

Foundations of Reinforcement Learning and Interactive Decision Making

Dylan J. Foster

University of California, Berkeley
dylanf@mit.edu

Alexander Rakhlin

Massachusetts Institute of Technology
rakhlin@mit.edu

Abstract

Interactive decision making is the problem of learning to act well in an unknown environment, using the data that one's own actions generate to continuously improve, and arises in situations ranging from online platforms and robotics to medical treatments. This monograph gives a *statistical perspective* on algorithm design and complexity for interactive decision making, building from multi-armed bandits through contextual and structured bandits to reinforcement learning with function approximation within a single, unified framework. Special attention is paid to function approximation and flexible models such as neural networks, and to the connection between supervised learning and decision making: the reader will learn how to turn any supervised learning method into a decision making algorithm, how to analyze the result, and how to determine whether a given problem can be solved with few interactions.

A unifying theme is that for interactive problems, unlike supervised learning, the question of how hard a problem is cannot be separated from the question of how to solve it. We develop this theme through two recent ideas, which complement classical approaches such as optimism and posterior sampling: the Estimation-to-Decisions principle, which reduces decision making to supervised estimation combined with an exploration rule, and the Decision-Estimation Coefficient, a complexity measure that both determines the exploration rule and lower bounds the regret of any algorithm. For contextual bandits, the resulting theory is essentially complete and its algorithms are widely deployed; for reinforcement learning, it represents the current frontier.

CONTENTS

1	Introduction	4
1.1	Interactive Decision Making	4
1.2	A Spectrum of Decision Making Problems	5
1.3	Scope of This Monograph	7
1.4	The Interplay Between Statistical Complexity and Algorithm Design	8
2	Background: Statistical and Online Learning	10
2.1	Statistical Learning and Generalization	10
2.2	Online Learning and Prediction	15
2.3	Online Learning Algorithms	19
2.4	Additional Background: Random Variables and Tail Bounds	22
2.5	Bibliographic Remarks and Further Reading	23
2.6	Deferred Proofs: Statistical Learning	24
2.7	Deferred Proofs: Online Learning	26
3	Multi-Armed Bandits: The Exploration–Exploitation Tradeoff	30
3.1	The Multi-Armed Bandit Problem	30
3.2	The Need for Exploration	32
3.3	The ε -Greedy Algorithm	33

3.4	Optimism and the Upper Confidence Bound (UCB) Algorithm	36
3.5	Bayesian Bandits and the Posterior Sampling Algorithm*	40
3.6	Adversarial Bandits and the Exp3 Algorithm*	44
3.7	Bibliographic Remarks and Further Reading	46
3.8	Deferred Proofs	46
4	Contextual Bandits: Exploration and Generalization	49
4.1	The Contextual Bandit Problem	49
4.2	Optimism for Contextual Bandits: A Generic Template	52
4.3	Optimism for Linear Models: The LinUCB Algorithm	55
4.4	Failure of Optimism Beyond Linear Classes	59
4.5	Reduction to Online Regression	60
4.6	A First Attempt: The ε -Greedy Algorithm	61
4.7	Inverse Gap Weighting and the Estimation-to-Decisions Principle	63
4.8	Bibliographic Remarks and Further Reading	68
4.9	Deferred Proofs	69
5	Structured Bandits: Exploration in Large Decision Spaces	72
5.1	The Structured Bandit Problem	72
5.2	Warm-Up: Optimism and the Eluder Dimension	74
5.3	The Decision-Estimation Coefficient: Estimation-to-Decisions for Structured Bandits	80
5.4	Examples: Computing the Decision-Estimation Coefficient	85
5.5	Relationship to Optimism and Posterior Sampling	91
5.6	Incorporating Contexts*	95
5.7	Additional Properties of the Decision-Estimation Coefficient*	96
5.8	Bibliographic Remarks and Further Reading	97
6	Reinforcement Learning: Basic Principles and Finite State Spaces	99
6.1	The Episodic Reinforcement Learning Problem	99
6.2	Planning via Dynamic Programming	101
6.3	The Need for Strategic Exploration	103
6.4	Analysis Tools: Performance Difference, Bellman Residuals, and Simulation	104
6.5	Optimism in Reinforcement Learning	107
6.6	Optimism for Tabular MDPs: The UCB-VI Algorithm	109
6.7	Estimation-to-Decisions for Tabular Reinforcement Learning	114
6.8	Bibliographic Remarks and Further Reading	119
6.9	Deferred Proofs	120
7	General Decision Making: A Unified Framework	123
7.1	General Decision Making Framework	123
7.2	Refresher: Information-Theoretic Divergences	125
7.3	Decision-Estimation Coefficient for General Decision Making	127
7.4	Estimation-to-Decisions Algorithm for General Decision Making	131
7.5	The Constrained Decision-Estimation Coefficient	135
7.6	Bibliographic Remarks and Further Reading	137
7.7	Deferred Proofs	138
8	Lower Bounds: Necessity of the Decision-Estimation Coefficient	141

8.1	Minimax Regret and Optimality	141
8.2	Warm-Up: Lower Bound for the Multi-Armed Bandit	141
8.3	A First Lower Bound via the Constrained Decision-Estimation Coefficient	146
8.4	Main Lower Bound for the Constrained Decision-Estimation Coefficient	148
8.5	Examples	151
8.6	Tighter Regret Bounds for the Decision-Estimation Coefficient*	154
8.7	Bibliographic Remarks and Further Reading	160
9	Reinforcement Learning: Function Approximation and Large State Spaces	161
9.1	Setting and Realizability	161
9.2	Low-Rank MDPs and LSVI-UCB	163
9.3	Bellman Rank and BiLinUCB	170
9.4	Examples and Generalizations of Bellman Rank	178
9.5	Estimation-to-Decisions for Low Bellman Rank	181
9.6	Concluding Remarks and Broader Outlook	183
9.7	Bibliographic Remarks	185
A	Technical Tools	201
A.1	Probabilistic Inequalities	201
A.2	Information Theory	203
A.3	Minimax Theorem	204

1. INTRODUCTION

Interactive decision making is the problem of learning to act well in an unknown environment, using the data that one’s own actions generate to continuously improve. A doctor choosing treatments, a recommendation system choosing what to display, and a robot learning to navigate all face the same tension: to make good decisions, one must know the consequences of each option, yet consequences can only be learned by choosing, and there is no way to learn about the options one does not try. This monograph develops the statistical foundations of this problem: what makes a decision making problem easy or hard, and how to design algorithms that learn from as few interactions as possible.

We give a unified answer to both questions through *Estimation-to-Decisions*, a single principle for choosing decisions that encompasses settings ranging from classical bandit problems to general reinforcement learning. The principle splits decision making into two parts: estimation, which can be handed to whatever off-the-shelf machine learning model one prefers, and an exploration rule that turns the model’s predictions into decisions. A reader who works through the monograph will learn to build decision making algorithms out of any supervised learning method, analyze them, and tell, for a problem they have not seen before, whether it can be solved with few interactions and what the cost will be.

1.1 Interactive Decision Making

Interactive decision making is a ubiquitous problem, found in medicine, online platforms, robotics, and the sciences; the following examples illustrate the key challenges.

- ▷ Medical treatment: based on a patient’s medical history and vital signs, a doctor must decide which treatment will lead to the best outcome.
- ▷ Controlling a robot: based on sensor readings, a controller must decide what signals to send to the robot’s actuators so that it reaches a goal.

In both examples, the decision maker or *learner* (in the first example the doctor, in the second the controller) interacts with an *unknown environment*. The doctor does not know in advance how a given patient will respond to each treatment, and the controller does not know how the signals it sends change the robot’s configuration, or what the terrain looks like. But because the learner acts, it can learn: every decision produces data, the data reduces uncertainty about the environment, and less uncertainty means better decisions later. The crux of interactive decision making is that each decision plays two roles at once, and the learner must balance them: *exploring*, to reduce uncertainty about the environment, and *exploiting*, to perform well given what they know now (e.g., reaching the goal as fast as possible). In this monograph, the learner is an algorithm: our goal is to automate the treatment decisions or robot control signals, in a way that improves with experience.

More concretely, [Figure 1](#) depicts an idealized version of the medical treatment example, to which we return throughout the monograph. At each round t , the learner—here the doctor, who plays the role of the algorithm—observes a patient’s medical history and vital signs, summarized in a *context* x^t ; chooses a treatment, the *decision* π^t ; and then observes the outcome of the treatment, in the form of a *reward* r^t and an auxiliary *observation* o^t about, say, the progression of the illness. Over many rounds, we wish to use the data $\{(x^t, \pi^t, r^t, o^t)\}_{t=1}^T$ collected by repeating this process to learn a good mapping $x \mapsto \pi$ from contexts to decisions. How should such an algorithm choose

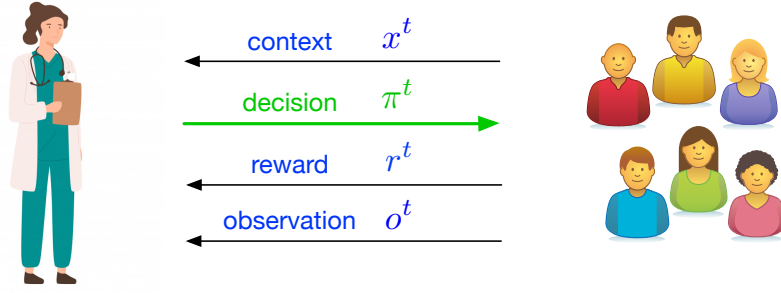


Figure 1: The general decision making problem. At each round, the learner (a doctor) selects a decision and observes a reward and an observation.

its decisions, so that it learns a good mapping from as few interactions as possible?

1.2 A Spectrum of Decision Making Problems

The problem in Figure 1 can be formalized through several settings, including multi-armed bandits, contextual bandits, and reinforcement learning, which differ in the assumptions they place on the environment and the decision making protocol. Each assumption corresponds to a challenge that an algorithm must overcome, and together they give rise to a *spectrum* of decision making problems. Figure 2 organizes this spectrum along three axes, one for each of the core challenges we focus on:

- ▷ **Interactivity.** Does the learner observe data passively, or do its decisions influence what data it collects? In Figure 1, the doctor observes the effects of the treatments prescribed, but not the *counterfactuals*: the effects of the treatments not given. The doctor’s decisions therefore shape the data, and with them the ability to estimate the effects of the different treatments. This contrasts with classical statistical learning (Section 2), where the dataset is given to the learner upfront, with no control over how it was collected. Multi-armed bandits (Section 3) isolate this challenge.
- ▷ **Function approximation and generalization.** Supervised learning uses *function approximation* (neural networks, kernels, forests) to generalize across the space of covariates. Interactive decision making algorithms can use function approximation in the same way, either to generalize across contexts (e.g., different patients) or to generalize across *decisions* (e.g., different treatments). In Figure 1, the context x^t summarizing the medical history and vital signs may be a highly structured object, and the treatment π^t may be a high-dimensional vector with interacting components, or a multi-stage treatment strategy. Contextual bandits (Section 4) and structured bandits (Section 5) isolate generalization across contexts and across decisions, respectively.
- ▷ **Data.** Is the data (rewards and observations) the learner collects generated from their decisions through a fixed process, or does the underlying process change over time, potentially adversarially, or in response to the learner’s actions? If the process is fixed, do we wish to model it directly, or proceed agnostically? Do we observe only labels, as in supervised learning, or a full trajectory of states, actions, and rewards, as for the robot control example? Online learning (Section 2) isolates the first of these questions in a non-interactive setting; reinforcement learning (Sections 6 and 9) raises the last.

