode* proceeds under the following protocol. At the beginning of the episode, the learner selects a randomized, non-stationary *policy*

$$\pi = (\pi_1, \ldots, \pi_H),$$

where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$; we let Π_{RNS} for "randomized, non-stationary" denote the set of all such policies. The episode then evolves through the following process, beginning from $s_1 \sim d_1$. For $h = 1, \ldots, H$:

- $a_h \sim \pi_h(s_h)$.
- $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(\cdot \mid s_h, a_h)$.

For notational convenience, we take s_{H+1} to be a deterministic terminal state. The *Markov* property refers to the fact that under this evolution,

$$\mathbb{P}^M(s_{t+1} = s' \mid s_t, a_t) = \mathbb{P}^M(s_{t+1} = s' \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_1, a_1).$$

The value for a policy π under M is given by

$$f^M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^H r_h\right], \tag{5.1}$$

where $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process above. We define an optimal policy for model M as

$$\pi_M \in \operatorname*{arg\,max}_{\pi \in \Pi_{\mathrm{RNS}}} f^M(\pi). \tag{5.2}$$

Value functions. Maximization in (5.2) is a daunting task, since each policy π is a complex multi-stage object. It is useful to define intermediate "reward-to-go" functions to start breaking this complex task into smaller sub-tasks. Specifically, for a given model M and policy π, we define the *state-action value function* and *state value function* via

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^H r_{h'} \mid s_h = s, a_h = a\right], \quad \text{and} \quad V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^H r_{h'} \mid s_h = s\right].$$

Hence, the definition in (5.1) reads

$$f^M(\pi) = \mathbb{E}_{s\sim d_1, a\sim\pi(s)}\left[Q_1^{M,\pi}(s,a)\right] = \mathbb{E}_{s\sim d_1}\left[V_1^{M,\pi}(s)\right]$$

Online RL. For reinforcement learning, our main focus will be on what is called the *online reinforcement learning problem*, in which we interact with an unknown MDP $M^\star$ for T episodes. For each episode $t = 1, \ldots, T$, the learner selects a policy $\pi^t \in \Pi_{\mathrm{RNS}}$. The policy is executed in the MDP $M^\star$, and the learner observes the resulting trajectory

$$\tau^t = (s_1^t, a_1^t, r_1^t), \ldots, (s_H^t, a_H^t, r_H^t).$$

The goal is to minimize the total regret

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] \tag{5.3}$$

against the optimal policy $\pi_{M^\star}$ for $M^\star$.

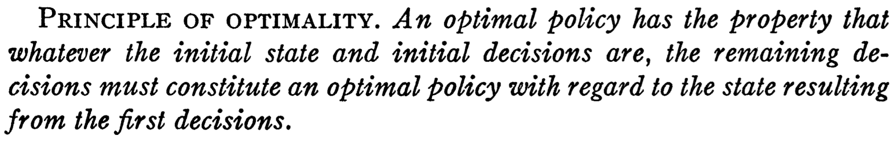
The online RL framework is a strict generalization of (structured) bandits and contextual bandits (with i.i.d. contexts). Indeed, if $\mathcal{S} = \{s_0\}$ and $H = 1$, each episode amounts to choosing an action $a^t \in \mathcal{A}$ and observing a reward r^t with mean $f^M(a^t)$, which is precisely a bandit problem. On the other hand, taking $\mathcal{S} = \mathcal{X}$ and $H = 1$ puts us in the setting of contextual bandits, with d_1 being the distribution of contexts. In both cases, the notion of regret (5.3) coincides with the notion of regret in the respective setting.

We mention in passing that many alternative formulations for Markov decision processes and for the reinforcement learning problem appear throughout the literature. For example, MDPs can be studied with infinite horizon (with or without discounting), and an alternative to minimizing regret is to consider *PAC-RL* which aims to minimize the sub-optimality of a final output policy produced after exploring for T rounds.

5.2 Planning via Dynamic Programming

In some reinforcement learning problems, it is natural to assume that the true MDP $M^\star$ is known. This may be the case with games, such as chess or backgammon, where transition probabilities are postulated by the game itself. In other settings, such as robotics or medical treatment, the agent interacts with an unknown $M^\star$ and needs to learn at least some aspects of this environment. The online reinforcement learning problem described above falls in the latter category. Before attacking the learning problem, we need to understand the structure of solutions to (5.2) in the case where $M^\star$ is known to the decision-maker. In this section, we show that the problem of maximizing $f^M(\pi)$ over $\pi \in \Pi$ in a known model M (known as *planning*) can be solved efficiently via the principle of *dynamic programming*. Dynamic programming can be viewed as solving the problem of credit assignment by breaking down a complex multi-stage decision (policy) into a sequence of small decisions.

We start by observing that the optimal policy π_M in (5.2) may not be uniquely defined. For instance, if d_1 assigns zero probability to some state s_1, the behavior of π_M on this state is immaterial. In what follows, we introduce a fundamental result, Proposition 25, which guarantees existence of an optimal policy $\pi_M = (\pi_{M,1}, \ldots, \pi_{M,H})$ that maximizes $V_1^{M,\pi}(s)$ over $\pi \in \Pi_{\mathrm{RNS}}$ for all states $s \in \mathcal{S}$ *simultaneously* (rather than just on average, as in (5.2)). The fact that such a policy exists may seem magical at first, but it is rather straightforward. Indeed, if $\pi_{M,h}(s)$ is defined for all $s \in \mathcal{S}$ and $h = 2, \ldots, H$, then defining the optimal $\pi_{M,1}(s)$ at any s is a matter of greedily choosing an action that maximizes the sum of the expected immediate reward and the remaining expected reward under the optimal policy. Indeed, this observation is Bellman's principle of optimality, stated more generally as follows [13]: To state the result formally, we introduce the *optimal value functions* as



Principle of optimality. ***An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.***

$$Q_h^{M,\star}(s,a) = \max_{\pi \in \Pi_{\mathrm{RNS}}} \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{T} r_{h'} \mid s_h = s, a_h = a\right] \quad \text{and} \quad V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s,a). \tag{5.4}$$

for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in [H]$; we adopt the convention that $V_{H+1}^{M,\star}(s) = Q_{H+1}^{M,\star}(s,a) = 0$. Since these optimal values are separate maximizations for each s, a, h, it is reasonable to ask whether there exists a single policy that maximizes all these value functions. Indeed, the following lemma shows that there exists π_M such that for all s, a, h,

$$Q_h^{M,\star}(s,a) := Q_h^{M,\pi_M}(s,a), \quad \text{and} \quad V_h^{M,\star}(s) = V_h^{M,\pi_M}(s). \tag{5.5}$$

Proposition 25 (Bellman Optimality): The optimal value function (5.4) for MDP M can be computed via $V_{H+1}^{M,\pi_M}(s) := 0$, and for each $s \in \mathcal{S}$,

$$V_h^{M,\pi_M}(s) = \max_{a \in \mathcal{A}} \mathbb{E}^M\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s, a_h = a\right]. \tag{5.6}$$

The optimal policy is given by

$$\pi_{M,h}(s) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}^M\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s, a_h = a\right]. \tag{5.7}$$

Equivalently, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$Q_h^{M,\pi_M}(s,a) = \mathbb{E}^M\left[r_h + \max_{a' \in \mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1}, a') \mid s_h = s, a_h = a\right], \tag{5.8}$$

and an the optimal policy is given by

$$\pi_{M,h}(s) \in \arg\max_{a \in \mathcal{A}} Q_h^{M,\pi_M}(s,a). \tag{5.9}$$

The update in (5.8) is referred to as *value iteration* (VI). It is useful to introduce a more succinct notation for this update. For a MDP M, define the *Bellman Operators* $\mathcal{T}_1^M, \ldots, \mathcal{T}_H^M$ via

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}_{s_{h+1} \sim P_h^M(\cdot|s,a), r_h \sim R_h^M(\cdot|s,a)}\left[r_h + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a')\right] \tag{5.10}$$

for any $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Going forward, we will write the expectation above more succinctly as

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}^M\left[r_h(s_h, a_h) + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a') \mid s_h = s, a_h = a\right] \tag{5.11}$$

In the language of Bellman operators, (5.8) can be written as

$$Q_h^{M,\pi_M} = \mathcal{T}_h^M Q_{h+1}^{M,\pi_M}. \tag{5.12}$$

5.3 Failure of Uniform Exploration

The task of planning using dynamic programming—which requires knowledge of the MDP—is fairly straightforward, at least if we disregard the computational concerns. In this course, however, we are interested in the problem of learning to make decisions in the face of an unknown environment. Minimizing regret in an unknown MDP requires exploration. As the next example shows, exploration in MDPs is a more delicate issue than in bandits.

Recall that ε-Greedy, a simple algorithm, is a reasonable solution for bandits and contextual bandits, albeit with a suboptimal rate ($T^{2/3}$ as opposed to $\sqrt{T}$). The next (classical) example, a so-called "combination lock," shows that such a strategy can be disastrous in reinforcement learning, as it leads to exponential (in the horizon H) regret.

Consider the MDP depicted in Figure 8, with $H+2$ states, and two actions a_g and a_b, and a starting state 1. The "good" action a_g deterministically leads to the next state in the chain, while the "bad" action deterministically leads to a terminal state. The only place where a non-zero reward can be received is the last state H, if the good action is chosen. The starting state is 1, and so the only way to receive non-zero reward is to select a_g for *all* the H time steps within the episode. Since the length of the episode is also H, selecting actions uniformly brings no information about the optimal sequence of actions decision-maker, unless by change all of the actions sampled happen to be good; the probability that this occurs is exponentially small in H.This means that T needs to be at least $O(2^H)$ to achieve nontrivial regret, and highlights the need for more strategic exploration.

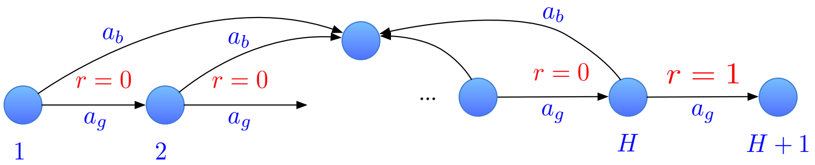


Figure 8: Combination Lock MDP.

Given the failure of ε-Greedy for this example, one can ask whether other algorithmic principles also fail. As we will show now, the principle of optimism succeeds, and an analogue of the UCB method yields a regret bound that is *polynomial* in the parameters $|\mathcal{S}|$, $|\mathcal{A}|$, and H. Before diving into the details, we present a collection of standard tools for analysis in MDPs, which will find use throughout the remainder of the lecture notes.

5.4 Analysis Tools

One of the most basic tools employed in the analysis of reinforcement learning algorithms is the *performance difference lemma*, which expresses the difference in values for two policies in terms of differences in *single-step decisions* made by the two policies. The simple proof, stated below, proceeds by successively changing one policy into another and keep track of the ensuing differences in expected rewards. One may also interpret this lemma as a version of the credit assignment mechanism.

Henceforth, we adopt the following simplified notation. When a policy π is applied to the random variable s_h, we drop the subscript h and write $\pi(s_h)$ instead of $\pi_h(s_h)$, whenever this does not cause confusion.

Lemma 14 (Performance Difference Lemma): For any $s \in \mathcal{S}$, and $\pi, \pi' \in \Pi_{\mathrm{RNS}}$,

$$V_1^{M,\pi'}(s) - V_1^{M,\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi}\Big[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, a_h) \mid s_1 = s\Big] \tag{5.13}$$

Proof of Lemma 14. Fix a pair of policies π, π' and define

$$\pi^h = (\pi_1, \ldots, \pi_{h-1}, \pi'_h, \ldots, \pi'_H),$$

with $\pi^1 = \pi'$ and $\pi^H = \pi$. By telescoping, we can write

$$V_1^{M,\pi'}(s) - V_1^{M,\pi}(s) = \sum_{h=1}^{H} V_1^{M,\pi^h}(s) - V_1^{M,\pi^{h+1}}(s). \tag{5.14}$$

Observe that for each h, we have

$$V_1^{M,\pi^h}(s) - V_1^{M,\pi^{h+1}}(s) = \mathbb{E}^{M,\pi^h}\left[\sum_{t=1}^{H} r_t \mid s_1 = s\right] - \mathbb{E}^{M,\pi^{h+1}}\left[\sum_{t=1}^{H} r_t \mid s_1 = s\right]. \tag{5.15}$$

Here, one process evolves according to (M, π^h) and the one evolves according to (M, π^{h+1}). The processes only differ in the action taken once the state s_h is reached. In the former, the action $\pi'(s_h)$ is taken, whereas in the latter it is $\pi(s_h)$. Hence, (5.15) is equal to

$$\mathbb{E}_{s_h \sim (M,\pi)}\, \mathbb{E}^{M,\pi}\Big[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, \pi(s_h)) \mid s_1 = s\Big] \tag{5.16}$$

which can be written as

$$\mathbb{E}_{(s_h,a_h) \sim (M,\pi)}\, \mathbb{E}^{M,\pi}\Big[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, a_h) \mid s_1 = s\Big]. \tag{5.17}$$

□

In contrast to the performance difference lemma, which relates the values of two policies under the same MDP, the next result relates the performance of the same policy under two different MDPs. Specifically, the difference in initial value for two MDPs is decomposed into a sum of errors between layer-wise value functions.

Lemma 15 (Bellman residual decomposition): For any pair of MDPs $M = (P^M, R^M)$ and $\widehat{M} = (P^{\widehat{M}}, R^{\widehat{M}})$, for any $s \in \mathcal{S}$, and policies $\pi \in \Pi_{\mathrm{RNS}}$,

$$V_1^{M,\pi}(s) - V^{\widehat{M},\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[Q_h^{M,\pi}(s_h, a_h) - \big[\mathcal{T}_h^M Q_{h+1}^{M,\pi}\big](s_h, a_h) \mid s_1 = s\big] \tag{5.18}$$

Hence, for $M, \widehat{M}$ with the same initial state distribution,

$$f^M(\pi) - f^{\widehat{M}}(\pi) = \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[Q_h^{M,\pi}(s_h, a_h) - \big[\mathcal{T}_h^M Q_{h+1}^{M,\pi}\big](s_h, a_h)\big]. \tag{5.19}$$

In addition, for any MDP M and function $Q = (Q_1, \ldots, Q_H, Q_{H+1})$ with $Q_{H+1} = 0$, letting $\pi_{Q,h}(s) = \arg\max_{a\in\mathcal{A}} Q_h(s,a)$, we have

$$\max_{a\in\mathcal{A}} Q_1(s,a) - V_1^{M,\pi_Q}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h,a_h) \mid s_1 = s\big]. \tag{5.20}$$

and, hence, under the assumption of the same initial distribution,

$$\mathbb{E}_{s_1\sim d_1}\big[\max_{a\in\mathcal{A}} Q_1(s_1,a)\big] - f^M(\pi_Q) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h,a_h)\big]. \tag{5.21}$$

Note that for the second part of Lemma 15 $Q = (Q_1,\ldots,Q_H)$ can be any sequence of functions, and need not be a value function corresponding to a particular policy or MDP. It is worth noting that Q gives rise to the greedy policy π_Q, which, in turn, gives rise to Q^{M,π_Q} (the value of π_Q in model M), but it may well be the case that $Q^{M,\pi_Q} \neq Q$.

Proof of Lemma 15. We will prove (5.19), and omit the proof for (5.18), which is similar but more verbose. We have

$$\begin{aligned}\sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1})\big] &= \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] - \mathbb{E}^{\widehat{M},\pi}\left[\sum_{h=1}^{H} r_h\right]\\ &= \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] - f^{\widehat{M}}(\pi).\end{aligned}$$

On the other hand, since $V_h^{M,\pi}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^{M,\pi}(s,a)]$, a telescoping argument yields

$$\begin{aligned}\sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] &= \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\big[V_h^{M,\pi}(s_h) - V_{h+1}^{M,\pi}(s_{h+1})\big]\\ &= \mathbb{E}^{\widehat{M},\pi}\big[V_1^{M,\pi}(s_1)\big] - \mathbb{E}^{\widehat{M},\pi}\big[V_{H+1}^{M,\pi}(s_{H+1})\big]\\ &= f^M(\pi),\end{aligned}$$

where we have used that $V_{H+1}^{M,\pi} = 0$, and that both MDPs have the same initial state distribution. We prove (5.21) (omitting the proof of (5.20)) using a similar argument. We have

$$\begin{aligned}&\sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - r_h - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},a)\big]\\ &= \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},a)\big] - \mathbb{E}^{M,\pi_Q}\left[\sum_{h=1}^{H} r_h\right]\\ &= \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},a)\big] - f^M(\pi_Q).\end{aligned}$$

Since $a_{h+1} = \pi_{Q,h}(s_{h+1}) = \arg\max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},a)$, we have $\mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},a)\big] = \mathbb{E}^{M,\pi_Q}\big[Q_h(s_h,a_h) - Q_{h+1}(s_{h+1},a_{h+1})\big]$, and the result follows by telescoping.

□

Another similar analysis tool for MDPs, the *simulation lemma*, is deferred to Section 6 (Lemma 24). This result can be proven as a consequence of Lemma 15.

5.5 Optimism

To develop algorithms for regret minimization in unknown MDPs, we turn to the principle of optimism, which we have seen is successful in tackling multi-armed bandits and linear bandits (in small dimension). Recall that for bandits, Lemma 7 gave a way to decompose the regret of optimistic algorithms into width of confidence intervals. What is the analogue of Lemma 7 for MDPs? Thinking of optimistic estimates at the level of expected rewards for policies π is unwieldy, and we need to dig into the structure of these multi-stage decisions. In particular, the approach we employ is to construct a sequence of optimistic *value functions* $\overline{Q}_1, \ldots, \overline{Q}_H$ which are guaranteed to over-estimate the optimal value function $Q^{M,\star}$. For multi-armed bandits, implementing optimism amounted to adding "bonuses", constructed from past data to estimates for the reward function. We will construct optimistic value functions in a similar fashion. Before giving the construction, we introduce a technical lemma, which quantifies the error in using such optimistic estimates in terms of *Bellman residuals*; Bellman residuals measure self-consistency of the optimistic estimates under the application of the Bellman operator.

Lemma 16 (Error decomposition for optimistic policies): Let $\{\overline{Q}_1, \ldots, \overline{Q}_H\}$ be a sequence of functions $\overline{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with the property that for all (s, a),

$$Q_h^{M,\star}(s,a) \le \overline{Q}_h(s,a) \tag{5.22}$$

and set $\overline{Q}_{H+1} \equiv 0$. Let $\widehat{\pi} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_H)$ be such that $\widehat{\pi}_h(s) = \arg\max_a \overline{Q}_h(s,a)$. Then for all $s \in \mathcal{S}$,

$$V_1^{M,\star}(s) - V_1^{M,\widehat{\pi}}(s) \le \sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}}\big[(\overline{Q}_h - \mathcal{T}_h^M \overline{Q}_{h+1})(s_h, \widehat{\pi}(s_h)) \mid s_1 = s\big]. \tag{5.23}$$

Lemma 16 tells us that closeness of $\overline{Q}_h$ to the Bellman backup $\mathcal{T}_h^M \overline{Q}_{h+1}$ implies closeness of $\widehat{\pi}$ to π_M in terms of the value. As a sanity check, if $\overline{Q}_h = Q_h^{M,\star}$, the right-hand side of (5.23) is zero, since $Q_h^{M,\star} = \mathcal{T}_h^M Q_{h+1}^{M,\star}$. Crucially, errors do not accumulate too fast as a function of the horizon. This fact should not be taken for granted: in general, if $\overline{Q}$ is not optimistic, it could have been the case that small changes in $\overline{Q}_h$ exponentially degrade the quality of the policy $\widehat{\pi}$.

Another important aspect of the decomposition (5.23) is the *on-policy* nature of the terms in the sum. Observe that the law of s_h for each of the terms is given by executing $\widehat{\pi}$ in model M. The distribution of s_h is often referred to as the *roll-in distribution*; when this distribution is induced by the policy executed by the algorithm, we may have a better control of the error than in the *off-policy* case when the roll-in distribution is given by π_M or another unknown policy.

Proof of Lemma 16. Let $\overline{V}_h(s) := \max_{a \in \mathcal{A}} \overline{Q}_h(s,a)$. Just as in the proof of Lemma 7, the assumption that $\overline{Q}_h$ is "optimistic" implies that

$$Q_h^{M,\star}(s_h, \pi_M(s_h)) \le \overline{Q}_h(s_h, \pi_M(s_h)) \le \overline{Q}_h(s_h, \widehat{\pi}(s_h))$$

and, hence, $V_1^{M,\star}(s) \le \overline{V}_1(s)$. Then, (5.20) applied to $Q = \overline{Q}$ and $\widehat{\pi} = \pi_Q$ states that

$$\overline{V}_1(s) - V_1^{M,\widehat{\pi}}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}}\big[\overline{Q}_h(s_h,a_h) - \big[\mathcal{T}_h^M \overline{Q}_{h+1}\big](s_h,a_h) \mid s_1 = s\big]. \tag{5.24}$$

□

Remark 17: In fact, the proof of Lemma 16 only uses that the initial value $\overline{Q}_1$ is optimistic. However, to construct a value function with this property, the algorithms we consider will proceed by backwards induction, producing optimistic estimates $\overline{Q}_1, \dots, \overline{Q}_H$ in the process.

5.6 The UCB-VI Algorithm for Tabular MDPs

We now instantiate the principle of optimism to give regret bounds for online reinforcement learning in *tabular MDPs.* Tabular RL may be thought of as an analogue of finite-armed bandits: we assume no structure across states and action, but require that the state and action spaces are small. The regret bounds we present will depend polynomially on $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, as well as the horizon H.

Preliminaries. For simplicity, we assume that the reward function is known to the learner, so that only the transition probabilities are unknown. This does not change the difficulty of the problem in a meaningful way, but allows us to keep notation light.

Assumption 6: Rewards are deterministic, bounded, and known to the learner: $R_h^M(\cdot \mid s,a) = \delta_{r_h(s,a)}$ for known $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$, for all M. In addition, assume for simplicity that $V_1^{M,\star}(s) \in [0,1]$ for any $s \in \mathcal{S}$.

Define, with a slight abuse of notation,

$$n_h^t(s,a) = \sum_{i=1}^{t-1} \mathbb{I}\left\{(s_h^i, a_h^i) = (s,a)\right\}, \quad \text{and} \quad n_h^t(s,a,s') = \sum_{i=1}^{t-1} \mathbb{I}\left\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\right\},$$

as the empirical state-action and state-action-next state frequencies. We can estimate the transition probabilities via

$$\widehat{P}_h^t(s' \mid s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)}. \tag{5.25}$$

The UCB-VI algorithm. The following algorithm, UCB-VI ("Upper Confidence Bound Value Iteration"), combines the notion of optimism with dynamic programming.

```
UCB-VI
for $t = 1, \dots, T$ do
    Let $\overline{V}_{H+1}^t \equiv 1$.
    for $h = H, \dots, 1$ do
```

```
        Update $n_h^t(s,a)$, $n_h^t(s,a,s')$, and $b_{h,\delta}^t(s,a)$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
                                                // $b_{h,\delta}^t(s,a)$ is a bonus computed in (5.27).
        Compute:
            $\overline{Q}_h^t(s,a) = \left\{ r_h(s,a) + \mathbb{E}_{s' \sim \widehat{P}_h^t(\cdot|s,a)} \overline{V}_{h+1}^t(s') + b_{h,\delta}^t(s,a) \right\} \wedge 1.$        (5.26)
        Set $\overline{V}_h^t(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^t(s,a)$ and $\widehat{\pi}_h^t(s) = \arg\max_{a \in \mathcal{A}} \overline{Q}_h^t(s,a)$.
    Collect trajectory $(s_1^t, a_1^t, r_1^t), \ldots, (s_H^t, a_H^t, r_H^t)$ according to $\widehat{\pi}^t$.
```

The UCB-VI algorithm will be analyzed using Lemma 16. In constructing functions $\overline{Q}_h$, we will need to satisfy two goals: (1) ensure that with high probability (5.22) is satisfied, i.e. $\overline{Q}_h$s are optimistic; and (2) that $\overline{Q}_h$s are "self-consistent," in the sense that the Bellman residuals in (5.23) are small. The second requirement already suggests that we should define $\overline{Q}_h$ approximately as a Bellman backup $\mathcal{T}_h^M \overline{Q}_{h+1}$, going backwards for $h = H+1, \ldots, 1$ as in dynamic programming, while ensuring the first requirement. In addition to these considerations, we will have to use a surrogate for the Bellman operator $\mathcal{T}_h^M$, since the model M is not known. This is achieved by estimating M using empirical transition frequencies. Putting these ideas together gives the update in (5.26). We apply the principle of value iteration, except that

1. For each episode t, we augment the rewards $r_h(s,a)$ with a "bonus" $b_{h,\delta}^t(s,a)$ designed to ensure optimism.

2. The Bellman operator is approximated using the estimated transition probabilities in (5.25).

The bonus functions play precisely the same role as the width of the confidence interval in (2.19): these bonuses ensure that (5.22) holds with high probability, as we will show below in Lemma 17.

The following theorem shows that with an appropriate choice of bonus, this algorithm achieves a polynomial regret bound.

Theorem 1: For any $\delta > 0$, UCB-VI with

$$b_{h,\delta}^t(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n_h^t(s,a)}} \tag{5.27}$$

guarantees that with probability at least $1-\delta$,

$$\mathbf{Reg} \lesssim HS\sqrt{AT} \cdot \sqrt{\log(SAHT/\delta)}$$

We mention that a slight variation on Lemma 19 below (using the Freedman inequality instead of the Azuma-Hoeffding inequality) yields an improved rate of $O(H\sqrt{SAT} + \mathrm{poly}(H,S,A)\log T)$, and the optimal rate can be shown to be $\Theta(\sqrt{HSAT})$; this is achieved through a more careful choice for the bonus $b_{h,\delta}^t$ and a more refined analysis. We remark that care should be taken in comparing results in the literature, as scaling conventions for the individual and cumulative rewards (as in Assumption 6) can vary.

5.6.1 Analysis for a Single Episode

Our aim is to bound the regret

$$\mathbf{Reg} = \sum_{t=1}^{T} f^M(\pi_{M^\star}) - f^M(\pi^t)$$

for UCB-VI. To do so, we first prove several helper lemmas concerning the performance within each episode t. In what follows, we fix t and drop the superscript t.

Given the estimated transitions $\left\{\widehat{P}_h(\cdot \mid s,a)\right\}_{h,s,a}$, define the estimated MDP $\widehat{M} = \left\{\mathcal{S}, \mathcal{A}, \{\widehat{P}_h\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\}$. The associated Bellman operator is

$$\mathcal{T}_h^{\widehat{M}} Q(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim \widehat{P}_h(\cdot|s,a)} \max_a Q(s',a) \tag{5.28}$$

for $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$. Consider the sequence of functions $\overline{Q}_h : \mathcal{S}\times\mathcal{A} \to [0,1], \overline{V}_h : \mathcal{S} \to [0,1]$, for $h = 1,\dots,H+1$, with $\overline{Q}_{H+1} \equiv 0$ and

$$\overline{Q}_h(s,a) = \left\{[\mathcal{T}_h^{\widehat{M}}\overline{Q}_{h+1}](s,a) + b_{h,\delta}(s,a)\right\} \wedge 1, \quad \text{and} \quad \overline{V}_h(s) = \max_a \overline{Q}_h(s,a) \tag{5.29}$$

for bonus functions $b_{h,\delta} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ to be chosen later. Henceforth, we follow the usual notation that for functions f,g over the same domain, $f \le g$ indicates pointwise inequality over the domain.

The first lemma we present shows that as long as the bonuses $b_{h,\delta}$ are large enough to bound the error between the estimated transition probabilities and true transition probabilities, the functions $\overline{Q}_1,\dots,\overline{Q}_H$ constructed above will be optimistic.

Lemma 17: Suppose we have estimates $\left\{\widehat{P}_h(\cdot \mid s,a)\right\}_{h,s,a}$ and a function $b_{h,\delta} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ with the property that

$$\left|\sum_{s'} \widehat{P}_h(s' \mid s,a) V_h^{M,\star}(s') - \sum_{s'} P_h^M(s' \mid s,a) V_h^{M,\star}(s')\right| \le b_{h,\delta}(s,a). \tag{5.30}$$

Then for all $h \in [H]$, we have

$$\overline{Q}_h \ge Q_h^{M,\star}, \quad \text{and} \quad \overline{V}_h \ge V_h^{M,\star} \tag{5.31}$$

for $\overline{Q}_h, \overline{V}_h$ defined in (5.29).

Proof of Lemma 17. The proof proceeds by backward induction on the statement

$$\overline{V}_h \ge V_h^{M,\star}$$

with $h = H+1$ down to $h = 1$. We start with the base case $h = H+1$, which is trivial because $\overline{V}_{H+1} = V_{H+1}^{M,\star} \equiv 0$. Now, assume $\overline{V}_{h+1} \ge V_{h+1}^{M,\star}$, and let us prove the induction step. Fix $(s,a) \in \mathcal{S}\times\mathcal{A}$. If $\overline{Q}_h(s,a) = 1$, then, trivially, $\overline{Q}_h(s,a) \ge Q_h^{M,\star}(s,a)$. Otherwise, $\overline{Q}_h(s,a) = \mathcal{T}_h^{\widehat{M}}\overline{Q}_{h+1}(s,a) + b_{h,\delta}(s,a)$, and thus

$$\begin{aligned}\overline{Q}_h(s,a) - Q_h^{M,\star}(s,a) &= b_{h,\delta}(s,a) + \mathbb{E}_{s'\sim\widehat{P}_h(\cdot|s,a)}\overline{V}_{h+1}(s') - \mathbb{E}_{s'\sim P_h^M(\cdot|s,a)} V_{h+1}^{M,\star}(s') \\ &\ge b_{h,\delta}(s,a) + \mathbb{E}_{s'\sim\widehat{P}_h(\cdot|s,a)} V_{h+1}^{M,\star}(s') - \mathbb{E}_{s'\sim P_h^M(\cdot|s,a)} V_{h+1}^{M,\star}(s') \ge 0.\end{aligned}$$

This, in turn, implies that $\overline{V}_h(s) = \max_a \bar{Q}_h(s,a) \ge \max_a Q_h^{M,\star}(s,a) = V_h^{M,\star}(s)$, concluding the induction step. □

We now analyze the effect of using the effect of using an estimated model $\widehat{M}$ for the Bellman operator rather than the true unknown $\mathcal{T}_h^M$.

Lemma 18: Suppose we have estimates $\left\{\widehat{P}_h(\cdot \mid s,a)\right\}_{h,s,a}$ and $b'_{h,\delta}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with the property that

$$\max_{V \in \{0,1\}^S} \left| \sum_{s'} \widehat{P}_h(s' \mid s,a) V(s') - \sum_{s'} P_h^M(s' \mid s,a) V(s') \right| \le b'_{h,\delta}(s,a) \tag{5.32}$$

Then the Bellman residual satisfies

$$\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1} \le (b_{h,\delta} + b'_{h,\delta}) \wedge 1. \tag{5.33}$$

for $\bar{Q}_h, \overline{V}_h$ defined in (5.29).

Proof of Lemma 18. That $\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1} \le 1$ is immediate. To prove the main result, observe that

$$\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1} = \left\{ \mathcal{T}_h^{\widehat{M}} \bar{Q}_{h+1} + b_{h,\delta} \right\} \wedge 1 - \mathcal{T}_h^M \bar{Q}_{h+1} \le (\mathcal{T}_h^{\widehat{M}} - \mathcal{T}_h^M) \bar{Q}_{h+1} + b_{h,\delta} \tag{5.34}$$

For any $Q \in \mathcal{S} \times \mathcal{A} \to [0,1]$,

$$\begin{aligned} (\mathcal{T}_h^{\widehat{M}} - \mathcal{T}_h^M) Q(s,a) &= \mathbb{E}_{s' \sim \widehat{P}_h(\cdot|s,a)} \max_a Q(s',a) - \mathbb{E}_{s' \sim P_h^M(\cdot|s,a)} \max_a Q(s',a) && (5.35) \\ &\le \max_{V \in [0,1]^S} |\mathbb{E}_{s' \sim \widehat{P}_h(\cdot|s,a)} V(s') - \mathbb{E}_{s' \sim P_h^M(\cdot|s,a)} V(s')|. && (5.36) \end{aligned}$$

Since the maximum is achieved at a vertex of $[0,1]^S$, the statement follows. □

5.6.2 Regret Analysis

We now bring back the time index t and show that the estimated transition probabilities in UCB-VI satisfy conditions of Lemma 17 and Lemma 18, ensuring that the functions $\bar{Q}_1^t, \ldots, \bar{Q}_H^t$ are optimistic.

Lemma 19: Let $\{\widehat{P}_h^t\}_{h \in [H], t \in [T]}$ be defined as in (5.25). Then with probability at least $1 - \delta$, the functions

$$b_{h,\delta}^t(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n_h^t(s,a)}}, \quad \text{and} \quad b'^{t}_{h,\delta}(s,a) = 8\sqrt{\frac{S \log(2SAHT/\delta)}{n_h^t(s,a)}}$$

satisfy the assumptions of Lemma 17 and Lemma 18, respectively, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$, and $t \in [T]$ simultaneously.

Proof of Lemma 19. We leave the proof as an exercise. □

Proof of Theorem 1. Putting everything together, we can now prove Theorem 1. Under the event in Lemma 19, the functions $\overline{Q}_1^t, \ldots, \overline{Q}_H^t$ are optimistic, which means that the conditions of Lemma 16 hold, and the instantaneous regret on round t (conditionally on $s_1 \sim d_1$) is at most

$$\sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}^t}\left[(\overline{Q}_h^t - \mathcal{T}_h^M \overline{Q}_{h+1}^t)(s_h^t, \widehat{\pi}_h^t(s_h^t)) \mid s_1 = s\right] \le \sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}^t}\left[(b_{h,\delta}(s_h^t, \widehat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \widehat{\pi}_h^t(s_h^t))) \wedge 1\right],$$

where the second inequality invokes Lemma 18. Summing over $t = 1, \ldots, T$, and applying the Azuma-Hoeffding inequality, we have that with probability at least $1 - \delta$, the regret of UCB-VI is bounded by

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}^t}\left[(b_{h,\delta}(s_h^t, \widehat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \widehat{\pi}_h^t(s_h^t))) \wedge 1\right] \\
&\lesssim \sum_{t=1}^{T}\sum_{h=1}^{H} (b_{h,\delta}(s_h^t, \widehat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \widehat{\pi}_h^t(s_h^t))) \wedge 1 + \sqrt{HT\log(1/\delta)}.
\end{aligned}$$

Using the bonus definition in (5.27), the bonus term above is bounded by

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \sqrt{\frac{S\log(2SAHT/\delta)}{n_h^t(s_h^t, \widehat{\pi}_h^t(s_h^t))}} \wedge 1 \le \sqrt{S\log(2SAHT/\delta)} \sum_{t=1}^{T}\sum_{h=1}^{H} \frac{1}{\sqrt{n_h^t(s_h^t, \widehat{\pi}_h^t(s_h^t))}} \wedge 1 \quad (5.37)$$

The double summation can be handled in the same fashion as Lemma 8:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{1}{\sqrt{n_h^t(s_h^t, \widehat{\pi}_h^t(s_h^t))}} \wedge 1 &= \sum_{h=1}^{H}\sum_{(s,a)}\sum_{t=1}^{T} \frac{\mathbb{I}\left\{(s_h^t, \widehat{\pi}_h^t(s_h^t)) = (s,a)\right\}}{\sqrt{n_h^t(s,a)}} \wedge 1 \\
&\lesssim \sum_{h=1}^{H}\sum_{(s,a)} \sqrt{n_h^T(s,a)} \le H\sqrt{SAT}.
\end{aligned}$$

□

6. GENERAL DECISION MAKING

So far, we have covered three general frameworks for interaction decision making: The contextual bandit problem, the structured bandit problem, and the episodic reinforcement learning problem; all of these frameworks generalize the classical multi-armed bandit problem in different directions. In the context of structured bandits, we introduced a complexity measure called the Decision-Estimation Coefficient (DEC 0, which gave a generic approach to algorithm design, and allowed us to reduce the problem of interactive decision making to that of supervised online estimation. In this section, we will build on this development on two fronts: First, we will introduce a unified framework for decision making, which subsumes all of the frameworks we have covered so far. Then, we will show that i) the Decision-Estimation Coefficient and its associated meta-algorithm (E2D) extend to the general decision making framework, and ii) boundedness of the DEC is not just sufficient, but actually *necessary* for low regret, and thus constitutes a fundamental limit. As an application the general tools we introduce, we will show how to use the (generalized) Decision-Estimation Coefficient to solve the problem of tabular reinforcement learning (Section 6.6), offering an alternative to the UCB-VI method we introduced in Section 5.

6.1 Setting

For the remainder of the course, we will focus on a framework called Decision Making with Structured Observations (DMSO), which subsumes all of the decision making frameworks we have encountered so far. The protocol proceeds in T rounds, where for each round $t = 1, \ldots, T$:

1. The learner selects a *decision* $\pi^t \in \Pi$, where Π is the *decision space.*
2. Nature selects a *reward* $r^t \in \mathcal{R}$ and *observation* $o^t \in \mathcal{O}$ based on the decision, where $\mathcal{R} \subseteq \mathbb{R}$ is the *reward space* and $\mathcal{O}$ is the *observation space.* The reward and observation are then observed by the learner.

We focus on a stochastic variant of the DMSO framework.

Assumption 7 (Stochastic Rewards): Rewards and observations are generated independently via

$$(r^t, o^t) \sim M^\star(\cdot \mid \pi^t), \tag{6.1}$$

where $M^\star : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ is the underlying *model.*

To facilitate the use of learning and function approximation, we assume the learner has access to a *model class* $\mathcal{M}$ that contains the model $M^\star$. Depending on the problem domain, $\mathcal{M}$ might consist of linear models, neural networks, random forests, or other complex function approximators; this generalizes the role of the reward function class $\mathcal{F}$ used in contextual/structured bandits. We make the following standard realizability assumption, which asserts that $\mathcal{M}$ is flexible enough to express the true model.

Assumption 8 (Realizability): The model class $\mathcal{M}$ contains the true model $M^\star$.

For a model $M \in \mathcal{M}$, let $\mathbb{E}^{M,\pi}[\cdot]$ denote the expectation under $(r, o) \sim M(\pi)$. Further, following the notation in Section 5, let

$$f^M(\pi) := \mathbb{E}^{M,\pi}[r]$$

denote the mean reward function, and let

$$\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$$

denote the optimal decision with maximal expected reward. Finally, define

$$\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\} \tag{6.2}$$

as the induced class of mean reward functions. We evaluate the learner's performance in terms of *regret* to the optimal decision for $M^\star$:

$$\mathbf{Reg} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\big], \tag{6.3}$$

where $p^t \in \Delta(\Pi)$ is the learner's distribution over decisions at round t. Going forward, we abbreviate $f^\star = f^{M^\star}$ and $\pi^\star = \pi_{M^\star},$.

The DMSO framework is general enough to capture most online decision making problems. Let us first see how it subsumes the structured bandit and contextual bandit problems.

Example 6.1 (Structured bandits)**.** When there are no observations (i.e., $\mathcal{O} = \{\varnothing\}$), the DMSO framework is equivalent to *structured bandits* studied earlier in Section 4. Therein, we defined a structured bandit instance by specifying a class $\mathcal{F}$ of mean reward functions and a general class of reward distributions, such as sub-Gaussian or bounded. In the DMSO framework, we may equivalently start with a set of models $\mathcal{M}$ and let $\mathcal{F}_{\mathcal{M}}$ be the induced class (6.2). By changing the class $\mathcal{F}$, this encompasses all of the concrete examples of structured bandit problems we studied in Section 4, including linear bandits, nonparametric bandits, and concave/convex bandits.

◁

Example 6.2 (Contextual bandits)**.** The DMSO framework readily captures contextual bandits (Section 3) with stochastic contexts (see Assumption 2). To make this precise, we will slightly abuse the notation and think of π^t as *functions* mapping the context x^t to an action in $\Pi = [A]$. To this end, on round t, the decision-maker selects a mapping $\pi^t : \mathcal{X} \to [A]$ from contexts to actions, and the context $o^t = x^t$ is observed at the end of the round. This is equivalent to first observing x^t and selecting $\pi^t(x^t) \in [A]$.

Formally, let $\mathcal{O} = \mathcal{X}$ be the space of contexts, $\Pi = [A]$ be the set of actions, and $\Pi : \mathcal{X} \to [A]$ be the space of decisions. The distribution $(r, x) \sim M(\pi)$ then has the following structure: $x \sim \mathcal{D}^M$ and $r \sim R^M(\cdot|x, \pi(x))$ for some context distribution $\mathcal{D}^M$ and reward distribution R^M. In other words, the distribution $\mathcal{D}^M$ for the context x (treated as an observation) is part of the model M.

We mention in passing that the DMSO framework also naturally extends to the case when contexts are adversarial rather than i.i.d., as in Section 4.5; see Foster et al. [35]. ◁

Example 6.3 (Online reinforcement learning)**.** The online reinforcement learning framework we introduced in Section 5 immediately falls into the DMSO framework by taking $\Pi = \Pi_{\mathrm{RNS}}$, $r^t = \sum_{h=1}^{H} r_h^t$, and $o^t = \tau^t$. While we have only covered tabular reinforcement learning so far, the literature on online reinforcement learning contains algorithms and sample complexity bounds for a rich and extensive collection of different MDP structures (e.g., Dean et al. [26], Yang and Wang [75], Jin et al. [42], Modi et al. [56], Ayoub et al. [12], Krishnamurthy et al. [48], Du et al. [29], Li [53], Dong et al. [27]). All of these settings correspond to specific choices for the model class $\mathcal{M}$ in the DMSO framework, and we will cover this topic in detail in Section 7. ◁

We adopt the DMSO framework because it gives simple, yet unified approach to describing and understanding what is—at first glance—a very general and seemingly complicated problem. Other examples that are covered by the DMSO framework include:

- Partially Observed Markov Decision Processes (POMDPs)
- Bandits with graph-structured feedback
- Partial monitoring*

6.2 Refresher: Information-Theoretic Divergences

To develop algorithms and complexity measures for general decision making, we need a way to measure the distance between distributions over abstract observations (this was not a concern for the structured and contextual bandit settings, where we only needed to consider the mean reward function). To do this, we will introduce the notion of the *Csiszar f-divergence*, which generalizes a number of familiar divergences including the Kullback-Leibler (KL) divergence, total variation distance, and Hellinger distance.

Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\Omega, \mathscr{F})$. We say that $\mathbb{P}$ is *absolutely continuous* with respect to $\mathbb{Q}$ if for all events $A \in \mathscr{F}$, $\mathbb{Q}(A) = 0 \implies \mathbb{P}(A) = 0$; we denote this by $\mathbb{P} \ll \mathbb{Q}$. For a convex function $f : (0, \infty) \to \mathbb{R}$, the associated f-divergence for $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$D_f(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right] \tag{6.4}$$

whenever $\mathbb{P} \ll \mathbb{Q}$. More generally, defining $p = \frac{d\mathbb{P}}{d\nu}$ and $q = \frac{d\mathbb{Q}}{d\nu}$ for a common dominating measure ν, we have

$$D_f(\mathbb{P} \,\|\, \mathbb{Q}) := \int_{q>0} q f\left(\frac{p}{q}\right) d\nu + \mathbb{P}(q = 0) \cdot f'(\infty), \tag{6.5}$$

where $f'(\infty) := \lim_{x \to 0^+} x f(1/x)$.

We will make use of the following f-divergences, all of which have unique properties that make them useful in different contexts.

- Choosing $f(t) = \frac{1}{2}|t - 1|$ gives the *total variation (TV) distance*

$$D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int \left|\frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{Q}}{d\nu}\right| d\nu,$$

  which can also be written as

$$D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

- Choosing $f(t) = (1 - \sqrt{t})^2$ gives *squared Hellinger distance*

$$D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu.$$

- Choosing $f(t) = t \log(t)$ gives the *Kullback-Leibler divergence*:

$$D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \begin{cases} \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Note that for TV distance and Hellinger distance, we use the notation $D(\cdot, \cdot)$ rather than $D(\cdot \,\|\, \cdot)$ to emphasize that the divergence is symmetric. Other standard examples include the Neyman-Pearson χ^2-divergence.

Lemma 20: For all distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$D^2_{\mathsf{TV}}(\mathbb{P},\mathbb{Q}) \leq D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}) \leq D_{\mathsf{KL}}(\mathbb{P}\,\|\,\mathbb{Q}). \tag{6.6}$$

It is known that $D_{\mathsf{TV}}(\mathbb{P},\mathbb{Q}) = 1$ if and only if $D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}) = 2$, and $D_{\mathsf{TV}}(\mathbb{P},\mathbb{Q}) = 0$ if and only if $D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}) = 0$ (more generally, $D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}) \leq 2D_{\mathsf{TV}}(\mathbb{P},\mathbb{Q})$). Moreover, they induce same topology, i.e. a sequence converges in one distance if and only if it converges the other. KL divergence cannot be bounded by TV distance or Hellinger distance in general, but the following lemma shows that it is possible to relate these quantities if the density ratios under consideration are bounded.

Lemma 21: Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\Omega,\mathscr{F})$. If $\sup_{F\in\mathscr{F}} \frac{\mathbb{P}(F)}{\mathbb{Q}(F)} \leq V$, then

$$D_{\mathsf{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) \leq (2+\log(V))D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}). \tag{6.7}$$

Other properties we will use include:

- Boundedness of TV (by 1) and Hellinger (by 2).
- Triangle inequality for TV and Hellinger distance.
- The *data-processing inequality*, which is satisfied by all f-divergences.
- The *chain rule* for KL divergence (see Lemma 39).
- A variational representation for TV distance:

$$D_{\mathsf{TV}}(\mathbb{P},\mathbb{Q}) = \sup_{g:\Omega\to[0,1]} |\mathbb{E}_{\mathbb{P}}[g] - \mathbb{E}_{\mathbb{Q}}[g]| \tag{6.8}$$

See Polyanskiy [59] for further background.

6.3 The Decision-Estimation Coefficient for General Decision Making

Developing algorithms for the general decision making framework poses a number of additional challenges compared to the basic bandit frameworks we have studied so far. The problem of understanding how to optimally explore and make decisions for a given model class $\mathcal{M}$ is deeply connected to the problem of understanding the optimal statistical complexity (i.e., minimax regret) for $\mathcal{M}$. Any notion of problem complexity needs to capture both i) simple problems like the multi-armed bandit, where the mean rewards serve as a sufficient statistic, and ii) problems with rich, structured feedback (e.g., reinforcement learning), where observations, or even structure in the noise itself, can provide non-trivial information about the underlying problem instance. In spite of these apparent difficulties, we will show that by incorporating an appropriate information-theoretic divergence, we can use the Decision-Estimation Coefficient to address these challenges, in a similar fashion to Section 4.

For a model class $\mathcal{M}$, reference model $\widehat{M} \in \mathcal{M}$, and scale parameter $\gamma > 0$, the Decision-Estimation Coefficient for general decision making is defined via

$$\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[\underbrace{f^{M}(\pi_M) - f^{M}(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{D^2_{\mathsf{H}}\big(M(\pi), \widehat{M}(\pi)\big)}_{\text{information gain for obs.}}\Big]. \tag{6.9}$$

We further define

$$\mathsf{dec}_{\gamma}(\mathcal{M}) = \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}). \tag{6.10}$$

The DEC in (6.9) should look familiar to the definition we used for structured bandits in Section 4 (Eq. (4.15)). The main difference is that instead of being defined over a class $\mathcal{F}$ of reward functions, the general DEC is defined over the class of *models* $\mathcal{M}$, and the notion of estimation error/information gain has changed to account for this. In particular, rather than measuring information gain via the distance between mean reward functions, we now consider the information gain

$$\mathbb{E}_{\pi \sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big],$$

which measures the distance between the *distributions* over rewards and observations under the models M and $\widehat{M}$ (for the learner's decision π). This is a stronger notion of distance since i) it incorporates observations (e.g., trajectories for reinforcement learning), and ii) even for bandit problems, we consider distance between distributions as opposed to distance between means; the latter feature means that this notion of information gain can capture fine-grained properties of the models under consideration, such as noise in the reward distribution.

6.3.1 Basic Examples

To build intuition as to how the general Decision-Estimation Coefficient adapts to the structure of the model class $\mathcal{M}$, let us review a few examples—some familiar, and some new.

Example 6.4 (Multi-armed bandit with Gaussian rewards)**.** Let $\Pi = [A]$, $\mathcal{R} = \mathbb{R}$, $\mathcal{O} = \{\varnothing\}$. We define

$$\mathcal{M}_{\mathsf{MAB\text{-}G}} = \{M : M(\pi) = \mathcal{N}(f(\pi), 1), f : \Pi \to [0, 1]\}.$$

We claim that

$$\mathsf{dec}_{\gamma}(\mathcal{M}_{\mathsf{MAB\text{-}G}}) \propto \frac{A}{\gamma}. \tag{6.11}$$

To prove this, consider the case where $\widehat{M} \in \mathcal{M}$ for simplicity. Recall that we have previously shown that this behavior holds for the squared error version of the DEC defined in (4.15). Thus, it is sufficient to argue that the squared Hellinger divergence for Gaussian distributions reduces to square difference between the means:

$$D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big) \propto (f^{M}(\pi) - f^{\widehat{M}}(\pi))^2.$$

The claim will then follow from Proposition 14. To prove this, first note that

$$D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big) \leq D_{\mathsf{KL}}\Big(M(\pi) \parallel \widehat{M}(\pi)\Big) = \frac{1}{2}(f^{M}(\pi) - f^{\widehat{M}}(\pi))^2. \tag{6.12}$$

In the other direction, one can directly compute

$$D^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big) = 1 - \exp\bigg\{-\frac{1}{8}(f^M(\pi) - f^{\widehat{M}}(\pi))^2\bigg\}$$

and using that $1-\exp\{-x\} \geq (1-e^{-1})x$ for $x \in [0,1]$, we establish

$$D^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big) \geq c \cdot (f^M(\pi) - f^{\widehat{M}}(\pi))^2$$

for $c = \frac{1-1/e}{8}$. ◁

In fact, one can show that the general DEC in (6.9) coincides with the basic squared error version from Section 4 for general structured bandit problems, not just multi-armed bandits; see Proposition 41.

Let us next consider a twist on the bandit problem that is more information-theoretic in nature, and highlights the need to work with information-theoretic divergences if we want to handle general decision making problems.

Example 6.5 (Bandits with structured noise)**.** Let $\Pi = [A]$, $\mathcal{R} = \mathbb{R}$, $\mathcal{O} = \{\varnothing\}$. We define

$$\mathcal{M}_{\mathsf{MAB\text{-}SN}} = \{M_1, \ldots, M_A\} \cup \Big\{\widehat{M}\Big\}$$

where $M_i(\pi) := \mathcal{N}(1/2, 1)$ for $\pi \neq i$ and $M_i(\pi) := \mathrm{Ber}(3/4)$ for $\pi = i$; we further define $\widehat{M}(\pi) := \mathcal{N}(1/2, 1)$ for all $\pi \in \Pi$. Before proceeding with the calculations, observe that we can solve the general decision making problem when the underlying model is $M^\star \in \mathcal{M}$ with a simple algorithm. It is sufficient to select every action in $[A]$ only once: all suboptimal actions have Bernoulli rewards and give $r \in \{0,1\}$ almost surely, while the optimal action has Gaussian rewards, and gives $r \notin \{0,1\}$ almost surely. Thus, if we select an action and observe a reward $r \notin \{0,1\}$, we know that we have identified the optimal action.

The valuable information contained in the reward distribution is reflected in the Hellinger divergence, which attains its maximum value when comparing a continuous distribution to a discrete one:

$$D^2_{\mathsf{H}}\Big(M_i(\pi),\widehat{M}(\pi)\Big) = 2\mathbb{I}\{\pi = i\}.$$

To use this property to derive the upper bound on $\mathsf{dec}_\gamma(\mathcal{M}_{\mathsf{MAB\text{-}SN}}, \widehat{M})$, first note that the maximum over M in the definition of $\mathsf{dec}_\gamma(\mathcal{M}_{\mathsf{MAB\text{-}SN}}, \widehat{M})$ is not attained at $M = \widehat{M}$, since in that case both the divergence and regret terms are zero, irrespective of p. Now, take $p = \mathrm{unif}[A]$. Then for any $M \in \{M_1, \ldots, M_A\}$,

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] = (1-1/A)(3/4 - 1/2),$$

and

$$\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \lesssim (1-1/A)(3/4 - 1/2) - \gamma\frac{2}{A} \lesssim \mathbb{I}\{\gamma \leq A/4\}.$$

This leads to an upper bound

$$\mathsf{dec}_\gamma(\mathcal{M}_{\mathsf{MAB\text{-}SN}}, \widehat{M}) \lesssim \mathbb{I}\{\gamma \leq A/4\} \tag{6.13}$$

which can also be shown to be tight. ◁

Example 6.6 (Bandits with Full Information)**.** Consider a "full-information" learning setting. We have $\Pi = [A]$ and $\mathcal{R} = [0,1]$, and for a given decision π we observe a reward r as in the standard multi-armed bandit, but also receive an observation $o = (r(\pi'))_{\pi' \in [A]}$ consisting of (counterfactual) rewards for *every* action.

For a given model M, let $M_{\mathcal{R}}(\pi)$ denote the distribution over the reward r for π, and let $M_{\mathcal{O}}(\pi)$ denote the distribution of o. Then for any decision π, since all rewards are observed, the data processing inequality implies that for all $M, \widehat{M} \in \mathcal{M}$ and $\pi' \in \Pi$,

$$D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big) \geq D^2_{\mathsf{H}}\Big(M_{\mathcal{O}}(\pi), \widehat{M}_{\mathcal{O}}(\pi)\Big) \tag{6.14}$$

$$= D^2_{\mathsf{H}}\Big(M_{\mathcal{O}}(\pi'), \widehat{M}_{\mathcal{O}}(\pi')\Big) \geq D^2_{\mathsf{H}}\Big(M_{\mathcal{R}}(\pi'), \widehat{M}_{\mathcal{R}}(\pi')\Big). \tag{6.15}$$

Using this property, we will show that for any $\widehat{M} \in \mathcal{M}$,

$$\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \leq \frac{1}{\gamma}. \tag{6.16}$$

Comparing to the finite-armed bandit, we see that the DEC for this example is independent of A, which reflects the extra information contained in the observation o.

To prove (6.16), for a given model $\widehat{M} \in \mathcal{M}$ we choose $p = \mathbb{I}_{\pi_{\widehat{M}}}$ (i.e. the decision maker selects $\pi_{\widehat{M}}$ deterministically), and bound $\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)]$ by

$$\begin{aligned} f^M(\pi_M) - f^M(\pi_{\widehat{M}}) &\leq f^M(\pi_M) - f^M(\pi_{\widehat{M}}) + f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M) \\ &\leq 2 \cdot \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} |f^M(\pi) - f^{\widehat{M}}(\pi)| \\ &\leq 2 \cdot \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} D_{\mathsf{H}}\Big(M_{\mathcal{R}}(\pi), \widehat{M}_{\mathcal{R}}(\pi)\Big). \end{aligned}$$

We then use the AM-GM inequality, which implies that for any $\gamma > 0$,

$$\begin{aligned} \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} D^2_{\mathsf{H}}\Big(M_{\mathcal{R}}(\pi), \widehat{M}_{\mathcal{R}}(\pi)\Big) &\lesssim \gamma \cdot \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} D^2_{\mathsf{H}}\Big(M_{\mathcal{R}}(\pi), \widehat{M}_{\mathcal{R}}(\pi)\Big) + \frac{1}{\gamma} \\ &\leq \gamma \cdot D^2_{\mathsf{H}}\Big(M(\pi_{\widehat{M}}), \widehat{M}(\pi_{\widehat{M}})\Big) + \frac{1}{\gamma}, \end{aligned}$$

where the final inequality uses (6.14). This certifies that for all $M \in \mathcal{M}$, the choice for p above satisfies

$$\mathbb{E}_{\pi \sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big] \lesssim \frac{1}{\gamma},$$

so we have $\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \lesssim \frac{1}{\gamma}$. ◁

In what follows, we will show that the different behavior for the DEC for these examples reflects the fact that the optimal regret is fundamentally different.

6.4 E2D Algorithm for General Decision Making

Estimation-to-Decisions (E2D), the meta-algorithm based on the DEC that we gave for structured bandits in Section 4, readily extends to general decision making. The general version of the meta-algorithm is given in Algorithm 1. Compared to structured bandits,

Algorithm 1 Estimation to Decision-Making (E2D) for General Decision Making

```
parameters: Exploration parameter $\gamma > 0$.
for $t = 1, 2, \cdots, T$ do
    Obtain $\widehat{M}^t$ from online estimation oracle with $(\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$.
    Compute                                    // Minimizer for $\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^t)$.
        $p^t = \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[ f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \widehat{M}^t(\pi)\big) \Big].$
    Sample decision $\pi^t \sim p^t$ and update estimation algorithm with $(\pi^t, r^t, o^t)$.
```

the main difference is that rather than trying to estimate the reward function $f^\star$, we now estimate the underlying model $M^\star$. To do so, we appeal once again to the notion of an *online estimation oracle*, but this time for *model estimation.*

At each timestep t, the algorithm calls invokes an online estimation oracle to obtain an estimate $\widehat{M}^t$ for $M^\star$ using the data $\mathcal{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$ observed so far. Using this estimate, E2D proceeds by computing the distribution p^t that achieves the value $\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^t)$ for the Decision-Estimation Coefficient. That is, we set

$$p^t = \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \widehat{M}^t(\pi)\big) \Big]. \tag{6.17}$$

E2D then samples the decision π^t from this distribution and moves on to the next round.

Like structured bandits, one can show that by running Estimation-to-Decisions in the general decision making setting, the regret for decision making is bounded in terms of the DEC and a notion of estimation error for the estimation oracle. The main difference is that for general decision making, the notion of estimation error we need to control is the sum of *Hellinger distances* between the estimates from the supervised estimation oracle $M^\star$, which we define via

$$\mathbf{Est}_{\mathsf{H}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big) \Big]. \tag{6.18}$$

With this definition, we can show that E2D enjoys the following bound on regret, analogous to Proposition 13.