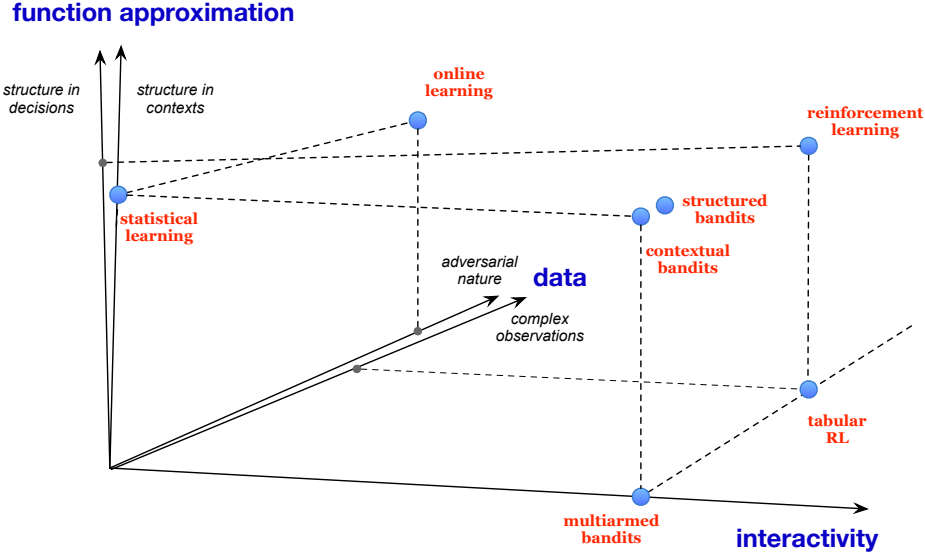


Figure 2: The landscape of decision making problems, organized along three axes: interactivity, function approximation, and data.

As Figure 2 shows, the basic frameworks—statistical learning, online learning, contextual bandits, structured bandits—are idealized problems that each capture one or two of these challenges, while richer settings such as reinforcement learning involve all of them.

Given this diversity, one might worry that every point on the spectrum calls for its own algorithm, designed from scratch. Supervised learning does not have this problem: there is essentially a single guiding principle—fit a model to the data—and it applies whether the model is a linear function or a neural network. For decision making, the situation was far less clear until recently. The body of theory this monograph presents, developed by many researchers over the past two decades, has played a key role in clarifying it: it tells us what each position on the spectrum demands of an algorithm, and it has produced algorithms that are now widely used in practice, quite apart from their theoretical interest. What a reader working through this monograph can hope to take away depends on where on the spectrum one looks.

Short-horizon problems: A settled theory. At the near end of the spectrum, for problems in which each round involves a single decision, such as contextual bandits, the reader will find a theory of exploration that is, at least for the minimax questions we ask, essentially complete, and that has been validated in practice. The first generation of contextual bandit algorithms with provable guarantees—LinUCB and its relatives [21, 145, 57, 1], Thompson sampling [49, 14], and the agnostic methods of Langford and Zhang [133], Dudik et al. [72], Agarwal et al. [5]—was developed in the 2000s and 2010s, and deployed in applications from news recommendation [145] to mobile health [197]. The regression-based methods that grew out of this line of work a decade later [80], SquareCB and the Estimation-to-Decisions principle behind it [79, 84], are now standard, and are deployed across industry.¹ Sections 3 to 5 present this theory.

¹SquareCB ships in Vowpal Wabbit [96], the engine behind Azure Personalizer, and in Meta’s Pearl library [230]. Documented deployments of SquareCB and its inverse gap weighting rule include Uber [104], Adyen [203], and Expedia [35]. See Agarwal et al. [6] for the Decision Service, a canonical example of an early deployment, and Bietti et al. [32] for an empirical comparison of exploration algorithms.

Long-horizon problems: The frontier. At the far end of the spectrum, for problems that unfold over many steps, as in reinforcement learning, the reader will find the current frontier. Exploration was a central theme of the deep reinforcement learning era of artificial intelligence that began with DQN and AlphaGo [153, 188], which produced a range of methods aimed at exploring more deliberately than the random perturbations used by default: bootstrapped value functions, count-based and curiosity-driven bonuses, random network distillation, and Go-Explore [162, 30, 196, 164, 43, 73]. Yet how much deliberate exploration matters in practice, and in which domains, remains unclear, as does the role it will play as AI systems increasingly build on powerful pre-trained models. We expect exploration to remain a fundamental capability, but it may take a form quite different from the one studied classically. For now, what we give is our best account of the field’s frontier of understanding for when, and how, one can explore better than at random. Sections 7 to 9 present this account as it stands today, after Section 6 lays the groundwork in the tabular setting.

In both regimes, the lesson as to what the theory can offer is the same: reasoning about *statistical complexity*, that is, about how many interactions a problem fundamentally requires, tells us how to use what we know about the problem—whether stated explicitly, through the decision making protocol, or implicitly, through the choice of a model such as a neural network—to decide how to explore, adaptively, as data come in. This kind of reasoning has produced methods, SquareCB among them, that are useful in practice and yet explore in ways that were far from obvious a priori. Underlying this is a perspective that Section 1.4 develops: in interactive decision making, the question of how hard a problem is cannot be separated from the question of how to solve it.

1.3 Scope of This Monograph

We aim to do one thing well: to give the reader a working understanding of the statistical complexity of interactive decision making. By this we mean an understanding of what determines how many interactions a problem requires, how this depends on the structure of the problem and on the model class, and how to design algorithms that come as close as possible to the best achievable rate. We build up this understanding one setting at a time, using classical tools such as optimism, posterior sampling, and concentration inequalities alongside the more recent Estimation-to-Decisions principle and Decision-Estimation Coefficient, which tie the settings together into a single framework.

Roadmap. Sections 2 to 6 add one form of structure at a time, following the axes of Figure 2; Sections 7 to 9 give the general theory and apply it to reinforcement learning.

- ▷ **Background.** Section 2 reviews statistical and online learning, and states the estimation guarantees that the algorithms in later sections use as subroutines.
- ▷ **Multi-armed bandits.** Section 3 studies the simplest setting, with finitely many decisions and no contexts. It shows that the greedy strategy fails, and introduces the principle of optimism through the UCB algorithm.
- ▷ **Contextual bandits.** Section 4 adds contexts and function approximation. It shows that optimism can fail for general function classes, and introduces the Estimation-to-Decisions principle through SquareCB, which reduces contextual bandits to supervised regression.
- ▷ **Structured bandits.** Section 5 adds structure across decisions, as in linear bandits. It introduces the Decision-Estimation Coefficient, which determines how much exploration a class of problems requires, and the E2D algorithm, which achieves it.

- ▷ **Tabular reinforcement learning.** [Section 6](#) adds multi-step decisions and credit assignment, in the “tabular” (finite state/action) setting. It gives the UCB-VI algorithm, based on optimism, and a first Estimation-to-Decisions algorithm for reinforcement learning.
- ▷ **General decision making.** [Section 7](#) brings the preceding sections together. It introduces a general decision making framework that contains every setting above, and extends the Decision-Estimation Coefficient and E2D to it, so that a single algorithm and a single complexity measure cover all of them.
- ▷ **Lower bounds.** [Section 8](#) is the punchline of this development. It shows that the Decision-Estimation Coefficient also *lower bounds* the regret of any algorithm—whether or not it follows the Estimation-to-Decisions principle—quantifying the limit of sample-efficient learning. The upper bounds of [Section 7](#) are therefore optimal up to the cost of estimation.
- ▷ **Reinforcement learning with function approximation.** [Section 9](#) takes the theory to its current frontier. It shows that realizability alone is not enough for sample-efficient reinforcement learning, and gives algorithms for low-rank MDPs and for MDPs with low Bellman rank, and surveys the landscape of what is known to be tractable and what is not.

Subsections marked with (★) contain optional material that can be skipped on a first reading. Readers interested in contextual bandits as deployed in practice need only [Sections 2](#) to [4](#).

What we do not cover. Our focus on the statistical complexity of exploration, and on the algorithms that address it, leaves out several topics that are important in their own right. On the statistical side, we do not cover offline reinforcement learning, in which the learner receives a fixed dataset and never interacts [\[143\]](#), or imitation learning, in which it learns from demonstrations [\[168, 179, 91\]](#); the latter in particular is of great practical importance. We also do not cover active learning or experimental design, in which the learner chooses what data to collect but earns no reward along the way, nor off-policy evaluation or non-stationary environments, all of which matter in deployment. We treat adversarial environments and Bayesian methods only briefly, and, beyond passing examples, we do not cover control or other topics such as partial monitoring. Finally, while we cover many efficient and practically successful algorithms, we do not systematically address computational complexity in reinforcement learning, a fascinating topic in its own right [\[99, 178\]](#).

On the practical side, we do not cover policy gradient methods [\[211, 195, 184, 185\]](#), which are the workhorse of deep reinforcement learning and widely used in practice (see Agarwal et al. [\[9\]](#) for their theory). Notably, existing guarantees for these methods require the data distribution to already cover the relevant states, and little is known about exploration with policy gradient methods beyond tabular and linear settings [\[44, 7\]](#). A related emerging topic we do not cover is the role of reinforcement learning in foundation models, where one starts from a powerful pre-trained model rather than from scratch [\[56, 192, 163, 28, 170, 216, 106, 107, 93, 54\]](#).

For the broader picture, we recommend Agarwal et al. [\[10\]](#) for theoretical guarantees for common reinforcement learning recipes; see also Lattimore and Szepesvári [\[139\]](#) and Slivkins [\[191\]](#) for bandits, and Cesa-Bianchi and Lugosi [\[47\]](#) and Hazan [\[102\]](#) for online learning.

1.4 The Interplay Between Statistical Complexity and Algorithm Design

Two things remain before we begin. The first is to say precisely what we mean by *statistical complexity*, since the term motivates the questions asked in every later section. The second is to expand a claim above: that in interactive decision making, the question of how hard a problem is

cannot be separated from the question of how to solve it. The points are related, and together they explain why a unified treatment is possible at all.

We begin with the definition. We take a *minimax* point of view. Let $\mathcal{M}$ be a set of possible models, or environments, that the learner may encounter; $\mathcal{M}$ represents the learner’s prior knowledge about the environment. Let Alg denote an algorithm, and $\text{Perf}_T(\text{Alg}, M)$ a measure of its performance on the model $M \in \mathcal{M}$ after T rounds of interaction (or, in passive learning, after observing T data points). We want algorithms that perform well for any model in $\mathcal{M}$, in the sense that Alg approximately solves the minimax problem

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

Understanding the *statistical complexity* of a 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). Notice what this definition demands. An upper bound requires exhibiting an algorithm; a lower bound requires reasoning about every algorithm at once. Whether these two tasks can be separated is the question we turn to next.

In statistical learning, to a first approximation, there is only a single algorithm. Empirical risk minimization, which fits the model that best explains the data, is the default for nearly every problem, and theory is largely concerned with understanding how well it does: how the performance is determined by the richness of the model class and data. Thus, the question of what makes a problem hard is largely separate from the question of how to solve it. Interactive decision making is different. No single exploration strategy is optimal across problems; the right one depends on the structure of the model class, often in subtle ways; and the analysis that reveals what makes a problem hard is the same analysis that produces the algorithm for it. The central perspective of this monograph is that statistical complexity and algorithm design must be studied together. A theory of statistical complexity for decision making therefore cannot avoid the question: for a given class $\mathcal{M}$, which algorithms solve (1.1)?

For many years, the answer was given one class at a time: pick a model class $\mathcal{M}$ (linear, generalized linear, and so on), design an algorithm for it, and prove a guarantee. General techniques pointing toward a deeper understanding existed (the eluder dimension, information-directed sampling, Bellman rank), but an answer for *any* class $\mathcal{M}$ did not. Part of the difficulty is that the minimax problem (1.1) ranges over the space of all algorithms, which is far too large to search directly.

The key insight, used throughout the monograph, is that for the general problem of Figure 1 we can restrict attention to algorithms that interleave *online estimation*, a form of supervised learning described in Sections 2.1 and 2.2, with an *exploration strategy* that decides how to balance exploration against exploitation. This reduction is the *Estimation-to-Decisions* principle. As we show, algorithms built this way achieve or nearly achieve optimality in (1.1) for a surprisingly wide range of problems. The performance of this method is quantified by the *Decision-Estimation Coefficient* (DEC), a central object that provides a link between complexity and algorithm design: the same quantity both (1) bounds the minimax value from below and (2) tells the algorithm how much to explore. This is what allows the theory we develop to treat every class at once.

2. BACKGROUND: STATISTICAL AND ONLINE LEARNING

This section reviews the two non-interactive learning problems on which the rest of the monograph builds: statistical learning, in which the learner receives an i.i.d. dataset and aims to generalize to unseen data, and online learning, in which the data arrive sequentially and may be chosen adversarially. In the language of [Figure 2](#), these are the problems obtained by switching off interactivity. Our interest in these problems is not just for their own sake, but in the *estimation guarantees* they provide. Every decision making algorithm in later sections reduces part of its problem to supervised or online estimation, and inherits its sample complexity from the results stated here.

Organization. [Section 2.1](#) introduces statistical learning and empirical risk minimization. [Section 2.2](#) introduces online learning and regret, and shows that regret controls estimation error for realizable sequential estimation problems; this is the form in which later sections use online learning. [Section 2.3](#) presents algorithms that attain these guarantees, chiefly exponential weights, and [Section 2.4](#) collects the concentration inequalities behind the proofs. Readers familiar with the material can skim [Sections 2.1](#) and [2.2](#) for the main statements and return to the rest as needed.

2.1 Statistical Learning and Generalization

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 it captures the challenge of *generalization and function approximation* in the context of [Figure 2](#).

In the statistical learning problem, we receive examples

$$(x^1, y^1), \dots, (x^T, y^T) \in \mathcal{X} \times \mathcal{Y},$$

i.i.d. from an unknown distribution $M^* = M_x^* \times M_{y|x}^*$, where M_x^* is the marginal distribution of the x variable and $M_{y|x}^*$ is the conditional distribution of y given x . 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 the data $(x^1, y^1), \dots, (x^T, y^T)$, the goal is to produce a model (or, estimator) $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}^{\text{pred}}$ that will do a good job *predicting* outcomes from features for future examples (x, y) drawn from M^* . As discussed below, there are a number of settings where the outcome space $\mathcal{Y}$ is different from the prediction space $\mathcal{Y}^{\text{pred}}$, and thus we give them a separate notation.

To measure prediction performance, we decide on a *loss function* $\ell : \mathcal{Y}^{\text{pred}} \times \mathcal{Y} \rightarrow \mathbb{R}$. Standard examples include:

▷ Regression, where $\mathcal{Y} = \mathcal{Y}^{\text{pred}} = \mathbb{R}$ and common losses include the *square loss*

$$\ell_{\text{sq}}(a, b) = (a - b)^2.$$

▷ Classification, where $\mathcal{Y} = \mathcal{Y}^{\text{pred}} = \{-1, 1\}$ and we may consider the indicator (or 0-1) loss

$$\ell_{01}(a, b) = \mathbb{I}\{a \neq b\}.$$

Alternatively, if $\mathcal{Y} = \{-1, 1\}$ but $\mathcal{Y}^{\text{pred}} = \mathbb{R}$ (i.e. prediction is a real-valued “score”; large means “confident”), we may consider the logistic loss

$$\ell_{\text{logistic}}(a, b) = \log(1 + \exp\{-ab\}).$$

▷ Conditional density estimation with the *logarithmic loss* (log loss). Here $\mathcal{Y}^{\text{pred}} = \Delta(\mathcal{Y})$, the set of distributions on $\mathcal{Y}$, and for $p \in \mathcal{Y}^{\text{pred}}$,

$$\ell_{\log}(p, y) = -\log p(y). \quad (2.1)$$

The choice of the loss function is typically far from unique, and may lead to different interpretations of the results and different computational properties of the ensuing algorithms. For instance, we rarely optimize a model to fit the indicator loss directly because of its discontinuous behavior, and instead opt for its “softmax” relaxation, the logistic loss.

For a function $f : \mathcal{X} \rightarrow \mathcal{Y}^{\text{pred}}$, we measure the prediction performance via the *population* (or, “test”, “out of sample”) loss

$$\mathbf{L}(f) := \mathbb{E}_{(x,y) \sim M^*} [\ell(f(x), y)]. \quad (2.2)$$

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:

$$\hat{f}(\cdot; \mathcal{H}^T) : \mathcal{X} \rightarrow \mathcal{Y}^{\text{pred}}. \quad (2.3)$$

A classical goal in designing algorithms is to ensure that $\mathbb{E}[\mathbf{L}(\hat{f})]$ 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

$$\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathcal{Y}^{\text{pred}}\}$$

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. For language modeling, recent practice has shown that autoregressive models (such as the Transformer architecture) effectively encode temporal structure through attention mechanisms.

The setting of conditional density estimation may be less familiar to the reader, so let us elaborate. In this setting, the learner is required to compute, for each x , a distribution rather than a point estimate (see [Figure 3](#) for a cartoon illustration). We write this conditional distribution as $f(\cdot | x)$ or $f(x)$ interchangeably. For example, given an x containing radar data, a probabilistic forecast could be: probability of “rain” is 30% and “sunshine” 70%; then, for a given realized outcome “rain”, the assessed quality of the probabilistic prediction is the negative logarithm of the mass placed on that outcome, i.e. $-\log 0.3$.

In the context of language modeling, $\mathcal{X} = \mathcal{Y} = \Sigma^*$, the set of sequences of tokens in a dictionary Σ , and $\mathcal{Y}^{\text{pred}} = \Delta(\Sigma^*)$. We may then view $f(\cdot | x)$ as a generative model of (say) an answer y to a question x .

The aforementioned loss functions induce (or suggest) corresponding notions of “discrepancy” between functions in $\mathcal{F}$. Since each function maps $\mathcal{X}$ to $\mathcal{Y}^{\text{pred}}$, we can measure the discrepancy between $f, g \in \mathcal{F}$ by comparing the predictions of $f(x)$ and $g(x)$ for each $x \in \mathcal{X}$ and aggregating; it is natural to average these differences over x with respect to the marginal distribution

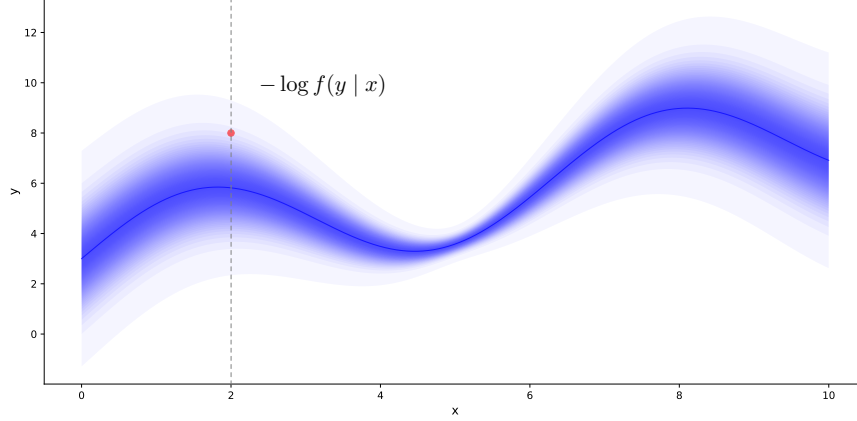


Figure 3: Conditional density estimation. Each $f(x)$ is a distribution $f(\cdot | x)$ over outcomes, and the loss on (x, y) is $-\log f(y | x)$.

M_x^* on $\mathcal{X}$. For instance, for regression problems, we can measure the (squared) distance between $f, g \in \mathcal{F}$ as $\mathbb{E}_x[(f(x) - g(x))^2]$. For classification, the corresponding expected indicator becomes $\mathbb{P}(f(x) \neq g(x))$, and for conditional density estimation, logarithmic loss can be related to the expectation over $x \sim M_x^*$ of the Kullback-Leibler divergence

$$D_{\text{KL}}(f(\cdot | x) \| g(\cdot | x)) = \mathbb{E}_{y \sim f(\cdot | x)}[\log \{f(y | x)/g(y | x)\}]$$

or the squared Hellinger distance

$$D_{\text{H}}^2(f(\cdot | x), g(\cdot | x)) = 2 \mathbb{E}_{y \sim g(\cdot | x)} \left[1 - \sqrt{\frac{f(y | x)}{g(y | x)}} \right].$$

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

$$\widehat{\mathcal{L}}(f) = \frac{1}{T} \sum_{i=1}^T \ell(f(x^i), y^i). \quad (2.4)$$

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

$$\widehat{f} \in \arg \min_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f). \quad (2.5)$$

In the setting of conditional density estimation, (2.5) is referred to as the Maximum Likelihood Estimate (MLE); Figure 4 illustrates the two cases.

To measure the performance of ERM and other algorithms that use $\mathcal{F}$ as encoding prior belief about a good set of models, we consider *excess loss* (or, regret)

$$\mathcal{E}(f) = \mathcal{L}(f) - \min_{g \in \mathcal{F}} \mathcal{L}(g). \quad (2.6)$$

Intuitively, the quantity $\min_{g \in \mathcal{F}} \mathcal{L}(g)$ in (2.6) captures the best prediction performance any function in $\mathcal{F}$ can achieve, even with the knowledge of the *true distribution* M^* . 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 and other algorithms can ensure that $\mathcal{E}(\widehat{f})$ is small in expectation or with high probability over draw of the dataset $\mathcal{H}^T$.