Proposition 26: E2D (Algorithm 1) with exploration parameter $\gamma > 0$ guarantees that

$$\mathbf{Reg} \le \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{H}}, \tag{6.19}$$

almost surely, where $\widehat{\mathcal{M}}$ is any set such that $\widehat{M}^t \in \widehat{\mathcal{M}}$ for all $t \in [T]$.

Note that we can optimize over the parameter γ in the result above, which yields

$$\mathbf{Reg} \le \inf_{\gamma > 0} \Big\{ \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{H}} \Big\} \le 2 \cdot \inf_{\gamma > 0} \max \Big\{ \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \cdot T, \gamma \cdot \mathbf{Est}_{\mathsf{H}} \Big\}.$$

We will show in the sequel that for any finite class $\mathcal{M}$, the averaged exponential weights algorithm with the logarithmic loss achieves $\mathbf{Est}_{\mathsf{H}} \lesssim \log(|\mathcal{M}|/\delta)$ with probability at least $1-\delta$. For this algorithm, and most others we will consider, one can take $\widehat{\mathcal{M}} = \mathrm{co}(\mathcal{M})$. In fact, one can show (see **??**) that for any $\widehat{M}$, even if $\widehat{M} \notin \mathrm{co}(\mathcal{M})$, we have $\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \leq \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_{c\gamma}(\mathcal{M}, \widehat{M}) \leq \mathsf{dec}_{c\gamma}(\mathcal{M})$ for any absolute constant $c > 0$. This means we can restrict our attention to the convex hull without loss of generality. Putting these facts together, we see that for any finite class, it is possible to achieve

$$\mathbf{Reg} \leq \mathsf{dec}_{\gamma}(\mathcal{M}) \cdot T + \gamma \cdot \log(|\mathcal{M}|/\delta) \tag{6.20}$$

with probability at least $1-\delta$.

Proof of Proposition 26. We write

$$\begin{aligned}
\mathbf{Reg} &= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\big)\right] + \gamma \cdot \mathbf{Est}_{\mathsf{H}}.
\end{aligned}$$

For each t, since $M^\star \in \mathcal{M}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\big)\right] \\
&\leq \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi^t \sim p^t}[f^{M}(\pi_{M}) - f^{M}(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\big(M(\pi^t), \widehat{M}^t(\pi^t)\big)\right] \\
&= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \widehat{M}^t(\pi)\big)\right] \\
&= \mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}^t).
\end{aligned} \tag{6.21}$$

Summing over all rounds t, we conclude that

$$\mathbf{Reg} \leq \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{H}}.$$

□

Examples for the upper bound. We now revisit the examples from Section 6.3 and use $\mathsf{E2D}$ and Proposition 26 to derive regret bounds for them.

Example 6.4 (cont'd). For the Gaussian bandit problem from Example 6.4, plugging the bound $\mathsf{dec}_{\gamma}(\mathcal{M}_{\mathsf{MAB\text{-}G}}) \lesssim A/\gamma$ into Proposition 26 yields

$$\mathbf{Reg} \lesssim \frac{AT}{\gamma} + \gamma \cdot \mathbf{Est}_{\mathsf{H}},$$

Choosing $\gamma = \sqrt{AT/\mathbf{Est}_{\mathsf{H}}}$ balances the terms above and gives

$$\mathbf{Reg} \lesssim \sqrt{AT \cdot \mathbf{Est}_{\mathsf{H}}}.$$

◁

Example 6.5 (cont'd). For the bandit-type problem with structured noise from Example 6.5, the bound $\mathsf{dec}_{\gamma}(\mathcal{M}_{\mathsf{MAB\text{-}SN}}) \lesssim \mathbb{I}\{\gamma \leq A/4\}$ yields

$$\mathbf{Reg} \lesssim \mathbb{I}\{\gamma \leq A/4\} \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{H}}.$$

We can choose $\gamma = A$, which gives

$$\mathbf{Reg} \lesssim A \cdot \mathbf{Est}_{\mathsf{H}}.$$

◁

6.4.1 Online Estimation with Hellinger Distance

Let us now give some more detail as to how to perform the online model estimation required by **??**. Model estimation is a more challenging problem than regression, since we are estimating the underlying *condition distribution* rather than just the conditional mean. In spite of this difficulty, estimating the model $M^{\star}$ with respect to Hellinger distance is a classical problem that we can solve using the online learning tools introduced in Section 1; in particular, online *conditional density estimation* with the log loss. This generalizes the method of online regression employed in Sections 3 and 4.

Instead of directly performing estimation with respect to Hellinger distance, the simplest way to develop conditional density estimation algorithms is to work with the logarithmic loss. Given a tuple (π^t, r^t, o^t), define the logarithmic loss for a model M as

$$\ell^t_{\log}(M) = \log\left(\frac{1}{m^M(r^t, o^t \mid \pi^t)}\right), \tag{6.22}$$

where we define $m^M(\cdot, \cdot \mid \pi)$ as the conditional density for (r, o) under M. We define regret under the logarithmic loss as:

$$\mathbf{Reg}_{\mathsf{KL}} = \sum_{t=1}^{T} \ell^t_{\log}(\widehat{M}^t) - \inf_{M \in \mathcal{M}} \sum_{t=1}^{T} \ell^t_{\log}(M). \tag{6.23}$$

The following result shows that a bound on the log-loss regret immediately yields a bound on the Hellinger estimation error.

Lemma 22: For any online estimation algorithm, whenever Assumption 8 holds, we have

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{KL}}] \geq \mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{KL}}\left(M^{\star}(\pi^t) \,\|\, \widehat{M}^t(\pi^t)\right)\right], \tag{6.24}$$

so that

$$\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq \mathbb{E}[\mathbf{Reg}_{\mathsf{KL}}]. \tag{6.25}$$

Furthermore, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\mathbf{Est}_{\mathsf{H}} \leq \mathbf{Reg}_{\mathsf{KL}} + 2\log(\delta^{-1}). \tag{6.26}$$

This result is desirable because regret minimization with the logarithmic loss is a well-studied problem in online learning. Efficient algorithms are known for model classes of interest [23, 72, 44, 39, 58, 61, 33, 54], and this is complemented by theory which provides minimax rates for generic model classes [67, 57, 20, 14]. One example we have already seen (Section 1) is the averaged exponential weights method, which guarantees

$$\mathbf{Reg}_{\mathsf{KL}} \leq \log|\mathcal{M}|$$

for finite classes $\mathcal{M}$. Another example is that for linear models, where (i.e., $m^M(r, o \mid \pi) = \langle \phi(r, o, \pi), \theta \rangle$ for a fixed feature map in $\phi \in \mathbb{R}^d$), algorithms with $\mathbf{Reg}_{\mathsf{KL}} = O(d \log(T))$ are known [62, 67]. All of these algorithms satisfy $\widehat{\mathcal{M}} = \mathrm{co}(\mathcal{M})$. We refer the reader to Chapter 9 of [21] for further examples and discussion.

While (6.25) is straightforward, (6.26) is rather remarkable, as the remainder term does not scale with T. Indeed, a naive attempt at applying concentration inequalities to control the deviations of the random quantities $\mathbf{Est}_{\mathsf{H}}$ and $\mathbf{Reg}_{\mathsf{KL}}$ would require boundedness of the loss function, which is problematic because the logarithmic loss can be unbounded. The proof exploits unique properties of the moment generating function for the log loss.

6.5 Decision-Estimation Coefficient: Lower Bound on Regret

Up to this point, we have been focused on developing algorithms that lead to upper bounds on regret for specific model classes. We now turn our focus to lower bounds, and the question of optimality: That is, for a given class of models $\mathcal{M}$, what is the best regret that can be achieved by any algorithm? We will show that in addition to upper bounds, the Decision-Estimation Coefficient actually leads to *lower bounds* on the optimal regret.

Background: Minimax regret. What does it mean to say that an algorithm is *optimal* for a model class $\mathcal{M}$? There are many notions of optimality, but in this course we will focus on *minimax optimality*, which is one of the most basic and well-studied notions.

For a model class $\mathcal{M}$, we define the minimax regret via[14]

$$\mathfrak{M}(\mathcal{M}, T) = \inf_{p^1, \ldots, p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}(T)], \tag{6.27}$$

where $p^t = p^t(\cdot \mid \mathcal{H}^{t-1})$ is the algorithm's strategy for step t (a function of the history $\mathcal{H}^{t-1}$), and where we write regret as $\mathbf{Reg}(T)$ to make the dependence on T explicit. Intuitively, minimax regret asks what is the best any algorithm can perform on a worst-case model (in $\mathcal{M}$) possibly chosen with the algorithm in mind. Another way to say this is: For any algorithm, there exists a model in $\mathcal{M}$ for which $\mathbb{E}[\mathbf{Reg}(T)] \geq \mathfrak{M}(\mathcal{M}, T)$. We will say that an algorithm is *minimax optimal* if it achieves (6.27) up to absolute constants that do not depend on $\mathcal{M}$ or T.

6.5.1 The Constrained Decision-Estimation Coefficient

We now show how to lower bound the minimax regret for any model class $\mathcal{M}$ in terms of the DEC for $\mathcal{M}$. Instead of working with the quantity $\mathsf{dec}_\gamma(\mathcal{M})$ appearing in Proposition 26 directly, it will be more convenient to work with a related quantity called the *constrained*

[14] Here, for any algorithm $p = p^1, \ldots, p^T$, $\mathbb{E}^{M^\star, p}$ denotes the expectation with respect to the observation process $(r^t, o^t) \sim M^\star(\pi^t)$ and any randomization used by the algorithm, when $M^\star$ is the true model.

Decision-Estimation Coefficient, which we define for a parameter $\varepsilon > 0$ as[15]

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}(\pi)\right)\right] \leq \varepsilon^2 \right\},$$

with

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) := \sup_{\widehat{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\widehat{M}\}, \widehat{M}).$$

This is similar to the definition for the DEC we have been working with so far— which we will call the *offset DEC* going forward—-except that it places a hard constraint on the information gain as opposed to subtracting the information gain. Both quantities have a similar interpretation, since subtracting the information gain implicitly biases the max player towards model where the gain is small. Indeed, the offset DEC can be thought of as a Lagrangian relaxation of the constrained DEC, and always upper bounds it via

$$\begin{aligned}
\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) &= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}(\pi)\right)\right] \leq \varepsilon^2 \right\} \\
&= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \inf_{\gamma\geq 0} \left\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma\left(\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}(\pi)\right)\right] - \varepsilon^2\right) \right\} \vee 0 \\
&\leq \inf_{\gamma\geq 0} \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma\left(\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}(\pi)\right)\right] - \varepsilon^2\right) \right\} \vee 0 \\
&= \inf_{\gamma\geq 0} \left\{ \mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) + \gamma\varepsilon^2 \right\} \vee 0.
\end{aligned}$$

For the opposite direction, it is straightforward to show that

$$\mathsf{dec}_{\gamma}(\mathcal{M}) \lesssim \mathsf{dec}^{\mathrm{c}}_{\gamma^{-1/2}}(\mathcal{M}).$$

This inequality is lossy, but cannot be improved in general. That is, there some classes for which the constrained DEC is meaningfully smaller than the offset DEC. However, it is possible to relate the two quantities if we restrict to a "localized" sub-class of models that are not "too far" from the reference model $\widehat{M}$.

Proposition 27: Given a model $\widehat{M}$ and parameter α, define the localized subclass around $\widehat{M}$ via

$$\mathcal{M}_{\alpha}(\widehat{M}) = \left\{ M \in \mathcal{M} : f^{\widehat{M}}(\pi_{\widehat{M}}) \geq f^M(\pi_M) - \alpha \right\}. \tag{6.28}$$

For all $\varepsilon > 0$ and $\gamma \geq c_1 \cdot \varepsilon^{-1}$, we have

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = c_3 \cdot \sup_{\gamma\geq c_1\varepsilon^{-1}} \sup_{\widehat{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}_{\alpha(\varepsilon,\gamma)}(\widehat{M}), \widehat{M}), \tag{6.29}$$

where $\alpha(\varepsilon, \gamma) := c_2 \cdot \gamma\varepsilon^2$, and $c_1, c_2, c_3 > 0$ are absolute constants.

For many "well-behaved" classes one can consider (e.g., multi-armed bandits and linear bandits), one has $\mathsf{dec}_{\gamma}(\mathcal{M}_{\alpha(\varepsilon,\gamma)}(\widehat{M}), \widehat{M}) \approx \mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M})$ whenever $\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \approx \gamma\varepsilon^2$ (that

[15] We adopt the convention that the value of $\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M})$ is zero is there exists p such that the set of $M \in \mathcal{M}$ with $\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}(\pi)\right)\right] \leq \varepsilon^2$ is empty.

is, localization does not change the complexity), so that lower bounds in terms of the constrained DEC immediately imply lower bounds in terms of the offset DEC. In general, this is not the case, and it turns out that it is possible to obtain tighter *upper bounds* that depend on the constrained DEC by using a refined version of the E2D algorithm. We refer to Foster et al. [38] for details and further background on the constrained DEC.

6.5.2 Lower Bound

The main lower bound based on the constrained DEC is as follows.

Proposition 28 (DEC Lower Bound [38]): Let $\underline{\varepsilon}_T := c \cdot \frac{1}{\sqrt{T}}$, where $c > 0$ is a sufficiently small numerical constant. For all T such that the condition[a]

$$\mathsf{dec}^{\mathrm{c}}_{\underline{\varepsilon}_T}(\mathcal{M}) \geq 10\underline{\varepsilon}_T \tag{6.30}$$

is satisfied, it holds that for any algorithm, there exists a model $M \in \mathcal{M}$ for which

$$\mathbb{E}[\mathbf{Reg}(T)] \gtrsim \mathsf{dec}^{\mathrm{c}}_{\underline{\varepsilon}_T}(\mathcal{M}) \cdot T. \tag{6.31}$$

[a] The numerical constant here is not important.

Proposition 28 shows that for any algorithm and model class $\mathcal{M}$, the optimal regret must scale with the constrained DEC in the worst-case. As a concrete example, we will show in the sequel that for the multi-armed bandit with A actions, $\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon\sqrt{A}$, which leads to

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{AT}.$$

We mention in passing that by combining Proposition 28 with Proposition 27, we obtain the following lower bound based on the (localized) offset DEC.

Corollary 1: Fix $T \in \mathbb{N}$. Then for any algorithm, there exists a model $M \in \mathcal{M}$ for which

$$\mathbb{E}[\mathbf{Reg}(T)] \gtrsim \sup_{\gamma \gtrsim \sqrt{T}} \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}_{\alpha(T,\gamma)}(\widehat{M}), \widehat{M}), \tag{6.32}$$

where $\alpha(T, \gamma) := c \cdot \gamma / T$ for an absolute constant $c > 0$

The DEC is necessary and sufficient. To understand the significance of Proposition 28 more broadly, we state but do not prove the following *upper* bound on regret based on the constrained DEC, which is based on a refined variant of E2D.

Proposition 29 (Upper bound for constrained DEC [38]): Let $\mathcal{M}$ be a finite class, and set $\bar{\varepsilon}_T := c \cdot \sqrt{\frac{\log(|\mathcal{M}|/\delta)}{T}}$, where $c > 0$ is a sufficiently large numerical constant.

Under appropriate technical conditions, there exists an algorithm that achieves

$$\mathbb{E}[\mathbf{Reg}(T)] \lesssim \mathsf{dec}^{\mathrm{c}}_{\bar{\varepsilon}_T}(\mathcal{M}) \cdot T \tag{6.33}$$

with probability at least $1 - \delta$.

This matches the lower boud in Proposition 28 upper to a difference in the radius: we have $\underline{\varepsilon}_T \propto \sqrt{\frac{1}{T}}$ for the lower bound, and $\bar{\varepsilon}_T \propto \sqrt{\frac{\log(|\mathcal{M}|/\delta)}{T}}$ for the upper bound. This implies that for any class where $\log|\mathcal{M}| < \infty$, the constrained DEC is *necessary and sufficient* for low regret. By the discussion in the prequel, a similar conclusion holds for the offset DEC (albeit, with a polynomial loss in rate). The interpretation of the $\log|\mathcal{M}|$ gap between the upper and lower bounds is that the DEC is capturing the complexity of exploring the decision space, but the statistical capacity required to estimate the underlying model is a separate issue which is not captured.

6.5.3 Proof of Proposition 28

Before proving Proposition 28, let us give some background on a typical approach to proving lower bounds on the minimax regret for a decision making problem.

Anatomy of a lower bound. How should one go about proving a lower bound on the minimax regret in (6.27)? We will follow a general recipe which can be found throughout statistics, information theory, and decision making [28, 76, 71]. The approach will be to find a pair of models M and $\widehat{M}$ that satisfy the following properties:

1. Any algorithm with regret much smaller than the DEC must query substantially different decisions in Π depending on whether the underlying model is M or $\widehat{M}$. Intuitively, this means that any algorithm that achieves low regret must be able to distinguish between the two models.

2. M and $\widehat{M}$ are "close" in a statistical sense (typically via total variation distance or another f-divergence), which implies via change-of-measure arguments that the decisions played by any algorithm which interacts with the models only via observations (in our case, (π^t, r^t, o^t)) will be similar for both models. In other words, the models are difficult to distinguish.

One then concludes that the algorithm must have large regret on either M or $\widehat{M}$.

To make this approach concrete, classical results in statistical estimation and supervised learning choose the models M and $\widehat{M}$ in a way that is *oblivious* to the algorithm under consideration [28, 76, 71]. However, due to the interactive nature of the decision making problem, the lower bound proof we present now will choose the models in an *adaptive* fashion.

Simplifications. Rather than proving the full result in Proposition 28, we will make the following simplifying assumptions:

- There exists a constant C such that

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \le C \cdot D^2_{\mathsf{H}}\big(M(\pi), M'(\pi)\big) \tag{6.34}$$

  for all $M, M' \in \mathcal{M}$ and $\pi \in \Pi$.

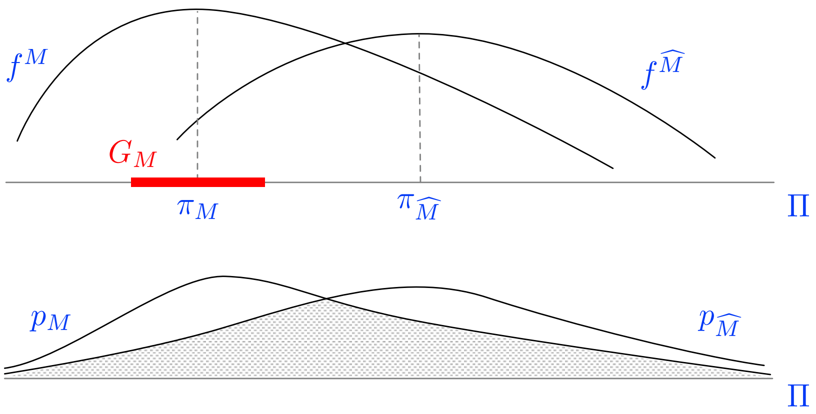


Figure 9: Models M and $\widehat{M}$ with corresponding mean rewards and average action distributions. The overlap between the action distributions is at least 0.9, while near-optimal choices for one model incur large regret for the other.

- Rather than proving a lower bound that scales with $\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\widehat{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\widehat{M}\}, \widehat{M})$, we will prove a weaker lower bound that scales with $\sup_{\widehat{M}\in\mathcal{M}} \mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M})$.

We refer to Foster et al. [38] for a full proof that removes these restrictions.

Preliminaries. We use the following technical lemma for the proof of Proposition 28.

Lemma 23 (Chain Rule for KL Divergence): Let $(\mathcal{X}_1, \mathscr{F}_1), \ldots, (\mathcal{X}_n, \mathscr{F}_n)$ be a sequence of measurable spaces, and let $\mathcal{X}^i = \prod_{i=t}^{i} \mathcal{X}_t$ and $\mathscr{F}^i = \bigotimes_{t=1}^{i} \mathscr{F}_t$. For each i, let $\mathbb{P}^i(\cdot \mid \cdot)$ and $\mathbb{Q}^i(\cdot \mid \cdot)$ be probability kernels from $(\mathcal{X}^{i-1}, \mathscr{F}^{i-1})$ to $(\mathcal{X}_i, \mathscr{F}_i)$. Let $\mathbb{P}$ and $\mathbb{Q}$ be the laws of $X_1, \ldots, X_n$ under $X_i \sim \mathbb{P}^i(\cdot \mid X_{1:i-1})$ and $X_i \sim \mathbb{Q}^i(\cdot \mid X_{1:i-1})$ respectively. Then it holds that

$$D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D_{\mathsf{KL}}(\mathbb{P}^i(\cdot \mid X_{1:i-1}) \,\|\, \mathbb{Q}^i(\cdot \mid X_{1:i-1}))\right]. \tag{6.35}$$

Proof of Proposition 28. Fix $T \in \mathbb{N}$ and consider any fixed algorithm, which we recall is defined by a sequence of mappings $p^1, \ldots, p^T$, where $p^t = p^t(\cdot \mid \mathcal{H}^{t-1})$. Let $\mathbb{P}^M$ denote the distribution over $\mathcal{H}^T$ for this algorithm when M is the true model, and let $\mathbb{E}^M$ denote the corresponding expectation.

Viewed as a function of the history $\mathcal{H}^{t-1}$, each p^t is a random variable, and we can consider its expected value under the model M. To this end, for any model $M \in \mathcal{M}$, let

$$p_M := \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^{T} p^t\right] \in \Delta(\Pi)$$

be the algorithm's average action distribution when M is the true model. Our aim is to show that we can find a model in $\mathcal{M}$ for which the algorithm's regret is at least as large as the lower bound in (6.32).

Let $T \in \mathbb{N}$, and fix a value $\varepsilon > 0$ to be chosen momentarily. Fix an arbitrary model $\widehat{M} \in \mathcal{M}$ and set

$$M = \underset{M \in \mathcal{M}}{\arg\max} \Big\{ \mathbb{E}_{\pi \sim p_{\widehat{M}}}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim p_{\widehat{M}}}\Big[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big] \le \varepsilon^2 \Big\}, \tag{6.36}$$

The model M should be thought of as a "worst-case alternative" g to $\widehat{M}$, but only for *the specific algorithm under consideration*. We will show that the algorithm needs to have large regret on either M or $\widehat{M}$. To this end, we establish some basic properties; let us abbreviate $g^M(\pi) = f^M(\pi_M) - f^M(\pi)$ going forward:

- For all models M, we have

$$\frac{1}{T}\,\mathbb{E}^M[\mathbf{Reg}(T)] = \mathbb{E}_{\pi \sim p_M}[g^M(\pi)]. \tag{6.37}$$

  So, to prove the desired lower bound, we need to show that either $\mathbb{E}_{\pi \sim p_M}[g^M(\pi)]$ or $\mathbb{E}_{\pi \sim p_{\widehat{M}}}\big[g^{\widehat{M}}(\pi)\big]$ is large.

- By the definition of the constrained DEC, we have

$$\mathbb{E}_{\pi \sim p_{\widehat{M}}}[g^M(\pi)] \ge \mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) =: \Delta, \tag{6.38}$$

  since by (6.36), the model M is the best response to a potentially suboptimal choice $p_{\widehat{M}}$. This is almost what we want, but there is a mismatch in models, since g^M considers the model M while $p_{\widehat{M}}$ considers the model $\widehat{M}$.

- Using the chain rule for KL divergence, we have

$$\begin{aligned} D_{\mathsf{KL}}\Big(\mathbb{P}^{\widehat{M}} \,\|\, \mathbb{P}^M\Big) &= \mathbb{E}^{\widehat{M}}\Bigg[\sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t}\, D_{\mathsf{KL}}\Big(\widehat{M}(\pi^t) \,\|\, M(\pi^t)\Big)\Bigg] \\ &\le C \cdot \mathbb{E}^{\widehat{M}}\Bigg[\sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t}\, D^2_{\mathsf{H}}\Big(\widehat{M}(\pi^t), M(\pi^t)\Big)\Bigg] \quad = CT \cdot \mathbb{E}_{\pi \sim p_{\widehat{M}}}\Big[D^2_{\mathsf{H}}\Big(\widehat{M}(\pi), M(\pi)\Big)\Big]. \end{aligned}$$

  To see why the first equality holds, we apply the chain rule to the sequence $\pi^1, z^1, \ldots, \pi^T, z^T$ with $z^t = (r^t, o^t)$. Let us use the bold notation $\mathbf{z}^t$ to refer to a random variable under consideration, and let z^t refer to its realization. Then we have

$$\begin{aligned} &D_{\mathsf{KL}}\Big(\mathbb{P}^{\widehat{M}} \,\|\, \mathbb{P}^M\Big) \\ &= \mathbb{E}^{\widehat{M}}\Bigg[\sum_{t=1}^T D_{\mathsf{KL}}\Big(\mathbb{P}^{\widehat{M}}(\mathbf{z}^t|\mathcal{H}^{t-1}, \pi^t) \,\|\, \mathbb{P}^M(\mathbf{z}^t|\mathcal{H}^{t-1}, \pi^t)\Big) + D_{\mathsf{KL}}\Big(\mathbb{P}^{\widehat{M}}(\boldsymbol{\pi}^t|\mathcal{H}^{t-1} \,\|\, \mathbb{P}^M(\boldsymbol{\pi}^t|\mathcal{H}^{t-1})\Big)\Bigg] \\ &= \mathbb{E}^{\widehat{M}}\Bigg[\sum_{t=1}^T D_{\mathsf{KL}}\Big(\widehat{M}(\pi^t) \,\|\, M(\pi^t)\Big)\Bigg] \end{aligned}$$

  since conditionally on $\mathcal{H}^{t-1}$, the law of π^t does not depend on the model.

  We can now choose $\varepsilon = c_1 \cdot \frac{1}{\sqrt{CT}}$, where $c_1 > 0$ is a sufficiently small numerical constant, to ensure that

$$D^2_{\mathsf{TV}}\Big(\mathbb{P}^{\widehat{M}}, \mathbb{P}^M\Big) \le D_{\mathsf{KL}}\Big(\mathbb{P}^{\widehat{M}} \,\|\, \mathbb{P}^M\Big) \le 1/100. \tag{6.39}$$

  In other words, with constant probability, the algorithm can fail to distinguish M and $\widehat{M}$.

Finally, we will make use of the fact that since rewards are in $[0,1]$, we have

$$\mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[f^{M}(\pi)-f^{\widehat{M}}(\pi)\right] \leq \mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[D_{\mathsf{TV}}\left(M(\pi),\widehat{M}(\pi)\right)\right] \leq \sqrt{\mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[D^2_{\mathsf{H}}\left(M(\pi),\widehat{M}(\pi)\right)\right]} \leq \varepsilon. \tag{6.40}$$

Step 1. Define $G_M = \{\pi\in\Pi \mid g^M(\pi)\leq\Delta/10\}$. Observe that

$$\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] \geq \frac{\Delta}{10}\cdot p_M(\pi\notin G_M) \geq \frac{\Delta}{10}\cdot(p_{\widehat{M}}(\pi\notin G_M) - D_{\mathsf{TV}}(p_M,p_{\widehat{M}})) \tag{6.41}$$

$$\geq \frac{\Delta}{10}\cdot(p_{\widehat{M}}(\pi\notin G_M) - 1/10), \tag{6.42}$$

since $D_{\mathsf{TV}}(p_M,p_{\widehat{M}}) \leq D_{\mathsf{TV}}\big(\mathbb{P}^M,\mathbb{P}^{\widehat{M}}\big)\leq 1/10$ by the data-processing inequality and (6.39). Going forward, let us assume that

$$\mathbb{E}_{\pi\sim p_{\widehat{M}}}\big[g^{\widehat{M}}(\pi)\big]\leq\Delta/10, \tag{6.43}$$

or else we are done, by (6.37). Our aim is to show that under this assumption, $p_{\widehat{M}}(\pi\notin G_M)\geq 1/2$, which will imply that $\mathbb{E}_{\pi\sim p_M}[g^M(\pi)]\gtrsim\Delta$ via (6.42).

Step 2. By adding the inequalities (6.43) and (6.38), we have that

$$\begin{aligned} f^M(\pi_M)-f^{\widehat{M}}(\pi_{\widehat{M}}) &\geq \mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[g^M(\pi)-g^{\widehat{M}}(\pi)\right] - \mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[|f^M(\pi)-f^{\widehat{M}}(\pi)|\right] \\ &\geq \frac{9}{10}\Delta - \mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[|f^M(\pi)-f^{\widehat{M}}(\pi)|\right]. \end{aligned}$$

In addition, by (6.40), we have $\mathbb{E}_{\pi\sim p_{\widehat{M}}}\big[|f^M(\pi)-f^{\widehat{M}}(\pi)|\big]\leq\varepsilon$, so that

$$f^M(\pi_M)-f^{\widehat{M}}(\pi_{\widehat{M}}) \geq \frac{9}{10}\Delta-\varepsilon. \tag{6.44}$$

Hence, as long as $\varepsilon\leq\frac{1}{10}\Delta$, which is implied by (6.30), we have

$$f^M(\pi_M)-f^{\widehat{M}}(\pi_{\widehat{M}}) \geq \frac{4}{5}\Delta. \tag{6.45}$$

Step 3. Observe that if $\pi\in G_M$, then

$$|f^M(\pi)-f^{\widehat{M}}(\pi)|_+ \geq |f^M(\pi_M)-f^{\widehat{M}}(\pi)-\Delta/10|_+ \geq |f^M(\pi_M)-f^{\widehat{M}}(\pi_{\widehat{M}})-\Delta/10|_+ \geq \frac{7}{10}\Delta,$$

where we have used (6.45). As a result, using (6.40),

$$\varepsilon \geq \mathbb{E}_{\pi\sim p_{\widehat{M}}}\left[|f^M(\pi)-f^{\widehat{M}}(\pi)|_+\right] \geq \frac{7}{10}\Delta\cdot p_{\widehat{M}}(\pi\in G_M).$$

Hence, since $\varepsilon\leq\Delta/10$ by (6.30), we have

$$\frac{\Delta}{10}\geq\frac{7}{10}\Delta\cdot p_{\widehat{M}}(\pi\in G_M),$$

or $p_{\widehat{M}}(\pi\in G_M)\leq 1/7$. Combining this with (6.42) gives

$$\frac{1}{T}\,\mathbb{E}^M[\mathbf{Reg}(T)] = \mathbb{E}_{\pi\sim p_M}[g^M(\pi)] \geq \frac{\Delta}{10}\cdot(1-1/7-1/10)\geq\frac{\Delta}{20}.$$

Finishing up. Note that since the choice of $\widehat{M} \in \mathcal{M}$ for this lower bound was arbitrary, we are free to choose $\widehat{M}$ to maximize $\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M})$. □

6.5.4 Examples of the Lower Bound

We now instantiate the lower bound in Proposition 28 for concrete model classes of interest. We begin by revisiting the examples at the beginning of the section.