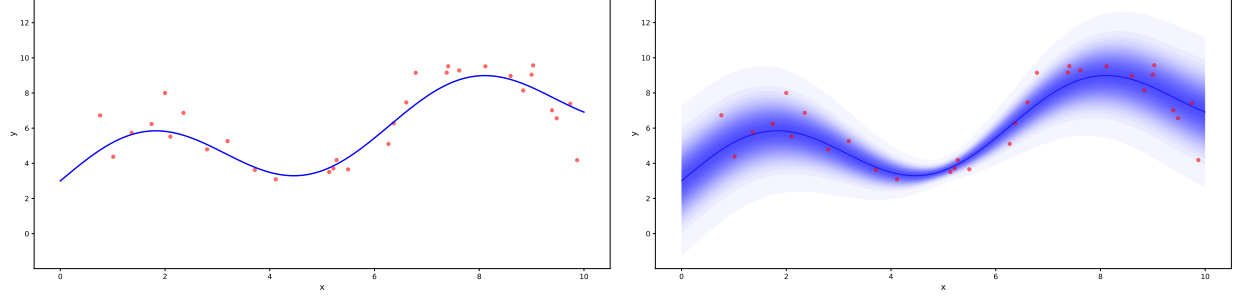


Figure 4: Least squares fit for a regression problem (left) and maximum likelihood estimate for conditional density estimation (right).

Connection to estimation. An appealing feature of the formulation in (2.6) 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 2.1 (Realizability): For regression problems, we say that the problem is *realizable* (or, *well-specified*)^a if the regression function $f^*(x) := \mathbb{E}[y | x]$ is in $\mathcal{F}$; we refer to this condition as *realizability*, and use the same term for the analogous assumption in later sections. For conditional density estimation, we say that the problem is realizable if the conditional distribution $M_{y|x}^*$ is in $\mathcal{F}$.

^aThe term “realizability” comes from statistical and computational learning theory, whereas “well-specified” is the more common term in statistics. This monograph draws heavily on both bodies of literature; we adopt the former term throughout, somewhat arbitrarily, for the sake of consistency.

For the case of square loss, the regression function f^* can also be seen as a minimizer of $L(f)$ over measurable functions f , for the same reason that $\mathbb{E}_z[(z - b)^2]$ is minimized at $b = \mathbb{E}[z]$. Similarly, for the logarithmic loss, $f = M_{y|x}^*$ is a minimizer of $L(f)$ over all conditional distributions coupled with the marginal distribution M_x^* .

Lemma 2.1 (Excess risk under realizability: square loss): For regression, if the problem is *realizable*, then for all $f : \mathcal{X} \rightarrow \mathcal{Y}$,

$$\mathcal{E}(f) = \mathbb{E}_x[(f(x) - f^*(x))^2] \quad (2.7)$$

Proof of Lemma 2.1. Adding and subtracting f^* in the first term of (2.6), we have

$$\mathbb{E}[(f(x) - y)^2] - \mathbb{E}[(f^*(x) - y)^2] = \mathbb{E}[(f(x) - f^*(x))^2] + 2\mathbb{E}[(f^*(x) - y)(f(x) - f^*(x))].$$

□

Inspecting (2.7), we see that any f achieving low excess loss necessarily estimates the true regression function f^* ; hence, the goals of prediction and estimation coincide.

Lemma 2.2 (Excess risk under realizability: log loss): For conditional density estimation, if the problem is *realizable* with $M_{y|x}^* = f^* \in \mathcal{F}$, then for all $f : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$,

$$\mathcal{E}(f) = \mathbb{E}_x[D_{\text{KL}}(f^* \parallel f)]. \quad (2.8)$$

Section 2.2 uses analogous quantities to measure the cumulative estimation error of a sequence of online predictors.

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 2.1 (ERM for finite classes): For any finite class $\mathcal{F}$, empirical risk minimization satisfies

$$\mathbb{E}[\mathcal{E}(\hat{f})] \lesssim \text{comp}(\mathcal{F}, T), \quad (2.9)$$

where

1. For any bounded loss (including classification), $\text{comp}(\mathcal{F}, T) = \sqrt{\frac{\log |\mathcal{F}|}{T}}$.
2. For square loss regression, if the problem is realizable, $\text{comp}(\mathcal{F}, T) = \frac{\log |\mathcal{F}|}{T}$, assuming Y and all predictions $f(x)$, $f \in \mathcal{F}$, lie in a common bounded interval.
3. For conditional density estimation, if the problem is realizable with $M_{y|x}^* = f^* \in \mathcal{F}$, then the maximum likelihood estimate satisfies

$$\mathbb{E}[D_{\text{H}}^2(\hat{f}, f^*)] \lesssim \frac{\log |\mathcal{F}|}{T},$$

where D_{H}^2 is the squared Hellinger distance, integrated over $x \sim M_x^*$.

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

Henceforth, we use the symbol $\lesssim$ to indicate an inequality that holds up to constants, or other problem parameters deemed less important for the present discussion. As an example, the range of losses for the first part of Proposition 2.1 is hidden in this notation, and we only focus on the dependence of the right-hand side on $\mathcal{F}$ and T .

The rate $\text{comp}(\mathcal{F}, T) = \sqrt{\frac{\log |\mathcal{F}|}{T}}$ above is sometimes referred to as a *slow rate*, and is optimal for generic losses. The rate $\text{comp}(\mathcal{F}, T) = \frac{\log |\mathcal{F}|}{T}$ 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 (2.6) more meaningful, since the benchmark will be stronger.

Remark 2.1 (From finite to infinite classes): Throughout this monograph, 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 $\text{comp}(\mathcal{F}, T)$ 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. [37], Boucheron et al. [36], Anthony and Bartlett [19], Shalev-Shwartz and Ben-David [186].

Note that in statistical learning, the most naive strategy, fitting the data as well as possible, is already near-optimal. This is not the case for decision making, as Section 3.2 shows.