Example 6.4 (cont'd). Let us lower bound the constrained DEC for the Gaussian bandit problem from Example 6.4. Set $\widehat{M}(\pi) = \mathcal{N}(1/2, 1)$, and let $\{M_1, \ldots, M_A\} \subseteq \mathcal{M}$ be a sub-family of models with $M_i(\pi) = \mathcal{N}(f^{M_i}(\pi), 1)$, where $f^{M_i}(\pi) := \frac{1}{2} + \Delta\mathbb{I}\{\pi = i\}$ for a parameter Δ whose value will be chosen in a moment. Observe that for all i, $\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M_i(\pi), \widehat{M}(\pi)\Big)\Big] \leq \frac{1}{2}\Delta^2 p(i)$ by (6.12), and $\mathbb{E}_{\pi\sim p}\big[f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi)\big] = (1 - p(i))\Delta$, so we can lower bound

$$\begin{aligned}\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) &= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \Big\{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big] \leq \varepsilon^2\Big\} \\ &\geq \inf_{p\in\Delta(\Pi)} \max_{i} \bigg\{(1 - p(i))\Delta \mid p(i)\frac{\Delta^2}{2} \leq \varepsilon^2\bigg\}\end{aligned}$$

For any p, there exists i such that $p(i) \leq 1/A$. If we choose $\Delta = \varepsilon \cdot \sqrt{2A}$, this choice for i will satisfy the constraint $p(i)\frac{\Delta^2}{2} \leq \varepsilon^2$, and we will be left with

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) \geq (1 - p(i))\Delta \geq \varepsilon\sqrt{A/2},$$

since $1 - p(i) \geq 1/2$.

Plugging this lower bound on the constrained Decision-Estimation Coefficient into Proposition 28 yields

$$\mathbb{E}[\mathbf{Reg}] \geq \widetilde{\Omega}(\sqrt{AT}).$$

◁

Generalizing the argument above, we can prove a lower bound on the Decision-Estimation Coefficient for any model class $\mathcal{M}$ that "embeds" the multi-armed bandit problem in a certain sense.

Proposition 30: Let a reference model $\widehat{M}$ be given, and suppose that a class $\mathcal{M}$ contains a sub-class $\{M_1, \ldots, M_N\}$ and collection of decisions $\pi_1, \ldots, \pi_N$ with the property that for all i:

1. $D^2_{\mathsf{H}}\Big(M_i(\pi), \widehat{M}(\pi)\Big) \leq \beta^2 \cdot \mathbb{I}\{\pi = \pi_i\}$.

2. $f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \alpha \cdot \mathbb{I}\{\pi \neq \pi_i\}$.

Then

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) \gtrsim \alpha \cdot \mathbb{I}\Big\{\varepsilon \geq \beta/\sqrt{N}\Big\}.$$

The examples that follow can be obtained by applying this result with an appropriate sub-family.

Example 6.5 (cont'd). Recall the bandit-type problem with structured noise from Example 6.5, where we have $\mathcal{M} = \{M_1, \ldots, M_A\}$, with $M_i(\pi) = \mathcal{N}(1/2, 1)\mathbb{I}\{\pi \neq i\} + \mathrm{Ber}(3/4)\mathbb{I}\{\pi = i\}$. If we set $\widehat{M}(\pi) = \mathcal{N}(1/2, 1)$, then this family satisfies the conditions of Proposition 30 with $\alpha = 1/4$ and $\beta^2 = 2$. As a result, we have $\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\mathsf{MAB\text{-}SN}}) \gtrsim \mathbb{I}\left\{\varepsilon \geq \sqrt{2/A}\right\}$, which yields

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \widetilde{O}(A)$$

if we apply Proposition 28.

◁

Example 6.6 (cont'd). Consider the full-information variant of the bandit setting in Example 6.6. By adapting the argument in Example 6.4, one can show that

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \gtrsim \varepsilon,$$

which leads to a lower bound of the form

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{T}.$$

◁

Next, we revisit some of the structured bandit classes considered in Section 4.

Example 6.7. Consider the linear bandit setting in Section 4.3.2, with $\mathcal{F} = \{\pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta\}$, where $\Theta \subseteq \mathsf{B}^d_2(1)$ is a parameter set and $\phi : \Pi \to \mathbb{R}^d$ is a fixed feature map that is known to the learner. Let $\mathcal{M}$ be the set of all reward distributions with $f^M \in \mathcal{F}$ and 1-sub-Gaussian noise. Then

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \gtrsim \varepsilon\sqrt{d},$$

which gives

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{dT}.$$

◁

Example 6.8. Consider the Lipschitz bandit setting in Section 4.3.3, where Π is a metric space with metric ρ, and

$$\mathcal{F} = \{f : \Pi \to [0,1] \mid f \text{ is 1-Lipschitz w.r.t } \rho\}.$$

Let $\mathcal{M}$ be the set of all reward distributions with $f^M \in \mathcal{F}$ and 1-sub-Gaussian noise. Let $d > 0$ be such that the covering number for Π satisfies

$$\mathcal{N}_{\rho}(\Pi, \varepsilon) \geq \varepsilon^{-d}.$$

Then

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \gtrsim \varepsilon^{\frac{2}{d+2}},$$

which leads to $\mathbb{E}[\mathbf{Reg}] \gtrsim T^{\frac{d+1}{d+2}}$. ◁

See Foster et al. [35, 38] for further details.

6.6 Decision-Estimation Coefficient and E2D: Application to Tabular RL

In this section, we use the Decision-Estimation Coefficient and E2D meta-algorithm to provide regret bounds for the tabular reinforcement learning. This will be the most complex example we consider in this section, and showcases the full power of DEC for general decision making. In particular, the example will show how the DEC can take advantage of the observations o^t, in the form of trajectories. This will provide an alternative to the optimistic algorithm (UCB-VI) we introduced in Section 5, and we will build on this approach to give guarantees for reinforcement learning with function approximation in Section 7.

Tabular reinforcement learning. When we view tabular reinforcement learning as a special case of the general decision making framework, $\mathcal{M}$ is the collection of all non-stationary MDPs $M = \left\{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\}$ (cf. Section 5), with state space $\mathcal{S} = [S]$, action space $\mathcal{A} = [A]$, and horizon H. The decision space $\Pi = \Pi_{\mathrm{RNS}}$ is the collection of all randomized, non-stationary Markov policies (cf. Example 6.3). We assume that rewards are normalized such that $\sum_{h=1}^H r_h \in [0,1]$ almost surely (so that $\mathcal{R} = [0,1]$). Recall that for each $M \in \mathcal{M}$, $\{P_h^M\}_{h=1}^H$ and $\{R_h^M\}_{h=1}^H$ denote the associated transition kernels and reward distributions, and d_1 is the initial state distribution.

Occupancy measures. The results we present make use of the notion of *occupancy measures* for an MDP M. Let $\mathbb{P}^{M,\pi}(\cdot)$ denote the law of a trajectory evolving under MDP M and policy π. We define state occupancy measures via

$$d_h^{M,\pi}(s) = \mathbb{P}^{M,\pi}(s_h = s)$$

and state-action occupancy measures via

$$d_h^{M,\pi}(s,a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a).$$

Note that we have $d_1^{M,\pi}(s) = d_1(s)$ for all M and π.

Bounding the DEC for tabular RL. Recall, that to certify a bound on the DEC, we need to—given any parameter $\gamma > 0$ and estimator $\widehat{M}$, exhibit a distribution (or, "strategy") p such that

$$\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma\cdot D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}(\pi)\right)\right] \le \overline{\mathsf{dec}}_\gamma(\mathcal{M}, \widehat{M})$$

for some upper bound $\overline{\mathsf{dec}}_\gamma(\mathcal{M}, \widehat{M})$. For tabular RL, we will choose p using an algorithm called *Policy Cover Inverse Gap Weighting*, which is displayed in Algorithm 2. As the name suggests, the approach combines the inverse gap weighting technique introduced in the multi-armed bandit setting with the notion of a *policy cover*—that is, a collection of policies that ensures good coverage on every state [29, 55, 41].

Algorithm 2 consists of two steps. First, in (6.46), we compute the collection of policies $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ that constitutes the policy cover. The basic idea here is that each policies in the policy cover should balance (i) regret and (ii) *coverage*—that is—ensure that all the states are sufficiently reached, which means we are exploring. We accomplish this by using policies of the form

$$\pi_{h,s,a} := \mathop{\arg\max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_h^{\widehat{M},\pi}(s,a)}{2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}$$

Algorithm 2 Policy Cover Inverse Gap Weighting (PC-IGW)

```
parameters:
    Estimated model $\widehat{M}$.
    Exploration parameter $\eta > 0$.
Define inverse gap weighted policy cover $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ via
    $\pi_{h,s,a} = \arg\max_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_h^{\widehat{M},\pi}(s,a)}{2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}.$   (6.46)
For each policy $\pi \in \Psi \cup \{\pi_{\widehat{M}}\}$, define
    $p(\pi) = \frac{1}{\lambda + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))},$   (6.47)
    where $\lambda \in [1, 2HSA]$ is chosen such that $\sum_\pi p(\pi) = 1$.
return $p$.
```

which—for each (s,a,h) tuple—maximize the ratio of the occupancy measure for (s,a) at layer h to the regret gap under $\widehat{M}$. This *inverse gap weighted policy cover* balances exploration and exploration by trading off coverage with suboptimality. With the policy cover in hand, the second step of Algorithm 2 computes the exploratory distribution p by simply applying inverse gap weighting to the elements of the cover and the greedy policy $\pi_{\widehat{M}}$.

The bound on the Decision-Estimation Coefficient for the PC-IGW algorithm is as follows.

Proposition 31: Consider the tabular reinforcement learning setting with $\sum_{h=1}^{H} r_h \in \mathcal{R} := [0,1]$. For any $\gamma > 0$ and $\widehat{M} \in \mathcal{M}$, the PC-IGW strategy in Algorithm 2, with $\eta = \frac{\gamma}{21H^2}$, ensures that

$$\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \widehat{M}(\pi)\big)\Big] \lesssim \frac{H^3SA}{\gamma},$$

and consequently certifies that $\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \lesssim \frac{H^3SA}{\gamma}$.

We remark that it is also possible to prove this bound non-constructively, by moving to the Bayesian DEC and adapting the posterior sampling approach described in Section 4.4.2.

Remark 18 (Computational efficiency): The PC-IGW strategy can be implemented in a computationally efficient fashion. Briefly, the idea is to solve (6.46) by taking a dual approach and optimizing over occupancy measures rather than policies. With this parameterization, (6.46) becomes a linear-fractional program, which can then be transformed into a standard linear program using classical techniques.

How to estimate the model. The bound on the DEC we proved using the PC-IGW algorithm assumes that $\widehat{M} \in \mathcal{M}$, but in general, estimators from online learning algorithm such as exponential weights will produce $\widehat{M}^t \in \mathrm{co}(\mathcal{M})$. While it is possible to show that the same bound on the DEC holds for $\widehat{M} \in \mathrm{co}(\mathcal{M})$, a slightly more complex version of the algorithm is required to certify such a bound. To run the PC-IGW algorithm as-is, we can use a simple approach to obtain a proper estimator $\widehat{M} \in \mathcal{M}$.

Assume for simplicity that rewards are known, i.e. $R_h^M(s,a) = R_h(s,a)$ for all $M \in \mathcal{M}$. Instead of directly working with an estimator for the entire model M, we work with layer-wise estimators $\mathbf{Alg}_{\mathsf{Est};1}, \ldots, \mathbf{Alg}_{\mathsf{Est};H}$. At each round t, given the history $\{(\pi^i, r^i, o^i)\}_{i=1}^{t-1}$, the layer-$h$ estimator $\mathbf{Alg}_{\mathsf{Est};h}$ produces an estimate $\widehat{P}_h^t$ for the true transition kernel $P_h^{M^\star}$. We measure performance of the estimator via layer-wise Hellinger error:

$$\mathbf{Est}_{\mathsf{H};h} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}^{M^\star, \pi^t}\Big[D_{\mathsf{H}}^2\Big(P_h^{M^\star}(s_h, a_h), \widehat{P}_h^t(s_h, a_h)\Big)\Big]. \tag{6.48}$$

We obtain an estimation algorithm for the full model $M^\star$ by taking $\widehat{M}^t$ as the MDP that has $\widehat{P}_h^t$ as the transition kernel for each layer h. This algorithm has the following guarantee.

Proposition 32: The estimator described above has

$$\mathbf{Est}_{\mathsf{H}} \leq O(\log(H)) \cdot \sum_{h=1}^{H} \mathbf{Est}_{\mathsf{H};h}.$$

In addition, $\widehat{M}^t \in \mathcal{M}$.

For each layer, we can obtain $\mathbf{Est}_{\mathsf{H};h} \leq \widetilde{O}(S^2 A)$ using the averaged exponential weights algorithm, by applying the approach described in Section 6.4.1 to each layer. That is, for each layer, we obtain $\widehat{P}_h^t$ by running averaged exponential weights with the loss $\ell_{\log}^t(P_h) = -\log(P_h(s_{h+1} \mid s_h, a_h))$. We obtain $\mathbf{Est}_{\mathsf{H};h} \leq \widetilde{O}(S^2 A)$ with this approach because there are $S^2 A$ parameters for the transition distribution at each layer.

A lower bound on the DEC. We state, but do not prove a complementary lower bound on the DEC for tabular RL.

Proposition 33: Let $\mathcal{M}$ be the class of tabular MDPs with $S \geq 2$ states, $A \geq 2$ actions, and $\sum_{h=1}^{H} r_h \in \mathcal{R} := [0,1]$. If $H \geq 2\log_2(S/2)$, then

$$\mathsf{dec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M}) \gtrsim \varepsilon\sqrt{HSA}.$$

Using Proposition 28, this gives $\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{HSAT}$.

6.6.1 Proof of Proposition 31

Toward proving Proposition 31, we provide some general-purpose technical lemmas which will find further use in Section 7 . First, we provide a *simulation lemma*, which allow us to decompose the difference in value functions for two MDPs into errors between their per-layer reward functions and transition probabilities.

Lemma 24 (Simulation lemma): For any pair of MDPs $M = (P^M, R^M)$ and $\widehat{M} = (P^{\widehat{M}}, R^{\widehat{M}})$ with the same initial state distribution and $\sum_{h=1}^H r_h \in [0,1]$, we have

$$\left|f^M(\pi) - f^{\widehat{M}}(\pi)\right| \leq D_{\mathsf{TV}}\Big(M(\pi), \widehat{M}(\pi)\Big) \tag{6.49}$$

$$\leq D_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big) \leq \frac{1}{2\eta} + \frac{\eta}{2} D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big) \quad \forall \eta > 0, \tag{6.50}$$

and

$$f^M(\pi) - f^{\widehat{M}}(\pi)$$

$$= \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi}\Big[\Big[(P^M_h - P^{\widehat{M}}_h)V^{M,\pi}_{h+1}\Big](s_h,a_h)\Big] + \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi}\Big[\mathbb{E}_{r_h \sim R^M_h(s_h,a_h)}[r_h] - \mathbb{E}_{r_h \sim R^{\widehat{M}}_h(s_h,a_h)}[r_h]\Big] \tag{6.51}$$

$$\leq \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi}\Big[D_{\mathsf{TV}}\Big(P^M_h(s_h,a_h), P^{\widehat{M}}_h(s_h,a_h)\Big) + D_{\mathsf{TV}}\Big(R^M_h(s_h,a_h), R^{\widehat{M}}_h(s_h,a_h)\Big)\Big]. \tag{6.52}$$

Next, we provide a "change-of-measure" lemma, which allows one to move from between quantities involving an estimator $\widehat{M}$ and those involving another model M.

Lemma 25 (Change of measure for RL): Consider any MDP M and reference MDP $\widehat{M}$ which satisfy $\sum_{h=1}^H r_h \in [0,1]$. For all $p \in \Delta(\Pi)$ and $\eta > 0$ we have

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]$$

$$\leq \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^{\widehat{M}}(\pi)\Big] + \eta\,\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big] + \frac{1}{4\eta}. \tag{6.53}$$

and

$$\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\left[\sum_{h=1}^H D^2_{\mathsf{TV}}\Big(P^M(s_h,a_h), P^{\widehat{M}}(s_h,a_h)\Big) + D^2_{\mathsf{TV}}\Big(R^M(s_h,a_h), R^{\widehat{M}}(s_h,a_h)\Big)\right]$$

$$\leq 8H\,\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big]. \tag{6.54}$$

Proof of Proposition 31. Let $M \in \mathcal{M}$ be fixed. The main effort in the proof will be to bound the quantity

$$\mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^{\widehat{M}}(\pi)\Big]$$

in terms of the quantity on the right-hand side of (6.54), then apply change of measure (Lemma 25). We begin with the decomposition

$$\mathbb{E}_{\pi\sim p}\Big[f^{M}(\pi_{M}) - f^{\widehat{M}}(\pi)\Big] = \underbrace{\mathbb{E}_{\pi\sim p}\Big[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)\Big]}_{\text{(I)}} + \underbrace{f^{M}(\pi_{M}) - f^{\widehat{M}}(\pi_{\widehat{M}})}_{\text{(II)}}. \tag{6.55}$$

For the first term (I), which may be thought of as exploration bias, we have

$$\mathbb{E}_{\pi\sim p}\Big[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)\Big] = \sum_{\pi\in\Psi\cup\{\pi_{\widehat{M}}\}} \frac{f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)}{\lambda + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))} \leq \frac{2HSA}{\eta}, \tag{6.56}$$

where we have used that $\lambda \geq 0$. We next bound the second term (II), which entails showing that the PC-IGW distribution *explores enough.* We have

$$f^{M}(\pi_{M}) - f^{\widehat{M}}(\pi_{\widehat{M}}) = f^{M}(\pi_{M}) - f^{\widehat{M}}(\pi_{M}) - (f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_{M})). \tag{6.57}$$

We use the simulation lemma to bound

$$\begin{aligned} f^{M}(\pi_{M}) - f^{\widehat{M}}(\pi_{M}) &\leq \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi_{M}}\Big[D_{\mathsf{TV}}\Big(P_h^{M}(s_h,a_h), P_h^{\widehat{M}}(s_h,a_h)\Big) + D_{\mathsf{TV}}\Big(R_h^{M}(s_h,a_h), R_h^{\widehat{M}}(s_h,a_h)\Big)\Big] \\ &= \sum_{h=1}^{H}\sum_{s,a} d_h^{\widehat{M},\pi_{M}}(s,a)\mathrm{err}_h^{M}(s,a), \end{aligned}$$

where $\mathrm{err}_h^{M}(s,a) := D_{\mathsf{TV}}\big(P^{M}(s,a), P^{\widehat{M}}(s,a)\big) + D_{\mathsf{TV}}\big(R^{M}(s,a), R^{\widehat{M}}(s,a)\big)$. Define $\bar{d}_h(s,a) = \mathbb{E}_{\pi\sim p}\big[d_h^{\widehat{M},\pi}(s,a)\big]$. Then, using the AM-GM inequality, we have that for any $\eta' > 0$,

$$\begin{aligned} \sum_{h=1}^{H}\sum_{s,a} d_h^{\widehat{M},\pi_{M}}(s,a)[\mathrm{err}_h^{M}(s,a)] &= \sum_{h=1}^{H}\sum_{s,a} d_h^{\widehat{M},\pi_{M}}(s,a)\left(\frac{\bar{d}_h(s,a)}{\bar{d}_h(s,a)}\right)^{1/2}(\mathrm{err}_h^{M}(s,a))^2 \\ &\leq \frac{1}{2\eta'}\sum_{h=1}^{H}\sum_{s,a}\frac{(d_h^{\widehat{M},\pi_{M}}(s,a))^2}{\bar{d}_h(s,a)} + \frac{\eta'}{2}\sum_{h=1}^{H}\sum_{s,a}\bar{d}_h(s,a)(\mathrm{err}_h^{M}(s,a))^2 \\ &= \frac{1}{2\eta'}\sum_{h=1}^{H}\sum_{s,a}\frac{(d_h^{\widehat{M},\pi_{M}}(s,a))^2}{\bar{d}_h(s,a)} + \frac{\eta'}{2}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\big[(\mathrm{err}_h^{M}(s_h,a_h))^2\big]. \end{aligned}$$

The second term is exactly the upper bound we want, so it remains to bound the ratio of occupancy measures in the first term. Observe that for each (h,s,a), we have

$$\frac{d_h^{\widehat{M},\pi_{M}}(s,a)}{\bar{d}_h(s,a)} \leq \frac{d_h^{\widehat{M},\pi_{M}}(s,a)}{d_h^{\widehat{M},\pi_{h,s,a}}(s,a)}\cdot\frac{1}{p(\pi_{h,s,a})} \leq \frac{d_h^{\widehat{M},\pi_{M}}(s,a)}{d_h^{\widehat{M},\pi_{h,s,a}}(s,a)}\Big(2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_{h,s,a})\Big),$$

where the second inequality follows from the definition of p and the fact that $\lambda \leq 2HSA$. Furthermore, since

$$\pi_{h,s,a} = \operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_h^{\widehat{M},\pi}(s,a)}{2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))},$$

and since $\pi_M \in \Pi_{\mathrm{RNS}}$, we can upper bound by

$$\frac{d_h^{\widehat{M},\pi_M}(s,a)}{d_h^{\widehat{M},\pi_M}(s,a)}\Big(2HSA+\eta(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi_M)\Big)=2HSA+\eta(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi_M). \tag{6.58}$$

As a result, we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{s,a}\frac{(d_h^{\widehat{M},\pi_M}(s,a))^2}{\bar{d}_h(s,a)} &\leq \sum_{h=1}^{H}\sum_{s,a}d_h^{\widehat{M},\pi_M}(s,a)(2HSA+\eta(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi_M))\\ &= 2H^2SA+\eta H(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi_M)).\end{aligned}$$

Putting everything together and returning to (6.57), this establishes that

$$\begin{aligned}&f^M(\pi_M)-f^{\widehat{M}}(\pi_{\widehat{M}})\\ &\leq \frac{H^2SA}{\eta'}+\frac{\eta'}{2}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\big[(\mathrm{err}_h^M(s_h,a_h))^2\big]+\frac{\eta H}{2\eta'}(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi_M))-(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi_M)).\end{aligned}$$

We set $\eta'=\frac{\eta H}{2}$ so that the latter terms cancel and we are left with

$$f^M(\pi_M)-f^{\widehat{M}}(\pi_{\widehat{M}})\leq\frac{2HSA}{\eta}+\frac{\eta H}{4}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\big[(\mathrm{err}_h^M(s_h,a_h))^2\big].$$

Combining this with (6.55) and (6.56) gives

$$\begin{aligned}&\mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M)-f^{\widehat{M}}(\pi)\Big]\\ &\leq\frac{4HSA}{\eta}+\frac{\eta H}{4}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\big[(\mathrm{err}_h^M(s_h,a_h))^2\big]\\ &\leq\frac{4HSA}{\eta}+\frac{\eta H}{2}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\Big[D^2_{\mathsf{TV}}\Big(P^M(s_h,a_h),P^{\widehat{M}}(s_h,a_h)\Big)+D^2_{\mathsf{TV}}\Big(R^M(s_h,a_h),R^{\widehat{M}}(s_h,a_h)\Big)\Big].\end{aligned}$$

We conclude by applying the change-of-measure lemma (Lemma 25), which implies that for any $\eta'>0$,

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]\leq\frac{4HSA}{\eta}+(4\eta')^{-1}+(4H^2\eta+\eta')\cdot\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big)\Big].$$

The result follows by choosing $\eta=\eta'=\frac{\gamma}{21H^2}$ (we have made no effort to optimize the constants here). □

6.7 Tighter Regret Bounds for the Decision-Estimation Coefficient

To close this section, we provide a number of refined regret bounds based on the Decision-Estimation Coefficient, which improve upon Proposition 26 in various situations.

Algorithm 3 E2D for General Divergences and Randomized Estimators

```
parameters: Exploration parameter $\gamma > 0$, divergence $D(\cdot \,\|\, \cdot)$.
for $t = 1, 2, \cdots, T$ do
    Obtain randomized estimate $\nu^t \in \Delta(\mathcal{M})$ from estimation oracle with $\{(\pi^i, r^i, o^i)\}_{i<t}$.
    Compute                                                    // Eq. (6.61).
        $$p^t = \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[ f^M(\pi_M) - f^M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M} \sim \nu^t}\left[ D^\pi\left( \widehat{M} \,\|\, M \right) \right] \right].$$
    Sample decision $\pi^t \sim p^t$ and update estimation algorithm with $(\pi^t, r^t, o^t)$.
```

6.7.1 Guarantees Based on Decision Space Complexity

In general, low estimation complexity (i.e., a small bound on $\mathbf{Est}_{\mathsf{H}}$ or $\log|\mathcal{M}|$) is not required to achieve low regret for decision making. This is because our end goal is to make good *decisions*, so we can give up on accurately estimating the model in regions of the decision space that do not help to distinguish the relative quality of decisions. The following result provides a tighter bound that scales only with $\log|\Pi|$, at the cost of depending on the DEC for a larger model class: $\mathrm{co}(\mathcal{M})$ rather than $\mathcal{M}$.

Proposition 34: There exists an algorithm that for any $\delta > 0$, ensures that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \inf_{\gamma > 0} \{ \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log(|\Pi|/\delta) \}. \tag{6.59}$$

Compared to (6.20), this replaces the estimation term $\log|\mathcal{M}|$ with the smaller quantity $\log|\Pi|$, replaces $\mathsf{dec}_\gamma(\mathcal{M})$ with the potentially larger quantity $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$. Whether or not this leads to an improvement depends on the class $\mathcal{M}$. For multi-armed bandits, linear bandits, and convex bandits, $\mathcal{M}$ is already convex, so this offers strict improvement. For MDPs though, $\mathcal{M}$ is not convex: Even for the simple tabular MDP setting where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$, grows exponentially $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$ in either H or S, whereas $\mathsf{dec}_\gamma(\mathcal{M})$ is polynomial in all parameters.

We mention in passing that this result is proven using a different algorithm from E2D; see Foster et al. [35, 37] for more background.

6.7.2 General Divergences and Randomized Estimators

In this section we give a generalization of the E2D algorithm that incorporates two extra features: *general divergences* and *randomized estimators*.

General divergences. The Decision-Estimation Coefficient measures estimation error via the Hellinger distance $D^2_{\mathsf{H}}\big(M(\pi), \widehat{M}(\pi)\big)$, which is fundamental in the sense that it leads to lower bounds on the optimal regret (Proposition 28). Nonetheless, for specific applications and model classes, it can be useful to work with alternative distance measures and divergences. For a non-negative function ("divergence") $D^{\pi}(\cdot \,\|\, \cdot)$, we define

$$\mathsf{dec}_{\gamma}^{D}(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D^{\pi}\left(\widehat{M} \,\|\, M\right)\right]. \tag{6.60}$$

This variant of the DEC naturally leads to regret bounds in terms of estimation error under $D^{\pi}(\cdot \,\|\, \cdot)$. Note that we use notation $D^{\pi}\left(\widehat{M} \,\|\, M\right)$ instead of say, $D\big(\widehat{M}(\pi), M(\pi)\big)$, to reflect that fact that the divergence may depend on M (resp. $\widehat{M}$) and π through properties other than $M(\pi)$ (resp. $\widehat{M}(\pi)$).

Randomized estimators. The basic version of $\mathsf{E2D}$ assumes that at each round, the online estimation oracle provides a point estimate $\widehat{M}^{t}$. In some settings, it useful to consider *randomized estimators* that, at each round, produce a distribution $\nu^{t} \in \Delta(\mathcal{M})$ over models. For this setting, we further generalize the DEC by defining

$$\overline{\mathsf{dec}}_{\gamma}^{D}(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M} \sim \nu}\left[D^{\pi}\left(\widehat{M} \,\|\, M\right)\right]\right] \tag{6.61}$$

for distributions $\nu \in \Delta(\mathcal{M})$. We additionally define $\overline{\mathsf{dec}}_{\gamma}^{D}(\mathcal{M}) = \sup_{\nu \in \Delta(\mathcal{M})} \overline{\mathsf{dec}}_{\gamma}^{D}(\mathcal{M}, \nu)$.

Algorithm. A generalization of $\mathsf{E2D}$ that incorporates general divergences and randomized estimators is given in Algorithm 3. The algorithm is identical to $\mathsf{E2D}$ with Option I, with the only differences being that i) we play the distribution that solves the minimax problem (6.61) with the user-specified divergence $D^{\pi}(\cdot \,\|\, \cdot)$ rather than squared Hellinger distance, and ii) we use the randomized estimate ν^{t} rather than a point estimate. Our performance guarantee for this algorithm depends on the estimation performance of the oracle's randomized estimates $\nu^{1}, \ldots, \nu^{T} \in \Delta(\mathcal{M})$ with respect to the given divergence $D^{\pi}(\cdot \,\|\, \cdot)$, which we define as

$$\mathbf{Est}_{\mathsf{D}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{t} \sim p^{t}} \mathbb{E}_{\widehat{M}^{t} \sim \nu^{t}}\left[D^{\pi^{t}}\left(\widehat{M}^{t} \,\|\, M^{\star}\right)\right]. \tag{6.62}$$

We have the following guarantee.

Proposition 35: Algorithm 3 with exploration parameter $\gamma > 0$ guarantees that

$$\mathbf{Reg} \leq \overline{\mathsf{dec}}_{\gamma}^{D}(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{D}} \tag{6.63}$$

almost surely.

Sufficient statistics and benefits of general divergences. Many divergences of interest have the useful property that they depend on the estimated model $\widehat{M}$ only through a "sufficient statistic" for the model class under consideration. Formally, there exists a *sufficient statistic space* Ψ and *sufficient statistic* $\boldsymbol{\psi} : \mathcal{M} \to \Psi$ with the property that we can write (overloading notation)

$$D^{\pi}\big(M \,\|\, M'\big) = D^{\pi}\big(\boldsymbol{\psi}(M) \,\|\, M'\big), \quad f^{M}(\pi) = f^{\boldsymbol{\psi}(M)}(\pi), \quad \text{and} \quad \pi_{M} = \pi_{\boldsymbol{\psi}(M)}$$

for all models M, M'. In this case, it suffices for the online estimation oracle to directly estimate the sufficient statistic by producing a randomized estimator $\nu^t \in \Delta(\Psi)$, and we can write the estimation error as

$$\mathbf{Est}_{\mathsf{D}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}_{\widehat{\psi}^t \sim \nu^t}\left[D^{\pi^t}\left(\widehat{\psi}^t \parallel M^\star\right)\right]. \tag{6.64}$$

The benefit of this perspective is that for many examples of interest, since the divergence depends on the estimate only through ψ, we can derive bounds on $\mathbf{Est}$ that scale with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.

For example, in structured bandit problems, one can work with the divergence

$$D_{\mathrm{Sq}}\left(\widehat{M}(\pi), M(\pi)\right) := (f^M(\pi) - f^{\widehat{M}}(\pi))^2$$

which uses the mean reward function as a sufficient statistic, i.e. $\psi(M) = f^M$. Here, it is clear that one can achieve $\mathbf{Est}_{\mathsf{D}} \lesssim \log|\mathcal{F}|$, which improves upon the rate $\mathbf{Est}_{\mathsf{H}} \lesssim \log|\mathcal{M}|$ for Hellinger distance, and recovers the specialized version of the $\mathsf{E2D}$ algorithm we considered in Section 4. Analogously, for reinforcement learning, one can consider value functions as a sufficient statistic, and use an appropriate divergence based on Bellman residuals to derive estimation guarantees that scale with the complexity $\log|\mathcal{Q}|$ of a given value function class $\mathcal{Q}$; see Section 7 for details.

Does randomized estimation help? Note that whenever D is convex in the first argument, we have $\overline{\mathsf{dec}}^D_\gamma(\mathcal{M}) \leq \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}^D_\gamma(\mathcal{M}, \widehat{M}) = \mathsf{dec}^D_\gamma(\mathcal{M})$ (that is, the randomized DEC is never larger than the vanilla DEC), but it is not immediately apparent whether the opposite direction of this inequality holds, and one might hope that working with the randomized DEC in (6.61) would lead to improvements over the non-randomized counterpart. The next result shows that this is not the case: Under mild assumptions on the divergence D, randomization offers no improvement.

Proposition 36: Let D be any bounded divergence with the property that for all models $M, M', \widehat{M}$ and $\pi \in \Pi$,

$$D^\pi(M \parallel M') \leq C\left(D^\pi\left(\widehat{M} \parallel M\right) + D^\pi\left(\widehat{M} \parallel M'\right)\right). \tag{6.65}$$

Then for all $\gamma > 0$,

$$\sup_{\widehat{M}} \mathsf{dec}^D_\gamma(\mathcal{M}, \widehat{M}) \leq \overline{\mathsf{dec}}^D_{\gamma/(2C)}(\mathcal{M}). \tag{6.66}$$

Squared Hellinger distance is symmetric and satisfies Condition (6.65) with $C = 2$. Hence, writing $\overline{\mathsf{dec}}^{\mathsf{H}}_\gamma(\mathcal{M})$ as shorthand for $\overline{\mathsf{dec}}^D_\gamma(\mathcal{M})$ with $D = D^2_{\mathsf{H}}(\cdot,\cdot)$, we obtain the following corollary.

Proposition 37: Suppose that $\mathcal{R} \subseteq [0,1]$. Then for all $\gamma > 0$,

$$\overline{\mathsf{dec}}^{\mathsf{H}}_{\gamma}(\mathcal{M}) \leq \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}^{\mathsf{H}}_{\gamma}(\mathcal{M}, \widehat{M}) \leq \sup_{\widehat{M}} \mathsf{dec}^{\mathsf{H}}_{\gamma}(\mathcal{M}, \widehat{M}) \leq \overline{\mathsf{dec}}^{\mathsf{H}}_{\gamma/4}(\mathcal{M}).$$

This shows that for Hellinger distance—at least from a statistical perspective—there is no benefit to using the randomized DEC compared to the original version. In some cases, however, strategies p that minimize $\overline{\mathsf{dec}}^{\mathsf{H}}_{\gamma}(\mathcal{M}, \nu)$ can be simpler to compute than strategies that minimize $\mathsf{dec}^{\mathsf{H}}_{\gamma}(\mathcal{M}, \widehat{M})$ for $\widehat{M} \in \mathrm{co}(\mathcal{M})$.

6.7.3 Optimistic Estimation

To derive stronger regret bounds that allow for estimation with general divergences, we can combine Estimation-to-Decisions with a specialized estimation approach introduced by Zhang [77] (see also Dann et al. [25], Agarwal and Zhang [3], Zhong et al. [78]), which we refer to as *optimistic estimation*. The results we present are based on Foster et al. [36].

Let a divergence $D^{\pi}(\cdot \,\|\, \cdot)$ be fixed. An *optimistic estimation oracle* $\mathbf{Alg}_{\mathsf{Est}}$ is an algorithm which, at each step t, given $\mathcal{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$, produces a randomized estimator $\nu^t \in \Delta(\mathcal{M})$. Compared to the previous section, the only change is that for a parameter $\gamma > 0$, we will measure the performance of the oracle via *optimistic estimation error*, defined as

$$\mathbf{OptEst}^{D}_{\gamma} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}_{\widehat{M}^t \sim \nu^t}\Big[D^{\pi}\Big(\widehat{M}^t \,\|\, M^{\star}\Big) + \gamma^{-1}(f^{M^{\star}}(\pi_{M^{\star}}) - f^{\widehat{M}^t}(\pi_{\widehat{M}^t})\Big]. \quad (6.67)$$

This quantity is similar to (6.62), but incorporates a bonus term

$$\gamma^{-1}(f^{M^{\star}}(\pi_{M^{\star}}) - f^{\widehat{M}^t}(\pi_{\widehat{M}^t})),$$

which encourages the estimation algorithm to *over-estimate* the optimal value $f^{M^{\star}}(\pi_{M^{\star}})$ for the underlying model, leading to a form of optimism.

Example 6.9 (Structured bandits)**.** Consider any structured bandit problem with decision space Π, function class $\mathcal{F} \subseteq (\Pi \to [0,1])$, and $\mathcal{O} = \{\varnothing\}$. Let $\mathcal{M}_{\mathcal{F}}$ be the class

$$\mathcal{M}_{\mathcal{F}} = \{M \mid f^{M} \in \mathcal{F}, M(\pi) \text{ is 1-sub-Gaussian } \forall \pi\}.$$

To derive bounds on the optimistic estimation error, we can appeal to an augmented version of the (randomized) exponential weights algorithm which, for a learning rate parameter $\eta > 0$, sets

$$\nu^t(f^M) \propto \exp\left(-\eta\left(\sum_{i<t}(f^M(\pi^i) - r^i)^2 - \gamma^{-1} f^M(\pi_M)\right)\right).$$

For an appropriate choice of η, this method achieves $\mathbb{E}\big[\mathbf{OptEst}^{D}_{\gamma}\big] \lesssim \log|\mathcal{F}| + \sqrt{T \log|\mathcal{F}|}/\gamma$ for $D = D_{\mathrm{Sq}}(\cdot,\cdot)$ [77]. ◁

Optimistic E2D. Algorithm 4 provides an *optimistic* variant of E2D, which we refer to as E2D.Opt. At each timestep t, the algorithm calls the estimation oracle to obtain a randomized estimator ν^t using the data $(\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$ collected so far.

Algorithm 4 Optimistic E2D (E2D.Opt)

```
1: parameters: Exploration parameter $\gamma > 0$, divergence $D(\cdot \,\|\, \cdot)$.
2: for $t = 1, 2, \cdots, T$ do
3:     Obtain randomized estimate $\nu^t \in \Delta(\mathcal{M})$ from optimistic estimation oracle with $\{(\pi^i, r^i, o^i)\}_{i<t}$.
4:     Compute                                                    // Eq. (6.68).
           $p^t = \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \mathbb{E}_{\widehat{M} \sim \nu^t}\Big[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^M(\pi) - \gamma \cdot D^\pi\Big(\widehat{M} \,\|\, M\Big)\Big].$
5:     Sample decision $\pi^t \sim p^t$ and update estimation algorithm with $(\pi^t, r^t, o^t)$.
```

The algorithm then uses the estimator to compute a distribution $p^t \in \Delta(\Pi)$ and samples π^t from this distribution. The main change relative to the version of E2D in Algorithm 3 is that the minimax problem in Algorithm 4 is derived from an "optimistic" variant of the DEC tailored to the optimistic estimation error in (6.67). This quantity, which we refer to as the *Optimistic Decision-Estimation Coefficient*, is defined for $\nu \in \Delta(\mathcal{M})$ as

$$\textsf{o-dec}_\gamma^D(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \mathbb{E}_{\widehat{M} \sim \nu}\Big[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^M(\pi) - \gamma \cdot D^\pi\Big(\widehat{M} \,\|\, M\Big)\Big]. \tag{6.68}$$

and

$$\textsf{o-dec}_\gamma^D(\mathcal{M}) = \sup_{\nu \in \Delta(\mathcal{M})} \textsf{o-dec}_\gamma^D(\mathcal{M}, \nu). \tag{6.69}$$

The Optimistic DEC the same as the generalized DEC in (6.61), except that the optimal value $f^M(\pi_M)$ in (6.61) is replaced by the optimal value $f^{\widehat{M}}(\pi_{\widehat{M}})$ for the (randomized) reference model $\widehat{M} \sim \nu$. This seemingly small change is the main advantage of incorporating optimistic estimation, and makes it possible to bound the Optimistic DEC for certain divergences D for which the value of the generalized DEC in (6.61) would otherwise be unbounded.

Remark 19: When the divergence D admits a sufficient statistic $\boldsymbol{\psi} : \mathcal{M} \to \Psi$, for any distribution $\nu \in \Delta(\mathcal{M})$, if we define $\nu \in \Delta(\Psi)$ via $\nu(\psi) = \nu(\{M \in \mathcal{M} : \boldsymbol{\psi}(M) = \psi\})$, we have

$$\textsf{o-dec}_\gamma^D(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \mathbb{E}_{\psi \sim \nu}\Big[f^\psi(\pi_\psi) - f^M(\pi) - \gamma \cdot D^\pi(\psi \,\|\, M)\Big].$$

In this case, by overloading notation slightly, we may simplify the definition in (6.69) to

$$\textsf{o-dec}_\gamma^D(\mathcal{M}) = \sup_{\nu \in \Delta(\Psi)} \textsf{o-dec}_\gamma^D(\mathcal{M}, \nu).$$

Regret bound for optimistic E2D. The following result shows that the regret of Optimistic Estimation-to-Decisions is controlled by the Optimistic DEC and the optimistic estimation error for the oracle.

Proposition 38: Algorithm 4 ensures that

$$\mathbf{Reg} \leq \textsf{o-dec}_{\gamma}^{D}(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{OptEst}_{\gamma}^{D} \tag{6.70}$$

almost surely.

This regret bound has the same structure as that of Proposition 35, but the DEC and estimation error are replaced by their optimistic counterparts.

When does optimistic estimation help?. When does the regret bound in Proposition 38 improve upon its non-optimistic counterpart in Proposition 35? It turns out that for *asymmetric* divergences such as those found in the context of reinforcement learning, the regret bound in (6.70) can be much smaller than the corresponding bound in (6.63); see Section 7.3.5 for an example. However, for symmetric divergences such as Hellinger distance, we will show now that the result never improves upon Proposition 35.

Given a divergence D, we define the *flipped divergence*, which swaps the first and second arguments, by

$$\check{D}^{\pi}\left(\widehat{M} \parallel M\right) := D^{\pi}\left(M \parallel \widehat{M}\right).$$

Proposition 39 (Equivalence of optimistic DEC and randomized DEC): Assume that For all pairs of models $M, \widehat{M} \in \mathrm{co}(\mathcal{M})$, we have $(f^{\widehat{M}}(\pi) - f^{M}(\pi))^2 \leq L_{\mathrm{lip}}^2 \cdot D^{\pi}\left(\widehat{M} \parallel M\right)$ for a constant $L_{\mathrm{lip}} > 0$. Then for all $\gamma > 0$,

$$\overline{\textsf{dec}}_{3\gamma/2}^{\check{D}}(\mathcal{M}) - \frac{L_{\mathrm{lip}}^2}{2\gamma} \leq \textsf{o-dec}_{\gamma}^{D}(\mathcal{M}) \leq \overline{\textsf{dec}}_{\gamma/2}^{\check{D}}(\mathcal{M}) + \frac{L_{\mathrm{lip}}^2}{2\gamma}. \tag{6.71}$$

This result shows that the optimistic DEC with divergence D is equivalent to the generalized DEC in (6.61), but with the arguments to the divergence flipped. Thus, for symmetric divergences, the quantities are equivalent. In particular, we can combine Proposition 39 with Proposition 36 to derive the following corollary for Hellinger distance.

Proposition 40: Suppose that rewards are bounded in $[0, 1]$. Then for all $\gamma > 0$,

$$\textsf{o-dec}_{2\gamma}^{\mathsf{H}}(\mathcal{M}) - \frac{1}{\gamma} \leq \sup_{\overline{M}} \textsf{dec}_{\gamma}^{\mathsf{H}}(\mathcal{M}, \overline{M}) \leq \textsf{o-dec}_{\gamma/6}^{\mathsf{H}}(\mathcal{M}) + \frac{3}{\gamma}.$$

For asymmetric divergences, in settings where there exists an estimation oracle for which the flipped estimation error

$$\mathbf{Est}^{\check{D}} = \sum_{t=1}^{T} \mathbb{E}_{\pi \sim p^t} \mathbb{E}_{\widehat{M}^t \sim \nu^t}\left[D^{\pi^t}\left(M^{\star} \parallel \widehat{M}^t\right)\right]$$

is controlled, Proposition 39 shows that to match the guarantee in Proposition 38, optimism is not required, and it suffices to run the non-optimistic algorithm in Algorithm 3. However, we show in Section 7.3.5 that for certain divergences found in the context of reinforcement learning, estimation with respect to the flipped divergence is not feasible, yet working with the optimistic DEC E2D.Opt leads to meaningful guarantees.

[Note: This subsection will be expanded in the next version.]

6.8 Decision-Estimation Coefficient: Structural Properties*

In what follows, we state some structural properties of the Decision-Estimation Coefficient, which are useful for calculating the value for specific model classes of interest.

Proposition 41 (Square loss is sufficient for structured bandit problems): Consider any structured bandit problem with decision space Π, function class $\mathcal{F} \subseteq (\Pi \to [0,1])$, and $\mathcal{O} = \{\varnothing\}$. Let $\mathcal{M}_{\mathcal{F}}$ be the class

$$\mathcal{M}_{\mathcal{F}} = \{M \mid f^M \in \mathcal{F}, M(\pi) \text{ is 1-sub-Gaussian } \forall \pi\}.$$

Then, letting

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{F}, \widehat{f}) = \inf_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\Big[f(\pi_f) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\Big],$$

we have

$$\mathsf{dec}^{\mathrm{Sq}}_{c_1\gamma}(\mathcal{F}) \leq \mathsf{dec}_{\gamma}(\mathcal{M}_{\mathcal{F}}) \leq \mathsf{dec}^{\mathrm{Sq}}_{c_2\gamma}(\mathcal{F}),$$

where $c_1, c_2 \geq 0$ are numerical constants.

Proposition 42 (Filtering irrelevant information): Adding observations that are unrelated to the model under consideration never changes the value of the Decision-Estimation Coefficient. In more detail, consider a model class $\mathcal{M}$ with observation space $\mathcal{O}_1$, and consider a class of conditional distributions $\mathcal{D}$ over a secondary observation space $\mathcal{O}_2$, where each $D \in \mathcal{D}$ has the form $D(\pi) \in \Delta(\mathcal{O}_2)$. For $M \in \mathcal{M}$ and $D \in \mathcal{D}$, let $(M \otimes D)(\pi)$ be the model that, given $\pi \in \Pi$, samples $(r, o_1) \sim M(\pi)$ and $o_2 \sim D(\pi)$, then emits $(r, (o_1, o_2))$. Set

$$\mathcal{M} \otimes \mathcal{D} = \{M \otimes D \mid M \in \mathcal{M}, D \in \mathcal{D}\}.$$

Then for all $\widehat{M} \in \mathcal{M}$ and $\widehat{D} \in \mathcal{D}$,

$$\mathsf{dec}_{\gamma}(\mathcal{M} \otimes \mathcal{D}, \widehat{M} \otimes \widehat{D}) = \mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}).$$

This can be seen to hold by restricting the supremum in (6.9) to range over models of the form $M \otimes \widehat{D}$.

Proposition 43 (Data processing): Passing observations through a channel never decreases the Decision-Estimation Coefficient. Consider a class of models $\mathcal{M}$ with observation space $\mathcal{O}$. Let $\rho : \mathcal{O} \to \mathcal{O}'$ be given, and define $\rho \circ M$ to be the model that, given decision π, samples $(r, o) \sim M(\pi)$, then emits $(r, \rho(o))$. Let $\rho \circ \mathcal{M} := \{\rho \circ M \mid M \in \mathcal{M}\}$. Then for all $\widehat{M} \in \mathcal{M}$, we have

$$\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \leq \mathsf{dec}_\gamma(\rho \circ \mathcal{M}, \rho \circ \widehat{M}).$$

This is an immediate consequence of the data processing inequality for Hellinger distance, which implies that $D^2_{\mathsf{H}}\Big(\big(\rho \circ M\big)(\pi), \big(\rho \circ \widehat{M}\big)(\pi)\Big) \leq D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)$.

6.9 Deferred Proofs

Proof of Lemma 22. We first prove the in-expectation bound. By assumption, we have that

$$\sum_{t=1}^{T} \ell^t_{\log}(\widehat{M}^t) - \sum_{t=1}^{T} \ell^t_{\log}(M^\star) \leq \mathbf{Reg}_{\mathsf{KL}}.$$

Taking expectations, Assumption 8 implies that

$$\sum_{t=1}^{T} \mathbb{E}\Big[D_{\mathsf{KL}}\Big(M^\star(\pi^t) \,\|\, \widehat{M}^t(\pi^t)\Big)\Big] \leq \mathbb{E}[\mathbf{Reg}_{\mathsf{KL}}].$$

The bound now follows from Lemma 20.

We now prove the high-probability bound. We will use the following lemma.

Lemma 26: For any sequence of real-valued random variables $(Z_t)_{t \leq T}$, it holds that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} Z_t \leq \sum_{t=1}^{T} \log\big(\mathbb{E}_{t-1}\big[e^{Z_t}\big]\big) + \log(\delta^{-1}). \tag{6.72}$$

Define $Z_t = \frac{1}{2}(\ell^t_{\log}(\widehat{M}^t) - \ell^t_{\log}(M^\star))$. Applying Lemma 26 with the sequence $(-Z_t)_{t \leq T}$, we are guaranteed that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} -\log\big(\mathbb{E}_{t-1}\big[e^{-Z_t}\big]\big) \leq \sum_{t=1}^{T} Z_t + \log(\delta^{-1}) = \frac{1}{2}\sum_{t=1}^{T}\Big(\ell^t_{\log}(\widehat{M}^t) - \ell^t_{\log}(M^\star)\Big) + \log(\delta^{-1}).$$

Let t be fixed, and define abbreviate $z^t = (r^t, o^t)$. Let $\nu(\cdot \mid \pi)$ be any (conditional) dominating measure for $m^{\widehat{M}^t}$ and $m^{M^\star}$, and observe that

$$\begin{aligned}\mathbb{E}_{t-1}\big[e^{-Z_t} \mid \pi^t\big] &= \mathbb{E}_{t-1}\left[\sqrt{\frac{m^{\widehat{M}^t}(z^t \mid \pi^t)}{m^{M^\star}(z^t \mid \pi^t)}} \mid \pi^t\right] \\ &= \int m^{M^\star}(z \mid \pi^t)\sqrt{\frac{m^{\widehat{M}^t}(z \mid \pi^t)}{m^{M^\star}(z \mid \pi^t)}}\nu(dz \mid \pi^t) \\ &= \int \sqrt{m^{M^\star}(z \mid \pi^t)m^{\widehat{M}^t}(z \mid \pi^t)}\nu(dz \mid \pi^t) = 1 - \frac{1}{2}D^2_{\mathsf{H}}\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big).\end{aligned}$$

Hence,

$$\mathbb{E}_{t-1}\big[e^{-Z_t}\big] = 1 - \frac{1}{2}\,\mathbb{E}_{t-1}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big)\Big]$$

and, since $-\log(1-x) \geq x$ for $x \in [0,1]$, we conclude that

$$\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_{t-1}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big)\Big] \leq \frac{1}{2}\sum_{t=1}^{T}\Big(\ell^t_{\log}(\widehat{M}^t) - \ell^t_{\log}(M^\star)\Big) + \log(\delta^{-1}).$$

□

Proof of Lemma 24. We first prove (6.49). Let $X = \sum_{h=1}^{H} r_h$. Since $X \in [0,1]$ almost surely, we have

$$\Big|f^M(\pi) - f^{\widehat{M}}(\pi)\Big| = \Big|\mathbb{E}^{M,\pi}[X] - \mathbb{E}^{\widehat{M},\pi}[X]\Big| \leq D_{\mathsf{TV}}\Big(M(\pi), \widehat{M}(\pi)\Big) \leq D_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big).$$

The final result now follows from the AM-GM inequality.

We now prove (6.51). From Lemma 15, we have

$$\begin{aligned}f^M(\pi) - f^{\widehat{M}}(\pi) &= \sum_{h=1}^{H}\mathbb{E}^{\widehat{M},\pi}\big[Q^{M,\pi}_h(s_h,a_h) - r_h - V^{M,\pi}_{h+1}(s_{h+1})\big] \\ &= \sum_{h=1}^{H}\mathbb{E}^{\widehat{M},\pi}\Big[\big[P^M_h V^{M,\pi}_{h+1}\big](s_h,a_h) - V^{M,\pi}_{h+1}(s_{h+1}) + \mathbb{E}_{r_h\sim R^M_h(s_h,a_h)}[r_h] - \mathbb{E}_{r_h\sim R^{\widehat{M}}_h(s_h,a_h)}[r_h]\Big] \\ &= \sum_{h=1}^{H}\mathbb{E}^{\widehat{M},\pi}\Big[\Big[(P^M_h - P^{\widehat{M}}_h)V^{M,\pi}_{h+1}\Big](s_h,a_h)\Big] + \sum_{h=1}^{H}\mathbb{E}^{\widehat{M},\pi}\Big[\mathbb{E}_{r_h\sim R^M_h(s_h,a_h)}[r_h] - \mathbb{E}_{r_h\sim R^{\widehat{M}}_h(s_h,a_h)}[r_h]\Big] \\ &\leq \sum_{h=1}^{H}\mathbb{E}^{\widehat{M},\pi}\Big[D_{\mathsf{TV}}\Big(P^M_h(s_h,a_h), P^{\widehat{M}}_h(s_h,a_h)\Big) + D_{\mathsf{TV}}\Big(R^M_h(s_h,a_h), R^{\widehat{M}}_h(s_h,a_h)\Big)\Big],\end{aligned}$$

where we have used that $V^{M,\pi}_{h+1}(s) \in [0,1]$.

□

Proof of Lemma 25. We first prove (6.53). For all $\eta > 0$, we have

$$\begin{aligned}&\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\ &\leq \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^{\widehat{M}}(\pi)\Big] + \eta\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}(\pi)\Big)\Big] + \frac{1}{4\eta}.\end{aligned}$$

We now prove (6.54). Using Lemma 38, we have that for all h,

$$\mathbb{E}^{\widehat{M},\pi}\Big[D^2_{\mathsf{H}}\Big(P^M(s_h,a_h),P^{\widehat{M}}(s_h,a_h)\Big)\Big]+\mathbb{E}^{\widehat{M},\pi}\Big[D^2_{\mathsf{H}}\Big(R^M(s_h,a_h),R^{\widehat{M}}(s_h,a_h)\Big)\Big]\leq 8D^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big).$$

As a result,

$$\mathbb{E}^{\widehat{M},\pi}\left[\sum_{h=1}^{H}D^2_{\mathsf{H}}\Big(P^M(s_h,a_h),P^{\widehat{M}}(s_h,a_h)\Big)+D^2_{\mathsf{H}}\Big(R^M(s_h,a_h),R^{\widehat{M}}(s_h,a_h)\Big)\right]\leq 8HD^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big).$$

Since this holds uniformly for all π, we conclude that

$$\begin{aligned}&\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\widehat{M},\pi}\left[\sum_{h=1}^{H}D^2_{\mathsf{TV}}\Big(P^M(s_h,a_h),P^{\widehat{M}}(s_h,a_h)\Big)+D^2_{\mathsf{TV}}\Big(R^M(s_h,a_h),R^{\widehat{M}}(s_h,a_h)\Big)\right]\\&\leq 8H\,\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big)\Big].\end{aligned}$$

□

6.10 Exercises

Exercise 11: Prove Lemma 20.

Exercise 12: In this exercise, we will prove Proposition 37 as follows:

1. Prove the first two inequalities.
2. Use properties of the Hellinger distance to show that for any $\pi\in\Pi$, $\mu\in\Delta(\mathcal{M})$, and $\widehat{M}$,

$$\mathbb{E}_{M\sim\mu}\,D^2_{\mathsf{H}}\Big(M(\pi),\widehat{M}(\pi)\Big)\geq\frac{1}{4}\,\mathbb{E}_{M,M'\sim\mu}\,D^2_{\mathsf{H}}(M(\pi),M'(\pi)).$$

Hint: start with the right-hand side and use symmetry and triangle inequality for Hellinger distance

3. With the help of Part 2, show that for any $\widehat{M}$,

$$\mathsf{dec}_\gamma(\mathcal{M},\widehat{M})\leq\sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\Big[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{4}\mathbb{E}_{M'\sim\mu}\,D^2_{\mathsf{H}}(M(\pi),M'(\pi))\Big].$$

4. Argue that

$$\mathsf{dec}_\gamma(\mathcal{M},\widehat{M})\leq\sup_{\nu\in\Delta(\mathcal{M})}\sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\Big[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{4}\mathbb{E}_{M'\sim\nu}\,D^2_{\mathsf{H}}(M(\pi),M'(\pi))\Big].$$

and conclude the third inequality in Proposition 37.

5. Show that

$$\sup_{\widehat{M}}\mathsf{dec}_\gamma(\mathcal{M},\widehat{M})\leq\sup_{\widehat{M}\in\mathrm{co}(\mathcal{M})}\mathsf{dec}_{\gamma/4}(\mathcal{M},\widehat{M}).\tag{6.73}$$

In other words, the estimation oracle cannot significantly increase the value of the DEC by selecting models outside $\mathrm{co}(\mathcal{M})$.

Exercise 13 (Lower Bound on DEC for Tabular RL): We showed that for Gaussian bandits,

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) \geq \varepsilon\sqrt{A/2},$$

for all $\varepsilon \lesssim 1/\sqrt{A}$ by considering a small sub-family models and explicitly computing the DEC for this sub-family. Show that if $\mathcal{M}$ is the set of all tabular MDPs with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, and $\sum_{h=1}^{H} r_h \in [0, 1]$,

$$\mathsf{dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \widehat{M}) \gtrsim \varepsilon\sqrt{SA}$$

for all $\varepsilon \lesssim 1/\sqrt{SA}$, as long as $H \gtrsim \log_A(S)$.