2.2 Online Learning and 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  $t = 1, \dots, T$  do
  Compute predictor  $\hat{f}^t : \mathcal{X} \rightarrow \mathcal{Y}^{\text{pred}}$ .
  Observe  $(x^t, y^t) \in \mathcal{X} \times \mathcal{Y}$ .

```

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

$$\mathcal{H}^{t-1} = \{(x^1, y^1), \dots, (x^{t-1}, y^{t-1})\} \quad (2.10)$$

observed so far, the algorithm produces a predictor

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

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

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

for a known loss function $\ell : \mathcal{Y}^{\text{pred}} \times \mathcal{Y} \rightarrow \mathbb{R}$; the cumulative loss can be thought of as a sum of “out-of-sample” prediction errors. Since we place no 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 $\mathcal{F}$:

$$\text{Reg}_{\text{OL}} = \sum_{t=1}^T \ell(\hat{f}^t(x^t), y^t) - \min_{f \in \mathcal{F}} \sum_{t=1}^T \ell(f(x^t), y^t). \quad (2.11)$$

The subscript OL identifies online prediction regret; we reserve **Reg** for the decision-making regret studied in later sections. When specifying the loss is useful, we write $\mathbf{Reg}_{\text{OL}}^{\text{Sq}}$ for square loss and $\mathbf{Reg}_{\text{OL}}^{\text{Log}}$ for log loss. 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.

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. Interactive decision making is a case in point: the data a decision maker collects depend on its own past decisions, and do not look like i.i.d. or simple time-series data.

Remark 2.2 (Proper learning, improper learning, and randomization): The online learning protocol does not require that $\hat{f}^t$ lies in $\mathcal{F}$ ($\hat{f}^t \in \mathcal{F}$). A method that chooses functions from $\mathcal{F}$ is called *proper*, and one that may select predictors outside of $\mathcal{F}$ is called *improper*. It is also useful to allow for *randomized* predictions of the form

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

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

$$\mathbf{Reg}_{\text{OL}} = \sum_{t=1}^T \mathbb{E}_{\hat{f}^t \sim q^t} [\ell(\hat{f}^t(x^t), y^t)] - \min_{f \in \mathcal{F}} \sum_{t=1}^T \ell(f(x^t), y^t). \quad (2.12)$$

The algorithms introduced 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 $\hat{f}^t$ 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 moves away from statistical learning along the data axis: it drops the i.i.d. assumption but retains general-purpose generalization through the class $\mathcal{F}$. It is not an *interactive* problem in the sense of this monograph. The learner's predictions may influence what Nature reveals next, but the learner observes the outcome y^t regardless of what it predicts, so it never faces a tradeoff between predicting well and gathering information. Like statistical learning, online learning is a building block for decision making algorithms in later sections.

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 2.2 (Online-to-batch conversion): Let $(x^1, y^1), \dots, (x^T, y^T)$ be drawn i.i.d. from a distribution M^* , 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

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

we have

$$\mathbb{E}[\mathcal{E}(\hat{f})] \leq \frac{1}{T} \cdot \mathbb{E}[\mathbf{Reg}_{\text{OL}}].$$

Proof of Proposition 2.2. Let $(x, y) \sim M^*$ be a fresh sample which is independent of the history $\mathcal{H}^T$. First, by Jensen's inequality,

$$\mathbb{E}[\mathbf{L}(\hat{f})] = \mathbb{E}\left[\mathbb{E}_{(x,y)}\left[\ell\left(\frac{1}{T} \sum_{t=1}^T \hat{f}^t(x), y\right)\right]\right] \leq \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{(x,y)}\left[\ell\left(\hat{f}^t(x), y\right)\right]\right] \quad (2.13)$$

which is equal to

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{(x^t, y^t)}\left[\ell\left(\hat{f}^t(x^t), y^t\right)\right]\right] \quad (2.14)$$

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

$$\min_{f \in \mathcal{F}} \mathbf{L}(f) = \min_{f \in \mathcal{F}} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \ell(f(x^t), y^t)\right] \geq \mathbb{E}\left[\min_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \ell(f(x^t), y^t)\right] \quad (2.15)$$

□

In light of Proposition 2.2, one can interpret regret as generalizing the notion of excess risk from i.i.d. data to arbitrary sequences. To motivate the algorithms below, we observe that ERM is insufficient to derive this type of guarantee: Repeatedly applying ERM to past observations need not give small regret on arbitrary sequences.

Proposition 2.3 (Failure of ERM in online learning): For binary classification with indicator loss, there is a class of two constant predictors and a fixed sequence of outcomes on which online ERM, with a fixed deterministic tie-breaking rule, incurs regret at least $T/2$.

Proof of Proposition 2.3. Use a single feature value and the predictors $f_0(x) = 0$ and $f_1(x) = 1$. Break ties in favor of f_0 , and take $y^t = 1$ for odd t and $y^t = 0$ for even t . At odd rounds the past counts are tied, so ERM predicts zero; at even rounds ones are in the majority, so it predicts one. Hence ERM makes T mistakes. The better constant makes $\lfloor T/2 \rfloor$ mistakes, giving regret $\lceil T/2 \rceil$. The symmetric construction applies when ties favor f_1 . □

2.2.1 Online Learning and Estimation

Similar to Lemma 2.1 in the setting of statistical learning, the regret for online learning has an additional interpretation in terms of *estimation* if the problem is realizable. We measure this through cumulative estimation errors, which play a central role in the design of decision making algorithms from Section 4 onward.

Lemma 2.3 (Regret controls estimation error: square loss): Suppose that the features $x^1, \dots, x^T$ are generated in an arbitrary fashion, but that for all t , the variable y^t is random with mean given by a fixed function $f^* \in \mathcal{F}$:

$$\mathbb{E}[y^t \mid \mathcal{H}^{t-1}, x^t, \hat{f}^t] = f^*(x^t).$$

Define the cumulative squared estimation error by

$$\mathbf{Est}_{\text{Sq}} := \sum_{t=1}^T (\hat{f}^t(x^t) - f^*(x^t))^2. \quad (2.16)$$

Then for prediction with square loss,

$$\mathbb{E}[\mathbf{Est}_{\text{Sq}}] \leq \mathbb{E}[\mathbf{Reg}_{\text{OL}}^{\text{Sq}}].$$

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.

The analogous result for log loss relates regret to KL divergence.

Lemma 2.4 (Regret controls estimation error: log loss): Suppose that the features $x^1, \dots, x^T$ are generated in an arbitrary fashion, but that for a fixed $f^* \in \mathcal{F}$ and every t , the conditional distribution of y^t given the past $\mathcal{H}^{t-1}$, x^t , and the current predictor $\hat{f}^t$ is $f^*(\cdot \mid x^t)$. Assume the log losses appearing in the regret and comparison below are integrable. Define the cumulative KL estimation error by

$$\mathbf{Est}_{\text{KL}} := \sum_{t=1}^T D_{\text{KL}}(f^*(\cdot \mid x^t) \parallel \hat{f}^t(\cdot \mid x^t)). \quad (2.17)$$

Then for prediction with log loss, any online predictor satisfies

$$\mathbb{E}[\mathbf{Est}_{\text{KL}}] \leq \mathbb{E}[\mathbf{Reg}_{\text{OL}}^{\text{Log}}].$$

Proof of Lemma 2.4. Comparing with the fixed predictor f^* in the definition of regret gives

$$\mathbf{Reg}_{\text{OL}}^{\text{Log}} \geq \sum_{t=1}^T \log \frac{f^*(y^t \mid x^t)}{\hat{f}^t(y^t \mid x^t)}.$$

Conditional on $\mathcal{H}^{t-1}$, x^t , and $\hat{f}^t$, the expected t th summand is precisely $D_{\text{KL}}(f^*(\cdot | x^t) \| \hat{f}^t(\cdot | x^t))$. Taking expectations and summing gives $\mathbb{E}[\mathbf{Est}_{\text{KL}}] \leq \mathbb{E}[\mathbf{Reg}_{\text{OL}}^{\text{Log}}]$. $\square$

[Lemma 2.3](#) and [Lemma 2.4](#) are the form in which later sections use online learning. In particular, we design decision making algorithms from [Section 4](#) onward by calling an online learner as a subroutine, in a modular fashion that depends on the specific choice of online learner only through its cumulative estimation error. In what follows, we give concrete examples of online learning algorithms; subsequent sections will abstract such algorithms away as *estimation oracles* ([Section 4.5](#)).

2.3 Online Learning Algorithms

We now state our first online learning algorithm, the *Bayes mixture predictor*, and prove its regret bound for conditional density estimation with log loss. This is arguably the simplest result one can imagine in online learning.

Lemma 2.5 (Regret of the Bayes mixture predictor): Let $\mathcal{F}$ be a finite class of conditional distributions. At time t , use the prediction distribution

$$\hat{f}^t(y | x) = \sum_{f \in \mathcal{F}} q^t(f) f(y | x), \quad q^{t+1}(f) = \frac{q^t(f) f(y^t | x^t)}{\hat{f}^t(y^t | x^t)},$$

with q^1 uniform on $\mathcal{F}$. For any sequence, this algorithm satisfies

$$\mathbf{Reg}_{\text{OL}}^{\text{Log}} = - \sum_{t=1}^T \log \hat{f}^t(y^t | x^t) - \min_{f \in \mathcal{F}} \left\{ - \sum_{t=1}^T \log f(y^t | x^t) \right\} \leq \log |\mathcal{F}|. \quad (2.18)$$

Proof of Lemma 2.5. Write $Z_t = \sum_{f \in \mathcal{F}} \prod_{i=1}^t f(y^i | x^i)$, so $Z_0 = |\mathcal{F}|$. By definition, $\hat{f}^t(y^t | x^t) = Z_t / Z_{t-1}$. Hence, for every $f \in \mathcal{F}$,

$$\prod_{t=1}^T \hat{f}^t(y^t | x^t) = \prod_{t=1}^T \frac{Z_t}{Z_{t-1}} = \frac{Z_T}{|\mathcal{F}|} \geq \frac{1}{|\mathcal{F}|} \prod_{t=1}^T f(y^t | x^t).$$

Taking negative logarithms and minimizing over $f \in \mathcal{F}$ proves the claim. $\square$

In the realizable case (conditioning on the past as well as the current features), [Lemma 2.4](#) and (2.18) imply

$$\mathbb{E}[\mathbf{Est}_{\text{KL}}] \leq \mathbb{E}[\mathbf{Reg}_{\text{OL}}^{\text{Log}}] \leq \log |\mathcal{F}|. \quad (2.19)$$

In particular, this KL guarantee allows the features x^t to be chosen adaptively and arbitrarily. On the other hand, when the data are i.i.d. and the problem is realizable, the averaged predictor $\bar{f}(\cdot | x) = \frac{1}{T} \sum_{t=1}^T \hat{f}^t(\cdot | x)$ satisfies, in view of [Proposition 2.2](#),

$$\mathbb{E}[\mathbb{E}_{x \sim M_x^*} [D_{\text{KL}}(f^*(\cdot | x) \| \bar{f}(\cdot | x))]] \leq \frac{\log |\mathcal{F}|}{T},$$

where x is a fresh draw independent of the training data.

The average predictor $\bar{f}(\cdot | x)$ also enjoys the $\log |\mathcal{F}|/T$ expected excess log-loss bound in [Corollary 2.1](#), without requiring the problem to be realizable.

Corollary 2.1 (Excess log loss without realizability): Let $\mathcal{F}$ be a finite, nonempty class of conditional densities and let $(x^t, y^t)_{t=1}^T$ be i.i.d. Let $\hat{f}^t$ be the Bayes mixture predictor of Lemma 2.5 and set $\bar{f}(\cdot | x) = T^{-1} \sum_{t=1}^T \hat{f}^t(\cdot | x)$. For log loss, assuming the relevant log losses are integrable,

$$\mathbb{E}[\mathbb{L}(\bar{f})] - \min_{f \in \mathcal{F}} \mathbb{L}(f) \leq \frac{\log |\mathcal{F}|}{T}.$$

The data-generating conditional distribution need not belong to $\mathcal{F}$.

Proof of Corollary 2.1. For a fresh pair (x, y) , convexity of $-\log$ gives

$$-\log \bar{f}(y | x) \leq \frac{1}{T} \sum_{t=1}^T -\log \hat{f}^t(y | x).$$

Since $\hat{f}^t$ depends only on the first $t - 1$ observations, its expected loss on a fresh pair equals its expected loss on (x^t, y^t) . Fix a minimizer f_0 of $\mathbb{L}$ over $\mathcal{F}$. The regret bound therefore implies

$$\mathbb{E}[\mathbb{L}(\bar{f})] - \mathbb{L}(f_0) \leq \frac{1}{T} \mathbb{E} \left[\sum_{t=1}^T \log \frac{f_0(y^t | x^t)}{\hat{f}^t(y^t | x^t)} \right] \leq \frac{\log |\mathcal{F}|}{T}.$$