Exercise 14 (Structured Bandits with ReLU Rewards): We will show that structured bandits with ReLU rewards suffer from the curse of dimensionality. Let $\mathsf{relu}(x) = \max\{x, 0\}$ and take $\Pi = \mathsf{B}^d_2(1) = \{\pi \in \mathbb{R}^d \mid \|\pi\|_2 \leq 1\}$. Consider the class of value functions of the form

$$f_\theta(\pi) = \mathsf{relu}(\langle\theta, \pi\rangle - b), \tag{6.74}$$

where $\theta \in \Theta = \mathbb{S}^{d-1}$, is an unknown parameter vector and $b \in [0, 1]$ is a known bias parameter. Here $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d \mid \|v\| = 1\}$ denotes the unit sphere. Let $\mathcal{M} = \{M_\theta\}_{\theta\in\Theta}$, where for all π, $M_\theta(\pi) := \mathcal{N}(f_\theta(\pi), 1)$.

We will prove that for all $d \geq 16$, there exists $\overline{M} \in \mathcal{M}$ such that for all $\gamma > 0$,

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \gtrsim \frac{e^{d/8}}{\gamma} \wedge 1, \tag{6.75}$$

for an appropriate choice of bias b. By slightly strengthening this result and appealing to (6.32), it is possible to show that any algorithm must have $\mathbb{E}[\mathbf{Reg}] \gtrsim e^{d/8}$.

To prove (6.75), we will use the fact that for large d, a random vector v chosen uniformly from the unit sphere is nearly orthogonal to any direction π. This fact is quantified as follows (see Ball '97):

$$\mathbb{P}_{v\sim\mathrm{unif}(\mathbb{S}^{d-1})}(\langle\pi, v\rangle > \alpha) \leq \exp\left(-\frac{\alpha^2}{2}d\right). \tag{6.76}$$

for any π with $\|\pi\| = 1$.

1. Prove that for all $\pi \in \Pi$, $v \in \Theta$, and any choice of b,

$$\max_{\pi'\in\Pi} f_v(\pi') - f_v(\pi) \geq (1 - b)\mathbb{I}\{\langle v, \pi\rangle \leq b\}$$

In other words, instantaneous regret is at least $(1 - b)$ whenever the decision π does not align well with v.

2. Let $\overline{M}(\pi) = \mathcal{N}(0, 1)$. Show that for all $\pi \in \Pi$, $v \in \Theta$, and for any choice of b,

$$D^2_{\mathsf{H}}\big(M_v(\pi), \overline{M}(\pi)\big) \leq \frac{1}{2}f^2_v(\pi) \leq \frac{(1 - b)^2}{2}\mathbb{I}\{\langle v, \pi\rangle > b\},$$

i.e. information is obtained by the decision-maker only if the decision π aligns well with v in the model M_v.

3. Show that

$$\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) \geq \inf_{p\in\Delta(\Pi)} \mathbb{E}_{v\sim\mathrm{unif}(\mathbb{S}^{d-1})}\mathbb{E}_{\pi\sim p}\left[(1-b)-(1-b)\mathbb{I}\left\{\langle v,\pi\rangle > b\right\}-\gamma\frac{(1-b)^2}{2}\mathbb{I}\left\{\langle v,\pi\rangle > b\right\}\right].$$

4. Set $\varepsilon := 1-b$. Use (6.76) and Part 3 above to argue that

$$\mathsf{dec}_\gamma(\mathcal{M}',\overline{M}) \geq \varepsilon - \varepsilon\exp(-d/8) - \gamma\frac{\varepsilon^2}{2}\exp(-d/8).$$

Conclude that for $d\geq 8$,

$$\mathsf{dec}_\gamma(\mathcal{M}',\overline{M}) \geq \frac{\varepsilon}{2} - \gamma\frac{\varepsilon^2}{2}\exp(-d/8)$$

5. Show that by choosing $\varepsilon = \frac{e^{d/8}}{6\gamma}\wedge\frac{1}{2}$ and recalling that $b=1-\varepsilon$, we get (6.75).

7. REINFORCEMENT LEARNING: FUNCTION APPROXIMATION AND LARGE STATE SPACES

In this section, we consider the problem of online reinforcement learning with function approximation. The framework is the same as that of Section 5 but, in developing algorithms, we no longer assume that the state and action spaces are finite/tabular, and in particular we will aim for regret bounds that are independent of the number of states. To do this, we will make use of function approximation—either directly modeling the transition probabilities for the underlying MDP, or modeling quantities such as value functions—and our goal will be to design algorithms that are capable of generalizing across the state space as they explore. This will pose challenges similar to that of the structured and contextual bandit settings, but we now face the additional challenge of credit assignment. Note that the online reinforcement learning framework is a special case of the general decision making setting in Section 6, but the algorithms we develop in this section will be tailored to the MDP structure.

Recall (Section 5) that for reinforcement learning, each MDP M takes the form

$$M = \left\{\mathcal{S},\mathcal{A},\{P_h^M\}_{h=1}^H,\{R_h^M\}_{h=1}^H,d_1\right\},$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P_h^M:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ is the probability transition kernel at step h, $R_h^M:\mathcal{S}\times\mathcal{A}\to\Delta(\mathbb{R})$ is the reward distribution, and $d_1\in\Delta(\mathcal{S}_1)$ is the initial state distribution. All of the results in this section will take $\Pi=\Pi_{\mathrm{RNS}}$, and we will assume that $\sum_{h=1}^H r_h\in[0,1]$ unless otherwise specified.

7.1 Is Realizability Sufficient?

For the frameworks we have considered so far (contextual and structured bandits, general decision making), all of the algorithms we analyzed leveraged the assumption of *realizability*, which asserts that we have a function class that is capable of modeling the underlying environment well. For reinforcement learning, there are various realizability assumptions one can consider:

- *Model realizability*: We have a model class $\mathcal{M}$ of MDPs that contains the true MDP $M^\star$.

- *Value function realizability*: We have a class $\mathcal{Q}$ of state-action value functions (Q-functions) that contains the optimal function $Q^{M^\star,\star}$ for the underlying MDP.
- *Policy realizability*: We have a class Π of policies that contains the optimal policy $\pi_{M^\star}$.

Note that model realizability implies value function realizability, which in turn implies policy realizability. Ideally, we would like to be able to say that whenever one of these assumptions holds, we can obtain regret bounds that scale with the complexity of the function class (e.g., $\log|\mathcal{M}|$ for model realizability, or $\log|\mathcal{Q}|$ for value function realizability), but do not depend on the number of states $|\mathcal{S}|$ or other properties of the underlying MDP, analogous to the situation for statistical learning. Unfortunately, the following result shows that this is too much to ask for.

Proposition 44: For any $S \in \mathbb{N}$ and $H \in \mathbb{N}$, there exists a class of horizon-H MDPs $\mathcal{M}$ with $|\mathcal{S}| = S$, $|\mathcal{A}| = 2$, and $\log|\mathcal{M}| = \log(S)$, yet any algorithm must have

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{\min\{S, 2^H\} \cdot T}.$$

The interpretation of this result is that even if model realizability holds, any algorithm needs regret that scales with $\min\{|\mathcal{S}|, |\mathcal{M}|, 2^H\}$. This means additional structural assumptions on the underlying MDP $M^\star$—beyond realizability—are required if we want to obtain sample-efficient learning guarantees. Note that since this construction satisfies model realizability, the strongest form of realizability, it also rules out sample-efficient results for value function and policy realizability.

In what follows, we will explore different structural assumptions that facilitate low regret for reinforcement learning with function approximation. Briefly, the idea will be to make assumptions that either i) allow for extrapolation across the state space, or ii) control the number of "effective" state distributions the algorithm can encounter. We will begin by investigating reinforcement learning with linear models, then explore a general structural property known as Bellman rank.

Remark 20 (Comparison to structured bandits): Proposition 44 is is analogous to the impossibility result we proved for structured bandits (Example 4.1), which is subsumed by the RL framework. That result required a large number of actions, while Proposition 44 holds even when $|\mathcal{A}| = 2$.

Remark 21 (Further notions of realizability): There are many notions of realizability beyond those we consider above. For example, for value function approximation, one can assume that $Q^{M^\star,\pi} \in \mathcal{Q}$ *for all* π, or assume that the class $\mathcal{Q}$ obeys certain notions of consistency with respect to the Bellman operator for $M^\star$.

7.2 Linear Function Approximation

Toward understanding the complexity of RL with function approximation, let us consider perhaps the simplest possible modeling approach: Linear function approximation. A natural idea here is to assume linearity of the underlying Q-function, generalizing the linear bandit setting in Section 4.

$$Q_h^{M,\star}(s,a) = \langle \phi(s,a), \theta_h^M \rangle, \quad \forall h \in [H] \tag{7.1}$$

where $\phi(s,a) \in \mathsf{B}_2^d(1)$ is a feature map that is known to the learner and $\theta_h^M \in \mathsf{B}_2^d(1)$ is an unknown parameter vector. Equivalently, we can define

$$\mathcal{Q} = \Big\{ Q_h(s,a) = \langle \phi(s,a), \theta_h \rangle \mid \theta_h \in \mathsf{B}_2^d(1) \ \forall h \Big\}, \tag{7.2}$$

and assume that $Q^{M,\star} \in \mathcal{Q}$. This is called the *Linear-$Q^\star$* model.

Linearity is a strong assumption, and it is reasonable to imagine that this would be sufficient for low regret. Indeed, one might hope that using linearity, we can extrapolate the value of $Q^{M,\star}$ once we estimate it for a small number of states. Unfortunately, even for this very simple class of functions, it turns out that realizability is still insufficient.

Proposition 45 (Weisz et al. [74], Wang et al. [73]): For any $d \in \mathbb{N}$ and $H \in \mathbb{N}$ sufficiently large, any algorithm for the Linear-$Q^\star$ model must have

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \min\Big\{ 2^{\Omega(d)}, 2^{\Omega(H)} \Big\}.$$

This contrasts the situation for contextual bandits and linear bandits, where linear rewards were sufficient for low regret. The intuition is that, even though $Q^{M,\star}$ is linear, it might take a very long time to estimate the value for even a small number of states. That is, linearity of the optimal value function is not a useful assumption unless there is some kind of additional structure that can guide us toward the optimal value function to being with.

We mention in passing that Proposition 45 can be proven by lower bounding the Decision-Estimation Coefficient [35].

The Low-Rank MDP model. Proposition 45 implies that linearity of the optimal Q-function alone does not sufficient for sample-efficient RL. To proceed, we will make a stronger assumption, which asserts that the transition probabilities themselves have linear structure: For all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in [H]$, we have

$$P_h^M(s' \mid s,a) = \big\langle \phi(s,a), \mu_h^M(s') \big\rangle, \quad \text{and} \quad \mathbb{E}[r_h|s,a] = \langle \phi(s,a), w_h^M \rangle. \tag{7.3}$$

Here, $\phi(s,a) \in \mathsf{B}_2^d(1)$ is a feature map that is known to the learner, $\mu_h^M(s') \in \mathbb{R}^d$ is another feature map which is *unknown* to the learner, and $w_h^M \in \mathsf{B}_2^d(\sqrt{d})$ is an unknown parameter vector. Additionally, for simplicity, we assume that $\big\| \sum_{s' \in \mathcal{S}} |\mu_h^M(s')| \big\| \leq \sqrt{d}$, which in particular holds if $[\mu_h^M]_i \in \Delta(\mathcal{S})$. As before, assume that both cumulative and individual-step rewards are in $[0,1]$. For the remainder of the subsection, we let $\mathcal{M}$ denote the set of MDPs with these properties.

The linear structure in (7.3) implies that the transition matrix has rank at most d, thus facilitating (as we shall see shortly) information sharing and generalization across states, even when the cardinality of $\mathcal{S}$ and $\mathcal{A}$ is large or infinite. For this reason, the setting considered here is called *low rank MDPs.*

Just as linear bandits generalize unstructured multi-armed bandits, the low rank MDP model (7.3) generalizes tabular RL, which corresponds to the special case in which $d = |\mathcal{S}| \cdot |\mathcal{A}|$, $\phi(s, a) = \boldsymbol{e}_{s,a}$, and $(\mu_h(s'))_{s,a} = P_h^M(s' \mid s, a)$.

Properties of low-rank MDPs. The linear structure of the transition probabilities and mean rewards is a significantly more stringent assumption than linearity of $Q_h^{M,\star}(s, a)$ in (7.1). Notably, it implies that Bellman backups of *arbitrary functions* are linear.

Lemma 27: For any linear MDP $M \in \mathcal{M}$ and any $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and any $h \in [H]$, the Bellman operator is linear in ϕ:

$$[\mathcal{T}_h^M Q](s, a) = \left\langle \phi(s, a), \theta_Q^M \right\rangle$$

for some $\theta_Q^M \in \mathbb{R}^d$. In particular, this implies that for any policy $\pi = (\pi^1, \ldots, \pi^H)$, functions $Q_h^{M,\pi}$ are linear in ϕ for every h. Finally, for $Q : \mathcal{S} \times \mathcal{A} \to [0, 1]$, it holds that $\left\|\theta_Q^M\right\| \leq 2\sqrt{d}$.

As a special case, this lemma implies that for low rank MDPs, $Q_h^{M,\pi}$ is linear *for all* π.

Proof of Lemma 27. We have

$$[\mathcal{T}_h^M Q](s, a) = \langle \phi(s, a), w_h^M \rangle + \sum_{s'} P_h^M(s' \mid s, a) \max_a Q(s', a) \tag{7.4}$$

$$= \langle \phi(s, a), w_h^M \rangle + \sum_{s'} \left\langle \phi(s, a), \mu_h^M(s') \right\rangle \max_{a'} Q(s', a') \tag{7.5}$$

$$= \left\langle \phi(s, a), w_h^M + \sum_{s'} \mu_h^M(s') \max_{a'} Q(s', a') \right\rangle. \tag{7.6}$$

The second statement follows since $Q_h^{M,\pi} = \left[\mathcal{T}_h^M Q_{h+1}^{M,\pi}\right]$. For the last statement,

$$\left\|\theta_Q^M\right\| \leq \|w_h^M\| + \left\| \sum_{s'} \mu_h^M(s') Q(s') \right\| \leq 2\sqrt{d}, \tag{7.7}$$

since μ_h^M is a vector of distributions on $\mathcal{S}$. □

7.2.1 The LSVI-UCB Algorithm

To provide regret bounds for the low rank MDP model, we analyze an algorithm called LSVI-UCB ("Least Squares Value Iteration UCB"), which was introduced and analyzed in the influential paper of Jin et al. [42]. Similar to the UCB-VI algorithm we analyzed for tabular RL, the main idea behind the algorithm is to compute a state-action value $\bar{Q}^t$ with the optimistic property that

$$\bar{Q}_h^t(s, a) \geq Q_h^{M,\star}(s, a)$$

for all s, a, h. This is achieved by combining the principle of dynamic programming with an appropriate choice of bonus to ensure optimism. However, unlike UCB-VI, the algorithm does not directly estimate transition probabilities (which is not feasible when μ^M is unknown), and instead implements approximate value iteration by solving a certain least squares objective.

```
LSVI-UCB
Input: $R, \rho > 0$
for $t = 1, \ldots, T$ do
    Let $\bar{Q}^t_{H+1} \equiv 0$.
    for $h = H, \ldots, 1$ do
        Compute least-squares estimator
            $\widehat{\theta}^t_h = \arg\min_{\theta \in \mathsf{B}^d_2(\rho)} \sum_{i<t} \Big( \langle \phi(s^i_h, a^i_h), \theta \rangle - r^i_h - \max_a \bar{Q}^t_{h+1}(s^i_{h+1}, a) \Big)^2,$
        and let $\widehat{Q}^t_h(s,a) := \big\langle \phi(s,a), \widehat{\theta}^t_h \big\rangle$.
        Define
            $\Sigma^t_h = \sum_{i<t} \phi(s^i_h, a^i_h)\phi(s^i_h, a^i_h)^\top + I.$
        Compute bonus:
            $b^t_{h,\delta}(s,a) = \sqrt{R}\|\phi(s,a)\|_{(\Sigma^t_h)^{-1}}.$
        Compute optimistic value function:
            $\bar{Q}^t_h(s,a) = \Big\{ \widehat{Q}^t_h(s,a) + b^t_{h,\delta}(s,a) \Big\} \wedge 1.$
        Set $\bar{V}^t_h(s) = \max_{a \in \mathcal{A}} \bar{Q}^t_h(s,a)$ and $\widehat{\pi}^t_h(s) = \arg\max_{a \in \mathcal{A}} \bar{Q}^t_h(s,a)$.
    Collect trajectory $(s^t_1, a^t_1, r^t_1), \ldots, (s^t_H, a^t_H, r^t_H)$ according to $\widehat{\pi}^t$.
```

In more detail, for each episode t, the algorithm computes $\bar{Q}^t_1, \ldots, \bar{Q}^t_H$ through approximate dynamic programming. At layer h, given $\bar{Q}^t_{h+1}$, the algorithm computes a linear Q-function $\widehat{Q}^t_h(s,a) := \big\langle \phi(s,a), \widehat{\theta}^t_h \big\rangle$, by solving a least squares problem in which $X = \phi(s_h, a_h)$ is the feature vector and $Y = r_h + \max_a \bar{Q}^t_{h+1}(s_{h+1}, a)$ is the target/outcome. This is motivated by Lemma 27, which asserts that the Bellman backup $\big[\mathcal{T}^M_h \bar{Q}^t_{h+1}\big](s,a)$ is linear. Given $\widehat{Q}^t_h$, the algorithm forms the optimistic estimate $\bar{Q}^t_h$ via

$$\bar{Q}^t_h(s,a) = \Big\{ \widehat{Q}^t_h(s,a) + b^t_{h,\delta}(s,a) \Big\} \wedge 1,$$

where

$$b^t_{h,\delta}(s,a) = \sqrt{R}\|\phi(s,a)\|_{(\Sigma^t_h)^{-1}}, \quad \text{with} \quad \Sigma^t_h = \sum_{i<t} \phi(s^i_h, a^i_h)\phi(s^i_h, a^i_h)^\top + I,$$

is an elliptic bonus analogous to the bonus used within LinUCB. With this, the algorithm proceeds to the next layer $h-1$. Once $\bar{Q}^t$ is computed for every layer, the algorithm executes the optimistic policy $\widehat{\pi}^t$ given by $\widehat{\pi}^t_h(s) = \arg\max_{a \in \mathcal{A}} \bar{Q}^t_h(s,a)$.

The LSVI-UCB algorithm enjoys the following regret bound.

Proposition 46: If any $\delta > 0$, if we set $R = c \cdot d^2 \log(HT/\delta)$ for a sufficiently large numerical constant c and $\rho = 2\sqrt{d}$, LSVI-UCB has that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim H\sqrt{d^3 \cdot T \log(HT/\delta)}. \tag{7.8}$$

7.2.2 Proof of Proposition 46

The starting point of our analysis for UCB-VI was Lemma 16, which states that it is sufficient to construct optimistic estimates $\{\overline{Q}_1, \ldots, \overline{Q}_H\}$ (i.e. $Q_h^{M,\star} \le \overline{Q}_h$) such that the Bellman residuals $\mathbb{E}^{M,\widehat{\pi}}\big[(\overline{Q}_h - \mathcal{T}_h^M \overline{Q}_{h+1})(s_h, a_h)\big]$ are small under the greedy (with respect to $\overline{Q}$'s) policy $\widehat{\pi}$. In order to control these residuals, we constructed an estimated model $\widehat{M}$ and defined empirical Bellman operators $\mathcal{T}_h^{\widehat{M}}$ in terms of estimated transition kernels. We then set $\overline{Q}_h$ to be the empirical Bellman backup $\mathcal{T}_h^{\widehat{M}} \overline{Q}_{h+1}$, plus an optimistic bonus term. In contrast, LSVI-UCB does not directly estimate the model. Instead, it performs regression with a target that is an empirical Bellman backup. As we shall see shortly, subtleties arise in the analysis of this regression step due to lack of independence.

Technical lemmas for regression. Recall from Lemma 27 that for any fixed $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\mathbb{E}^M\Big[r_h^i + \max_a Q(s_{h+1}^i, a) \mid s_h^i, a_h^i\Big] = [\mathcal{T}_h^M Q](s_h^i, a_h^i). \tag{7.9}$$

However, for layer h, the regression problem within LSVI-UCB concerns a *data-dependent* function $Q = \overline{Q}_{h+1}^t$ (with $i < t$), which is chosen as a function of all the trajectories $\tau^1, \ldots, \tau^{t-1}$. This dependence implies that the regression problem solved by LSVI-UCB is not of the type studied, say, in Proposition 1. Instead, in the language of Section 1.4, the mean of the outcome variable is itself a function that depends on all the data. The saving grace here is that this dependence does not result in arbitrarily complex functions, which will allow us to appeal to uniform convergence arguments. In particular, for every h and t, $\overline{Q}_h^t$ belongs to the class

$$\mathcal{Q} := \Big\{(s,a) \mapsto \Big\{\langle \theta, \phi(s,a)\rangle + \sqrt{R}\|\phi(s,a)\|_{(\Sigma)^{-1}}\Big\} \wedge 1 : \|\theta\| \le 2\sqrt{d}, \sigma_{\mathsf{min}}(\Sigma) \ge 1\Big\}. \tag{7.10}$$

To make use of this fact, we first state an abstract result concerning regression with dependent outcomes.

Lemma 28: Let $\mathcal{G}$ be an abstract set with $|\mathcal{G}| < \infty$. Let $x_1, \ldots, x_T \in \mathcal{X}$ be fixed, and for each $g \in \mathcal{G}$, let $y_1(g), \ldots, y_T(g) \in \mathbb{R}$ be 1-subGaussian outcomes satisfying

$$\mathbb{E}[y_i(g) \mid x_i] = f_g(x_i)$$

for $f_g \in \mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$.[a] In addition, assume that $y_1(g), \ldots, y_T(g)$ are conditionally independent given $x_1, \ldots, x_T$. For any latent $g \in \mathcal{G}$, define the least-squares solution

$$\widehat{f}_g \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^T (y_i(g) - f(x_i))^2.$$

With probability at least $1-\delta$, simultaneously for all $g \in \mathcal{G}$,

$$\sum_{i=1}^{T}(\widehat{f}_g(x_i) - f_g(x_i))^2 \lesssim \log(|\mathcal{F}||\mathcal{G}|/\delta).$$

Proof of Lemma 28. Fix $g \in \mathcal{G}$. To shorten the notation, it is useful to introduce empirical norms $\|f\|_T^2 = \frac{1}{T}\sum_{i=1}^T f(x_i)^2$ and empirical inner product $\langle f, f' \rangle_T = \sum_{i=1}^T f(x_i)f'(x_i)$ for $f, f' \in \mathcal{F}$. Optimality of $\widehat{f}_g$ implies that

$$\sum_{i=1}^{T}(y_i(g) - \widehat{f}_g(x_i))^2 \le \sum_{i=1}^{T}(y_i(g) - f_g(x_i))^2$$

which can be written succinctly (with a slight abuse of notation) as $\big\|Y_g - \widehat{f}_g\big\|_T^2 \le \|Y_g - f_g\|_T^2$ for $Y_g = (y_1(g), \dots, y_T(g))$. This implies

$$\big\|\widehat{f}_g - f_g\big\|_T^2 \le 2\big\langle Y_g - f_g, \widehat{f}_g - f_g\big\rangle_T.$$

Dividing both sides by $\big\|\widehat{f}_g - f_g\big\|_T$ and taking supremum over $\widehat{f}_g \in \mathcal{F}$ leads to

$$\big\|\widehat{f}_g - f_g\big\|_T \le 2\max_{f\in\mathcal{F}}\Big\langle Y_g - f_g, \frac{f - f_g}{\big\|f - f_g\big\|_T}\Big\rangle_T. \tag{7.11}$$

The random vector $Y_g - f_g$ has independent zero-mean 1-subGaussian entries by assumption, while the multiplier $\frac{f-f_g}{\|f-f_g\|_T}$ is simply a T-dimensional vector of Euclidean length $\sqrt{T}$, for each $f \in \mathcal{F}$. Hence, each inner product in (7.11) is a sub-Gaussian vector with variance proxy $\frac{1}{T}$ (see Definition 2). Thus, with probability at least $1-\delta$, the maximum on the right-hand side does not exceed $C\sqrt{\log(|\mathcal{F}|/\delta)/T}$ for an appropriate constant C. Taking the union bound over g and squaring both sides of (7.11) yields the desired bound. □

We may now apply Lemma 28 to analyze the regression step of LSVI-UCB.

Lemma 29: With probability at least $1-\delta$, we have that for all t and h,

$$\sum_{i<t}\Big(\widehat{Q}_h^t(s_h^i, a_h^i) - \big[\mathcal{T}_h^M \overline{Q}_{h+1}^t\big](s_h^i, a_h^i)\Big)^2 \lesssim d^2\log(HT/\delta). \tag{7.12}$$

Proof sketch for Lemma 29. Let $t \in [T]$ and $h \in [H]$ be fixed. To make the correspondence with Lemma 28 explicit, for the data $(s_h^i, a_h^i, s_{h+1}^i, r_h^i)$, we define $x_i = \phi(s_h^i, a_h^i)$ and $y_i(Q) = r_h^i + \max_a Q(s_{h+1}^i, a)$, with $Q \in \mathcal{Q}$ playing the role of the index $g \in \mathcal{G}$. With this, we have

$$\mathbb{E}[y_i(Q) \mid x_i] = \mathbb{E}^M\Big[r_h^i + \max_a Q(s_{h+1}^i, a) \mid s_h^i, a_h^i\Big] = [\mathcal{T}_h^M Q](s_h^i, a_h^i) = \big\langle \phi(s_h^i, a_h^i), \theta_Q^M \big\rangle$$

which is linear in $x^i = \phi(s_h^i, a_h^i)$, with the vector of coefficients θ_Q^M depending on Q. The regression problem is well-specified as long as we choose

$$\mathcal{F} = \Big\{\phi(s,a) \mapsto \langle \phi(s,a), \theta \rangle : \|\theta\| \le 2\sqrt{d}\Big\}$$

[a] The random variables $\{y_i(g)\}_{g\in\mathcal{G}}$ may be correlated.

and $\mathcal{Q}$ as in (7.10). While both of these sets are infinite, we can to a standard covering number argument for an appropriate scale ε. The cardinalities of ε-discretized classes can be shown to be of size $\widetilde{O}(d)$ and $\widetilde{O}(d^2)$, respectively, up to factors logarithmic in $1/\varepsilon$ and d. The statement follows after checking that discretization incurs a small price due to Lipschitzness with respect to parameters. Finally, we union bound over t and h. □

Establishing optimism. The next lemma shows that closeness of the regression estimate to the Bellman backup on the data $\{(s^i_h, a^i_h)\}_{i<t}$ translates into closeness at an arbitrary (s,a) pair as long as $\phi(s,a)$ is sufficiently covered by the data collected so far. This, in turn, implies that $\overline{Q}^t_1, \ldots, \overline{Q}^t_H$ are optimistic.

Lemma 30: Whenever the event in Lemma 29 occurs, we have that for all (s,a,h) and $t \in [T]$,

$$\left|\widehat{Q}^t_h(s,a) - \left[\mathcal{T}^M_h \overline{Q}^t_{h+1}\right](s,a)\right| \lesssim \sqrt{d^2 \log(HT/\delta)} \cdot \|\phi(s,a)\|_{(\Sigma^t_h)^{-1}} =: b^t_{h,\delta}(s,a). \tag{7.13}$$

and

$$\overline{Q}^t_h(s,a) \geq Q^{M,\star}_h(s,a). \tag{7.14}$$

Proof of Lemma 30. Writing the Bellman backup, via Lemma 27, as

$$\left[\mathcal{T}^M_h \overline{Q}^t_{h+1}\right](s,a) = \langle \phi(s,a), \theta^t_h \rangle$$

for some $\theta^t_h \in \mathbb{R}^d$ with $\|\theta^t_h\|_2 \leq 2\sqrt{d}$, we have that

$$\begin{aligned}\left|\widehat{Q}^t_h(s,a) - \left[\mathcal{T}^M_h \overline{Q}^t_h\right](s,a)\right| &= \left|\left\langle \phi(s,a), \widehat{\theta}^t_h - \theta^t_h \right\rangle\right| \\ &= \left|\left\langle (\Sigma^t_h)^{-1/2}\phi(s,a), (\Sigma^t_h)^{1/2}(\widehat{\theta}^t_h - \theta^t_h) \right\rangle\right| \\ &\leq \left\|\phi(s,a)\right\|_{(\Sigma^t_h)^{-1}} \cdot \left\|\widehat{\theta}^t_h - \theta^t_h\right\|_{\Sigma^t_h}.\end{aligned}$$

Lemma 29 then implies (7.13), since

$$\left\|\widehat{\theta}^t_h - \theta^t_h\right\|^2_{\Sigma^t_h} = (\widehat{\theta}^t_h - \theta^t_h)^\top \left(\sum_{i<t} \phi(s^i_h, a^i_h)\phi(s^i_h, a^i_h)^\top + I\right)(\widehat{\theta}^t_h - \theta^t_h) \tag{7.15}$$

$$= \sum_{i<t}\left(\widehat{Q}^t_h(s^i_h, a^i_h) - \left[\mathcal{T}^M_h \overline{Q}^t_{h+1}\right](s^i_h, a^i_h)\right)^2 + \left\|\widehat{\theta}^t_h - \theta^t_h\right\|^2 \tag{7.16}$$

and $\left\|\widehat{\theta}^t_h - \theta^t_h\right\|^2 \lesssim d$ by (7.7).

To show (7.14), we proceed by induction on $\overline{V}^t_h \geq V^{M,\star}_h$, as in the proof of Lemma 17. We start with the base case $h = H+1$, which has $\overline{V}^t_{H+1} = V^{M,\star}_{H+1} \equiv 0$. Assuming $\overline{V}^t_{h+1} \geq V^{M,\star}_{h+1}$, we first observe that $\mathcal{T}^M_h$ is monotone and $\mathcal{T}^M_h \overline{V}^t_{h+1} \geq \mathcal{T}^M_h V^{M,\star}_{h+1} = Q^{M,\star}_h$. Hence,

$$\widehat{Q}^t_h = \widehat{Q}^t_h - \mathcal{T}^M_h \overline{V}_{h+1} + \mathcal{T}^M_h \overline{V}_{h+1} \tag{7.17}$$

$$\geq \widehat{Q}^t_h - \mathcal{T}^M_h \overline{V}_{h+1} + Q^{M,\star}_h \tag{7.18}$$

$$\geq -b^t_{h,\delta} + Q^{M,\star}_h \tag{7.19}$$

and thus $\widehat{Q}_h^t + b_{h,\delta}^t \geq Q_h^{M,\star}$. Since $Q_h^{M,\star} \leq 1$, the clipped version $\overline{Q}_h^t$ also satisfies $\overline{Q}_h^t \geq Q_h^{M,\star}$. This, in turn, implies $\overline{V}_h^t \geq V_h^{M,\star}$. □

Finishing the proof. With the technical results above established, the proof of Proposition 46 follows fairly quickly.

Proof of Proposition 46. Let M be the true model. . Condition on the event in Lemma 29. Then, since $\overline{Q}$ is optimistic by Lemma 30, we have that for each timestep t,

$$\begin{aligned} f^M(\pi_M) - f^M(\widehat{\pi}^t) &\leq \mathbb{E}_{s_1 \sim d_1}\left[\overline{V}_1^t(s_1)\right] - f^M(\widehat{\pi}^t) \\ &= \sum_{h=1}^H \mathbb{E}^{M,\widehat{\pi}^t}\left[\overline{Q}_h^t(s_h, a_h) - \left[\mathcal{T}_h^M \overline{Q}_{h+1}^t\right](s_h, a_h)\right] \end{aligned}$$

by Lemma 15. Using the definition of $b_{h,\delta}^t$ and Lemma 30, we have

$$\sum_{h=1}^H \mathbb{E}^{M,\widehat{\pi}^t}\left[\overline{Q}_h^t(s_h, a_h) - \left[\mathcal{T}_h^M \overline{Q}_{h+1}^t\right](s_h, a_h)\right] \lesssim \sqrt{R} \sum_{h=1}^H \mathbb{E}^{M,\widehat{\pi}^t}\left[\|\phi(s_h, a_h)\|_{(\Sigma_h^t)^{-1}}\right].$$

Summing over all timesteps t gives

$$\mathbf{Reg} \leq \sqrt{R} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M,\widehat{\pi}^t}\left[\|\phi(s_h, a_h)\|_{(\Sigma_h^t)^{-1}}\right].$$

By Hoeffding's inequality, we have that with probability at least $1 - \delta$, this is at most

$$\sqrt{R} \sum_{t=1}^T \sum_{h=1}^H \|\phi(s_h^t, a_h^t)\|_{(\Sigma_h^t)^{-1}} + \sqrt{RHT \log(1/\delta)}.$$

The elliptic potential lemma (Lemma 12) now allows us to bound

$$\sum_{t=1}^T \|\phi(s_h^t, a_h^t)\|_{(\Sigma_h^t)^{-1}} \lesssim \sqrt{dT \log(T/d)}$$

for each h, which gives the result.

□

7.3 Bellman Rank

In this section, we continue our study of value-based methods, which assume access to a class $\mathcal{Q}$ of state-action value functions such that $Q^{M^\star,\star} \in \mathcal{Q}$. In the prequel, we saw that the Low Rank MDP assumption facilitates sample-efficient reinforcement learning when $\mathcal{Q}$ is a class of linear functions, but what if we want to learn with *nonlinear* functions such as neural networks? To this end, we will introduce a new structural property, *Bellman rank*, which allows for sample-efficient learning with general classes $\mathcal{Q}$, and subsumes a number of well-studied MDP families, including:

- Low Rank MDPs [75, 42, 5]
- Block MDPs and reactive POMDPs [48, 29].
- MDPs with Linear $Q^\star$ and $V^\star$ [30].
- MDPs with low occupancy complexity [30].
- Linear mixture MDPs [56, 12].
- Linear dynamical systems (LQR) [26].

We will learn about these examples in Section 7.3.3.

Building intuition. Bellman rank is a property of the underlying MDP $M^\star$ which gives a way of controlling *distribution shift*—that is, how many times a deliberate algorithm can be surprised by a substantially new state distribution $d^{M,\pi}$ when it updates its policy. To motivate the property, let us revisit the low rank MDP model. Let M be a low rank MDP with feature map $\phi(s,a) \in \mathbb{R}^d$, and let $Q_h(s,a) = \left\langle \phi(s,a), \theta_h^Q \right\rangle$ be an arbitrary linear value function. Observe that since M is a Low-Rank MDP, we have $[\mathcal{T}_h^M Q](s,a) = \left\langle \phi(x,a), \tilde{\theta}_h^{M,Q} \right\rangle$, where $\tilde{\theta}_h^{M,Q} := w_h^M + \int \mu_h^M(s') \max_{a'} Q_{h+1}(s',a') ds'$. As a result, for any policy π, we can write the Bellman residual for Q as

$$\mathbb{E}^{M,\pi}\left[Q_h(s_h,a_h) - r_h - \max_a Q_{h+1}(s_{h+1},a)\right] = \left\langle \mathbb{E}^{M,\pi}[\phi(s_h,a_h)], \theta_h^Q - w_h^M - \tilde{\theta}_h^{M,Q} \right\rangle \quad (7.20)$$

$$= \langle X_h^M(\pi), W_h^M(Q) \rangle, \quad (7.21)$$

where $X_h^M(\pi) := \mathbb{E}^{M,\pi}[\phi(s_h,a_h)] \in \mathbb{R}^d$ is an "embedding" that depends on π but not Q, and $W_h^M(Q) := \theta_h^Q - w_h^M - \tilde{\theta}_h^{M,Q} \in \mathbb{R}^d$ is an embedding that depends on Q but not π (both embeddings depend on M). Notably, if we view the Bellman residual as a huge $\Pi \times \mathcal{Q}$ matrix $\mathcal{E}_h(\cdot,\cdot) \in \mathbb{R}^{\Pi \times \mathcal{Q}}$ with

$$\mathcal{E}_h(\pi, Q) := \mathbb{E}^{M,\pi}\left[Q_h(s_h,a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1},a)\right)\right], \quad (7.22)$$

then the property (7.21) implies that $\operatorname{rank}(\mathcal{E}_h(\cdot,\cdot)) \leq d$. Bellman rank is an abstraction of this property.

Definition 8 (Bellman rank): For an MDP M with value function class $\mathcal{Q}$ and policy class Π, the Bellman rank is defined as[a]

$$d_{\mathsf{B}}(M) = \max_{h \in [H]} \operatorname{rank}(\{\mathcal{E}_h(\pi, Q)\}_{\pi \in \Pi, Q \in \mathcal{Q}}). \quad (7.23)$$

Equivalently, Bellman rank is the smallest dimension d such that for all h, there exist embeddings $X_h^M(\pi), W_h^M(Q) \in \mathbb{R}^d$ such that

$$\mathcal{E}_h(\pi, Q) = \langle X_h^M(\pi), W_h^M(Q) \rangle \quad (7.24)$$

for all $\pi \in \Pi$ and $Q \in \mathcal{Q}$.

[a] Familiar readers will recognize this notion as *Q-type* Bellman rank.

The utility of Bellman rank is that the factorization in (7.24) gives a way of controlling distribution shift in the MDP M, which facilitates the application of standard generalization guarantees for supervised learning/estimation. Informally, there are only d effective directions in which we can be “surprised” by the state distribution induced by a policy π, to the extent that this matters for the class $\mathcal{Q}$ under consideration; this property was used implicitly in the proof of the regret bound for LSVI-UCB. As we will see, low Bellman rank is satisfied in many settings that go beyond the Low Rank MDP model.

7.3.1 The BiLinUCB Algorithm

We now present an algorithm, BiLinUCB [30], which attains low regret for MDPs with low Bellman rank under the realizability assumption that

$$Q^{M^\star,\star} \in \mathcal{Q}.$$

Like many of the algorithms we have covered, BiLinUCB is based on confidence sets and optimism, though the way we will construct the confidence sets and implement optimism is new.

PAC versus regret. For technical reasons, we will not directly give a regret bound for BiLinUCB. Instead, we will prove a *PAC* (“probably approximately correct”) guarantee. For PAC, the algorithm plays for T episodes, then outputs a final policy $\widehat{\pi}$, and its performance is measured via

$$f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi}). \tag{7.25}$$

That is, instead of considering cumulative performance as with regret, we are only concerned with final performance. For PAC, we want to ensure that $f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi}) \leq \varepsilon$ for some $\varepsilon \ll 1$ using a number of episodes that is polynomial in ε^{-1} and other problem parameters. This is an easier task than achieving low regret: If we have an algorithm that ensures that $\mathbb{E}[\mathbf{Reg}] \lesssim \sqrt{CT}$ for some problem-dependent constant C, we can turn this into an algorithm that achieves PAC error ε using $O\big(\frac{C}{\varepsilon^2}\big)$ episodes via online-to-batch conversion. In the other direction, if we have an algorithm that achieves PAC error ε using $O\big(\frac{C}{\varepsilon^2}\big)$ episodes, one can use this to achieve $\mathbb{E}[\mathbf{Reg}] \lesssim C^{1/3}T^{2/3}$ using a simple explore-then-commit approach; this is lossy, but is the best one can hope for in general.

Algorithm overview. BiLinUCB proceeds in K iterations, each of which consists of n episodes. The algorithm maintains a confidence set $\mathcal{Q}^k \subseteq \mathcal{Q}$ of value functions (generalizing the confidence sets we constructed for structured bandits in Section 4), with the property that $Q^{M^\star,\star} \in \mathcal{Q}$ with high probability. Each iteration k consists of two parts:

- Given the current confidence set $\mathcal{Q}^k$, the algorithm computes a value function Q^k and corresponding policy $\pi^k := \pi_{Q^k}$ that is *optimistic on average*

  $$Q^k = \arg\max_{Q \in \mathcal{Q}^k} \mathbb{E}_{s_1 \sim d_1}[Q_1(s_1, \pi_Q(s_1))].$$

  The main novelty here is that we are only aiming for optimism with respect to the initial state distribution.

- Using the new policy π^k, the algorithm gathers n episodes and uses these to compute estimators $\{\widehat{\mathcal{E}}_h^k(Q)\}_{h \in [H]}$ which approximate the Bellman residual $\mathcal{E}_h(\pi^k, Q)$ for all $Q \in \mathcal{Q}$. Then, in (7.26), the algorithm computes the new confidence set $\mathcal{Q}^{k+1}$ by restricting to value functions for which the estimated Bellman residual is small for $\pi^1, \ldots, \pi^k$. Eliminating value functions with large Bellman residual is a natural idea, because we know from the Bellman equation that $Q^{M^\star,\star}$ has zero Bellman residual.

```
BiLinUCB
Input: $\beta > 0$, iteration count $K \in \mathbb{N}$, batch size $n \in \mathbb{N}$.
$\mathcal{Q}^1 \leftarrow \mathcal{Q}$.
for iteration $k = 1, \ldots, K$ do
    Compute optimistic value function:
        $Q^k = \operatorname*{arg\,max}_{Q \in \mathcal{Q}^k} \mathbb{E}_{s_1 \sim d_1}[Q_1(s_1, \pi_Q(s_1))].$
    and let $\pi^k := \pi_{Q^k}$.
    for $l = 1, \ldots, n$ do
        Execute $\pi^k$ for an episode and observe trajectory $(s_1^{k,l}, a_1^{k,l}, r_1^{k,l}), \ldots, (s_H^{k,l}, a_H^{k,l}, r_H^{k,l})$.
    Compute confidence set
        $\mathcal{Q}^{k+1} = \left\{ Q \in \mathcal{Q} \mid \sum_{i \le k} (\widehat{\mathcal{E}}_h^i(Q))^2 \le \beta \quad \forall h \in [H] \right\}, \qquad (7.26)$
    where
        $\widehat{\mathcal{E}}_h^i(Q) := \frac{1}{n} \sum_{l=1}^{n} \left( Q_h(s_h^{i,l}, a_h^{i,l}) - r_h^{i,l} - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}^{i,l}, a) \right).$
Let $\widehat{k} = \operatorname{arg\,max}_{k \in [K]} \widehat{V}^k$, where $\widehat{V}^k := \frac{1}{n} \sum_{l=1}^{n} \sum_{h=1}^{H} r_h^{k,l}$.
Return $\widehat{\pi} = \pi^{\widehat{k}}$.
```

Main guarantee. The main result for this section is the following PAC guarantee for BiLinUCB.

Proposition 47: Suppose that $M^\star$ has Bellman rank d and $Q^{M^\star,\star} \in \mathcal{Q}$. For any $\varepsilon > 0$ and $\delta > 0$, if we set $n \gtrsim \frac{H^3 d \log(|\mathcal{Q}|/\delta)}{\varepsilon^2}$, $K \gtrsim Hd \log(1+n/d)$, and $\beta \propto c \cdot K \frac{\log|\mathcal{Q}| + \log(HK/\delta)}{n}$, then BiLinUCB learns a policy $\widehat{\pi}$ such

$$f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi}) \le \varepsilon$$

with probability at least $1 - \delta$, and does so using

$$\widetilde{O}\left(\frac{H^4 d^2 \log(|\mathcal{Q}|/\delta)}{\varepsilon^2} \right)$$

episodes.

This result shows that low Bellman rank suffices to learn a near-optimal policy, with sample complexity that only depends on the rank d, the horizon H, and the capacity $\log|\mathcal{Q}|$ for the value function class; this reflects that the algorithm is able to generalize across the state space, with d and $\log|\mathcal{Q}|$ controlling the degree of generalization. The basic principles at play are:

- By choosing Q^k optimistically, we ensure that the suboptimality of the algorithm is controlled by the Bellman residual for Q^k, *on-policy*, similar to what we saw for UCB-

VI and LSVI-UCB. An important difference compared to the LSVI-UCB algorithm we covered in the previous section is that BiLinUCB is only optimistic "on average" with respect to the initial state distribution, i.e.,

$$\mathbb{E}_{s_1 \sim d_1}\left[Q_1^k(s_1, \pi_{Q^k}(s_1))\right] \geq f^{M^\star}(\pi_{M^\star}),$$

while LSVI-UCB aims to find a value function that is uniformly optimistic for all states and actions.

- The confidence set construction (7.26) explicitly removes value functions that have large Bellman residual on the policies encountered so far. The key role of the Bellman rank property is to ensure that there are only $\widetilde{O}(d)$ "effective" state distributions that lead to substantially different values for the Bellman residual, which means that eventually, only value functions with low residual will remain.

Interestingly, the Bellman rank property is only used for analysis, and the algorithm does not explicitly compute or estimate the factorization.

Regret bounds. The BiLinUCB algorithm can be lifted to provide a regret guarantee via a explore-then-commit strategy: Run the algorithm for T_0 episodes to learn a ε-optimal policy, then commit to this policy for the remaining rounds. It is a simple exercise to show that by choosing T_0 appropriately, this approach gives

$$\mathbf{Reg} \leq \widetilde{O}\Big((H^4 d^2 \log(|\mathcal{Q}|/\delta))^{1/3} \cdot T^{2/3}\Big).$$

Under an additional assumption known as *Bellman completeness*, it is possible to attain $\sqrt{T}$ with a variant of this algorithm that uses a slightly different confidence set construction [43].

7.3.2 Proof of Proposition 47

Recall from the definition of Bellman rank that there exist embeddings $X_h^{M^\star}(\pi), W_h^{M^\star}(Q) \in \mathbb{R}^d$ such that for all $\pi \in \Pi$ and $Q \in \mathcal{Q}$,

$$\mathcal{E}_h(\pi, Q) = \left\langle X_h^{M^\star}(\pi), W_h^{M^\star}(Q)\right\rangle.$$

We assume throughout this proof that $\left\|X_h^{M^\star}(\pi)\right\|, \left\|W_h^{M^\star}(Q)\right\|_2 \leq 1$ for simplicity.

Technical lemmas. Before proceeding, we state two technical lemmas. The first lemma establishes validity for the confidence set $\mathcal{Q}^k$ constructed by BiLinUCB.

Lemma 31: For any $\delta > 0$, if we set $\beta = c \cdot K \frac{\log|\mathcal{Q}| + \log(HK/\delta)}{n}$, where $c > 0$ is sufficiently large absolute constant, then with probability at least $1 - \delta$, for all $k \in [K]$:

1. All $Q \in \mathcal{Q}^k$ have
$$\sum_{i<k} (\mathcal{E}_h(\pi^i, Q))^2 \lesssim \beta \quad \forall h \in [H]. \tag{7.27}$$

2. $Q^{M^\star,\star} \in \mathcal{Q}^k$.

Proof of Lemma 31. Using Hoeffding's inequality and a union bound (Lemma 3), we have that with probability at least $1-\delta$, for all $k \in [K]$, $h \in [H]$, and $Q \in \mathcal{Q}$,

$$\left|\widehat{\mathcal{E}}_h^k(Q) - \mathcal{E}_h(\pi^k, Q)\right| \leq C \cdot \sqrt{\frac{\log(|\mathcal{Q}|HK/\delta)}{n}}, \tag{7.28}$$

where c is an absolute constant.

To prove Part 1, we observe that for all k, using the AM-GM inequality, we have that for all $Q \in \mathcal{Q}$,

$$\sum_{i<k}(\mathcal{E}_h(\pi^i, Q))^2 \leq 2\sum_{i<k}(\widehat{\mathcal{E}}_h^i(Q))^2 + 2\sum_{i<k}(\mathcal{E}_h(\pi^i, Q) - \widehat{\mathcal{E}}_h^i(Q))^2.$$

For $Q \in \mathcal{Q}^k$, the definition of $\mathcal{Q}^k$ implies that $\sum_{i<k}(\widehat{\mathcal{E}}_h^i(Q))^2 \leq \beta$, while (7.28) implies that $\sum_{i<k}(\mathcal{E}_h(\pi^i, Q) - \widehat{\mathcal{E}}_h^i(Q))^2 \lesssim \beta$, which gives the result.

For Part 2, we similarly observe that for all k, h and $Q \in \mathcal{Q}$,

$$\sum_{i<k}(\widehat{\mathcal{E}}_h^i(Q))^2 \leq 2\sum_{i<k}(\mathcal{E}_h(\pi^i, Q))^2 + 2\sum_{i<k}(\mathcal{E}_h(\pi^i, Q) - \widehat{\mathcal{E}}_h^i(Q))^2.$$

Since $Q^{M^\star,\star}$ has $\mathcal{E}_h(\pi, Q^{M^\star,\star}) = 0 \ \forall \pi$ by Bellman optimality, we have

$$\sum_{i<k}(\widehat{\mathcal{E}}_h^i(Q^{M^\star,\star}))^2 \leq 2\sum_{i<k}(\mathcal{E}_h(\pi^i, Q^{M^\star,\star}) - \widehat{\mathcal{E}}_h^i(Q^{M^\star,\star}))^2 \leq 2C^2\frac{\log(|\mathcal{Q}|HK/\delta)}{n},$$

where the last inequality uses (7.28). It follows that as long as $\beta \geq 2C^2\frac{\log(|\mathcal{Q}|HK/\delta)}{n}$, we will have $Q^{M^\star,\star} \in \mathcal{Q}^k$ for all k. □

The next result shows that whenever the event in the previous lemma holds, the value functions constructed by BiLinUCB are optimistic.

Lemma 32: Whenever the event in Lemma 31 occurs, the following properties hold:

1. Define

$$\Sigma_h^k = \sum_{i<k} X_h^{M^\star}(\pi^i) X_h^{M^\star}(\pi^i)^\top. \tag{7.29}$$

For all $k \in [K]$, all $Q \in \mathcal{Q}^k$ satisfy

$$\left\|W_h^{M^\star}(Q)\right\|_{\Sigma_h^k}^2 \lesssim \beta. \tag{7.30}$$

2. For all k, Q^k is optimistic in the sense that

$$\mathbb{E}_{s_1\sim d_1}[Q_1^k(s_1, \pi_Q(s_1))] \geq \mathbb{E}_{s_1\sim d_1}\left[Q_1^{M^\star,\star}(s_1, \pi_{M^\star}(s_1))\right] = f^{M^\star}(\pi_{M^\star}). \tag{7.31}$$

Proof of Lemma 32. For Part 1, recall that by the Bilinear class property, we can write $\mathcal{E}_h(\pi^k, Q) = \langle X^{M^\star}(\pi^k), W^{M^\star}(Q)\rangle$, so that (7.27) implies that

$$\left\|W^{M^\star}(Q)\right\|_{\Sigma_h^k} = \sum_{i<k}\left\langle X^{M^\star}(\pi^i), W^{M^\star}(Q)\right\rangle^2 = \sum_{i<k}(\mathcal{E}_h(\pi^i, Q))^2 \lesssim \beta.$$

For Part 2, we observe that for all k, since $Q^{M^\star,\star} \in \mathcal{Q}^k$, we have

$$\begin{aligned}\mathbb{E}_{s_1\sim d_1}[Q_1^k(s_1, \pi_Q(s_1))] &= \sup_{Q\in\mathcal{Q}} \mathbb{E}_{s_1\sim d_1}\left[Q_1(s_1, \pi_Q(s_1))\right] \\ &\geq \mathbb{E}_{s_1\sim d_1}\left[Q_1^{M^\star,\star}(s_1, \pi_{M^\star}(s_1))\right] \\ &= f^{M^\star}(\pi_{M^\star}).\end{aligned}$$

□

Proving the main result. Equipped with the lemmas above, we prove Proposition 47.

Proof of Proposition 47. We first prove a generic bound on the suboptimality of each policy π^k for $k \in [K]$. Let us condition on the event in Lemma 31, which occurs with probability at least $1-\delta$. Whenever this event occurs, Lemma 32 implies that Q^k is optimistic, so we have can bound

$$\begin{aligned}f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^k) &\leq \mathbb{E}_{s_1\sim d_1}\left[Q_1^k(s_1, \pi_{Q^k}(s_1)\right] - f^{M^\star}(\pi^k) &(7.32)\\ &= \sum_{h=1}^H \mathbb{E}^{M^\star,\pi^k}\left[Q_h^k(s_h, a_h) - r_h - \max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}, a)\right] &(7.33)\\ &= \sum_{h=1}^H \left\langle X_h^{M^\star}(\pi^k), W_h^{M^\star}(Q^k)\right\rangle, &(7.34)\end{aligned}$$

where the first equality uses the Bellman residual decomposition (Lemma 15), and the second inequality uses the Bellman rank assumption. For any $\lambda \geq 0$, using Cauchy-Schwarz, we can bound

$$\sum_{h=1}^H \left\langle X_h^{M^\star}(\pi^k), W_h^{M^\star}(Q^k)\right\rangle \leq \sum_{h=1}^H \left\|X_h^{M^\star}(\pi^k)\right\|_{(\lambda I+\Sigma_h^k)^{-1}}\left\|W_h^{M^\star}(Q^k)\right\|_{\lambda I+\Sigma_h^k}$$

For each $h \in [H]$, applying the bound in (7.30) gives

$$\left\|W_h^{M^\star}(Q^k)\right\|_{\lambda I+\Sigma_h^k} \leq \sqrt{\lambda\left\|W_h^{M^\star}(Q^k)\right\|_2^2 + \beta} \leq \lambda^{1/2} + \beta^{1/2},$$

where we have used that $\left\|W_h^{M^\star}(Q^k)\right\|_2 \leq 1$ by assumption. This allows us to bound

$$\sum_{h=1}^H \left\langle X_h^{M^\star}(\pi^k), W_h^{M^\star}(Q^k)\right\rangle \lesssim (\lambda^{1/2} + \beta^{1/2}) \cdot \sum_{h=1}^H \left\|X_h^{M^\star}(\pi^k)\right\|_{(\lambda I+\Sigma_h^k)^{-1}}. \tag{7.35}$$

If we can find a policy π^k for which the right-hand side of (7.35) is small, this policy will be guaranteed to have low regret. The following lemma shows that such a policy is guaranteed to exist.

Lemma 33: For any $\lambda > 0$, as long as $K \geq Hd\log\big(1+\lambda^{-1}K/d\big)$, there exists $k \in [K]$ such that

$$\big\|X_h^{M^\star}(\pi^k)\big\|^2_{(\lambda I+\Sigma_h^k)^{-1}} \lesssim \frac{Hd\log\big(1+\lambda^{-1}K/d\big)}{K} \quad \forall h \in [H]. \tag{7.36}$$

We choose $\lambda = \beta$, which implies that it suffices to take $K \gtrsim Hd\log(1+n/d)$ to satisfy the condition in Lemma 33. By choosing k to satisfy (7.36) and plugging this bound into (7.35), we conclude that the policy π^k has

$$f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^k) \lesssim H\sqrt{\beta \cdot \frac{Hd\log(1+\beta^{-1}K/d)}{K}} \lesssim \widetilde{O}\left(H^{3/2}\sqrt{\frac{d\log(|\mathcal{Q}|/\delta)}{n}}\right) \lesssim \varepsilon \tag{7.37}$$

as desired.

Finally, we need to argue that the policy $\widehat{\pi}$ returned by the algorithm is at least as good as π^k. This is straightforward and we only sketch the argument: By Hoeffding's inequality and a union bound, we have that with probability at least $1-\delta$, for all k,

$$\Big|f^{M^\star}(\pi^k) - \widehat{V}^k\Big| \lesssim \sqrt{\frac{\log(K/\delta)}{n}},$$

which implies that $f^{M^\star}(\widehat{\pi}) \gtrsim f^{M^\star}(\pi^k) - \sqrt{\frac{\log(K/\delta)}{n}}$. The error term here is of lower order than (7.37).

□

Deferred proofs. To finish up, we prove Lemma 33.

Proof of Lemma 33. To prove the result, we need a variant of the elliptic potential lemma (Lemma 12).

Lemma 34 (e.g. Lemma 19.4 in Lattimore and Szepesvári [51]): Let $a_1, \ldots, a_T \in \mathbb{R}^d$ satisfy $\|a_t\|_2 \leq 1$ for all $t \in [T]$. Fix $\lambda > 0$, and let $V_t = \lambda I + \sum_{s<t} a_s a_s^\intercal$. Then

$$\sum_{t=1}^{T} \log(1 + \|a_t\|^2_{V_t^{-1}}) \leq d\log\big(1+\lambda^{-1}T/d\big). \tag{7.38}$$

For any $\lambda > 0$, applying this result for each $h \in [H]$ and summing gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \log\big(1 + \big\|X_h^{M^\star}(\pi^k)\big\|^2_{(\lambda I+\Sigma_h^k)^{-1}}\big) \leq Hd\log\big(1+\lambda^{-1}K/d\big).$$

This implies that there exists k such that

$$\sum_{h=1}^{H} \log\big(1 + \big\|X_h^{M^\star}(\pi^k)\big\|^2_{(\lambda I+\Sigma_h^k)^{-1}}\big) \leq \frac{Hd\log\big(1+\lambda^{-1}K/d\big)}{K},$$

which means that for all $h \in [H]$, $\log\left(1 + \left\|X_h^{M^\star}(\pi^k)\right\|^2_{(\lambda I + \Sigma_h^k)^{-1}}\right) \leq \frac{Hd\log\left(1+\lambda^{-1}K/d\right)}{K}$, or equivalently:

$$\left\|X_h^{M^\star}(\pi^k)\right\|^2_{(\lambda I + \Sigma_h^k)^{-1}} \leq \exp\left(\frac{Hd\log\left(1+\lambda^{-1}K/d\right)}{K}\right) - 1.$$

As long as $K \geq Hd\log\left(1+\lambda^{-1}K/d\right)$, using that $e^x \leq 1 + 2x$ for $0 \leq x \leq 1$, we have

$$\left\|X_h^{M^\star}(\pi^k)\right\|^2_{(\lambda I + \Sigma_h^k)^{-1}} \leq 2\frac{Hd\log\left(1+\lambda^{-1}K/d\right)}{K}.$$

□

7.3.3 Bellman Rank: Examples

We now highlight concrete examples of models with low Bellman rank. We start with familiar examples, the introduce new models that allow for nonlinear function approximation.