□

The simple strategy in Lemma 2.5 can be extended to other losses, as we show next.

The exponential weights algorithm. Arguably the most important online learning algorithm is *Exponential Weights*, which generalizes the Bayes mixture predictor from log loss to arbitrary losses. The posterior weights $q^t(f) \propto \prod_{i < t} f(y^i | x^i)$ of the Bayes mixture predictor are replaced by weights that decay exponentially in the cumulative loss of each f . At each time t , the algorithm computes a distribution $q^t \in \Delta(\mathcal{F})$ via

$$q^t(f) \propto \exp \left\{ -\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i) \right\}, \quad (2.20)$$

where $\eta > 0$ is a learning rate. Based on q^t , the algorithm forms the prediction $\hat{f}^t$ as specified below; for reasons explained in a moment, we give two variants of the method.

Algorithm 2.1 Exp. Weights (averaged)	Algorithm 2.2 Exp. Weights (randomized)
input: Learning rate $\eta > 0$; finite class $\mathcal{F}$. for $t = 1, \dots, T$ do Compute q^t in (2.20). Let $\hat{f}^t = \mathbb{E}_{f \sim q^t}[f]$. Observe (x^t, y^t) , incur $\ell(\hat{f}^t(x^t), y^t)$.	input: Learning rate $\eta > 0$; finite class $\mathcal{F}$. for $t = 1, \dots, T$ do Compute q^t in (2.20). Sample $\hat{f}^t \sim q^t$. Observe (x^t, y^t) , incur $\ell(\hat{f}^t(x^t), y^t)$.

The only difference between these two forms is whether we compute the prediction $\hat{f}^t$ from q^t via

$$\hat{f}^t = \mathbb{E}_{f \sim q^t}[f], \quad \text{or} \quad \hat{f}^t \sim q^t. \quad (2.21)$$

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 $\mathcal{Y}^{\text{pred}}$ 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 (2.12)). The averaged version can therefore take advantage of the structure of the loss function, such as strong convexity, leading to faster rates. For log loss and $\eta = 1$, the averaged variant is exactly the Bayes mixture predictor of [Lemma 2.5](#).

The following result shows that Exponential Weights leads to regret bounds for online learning, with rates that parallel those in [Proposition 2.1](#).

Proposition 2.4 (Exponential weights for finite classes): For any finite class $\mathcal{F}$, the Exponential Weights algorithm (with appropriate choice of η) satisfies

$$\frac{1}{T} \text{Reg}_{\text{OL}} \lesssim \text{comp}(\mathcal{F}, T) \quad (2.22)$$

for any sequence, where:

1. For arbitrary bounded losses (including classification), $\text{comp}(\mathcal{F}, T) = \sqrt{\frac{\log |\mathcal{F}|}{T}}$. This is achieved by the randomized variant.
2. For regression with the square loss (under boundedness of $\mathcal{Y}$) and conditional density estimation with the log loss, $\text{comp}(\mathcal{F}, T) = \frac{\log |\mathcal{F}|}{T}$. This is achieved by the averaged variant.

The losses are those introduced in [Section 2.1](#). The bound holds for every sequence: deterministically for the averaged variant, and in expectation over the algorithm’s own randomization for the randomized variant, in the sense of (2.12). As in [Proposition 2.1](#), the bound for bounded losses hides the range of the loss, which enters linearly; the fast rates take $\mathcal{Y} = \mathcal{Y}^{\text{pred}} = [0, 1]$.

Remark 2.3 (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 2.4](#) to infinite classes. In general, the optimal regret for a class $\mathcal{F}$ will reflect the statistical capacity of the class [\[171\]](#).

Remark 2.4 (Online linear optimization): For the slow rate in [Proposition 2.4](#), 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 $\ell_f^t = \ell(f(x^t), y^t)$ and $\ell^t = (\ell_{f_1}^t, \dots, \ell_{f_N}^t) \in \mathbb{R}^N$ for $\mathcal{F} = \{f_1, \dots, f_N\}$. For a randomized strategy $q^t \in \Delta([N])$, expected loss of the learner can be written as

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

and the expected regret can be written as

$$\mathbf{Reg}_{\text{OL}} = \sum_{t=1}^T \langle q^t, \ell^t \rangle - \min_{j \in \{1, \dots, N\}} \sum_{t=1}^T \langle e_j, \ell^t \rangle \quad (2.23)$$

where $e_j \in \mathbb{R}^N$ is the standard basis vector with 1 in j th position. In its most general form, the exponential weights algorithm gives bounds on the regret in (2.23) for any sequence of vectors $\ell^1, \dots, \ell^T$, and the update takes the form

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

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*.

To summarize, exponential weights attains, for arbitrary sequences, the same rates that ERM attains for i.i.d. data, and these rates cannot be improved in general for finite classes. ERM itself, applied to the data observed so far, does not have low regret in general (Proposition 2.3). For the purposes of this monograph, this is all we need from online learning over finite classes.

2.4 Additional Background: Random Variables and Tail Bounds

The main tools for proving Proposition 2.1 and similar generalization bounds are *concentration inequalities* (or, tail bounds) for random variables, which we review here with further background in Appendix A.1.

Definition 2.2 (Sub-Gaussian random variable): A random variable Z is sub-Gaussian with variance factor (or variance proxy) σ^2 if

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

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 2.6 (Sub-Gaussian tail bound): If $Z_1, \dots, Z_T$ are i.i.d. sub-Gaussian random variables with variance proxy σ^2 , then

$$\mathbb{P} \left(\frac{1}{T} \sum_{i=1}^T Z_i - \mathbb{E}[Z] \geq u \right) \leq \exp \left\{ -\frac{Tu^2}{2\sigma^2} \right\}. \quad (2.24)$$

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

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{i=1}^T Z_i - \mathbb{E}[Z]\right| \geq u\right) \leq 2 \exp\left\{-\frac{Tu^2}{2\sigma^2}\right\}. \quad (2.25)$$

Setting the right-hand side of (2.25) equal to δ and solving for u , we find that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|\frac{1}{T}\sum_{i=1}^T Z_i - \mathbb{E}[Z]\right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{T}}. \quad (2.26)$$

Remark 2.5 (Union bound): Observe that the factor 2 under the logarithm in (2.26) is the result of applying union bound to (2.24). Throughout this monograph, we frequently apply the union bound to multiple—say N —high probability events involving sub-Gaussian random variables. In this case, the union bound results in terms of the form $\log(N/\delta)$. The mild logarithmic dependence is due to the sub-Gaussian tail behavior of the averages.

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

Lemma 2.7 (Hoeffding’s inequality): Any random variable Z taking values in $[a, b]$ is sub-Gaussian with variance proxy $(b - a)^2/4$, i.e.

$$\forall \eta \in \mathbb{R}, \quad \ln \mathbb{E}[\exp\{\eta(Z - \mathbb{E}[Z])\}] \leq \frac{\eta^2(b - a)^2}{8}. \quad (2.27)$$

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

$$\frac{1}{T}\sum_{i=1}^T Z_i - \mathbb{E}[Z] \leq (b - a)\sqrt{\frac{\log(1/\delta)}{2T}}. \quad (2.28)$$

Lemma 2.8 (Bernstein’s inequality): Let $Z_1, \dots, Z_T, Z$ be i.i.d. with variance v^2 and range $|Z - \mathbb{E}[Z]| \leq B$ almost surely. Then with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{i=1}^T Z_i - \mathbb{E}[Z] \leq v\sqrt{\frac{2\log(1/\delta)}{T}} + \frac{B\log(1/\delta)}{3T}. \quad (2.29)$$

A proof of Lemma 2.8 is given in Appendix A.1.1. We refer the reader to Appendix A.1 for further background on tail bounds. In particular, Lemma A.2 in Appendix A.1 is a common starting point for many inequalities in this monograph, including Hoeffding.

2.5 Bibliographic Remarks and Further Reading

Takeaways. This section reviewed the two non-interactive learning problems on which the rest of the monograph builds. Over a finite class, empirical risk minimization attains the slow rate $\sqrt{\log|\mathcal{F}|/T}$ for any bounded loss and the fast rate $\log|\mathcal{F}|/T$ under realizability for square loss and for maximum likelihood in Hellinger distance (Proposition 2.1 and Proposition 2.5); exponential weights attains the same two rates for arbitrary sequences (Proposition 2.4). Later sections use these estimation guarantees as a starting point to design interactive decision making algorithms. Notably, the guarantees we will study in later sections do not depend on the finite-class assumption: they depend in a modular fashion on the estimation error of the online learner, so any estimator with a comparable guarantee can be substituted.

Bibliographic remarks. The statistical learning framework and empirical risk minimization go back to Vapnik and Chervonenkis [205], Vapnik [204]; the union-bound argument of Lemma 2.9 is standard. That ERM cannot attain the fast rate in Proposition 2.1 without realizability, whereas aggregation can, is due to Tsybakov [199], Juditsky et al. [118], Lecué and Mendelson [142]. Hellinger rates for maximum likelihood are classical [212, 34, 202]; the proof of Proposition 2.5 follows Zhang [225]. Prediction of individual sequences originates with Hannan [100] and Cover [58], the latter containing the observation behind Proposition 2.3. The Bayes mixture predictor of Lemma 2.5 takes its name from universal coding [187], and Corollary 2.1 is the progressive-mixture method of Yang and Barron [219]. Exponential weights (Algorithm 2.1 and Algorithm 2.2) and the slow rate in Proposition 2.4 are due to Littlestone and Warmuth [146], Vovk [206], Freund and Schapire [94]; the fast rates for square and log loss, and mixability more generally (Remark 2.6), are due to Vovk [206, 207]. Online-to-batch conversion (Proposition 2.2) in high-probability form is due to Cesa-Bianchi et al. [48].

Further reading. Textbook treatments of statistical learning include Vapnik [204], Devroye et al. [64], Shalev-Shwartz and Ben-David [186]; for fast rates beyond finite classes see Bartlett et al. [29], Koltchinskii [129]. Cesa-Bianchi and Lugosi [47] is the standard reference for online learning, and Hazan [102], Orabona [157] develop the online convex optimization viewpoint of Remark 2.4. Regret for infinite classes is governed by sequential complexities developed in Rakhlin and Sridharan [171], Rakhlin et al. [173].

2.6 Deferred Proofs: Statistical Learning

Using Hoeffding's inequality, we now prove Part 1 (the slow rate) of Proposition 2.1.

Lemma 2.9 (Slow rate for ERM (Proposition 2.1, Part 1)): Let $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathcal{Y}^{\text{pred}}\}$ be finite, and assume that $\ell(f(x), y) \in [0, 1]$ for any $f \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}$. Then with probability at least $1 - \delta$, ERM satisfies

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

Proof of Lemma 2.9. For any $f \in \mathcal{F}$, we can write

$$\mathcal{L}(\hat{f}) - \mathcal{L}(f) = \left[\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f}) \right] + \left[\hat{\mathcal{L}}(\hat{f}) - \hat{\mathcal{L}}(f) \right] + \left[\hat{\mathcal{L}}(f) - \mathcal{L}(f) \right]. \quad (2.30)$$

Further, by Hoeffding's inequality,

$$\text{for any } f \in \mathcal{F}, \text{ with prob. at least } 1 - \delta, \quad \left| \mathbb{L}(f) - \widehat{\mathbb{L}}(f) \right| \leq \sqrt{\frac{\log(2/\delta)}{2T}}. \quad (2.31)$$

By union bound, with probability at least $1 - |\mathcal{F}|\delta$, inequality (2.31) holds for any $f \in \mathcal{F}$ simultaneously; thus, for any $\delta \in (0, 1)$,

$$\text{with prob. at least } 1 - \delta, \text{ for any } f \in \mathcal{F}, \quad \left| \mathbb{L}(f) - \widehat{\mathbb{L}}(f) \right| \leq \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2T}}. \quad (2.32)$$

The middle term in (2.30) is nonpositive, and the proof is completed by applying (2.32) to the first and third terms.² $\square$

To deduce the in-expectation bound of Proposition 2.1 from the high-probability tail bound of Lemma 2.9, the standard technique of *integrating out the tail* can be employed. Specifically, for a nonnegative random variable U , it holds that $\mathbb{E}[U] \leq \tau + \int_{\tau}^{\infty} \mathbb{P}(U \geq z) dz$ for all $\tau > 0$; choosing $\tau \propto T^{-1/2}$ concludes the proof.³

Proof of Proposition 2.1, Part 2. By rescaling the common bounded interval, assume that Y and every prediction $f(x)$ lie in $[0, 1]$. For each $f \in \mathcal{F}$, write

$$Z_i(f) = (f(x^i) - y^i)^2 - (f^*(x^i) - y^i)^2, \quad R(f) = \mathbb{E}[(f(x) - f^*(x))^2].$$

Realizability gives $\mathbb{E}[Z_i(f)] = R(f) = \mathcal{E}(f)$. Factoring the difference of squares also gives

$$Z_i(f)^2 \leq 4(f(x^i) - f^*(x^i))^2, \quad |Z_i(f) - \mathbb{E}[Z_i(f)]| \leq 2.$$

Thus $\text{Var}(Z_i(f)) \leq 4R(f)$. Apply Lemma 2.8 to $-Z_i(f)$ and take a union bound over $\mathcal{F}$. With $u = \log(|\mathcal{F}|/\delta)$, with probability at least $1 - \delta$, simultaneously for all $f \in \mathcal{F}$,

$$R(f) \leq \frac{1}{T} \sum_{i=1}^T Z_i(f) + \sqrt{\frac{8R(f)u}{T}} + \frac{2u}{3T} \leq \frac{1}{T} \sum_{i=1}^T Z_i(f) + \frac{R(f)}{2} + \frac{14u}{3T}.$$

The last inequality uses $\sqrt{8rv} \leq r/2 + 4v$ for $r, v \geq 0$. Since the empirical minimizer satisfies $\sum_{i=1}^T Z_i(\widehat{f}) \leq 0$, it follows that

$$\mathcal{E}(\widehat{f}) \leq \frac{28 \log(|\mathcal{F}|/\delta)}{3T}. \quad (2.33)$$

Integrating this tail bound yields $\mathbb{E}[\mathcal{E}(\widehat{f})] \leq 28(\log |\mathcal{F}| + 1)/(3T)$. For $|\mathcal{F}| \geq 2$ this is $O(\log |\mathcal{F}|/T)$; for $|\mathcal{F}| = 1$ the excess risk is zero. Rescaling back multiplies the bound by the squared length of the original interval. $\square$

We finish this section with a proof of the Part 3 (the fast rate for conditional density estimation) of Proposition 2.1.

²It is important to understand the difference between (2.31) and (2.32). In the former, the function f cannot be chosen to in a data-dependent manner; in the latter, the high probability event holds for all $f \in \mathcal{F}$, and thus also holds for $\widehat{f}$.

³For the in-expectation bound, only the first term in (2.30) requires this argument: for a fixed f , the third term has mean zero.

Proposition 2.5 (Fast rate for MLE (Proposition 2.1, Part 3)): Let $\mathcal{F}$ be a finite class of conditional densities $f : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$, and M_x^* be a marginal distribution on $\mathcal{X}$. Suppose $M_{y|x}^* = f^*$ for some $f^* \in \mathcal{F}$, and the data $x^1, \dots, x^T$ are i.i.d. from M_x^* and $y^t \sim f^*(\cdot | x^t)$, $t = 1, \dots, T$, independently. Let the MLE be

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T -\log f(y^t | x^t).$$

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

$$D_H^2(\hat{f}, f^*) \leq \frac{2 \log(|\mathcal{F}|/\delta)}{T}$$

where $D_H^2(f, g) = \mathbb{E}_{x \sim M_x^*} [D_H^2(f(\cdot | x), g(\cdot | x))]$ is the integrated squared Hellinger distance.

Proof of Proposition 2.5. By definition of MLE,

$$\sum_{t=1}^T \log \hat{f}(y^t | x^t) \geq \sum_{t=1}^T \log f^*(y^t | x^t). \quad (2.34)$$

Fix $f \in \mathcal{F}$ and apply Lemma A.2 to independent random variables

$$X_t = \frac{1}{2} (\log f(y^t | x^t) - \log f^*(y^t | x^t)).$$

This yields that with probability at least $1 - \delta$,

$$\sum_{t=1}^T \frac{1}{2} (\log f(y^t | x^t) - \log f^*(y^t | x^t)) \leq \sum_{t=1}^T \log \mathbb{E}_{(x^t, y^t)} \left[\exp \left\{ \frac{1}{2} \log \frac{f(y^t | x^t)}{f^*(y^t | x^t)} \right\} \right] + \log(1/\delta).$$

Rearranging, taking a union bound over $f \in \mathcal{F}$, applying to $\hat{f}$ and using (2.34), we have that with probability at least $1 - \delta$,

$$-T \log \mathbb{E}_{(x, y)} \left[\sqrt{\frac{\hat{f}(y | x)}{f^*(y | x)}} \right] \leq \log(|\mathcal{F}|/\delta). \quad (2.35)$$

Using the inequality $-\log x \geq 1 - x$, for the natural logarithm, and the fact that, conditionally on x , $\mathbb{E}_{y \sim f^*(\cdot | x)} \left[1 - \sqrt{\frac{f(y | x)}{f^*(y | x)}} \right] = \frac{1}{2} D_H^2(f(\cdot | x), f^*(\cdot | x))$, we conclude the proof. $\square$

2.7 Deferred Proofs: Online Learning

We now prove Proposition 2.4. 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 2.4 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 $\eta > 0$:

$$\Phi_\eta^t = -\log \sum_{f \in \mathcal{F}} \exp \left\{ -\eta \sum_{i=1}^t \ell(f(x^i), y^i) \right\}. \quad (2.36)$$

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

Proof of Proposition 2.4. We first prove the second statement, focusing on conditional density with the logarithmic loss; for the square loss, see Remark 2.6 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

$$\hat{f}^t(y^t | x^t) = \sum_{f \in \mathcal{F}} q^t(f) f(y^t | x^t) = \sum_{f \in \mathcal{F}} f(y^t | x^t) \frac{\exp \left\{ -\sum_{i=1}^{t-1} \ell_{\log}(f(x^i), y^i) \right\}}{\sum_{f \in \mathcal{F}} \exp \left\{ -\sum_{i=1}^{t-1} \ell_{\log}(f(x^i), y^i) \right\}}, \quad (2.37)$$

and thus

$$\ell_{\log}(\hat{f}(x^t), y^t) = -\log \hat{f}(y^t | x^t) = \Phi_1^t - \Phi_1^{t-1}. \quad (2.38)$$

Hence, by telescoping

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

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

$$\Phi_1^T \leq -\log \exp \left\{ -\sum_{i=1}^T \ell_{\log}(g(x^i), y^i) \right\} = \sum_{i=1}^T \ell_{\log}(g(x^i), y^i), \quad (2.39)$$

for any $g \in \mathcal{F}$. 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 (2.38), and the cumulative loss of any expert is related to the final potential (2.39). 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 $\ell(f(x), y) \in [0, 1]$ for all $f \in \mathcal{F}$ and (x, y) , as we can always re-scale the problem. The randomized variant of exponential weights (2.21) satisfies

$$\mathbb{E}_{\hat{f}^t \sim q^t} [\ell(\hat{f}^t(x^t), y^t)] = \sum_{f \in \mathcal{F}} \ell(f(x^t), y^t) \frac{\exp \left\{ -\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i) \right\}}{\sum_{f \in \mathcal{F}} \exp \left\{ -\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i) \right\}}. \quad (2.40)$$

Hoeffding's inequality (Lemma 2.7) implies that

$$\eta \mathbb{E}_{\hat{f}^t \sim q^t}[\ell(\hat{f}^t(x^t), y^t)] \leq -\log \sum_{f \in \mathcal{F}} \frac{\exp\{-\eta \ell(f(x^t), y^t)\} \exp\left\{-\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i)\right\}}{\sum_{f \in \mathcal{F}} \exp\left\{-\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i)\right\}} + \frac{\eta^2}{8}. \quad (2.41)$$

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 (2.38). Summing over t , this gives

$$\eta \sum_{t=1}^T \mathbb{E}_{\hat{f}^t \sim q^t}[\ell(\hat{f}^t(x^t), y^t)] \leq \Phi_\eta^T - \Phi_\eta^0 + \frac{T\eta^2}{8}. \quad (2.42)$$

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

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

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

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

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

$$\sum_{t=1}^T \mathbb{E}_{\hat{f}^t \sim q^t}[\ell(\hat{f}^t(x^t), y^t)] - \ell(g(x^t), y^t) \leq \sqrt{\frac{T \log |\mathcal{F}|}{2}}. \quad (2.43)$$

□

Observe that Hoeffding's inequality was all that was needed for Lemma 2.9. Curiously enough, it was also the only nontrivial step in the proof of Proposition 2.4. In fact, the connection between probabilistic inequalities and online learning regret inequalities (that hold for arbitrary sequences) runs much deeper [172, 82].

Remark 2.6 (Mixable losses): We did not provide a proof of Proposition 2.4 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 $\exp\{-(\hat{f}^t(x^t) - y^t)^2\}$ is, up to constants, the desired square loss. However, the mixture in (2.37) 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 $\hat{f}(x) \in \mathcal{Y}^{\text{pred}}$

such that for all $y \in \mathcal{Y}$,

$$\ell(\hat{f}(x), y) \leq -\frac{c}{\eta} \log \left(\sum_{f \in \mathcal{F}} q(f) \exp\{-\eta \ell(f(x), y)\} \right). \quad (2.44)$$

If loss is mixable, then given the exponential weights distribution q^t , the best prediction $\hat{y}^t = \hat{f}^t(x^t)$ can be written (by bringing the right-hand side of (2.44) to the left side) as an optimization problem

$$\arg \min_{\hat{y}^t \in \mathcal{Y}^{\text{pred}}} \max_{y^t \in \mathcal{Y}} \left[\ell(\hat{y}^t, y^t) + \frac{c}{\eta} \log \left(\sum_{f \in \mathcal{F}} q^t(f) \exp\{-\eta \ell(f(x^t), y^t)\} \right) \right] \quad (2.45)$$

which is equivalent to

$$\arg \min_{\hat{y}^t \in \mathcal{Y}^{\text{pred}}} \max_{y^t \in \mathcal{Y}} \left[\ell(\hat{y}^t, y^t) + \frac{c}{\eta} \log \left(\sum_{f \in \mathcal{F}} \exp\{-\eta \sum_{i=1}^t \ell(f(x^i), y^i)\} \right) \right] \quad (2.46)$$

once we remove the normalization factor. With this choice, mixability allows one to replicate the proof of [Proposition 2.4](#) for the logarithmic loss, with the only difference being that (2.37) (after applying $-\log$ to both sides) becomes an inequality. It can be verified that square loss is mixable with parameter $\eta = 1/2$ and $c = 1$ when $\mathcal{Y} = \mathcal{Y}^{\text{pred}} = [-1, +1]$ [\[206\]](#), leading to the desired fast rate for square loss in [Proposition 2.4](#). The idea of translating the English statement “there exists a strategy such that for any outcome...” into a min-max inequality recurs in [Section 5](#), where the Decision-Estimation Coefficient is defined through a min-max problem of this type.

3. MULTI-ARMED BANDITS: THE EXPLORATION–EXPLOITATION TRADEOFF

Section 2 concerned passive learning: data is given, and the learner’s only task is to generalize. This section takes the first step along the interactivity axis of Figure 2. In the *multi-armed bandit* problem, the learner’s decisions determine which rewards are observed, so that collecting informative data and earning high reward become competing goals. This is the *exploration–exploitation tradeoff*, and the multi-armed bandit isolates it in the simplest possible form, with a finite set of decisions and no additional structure to generalize across. The section establishes the optimal $\sqrt{AT}$ regret rate for this problem and, more importantly, introduces an algorithm design principle that recurs throughout the monograph: optimism in the face of uncertainty.

Organization. Section 3.1 formalizes the multi-armed bandit problem and the regret criterion. Section 3.2 shows that the greedy strategy inherited from statistical learning fails, which motivates deliberate exploration. Section 3.3 analyzes the simplest remedy, ϵ -Greedy, which explores uniformly at random. Section 3.4 introduces the principle of optimism and the UCB algorithm, which explores adaptively and attains the optimal regret; its analysis is the template for a variety of optimistic algorithms throughout the monograph. Two optional extensions follow: Section 3.5 (marked $\star$) treats the Bayesian setting and posterior sampling, whose analysis is revisited in Sections 5 and 7; Section 3.6 (marked $\star$) treats adversarial (non-stochastic) rewards and the Exp3 algorithm.

3.1 The Multi-Armed Bandit Problem

The multi-armed bandit problem is defined by the following protocol.

Multi-Armed Bandit Protocol