Example 7.1 (Tabular MDPs)**.** If M is a tabular MDP with $|\mathcal{S}| \leq S$ and $|\mathcal{A}| \leq A$, we can write the Bellman residual for any function Q and policy π as

$$\begin{aligned}\mathcal{E}_h(\pi, Q) &= \mathbb{E}^{M,\pi}\left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a)\right)\right] \\ &= \sum_{s,a} d_h^{M,\pi}(s,a)\,\mathbb{E}^M\left[Q_h(s,a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a')\right) \mid s_h = s, a_h = a\right].\end{aligned}$$

It follows that if we define

$$X_h^M(\pi) = \left\{d_h^{M,\pi}(s,a)\right\}_{s\in\mathcal{S}, a\in\mathcal{A}} \in \mathbb{R}^{SA}$$

and

$$W_h^M(Q) = \left\{\mathbb{E}^M\left[Q_h(s,a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a')\right) \mid s_h = s, a_h = a\right]\right\}_{s\in\mathcal{S}, a\in\mathcal{A}} \in \mathbb{R}^{SA},$$

we have

$$\mathcal{E}_h(\pi, Q) = \langle X_h^M(\pi), W_h^M(Q)\rangle.$$

This shows that $d_{\mathsf{B}}(M) \leq SA$. ◁

Example 7.2 (Low Rank MDPs)**.** The calculation in (7.21) shows that by choosing $X_h^M(\pi) := \mathbb{E}^{M,\pi}[\phi(s_h, a_h)] \in \mathbb{R}^d$ and $W_h^M(Q) := \theta_h^Q - w_h^M - \tilde{\theta}_h^{M,Q} \in \mathbb{R}^d$, any Low Rank MDP M has $d_{\mathsf{B}}(M) \leq d$. When specialized to this setting, the regret of BiLinUCB is worse than that of LSVI-UCB (though still polynomial in all of the problem parameters). This is because BiLinUCB is a more general algorithm, and does not take advantage of an additional feature of the Low Rank MDP model known as *Bellman completeness*: If M is a Low Rank MDP, then for all $Q \in \mathcal{Q}$, we have $\mathcal{T}_h^M Q_{h+1} \in \mathcal{Q}_{h+1}$. By using a more specialized relative of BiLinUCB that incorporates a modified confidence set construction to exploit completeness, it is possible to match and actually improve upon the regret of LSVI-UCB [43].

◁

We now explore Bellman rank for some MDP families that have not already been covered.

Example 7.3 (Low Occupancy Complexity)**.** An MDP M is said to have *low occupancy complexity* if there exists a feature map $\phi^M(s,a) \in \mathbb{R}^d$ such that for all π, there exists $\theta_h^{M,\pi} \in \mathbb{R}^d$ such that

$$d_h^{M,\pi}(s,a) = \langle \phi^M(s,a), \theta_h^{M,\pi} \rangle. \tag{7.39}$$

Note that neither ϕ^M nor $\theta^{M,\pi}$ is assumed to be known to the learner. If M has low occupancy complexity, then for any value function Q and policy π, we have

$$\begin{aligned}
\mathcal{E}_h(\pi, Q) &= \mathbb{E}^{M,\pi}\Big[Q_h(s_h,a_h) - \Big(r_h + \max_a Q_{h+1}(s_{h+1},a)\Big)\Big] \\
&= \sum_{s,a} d_h^{M,\pi}(s,a)\,\mathbb{E}^M\Big[Q_h(s,a) - \Big(r_h + \max_{a'} Q_{h+1}(s_{h+1},a')\Big) \mid s_h = s, a_h = a\Big] \\
&= \sum_{s,a} \langle \phi^M(s,a), \theta_h^{M,\pi} \rangle\,\mathbb{E}^M\Big[Q_h(s,a) - \Big(r_h + \max_{a'} Q_{h+1}(s_{h+1},a')\Big) \mid s_h = s, a_h = a\Big] \\
&= \Big\langle \theta_h^{M,\pi}, \sum_{s,a} \phi^M(s,a)\,\mathbb{E}^M\Big[Q_h(s,a) - \Big(r_h + \max_{a'} Q_{h+1}(s_{h+1},a')\Big) \mid s_h = s, a_h = a\Big] \Big\rangle.
\end{aligned}$$

It follows that if we define

$$X_h^M(\pi) = \theta_h^{M,\pi}$$

and

$$W_h^M(Q) = \sum_{s,a} \phi^M(s,a)\,\mathbb{E}^M\Big[Q_h(s,a) - \Big(r_h + \max_{a'} Q_{h+1}(s_{h+1},a')\Big) \mid s_h = s, a_h = a\Big],$$

then $\mathcal{E}_h(\pi, Q) = \langle X_h^M(\pi), W_h^M(Q) \rangle$, which shows that $d_{\mathsf{B}}(M) \leq d$.

This setting subsumes tabular MDPs and low rank MDPs, but is substantially more general. Notably, low occupancy complexity allows for *nonlinear* function approximation: As long as the occupancies satisfy (7.39), the Bellman rank is at most d for *any class* $\mathcal{Q}$, which might consist of neural networks or other nonlinear models. ◁

We close with two more new examples.

Example 7.4 (LQR)**.** A classical problem in continuous control is the Linear Quadratic Regulator, or LQR. Here, we have $\mathcal{S} = \mathcal{A} = \mathbb{R}^d$, and states evolve via

$$s_{h+1} = A^M s_h + B^M a_h + \zeta_h,$$

where $\zeta_h \sim \mathcal{N}(0, I)$, and $s_1 \sim \mathcal{N}(0, I)$. We assume that rewards have the form[16]

$$r_h = -s_h^\top Q^M s_h - a_h^\top R^M a_h$$

for matrices $Q^M, R^M \succeq 0$. A classical result, dating back to Kalman, is that the optimal controller for this system is a linear mapping of the form

$$\pi_{M,h}(s) = K_h^M s,$$

and that the value function $Q^{M,\star}(s,a) = (s,a)^\top P_h^M (s,a)$ is quadratic. Hence, it suffices to take $\mathcal{Q}$ to be the set of all quadratic functions in (s,a). With this choice, it can be shown that $d_{\mathsf{B}}(M) \leq d^2 + 1$. The basic idea is to choose $X_h^M(\pi) = (\mathrm{vec}(\mathbb{E}^{M,\pi}[s_h s_h^\top]), 1)$ and use the quadratic structure of the value functions. ◁

[16] LQR is typically stated in terms of losses; we negate because we consider rewards.

Example 7.5 (Linear $Q^{\star}/V^{\star}$)**.** In Proposition 45, we showed that for RL with linear function approximation, assuming only that $Q^{M^{\star},\star}$ is linear is not enough to achieve low regret. It turns out that if we also assume in addition that $V^{M^{\star},\star}$ is linear, the situation improves.

Consider an MDP M. Assume that there known feature maps $\phi(s,a) \in \mathbb{R}^d$ and $\psi(s') \in \mathbb{R}^d$ such that

$$Q_h^{M,\star}(s,a) = \langle \phi(s,a), \theta_h^M \rangle, \quad \text{and} \quad V_h^{M,\star}(s) = \langle \psi(s), w_h^M \rangle.$$

Let

$$\mathcal{Q} = \left\{ Q \mid Q_h(s,a) = \langle \phi(s,a), \theta_h \rangle : \theta_h \in \mathbb{R}^d, \exists w \ \langle \phi(s,a), \theta_h \rangle = \langle \psi(s), w \rangle \ \forall s \right\}.$$

Then $d_{\mathsf{B}}(M) \leq 2d$. We will not prove this result, but the basic idea is to choose $X_h^M(\pi) = \mathbb{E}^{M,\pi}[(\phi(s_h,a_h), \psi(s_{h+1}))]$. ◁

See Jiang et al. [40], Du et al. [30] for further examples.

7.3.4 Generalizations of Bellman Rank

While we gave (7.24) as the definition for Bellman rank, there are many variations on the assumption that also lead to low regret. One well-known variant is *V-type* Bellman rank, which asserts that for all $\pi \in \Pi$ and $Q \in \mathcal{Q}$,

$$\mathbb{E}^{M,\pi}\, \mathbb{E}^{M}_{s_{h+1}|a_h, s_h \sim \pi_Q(s_h)} \left[Q_h(s_h, \pi_Q(s_h)) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, \pi_Q(s_{h+1})) \right] = \langle X_h^M(\pi), W_h^M(Q) \rangle. \tag{7.40}$$

This is the same as the definition (7.24) (which is typically referred to as *Q-type* Bellman rank), except that we take $a_h = \pi_Q(s_h)$ instead of $a_h = \pi(s_h)$.[17] With an appropriate modification, $\mathsf{BiLinUCB}$ can be shown to give sample complexity guarantees that scale with V-type Bellman rank instead of Q-type. This definition captures meaningful classes of tractable RL models that are not captured by the Q-type definition (7.24), with a canonical example being *Block MDPs*.

Example 7.6 (Block MDP)**.** The block MDP [40, 29, 55] is a model in which the ("observed") state space $\mathcal{S}$ is large/high-dimensional, but the dynamics are governed by a (small) latent state space $\mathcal{Z}$. Formally, a Block MDP $M = (\mathcal{S}, \mathcal{A}, P, R, H, d_1)$ is defined based on an (unobserved) *latent state space* $\mathcal{Z}$, with z_h denoting the latent state at layer h. We first describe the dynamics for the latent space. Given initial latent state z_1, the latent states evolve via

$$z_{h+1} \sim P_h^{M,\text{latent}}(z_h, a_h).$$

The latent state z_h is not observed. Instead, we observe

$$x_h \sim q_h^M(z_h),$$

where $q_h^M : \mathcal{Z} \to \Delta(\mathcal{S})$ is an *emission distribution* with the property that $\operatorname{supp}(q_h(z)) \cap \operatorname{supp}(q_h(z')) = \varnothing$ if $z \neq z'$. This property (*decodability*) ensures that there exists a unique

[17] The name "V-type" refers to the fact that (7.40) only depends on Q through the induced V-function $s \mapsto Q_h(s, \pi_Q(s))$, while (7.24) depends on the full Q-function, hence "Q-type".

mapping $\phi^M_h : \mathcal{S} \to \mathcal{Z}$ that maps the observed state x_h to the corresponding latent state s_h. We assume that $R^M_h(s,a) = R^M_h(\phi^M_h(s), a)$, which implies that optimal policy π_M depends only on the endogenous latent state, i.e. $\pi_{M,h}(s) = \pi_{M,h}(\phi^M_h(s))$.

The main challenge of learning in block MDPs is that the decoder ϕ^M is not known to the learner in advance. Indeed, given access to the decoder, one can obtain regret $\mathrm{poly}(H, |\mathcal{Z}|, |\mathcal{A}|) \cdot \sqrt{T}$ by applying tabular reinforcement learning algorithms to the latent state space. In light of this, the aim of the Block MDP setting is to obtain sample complexity guarantees that are independent of the size of the observed state space $|\mathcal{S}|$, and scale as $\mathrm{poly}(|\mathcal{Z}|, |\mathcal{A}|, H, \log|\mathcal{F}|)$, where $\mathcal{F}$ is an appropriate class of function approximators (typically either a value function class $\mathcal{Q}$ containing $Q^{M,\star}$ or a class of decoders Φ that attempts to model ϕ^M directly).

We now show that the Block MDP setting admits low V-type Bellman rank. Observe that we can write

$$\begin{aligned}&\mathbb{E}^{M,\pi}\,\mathbb{E}^M_{s_{h+1}|a_h,s_h\sim\pi_Q(s_h)}\left[Q_h(s_h,\pi_Q(s_h)) - r_h - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},\pi_Q(s_{h+1}))\right]\\&=\sum_{z\in\mathcal{Z}} d^{M,\pi}_h(z)\,\mathbb{E}_{s\sim q^M_h(z)}\,\mathbb{E}^M_{s_{h+1}|s_h,a_h\sim\pi_Q(s_h)}\left[Q_h(s_h,\pi_Q(s_h)) - r_h - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},\pi_Q(s_{h+1}))\right].\end{aligned}$$

This implies that we can take

$$X^M_h(\pi) = \left\{d^{M,\pi}_h(z)\right\}_{z\in\mathcal{Z}}$$

and

$$W^M_h(Q) = \left\{\mathbb{E}_{s\sim q^M_h(z)}\,\mathbb{E}^M_{s_{h+1}|s_h,a_h\sim\pi_Q(s_h)}\left[Q_h(s_h,\pi_Q(s_h)) - r_h - \max_{a\in\mathcal{A}} Q_{h+1}(s_{h+1},\pi_Q(s_{h+1}))\right]\right\}_{z\in\mathcal{Z}}$$

so that the V-type Bellman rank is at most $|\mathcal{Z}|$. This means that as long as $\mathcal{Q}$ contains $Q^{M,\star}$, we can obtain sample complexity guarantees that scale with $|\mathcal{Z}|$ rather than $|\mathcal{S}|$, as desired. $\triangleleft$

See Du et al. [30] for a more general framework, *Bilinear classes*, which subsumes both of these Bellman rank definitions.

7.3.5 Decision-Estimation Coefficient for Bellman Rank

An alternative to the BiLinUCB method is to appeal to the E2D meta-algorithm and the Decision-Estimation Coefficient. The following result shows that the Decision-Estimation Coefficient is always bounded for classes with low Bellman rank.

Proposition 48: For any class of MDPs $\mathcal{M}$ for which all $M \in \mathcal{M}$ have Bellman rank at most d, we have

$$\mathsf{dec}_\gamma(\mathcal{M}) \lesssim \frac{H^2 d}{\gamma}. \tag{7.41}$$

This implies that that E2D meta-algorithm has $\mathbb{E}[\mathbf{Reg}] \lesssim H\sqrt{dT \cdot \mathbf{Est}_{\mathsf{H}}}$ whenever we have access to a realizable model class with low Bellman rank. As a special case, for any finite class $\mathcal{M}$, using averaged exponential weights as an estimation oracle gives

$$\mathbb{E}[\mathbf{Reg}] \lesssim H\sqrt{dT\log|\mathcal{M}|}. \tag{7.42}$$

We will not prove Proposition 48, but interested readers can refer to Foster et al. [35]. The result can be proven using two approaches, both of which build on the techniques we have already covered. The first approach is to apply a more general version of the PC-IGW algorithm from Section 6.6, which incorporates optimal design in the space of policies. The second approach is to move to the Bayesian DEC and appeal to posterior sampling, as in Section 4.4.2.

Value-based guarantees via optimistic estimation. In general, the model estimation complexity $\log|\mathcal{M}|$ in (7.42) can be arbitrarily large compared to the complexity $\log|\mathcal{Q}|$ for a realizable value function class (consider the low rank MDP—since μ is unknown, it is not possible to construct a small model class $\mathcal{M}$). To derive value-based guarantees along the lines of what BiLinUCB achieves in Proposition 47, a natural approach is to replace the Hellinger distance appearing in the DEC with a divergence tailored to value function iteration, following the development in Sections 6.7.2 and 6.7.3. Once such choice is the divergence

$$D^{\pi}_{\mathsf{sbr}}(Q \parallel M) = \sum_{h=1}^{H}\left(\mathbb{E}^{M,\pi}\left[Q_h(s_h,a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1},a)\right)\right]\right)^2,$$

which measures the squared bellman residual for an estimated value function under M. With this choice, we appeal to the optimistic E2D algorithm (E2D.Opt) from Section 6.7.3. One can show that the optimistic DEC for this divergence is bounded as

$$\text{o-dec}^{D_{\mathsf{sbr}}}_{\gamma}(\mathcal{M}) \lesssim \frac{H\cdot d}{\gamma}.$$

This implies that E2D.Opt, with an appropriate choice of estimation algorithm tailored to $D^{\pi}_{\mathsf{sbr}}(\cdot \parallel \cdot)$, achieves

$$\mathbb{E}[\mathbf{Reg}] \lesssim (H^2 d\log|\mathcal{Q}|)^{1/2}T^{3/4}.$$

Note that due to the asymmetric nature of $D^{\pi}_{\mathsf{sbr}}(\cdot \parallel \cdot)$, it is critical to appeal to optimistic estimation to derive this result. Indeed, the non-optimistic generalized DEC $\mathsf{dec}^{D_{\mathsf{sbr}}}_{\gamma}$ does not enjoy a polynomial bound. See Foster et al. [36] for details.

[Note: This subsection will be expanded in the next version.]

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

[2] N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 3–11. Morgan Kaufmann Publishers Inc., 1999.

[3] A. Agarwal and T. Zhang. Model-based RL with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022.

[4] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

[5] A. Agarwal, S. Kakade, A. Krishnamurthy, and W. Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Neural Information Processing Systems (NeurIPS)*, 2020.

[6] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[7] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[8] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations.* Cambridge University Press, 2009.

[9] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[11] P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

[12] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[13] R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

[14] B. Bilodeau, D. J. Foster, and D. Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, 2020.

[15] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[16] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[17] S. Bubeck and R. Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589, 2016.

[18] S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Conference on Learning Theory*, pages 266–278, 2015.

[19] 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*, pages 72–85, 2017.

[20] N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Conference on Computational Learning Theory*, 1999.

[21] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

[22] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[23] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1991.

[24] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, 2008.

[25] C. Dann, M. Mohri, T. Zhang, and J. Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

[26] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[27] S. Dong, B. Van Roy, and Z. Zhou. Provably efficient reinforcement learning with aggregated states. *arXiv preprint arXiv:1912.06366*, 2019.

[28] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence. *Annals of Statistics*, 1987.

[29] S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudik, and J. Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[30] S. S. Du, S. M. Kakade, J. D. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in RL. *International Conference on Machine Learning*, 2021.

[31] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

[32] D. J. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.

[33] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. *Conference on Learning Theory*, 2018.

[34] D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[36] D. J. Foster, N. Golowich, J. Qian, A. Rakhlin, and A. Sekhari. A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*, 2022.

[37] D. J. Foster, A. Rakhlin, A. Sekhari, and K. Sridharan. On the complexity of adversarial decision making. *arXiv preprint arXiv:2206.13063*, 2022.

[38] D. J. Foster, N. Golowich, and Y. Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023.

[39] E. Hazan and S. Kale. An online portfolio selection algorithm with regret logarithmic in price variation. *Mathematical Finance*, 2015.

[40] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

[41] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

[42] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[43] C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Neural Information Processing Systems*, 2021.

[44] A. Kalai and S. Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 2002.

[45] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[46] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17:697–704, 2004.

[47] R. Kleinberg, A. Slivkins, and E. Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

[48] A. Krishnamurthy, A. Agarwal, and J. Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

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

[50] T. Lattimore and C. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR, 2019.

[51] T. Lattimore and C. Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

[52] G. Li, P. Kamath, D. J. Foster, and N. Srebro. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021.

[53] L. Li. *A unifying framework for computational reinforcement learning theory.* Rutgers, The State University of New Jersey—New Brunswick, 2009.

[54] H. Luo, C.-Y. Wei, and K. Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, 2018.

[55] D. Misra, M. Henaff, A. Krishnamurthy, and J. Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

[56] A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[57] M. Opper and D. Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction and Distribution*, 1999.

[58] L. Orseau, T. Lattimore, and S. Legg. Soft-bayes: Prod for mixtures of experts with log-loss. In *International Conference on Algorithmic Learning Theory*, 2017.

[59] Y. Polyanskiy. Information theoretic methods in statistics and computer science. 2020. URL `https://people.lids.mit.edu/yp/homepage/sdpi_course.html`.

[60] A. Rakhlin and K. Sridharan. Statistical learning and sequential prediction, 2012. Available at `http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf`.

[61] A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

[62] J. Rissanen. Complexity of strings in the class of markov sources. *IEEE Transactions on Information Theory*, 32(4):526–532, 1986.

[63] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

[64] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[65] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[66] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[67] Y. M. Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 1987.

[68] D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

[69] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958.

[70] W. 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.

[71] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 2008.

[72] V. Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on computational learning theory*, pages 51–60. ACM, 1995.

[73] Y. Wang, R. Wang, and S. M. Kakade. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. *Neural Information Processing Systems (NeurIPS)*, 2021.

[74] G. Weisz, P. Amortila, and C. Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

[75] L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[76] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

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

[78] H. Zhong, W. Xiong, S. Zheng, L. Wang, Z. Wang, Z. Yang, and T. Zhang. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

A. TECHNICAL TOOLS

A.1 Probabilistic Inequalities

A.1.1 Tail Bounds with Stopping Times

Lemma 35 (Hoeffding's inequality with adaptive stopping time): For i.i.d. random variables $Z_1, \ldots, Z_T$ taking values in $[a, b]$ almost surely, with probability at least $1 - \delta$,

$$\frac{1}{T'} \sum_{i=1}^{T'} Z_i - \mathbb{E}[Z] \leq (b - a)\sqrt{\frac{\log(T/\delta)}{2T'}} \qquad \forall 1 \leq T' \leq T. \tag{A.1}$$

As a consequence, for any random variable $\tau \in [T]$ with the property that for all $t \in [T]$, $\mathbb{I}\{\tau \leq t\}$ is a measurable function of $Z_1, \ldots, Z_{t-1}$ (τ is called a *stopping time*), we have that with probability at least $1 - \delta$,

$$\frac{1}{\tau} \sum_{i=1}^{\tau} Z_i - \mathbb{E}[Z] \leq (b - a)\sqrt{\frac{\log(T/\delta)}{2\tau}}. \tag{A.2}$$

Proof of Lemma 35. Lemma 3 states that for any fixed $T' \in [T]$, with probability at least $1 - \delta$,

$$\frac{1}{T'} \sum_{i=1}^{T'} Z_i - \mathbb{E}[Z] \leq (b - a)\sqrt{\frac{\log(T/\delta)}{2T'}}.$$

(A.1) follows by applying this result with $\delta' = \delta/T$ and taking a union bound over all T choices for $T' \in [T]$. For (A.2), we observe that

$$\frac{1}{\tau} \sum_{i=1}^{\tau} (Z_i - \mathbb{E}[Z]) - (b - a)\sqrt{\frac{\log(T/\delta)}{2\tau}} \leq \max_{T' \in [T]} \left\{ \frac{1}{T'} \sum_{i=1}^{T'} (Z_i - \mathbb{E}[Z]) - (b - a)\sqrt{\frac{\log(T/\delta)}{2T'}} \right\}.$$

The result now follows from (A.1). □

A.1.2 Tail Bounds for Martingales

The next result is a martingale counterpart to Bernstein's inequality (Lemma 5).

Lemma 36 (Freedman's inequality (Bernstein for martingales)): Let $(X_t)_{t \leq T}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathscr{F}_t)_{t \leq T}$. If $|X_t| \leq R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1 - \delta$, for all $T' \leq T$,

$$\sum_{t=1}^{T'} X_t \leq \eta \sum_{t=1}^{T'} \mathbb{E}_{t-1}\left[X_t^2\right] + \frac{\log(\delta^{-1})}{\eta}.$$

Proof of Lemma 36. Without loss of generality, let $R = 1$, and fix $\eta \in (0,1)$. The sequence

$$Z_\tau := \exp\left(\sum_{t=1}^{\tau} \eta X_t - \eta^2 \mathbb{E}_{t-1}\left[X_t^2\right]\right)$$

is a nonnegative supermartingale with respect to the filtration $(\mathscr{F}_\tau)_{\tau \leq T}$. Indeed, we have

$$\mathbb{E}_{\tau-1}[Z_\tau] = \exp\left(\sum_{t=1}^{\tau-1} \eta X_t - \eta^2 \mathbb{E}_{t-1}\left[X_t^2\right]\right) \cdot \mathbb{E}_{\tau-1}\left[\exp\left(\eta X_\tau - \eta^2 \mathbb{E}_{\tau-1}\left[X_\tau^2\right]\right)\right]$$

Recall that $\mathbb{E}_{\tau-1} X_\tau = 0$. Using the fact that $e^a \leq 1 + a + (e-2)a^2$ for $a \leq 1$, and $1 + b \leq e^b$ for all $b \in \mathbb{R}$,

$$\mathbb{E}_{\tau-1}\left[\exp\left(\eta X_\tau - \eta^2 \mathbb{E}_{\tau-1}\left[X_\tau^2\right]\right)\right] \leq 1.$$

Ville's inequality implies that for all $\lambda > 0$,

$$\mathbb{P}(\exists \tau : Z_\tau > \lambda) \leq \frac{1}{\lambda}.$$

The result now follows by the Chernoff method. □

The following result is an immediate consequence of Lemma 36.

Lemma 37: Let $(X_t)_{t \leq T}$ be a sequence of random variables adapted to a filtration $(\mathscr{F}_t)_{t \leq T}$. If $0 \leq X_t \leq R$ almost surely, then with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} X_t \leq \frac{3}{2} \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + 4R \log(2\delta^{-1}),$$

and

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \leq 2 \sum_{t=1}^{T} X_t + 8R \log(2\delta^{-1}).$$

A.2 Information Theory

A.2.1 Properties of Hellinger Distance

Lemma 38: For any distributions $\mathbb{P}$ and $\mathbb{Q}$ over a pair of random variables (X, Y),

$$\mathbb{E}_{X \sim \mathbb{P}_X}\left[D^2_{\mathsf{H}}\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right] \leq 4 D^2_{\mathsf{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}).$$

Proof of Lemma 38. Since squared Hellinger distance is an f-divergence, we have

$$\mathbb{E}_{X \sim \mathbb{P}_X}\left[D^2_{\mathsf{H}}\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right] = D^2_{\mathsf{H}}\left(\mathbb{P}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{P}_X\right).$$

Next, using that Hellinger distance satisfies the triangle inequality, along with the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have,

$$\begin{aligned}
\mathbb{E}_{X\sim\mathbb{P}_X}\left[D^2_{\mathsf{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\right] &\leq 2D^2_{\mathsf{H}}\big(\mathbb{P}_{Y|X}\otimes\mathbb{P}_X, \mathbb{Q}_{Y|X}\otimes\mathbb{Q}_X\big) + 2D^2_{\mathsf{H}}\big(\mathbb{Q}_{Y|X}\otimes\mathbb{P}_X, \mathbb{Q}_{Y|X}\otimes\mathbb{Q}_X\big) \\
&= 2D^2_{\mathsf{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}) + 2D^2_{\mathsf{H}}(\mathbb{P}_X, \mathbb{Q}_X) \\
&\leq 4D^2_{\mathsf{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}),
\end{aligned}$$

where the final line follows from the data processing inequality. □

Lemma 39 (Subadditivity for squared Hellinger distance): Let $(\mathcal{X}_1, \mathscr{F}_1), \ldots, (\mathcal{X}_n, \mathscr{F}_n)$ be a sequence of measurable spaces, and let $\mathcal{X}^i = \prod_{i=t}^{i} \mathcal{X}_t$ and $\mathscr{F}^i = \bigotimes_{t=1}^{i} \mathscr{F}_t$. For each i, let $\mathbb{P}^i(\cdot \mid \cdot)$ and $\mathbb{Q}^i(\cdot \mid \cdot)$ be probability kernels from $(\mathcal{X}^{i-1}, \mathscr{F}^{i-1})$ to $(\mathcal{X}_i, \mathscr{F}_i)$. Let $\mathbb{P}$ and $\mathbb{Q}$ be the laws of $X_1, \ldots, X_n$ under $X_i \sim \mathbb{P}^i(\cdot \mid X_{1:i-1})$ and $X_i \sim \mathbb{Q}^i(\cdot \mid X_{1:i-1})$ respectively. Then it holds that

$$D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) \leq 10^2 \log(n) \cdot \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D^2_{\mathsf{H}}(\mathbb{P}^i(\cdot \mid X_{1:i-1}), \mathbb{Q}^i(\cdot \mid X_{1:i-1}))\right]. \tag{A.3}$$

A.2.2 Change-of-Measure Inequalities

Lemma 40 (Pinsker for sub-Gaussian random variables): Suppose that $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$ are both σ^2-sub-Gaussian. Then

$$|\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[Y]| \leq \sqrt{2\sigma^2 \cdot D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q})}.$$

Lemma 41 (Multiplicative Pinsker-type inequality for Hellinger distance): lemmampmin Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{F})$. For all $h : \mathcal{X} \to \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) \cdot D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q})}. \tag{A.4}$$

In particular,

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leq 3\,\mathbb{E}_{\mathbb{Q}}[h(X)] + 4R \cdot D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{A.5}$$

Proof of Lemma 41. Let a measurable event A be fixed. Let $p = \mathbb{P}(A)$ and $q = \mathbb{Q}(A)$. Then we have

$$\frac{(p-q)^2}{2(p+q)} \leq (\sqrt{p} - \sqrt{q})^2 \leq D^2_{\mathsf{H}}((p, 1-p), (q, 1-q)) \leq D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}),$$

where the third inequality is the data-processing inequality. It follows that

$$|p - q| \leq \sqrt{2(p+q) D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q})},$$

To deduce the final result for $R = 1$, we observe that $\mathbb{E}_{\mathbb{P}}[h(X)] = \int_0^1 \mathbb{P}(h(X) > t)dt$ and likewise for $\mathbb{E}_{\mathbb{Q}}[h(X)]$, then apply Jensen's inequality. The result for general R follows by rescaling.

The inequality in (A.5) follows by applying the AM-GM inequality to (A.4) and rearranging.

□

A.3 Minimax Theorem

Lemma 42 (Sion's Minimax Theorem [69]): Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets in linear topological spaces, and assume $\mathcal{X}$ is compact. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be such that (i) $f(x, \cdot)$ is concave and upper semicontinuous over $\mathcal{Y}$ for all $x \in \mathcal{X}$ and (ii) $f(\cdot, y)$ is convex and lower semicontinuous over $\mathcal{X}$ for all $y \in \mathcal{Y}$. Then

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y). \tag{A.6}$$