```

for  $t = 1, \dots, T$  do
  Select decision  $\pi^t \in \Pi := \{1, \dots, A\}$ .
  Observe reward  $r^t \in \mathbb{R}$ .
  
```

The process proceeds in T rounds. At each round $t \in [T]$, the learning agent selects a discrete *decision*⁴ $\pi^t \in \Pi = \{1, \dots, A\}$ using the data

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

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. We allow the learner to randomize the decision at step t according to a distribution $p^t = p^t(\cdot \mid \mathcal{H}^{t-1})$, sampling $\pi^t \sim p^t$. 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 5.

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

⁴In the literature on bandits, decisions are often referred to as *actions*. We use these terms interchangeably throughout this section.

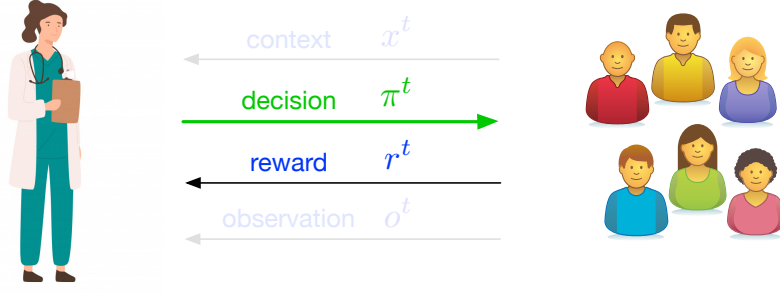


Figure 5: The multi-armed bandit problem. A doctor (the learner) selects a treatment (the decision) to improve a patient’s vital signs (the reward).

the vein of online learning (Section 2.2). We focus on the stochastic framework, and make the following assumption.

Assumption 3.1 (Stochastic Rewards): Rewards are generated independently via

$$r^t \sim M^*(\cdot \mid \pi^t), \quad (3.1)$$

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

Throughout this section, we also assume that rewards are bounded: $r^t \in [0, 1]$ almost surely.⁵ We define

$$f^*(\pi) := \mathbb{E}[r \mid \pi] \quad (3.2)$$

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

$$\mathbf{Reg} := \sum_{t=1}^T f^*(\pi^*) - \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^*(\pi^t)]. \quad (3.3)$$

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 5, 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.

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

Compared to the online regret (2.11), the definition in (3.3) has the same cumulative form, but the benchmark is the best fixed decision π^* rather than the best function in a class, and—as discussed

⁵All results in this section extend to sub-Gaussian rewards with means in $[0, 1]$, at the cost of constants.

next—only the reward of the chosen decision is revealed. Regret bounds in this section hold with probability at least $1 - \delta$ over the draw of the rewards unless stated otherwise.

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 $f^*(\pi^t)$, but reveals no information about the rewards for other decisions $\pi \neq \pi^t$. For example in Figure 5, 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 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).

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.⁶ As a result, the algorithms in this section have regret that scales with $\Omega(|\Pi|) = \Omega(A)$. This shortcoming is addressed by the *structured bandit* framework of Section 5, which allows for the use of function approximation to model structure in the decision space.⁷

Remark 3.1 (Alternative notions of regret): There are other, widely used notions of regret for the multi-armed bandit problem. One is the *empirical regret*, defined as

$$\max_{\pi \in \Pi} \sum_{t=1}^T r^t(\pi) - \sum_{t=1}^T r^t(\pi^t), \quad (3.4)$$

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 (3.3) up to additive $O(\sqrt{T})$ factors.

3.2 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^* via

$$\hat{f}^t(\pi) = \frac{1}{n^t(\pi)} \sum_{s < t} r^s \mathbb{I}\{\pi^s = \pi\}, \quad (3.5)$$

⁶Another way to say this is that we take $\mathcal{F} = \mathbb{R}^A$, so that $f^* \in \mathcal{F}$.

⁷Throughout this monograph, we use the term “multi-armed bandit” exclusively to refer to bandit problems with finite decision spaces, and the term “structured bandit” for problems with large decision spaces.

where $n^t(\pi)$ is the number of times π has been selected up to time t .⁸ Then, we can choose the greedy action

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

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 $\text{Ber}(3/4)$.

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

There are two lessons to carry from this example. First, the failure is caused by partial feedback, not by the sequential nature of the data. If the reward of every decision were revealed at each round, the greedy strategy would be an instance of empirical risk minimization on i.i.d. data, and [Proposition 2.1](#) would apply. Second, an algorithm should never discard a decision that might be optimal on the basis of a few samples; the confidence intervals of [Section 3.4](#) make this idea precise.

3.3 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 (e.g., Sutton and Barto [194], Auer et al. [22]).

Algorithm 3.1 ε -Greedy

input: Exploration parameter $\varepsilon \in [0, 1]$.
for $t = 1, \dots, T$ **do**
 Compute the empirical mean rewards $\hat{f}^t$ as in (3.5).
 With probability $1 - \varepsilon$, select $\pi^t = \arg \max_{\pi \in \Pi} \hat{f}^t(\pi)$; otherwise select $\pi^t \sim \text{unif}(\Pi)$.
 Observe reward r^t .

Let $\varepsilon \in [0, 1]$ be the *exploration parameter*. At each time $t \in [T]$, the ε -Greedy algorithm computes the estimated reward function $\hat{f}^t$ as in (3.5). With probability $1 - \varepsilon$, the algorithm chooses the greedy decision

$$\hat{\pi}^t = \arg \max_{\pi} \hat{f}^t(\pi), \tag{3.6}$$

and with probability ε it samples a uniform random action $\pi^t \sim \text{unif}(\{1, \dots, A\})$.

⁸If $n^t(\pi) = 0$, we set $\hat{f}^t(\pi) = 0$.

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 3.1 (Regret of ε -Greedy): For any T and $\delta \in (0, 1)$, the ε -Greedy algorithm with $\varepsilon \propto (A \log(AT/\delta)/T)^{1/3}$ ensures that with probability at least $1 - \delta$,

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

Note that the exploration parameter must decay with the horizon T : the algorithm explores less as it collects more data. This regret bound has $\frac{\mathbb{E}[\mathbf{Reg}]}{T} \rightarrow 0$ with $T \rightarrow \infty$ as desired, though more sophisticated strategies, beginning with [Section 3.4](#), attain improved regret bounds that scale with $\sqrt{AT}$.⁹

Proof of Proposition 3.1. Recall that $\hat{\pi}^t := \arg \max_{\pi} \hat{f}^t(\pi)$ 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:

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

In the last inequality, we have simply written off the contribution from exploring uniformly by using that $f^*(\pi) \in [0, 1]$. It remains to bound the regret we incur from playing the greedy action. Here, we bound the per-step regret in terms of estimation error using a similar decomposition to [Lemma 2.9](#) (note that we are now working with rewards rather than losses):

$$f^*(\pi^*) - f^*(\hat{\pi}^t) = [f^*(\pi^*) - \hat{f}^t(\pi^*)] + \underbrace{[\hat{f}^t(\pi^*) - \hat{f}^t(\hat{\pi}^t)]}_{\leq 0} + [\hat{f}^t(\hat{\pi}^t) - f^*(\hat{\pi}^t)] \quad (3.7)$$

$$\leq 2 \max_{\pi \in \{\pi^*, \hat{\pi}^t\}} |f^*(\pi) - \hat{f}^t(\pi)| \leq 2 \max_{\pi} |f^*(\pi) - \hat{f}^t(\pi)|. \quad (3.8)$$

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^* . In particular, we show that the event

$$\mathcal{E}_t = \left\{ \max_{\pi} |f^*(\pi) - \hat{f}^t(\pi)| \lesssim \sqrt{\frac{A \log(AT/\delta)}{\varepsilon t}} \right\} \quad (3.9)$$

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

⁹Note that $\sqrt{AT} \leq A^{1/3} T^{2/3}$ whenever $A \leq T$, and when $A \geq T$ both guarantees are vacuous.

To prove that (3.9) holds, we first apply Hoeffding's inequality. Because the number of times π has been pulled is random and depends on the rewards observed so far, the version for a fixed sample size does not apply directly; instead we use the version for adaptive stopping times (Lemma A.1), which gives that for any fixed π , with probability at least $1 - \delta$ over the draw of rewards,

$$|f^\star(\pi) - \hat{f}^t(\pi)| \leq \sqrt{\frac{2 \log(2T/\delta)}{n^t(\pi)}}. \quad (3.10)$$

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

$$|f^\star(\pi) - \hat{f}^t(\pi)| \leq \sqrt{\frac{2 \log(2AT^2/\delta)}{n^t(\pi)}} \quad (3.11)$$

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, $Z^t \sim \text{Ber}(\varepsilon/A)$, so we have $\mathbb{E}[m^t(\pi)] = \varepsilon(t-1)/A$. Using Bernstein's inequality (Lemma 2.8) with $Z^1, \dots, Z^{t-1}$, 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

$$m^t(\pi) \geq \frac{\varepsilon(t-1)}{2A} - \frac{4u}{3}. \quad (3.12)$$

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 $t \in [T]$

$$m^t(\pi) \geq \frac{\varepsilon(t-1)}{2A} - \frac{4 \log(2AT/\delta)}{3}. \quad (3.13)$$

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 (3.11), this implies that with probability at least $1 - \delta$, for all t ,

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

which leads to the overall regret bound

$$\begin{aligned} \mathbf{Reg} &\leq \sum_{t=1}^T \max_{\pi} |f^*(\pi) - \hat{f}^t(\pi)| + \varepsilon T \lesssim \sum_{t=1}^T \sqrt{\frac{A \log(AT/\delta)}{\varepsilon t}} + \varepsilon T \\ &\leq \sqrt{\frac{AT \log(AT/\delta)}{\varepsilon}} + \varepsilon T. \end{aligned} \quad (3.14)$$

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. $\square$

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 (3.14)). 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. Section 3.4 addresses this shortcoming with more adaptive strategies.

Remark 3.2 (Explore-then-commit): A relative of ε -Greedy is the explore-then-commit (ETC) algorithm (e.g., Robbins [177], Langford and Zhang [133]), 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.

3.4 Optimism and the Upper Confidence Bound (UCB) Algorithm

The next algorithm we study is the Upper Confidence Bound (UCB) algorithm [? 12, 22]. The UCB algorithm attains a regret bound of the order $\tilde{O}(\sqrt{AT})$, 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*, a general principle that recurs throughout this monograph 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”

$$\underline{f}^t, \bar{f}^t : \Pi \rightarrow \mathbb{R}, \quad (3.15)$$

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

$$\forall t \in [T], \pi \in \Pi, \quad f^*(\pi) \in [\underline{f}^t(\pi), \bar{f}^t(\pi)]. \quad (3.16)$$

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^* . Given confidence

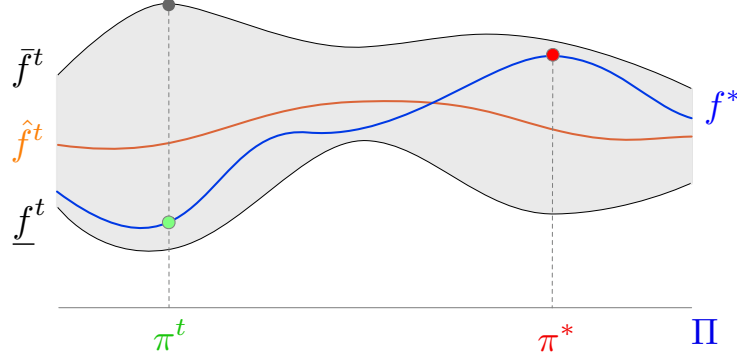


Figure 6: The UCB algorithm. Selecting π^t optimistically ensures that its suboptimality never exceeds the confidence width.

intervals, the UCB algorithm simply chooses π^t as the “optimistic” action that maximizes the upper confidence bound:

$$\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi). \quad (3.17)$$

Algorithm 3.2 Upper Confidence Bound (UCB)

input: Confidence parameter $\delta \in (0, 1)$.
for $t = 1, \dots, T$ **do**
 Compute upper confidence bound $\bar{f}^t$ as in (3.20).
 Select $\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi)$.
 Observe reward r^t .

The following lemma shows that the instantaneous regret for this strategy is bounded by the width of the confidence interval; see Figure 6 for an illustration.

Lemma 3.1 (Optimism bounds regret by confidence width): Fix t , and suppose that $f^*(\pi) \in [\underline{f}^t(\pi), \bar{f}^t(\pi)]$ for all π . Then the optimistic action

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

has

$$f^*(\pi^*) - f^*(\pi^t) \leq \bar{f}^t(\pi^t) - f^*(\pi^t) \leq \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t). \quad (3.18)$$

Proof of Lemma 3.1. The result follows immediate from the observation that for any $t \in [T]$ and any $\pi^* \in \Pi$, we have

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

□

[Lemma 3.1](#) 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 A.1](#)).¹⁰ Since $r^t \in [0, 1]$, a union bound gives that with probability at least $1 - \delta$, for all $t \in [T]$ and $\pi \in \Pi$,

$$|\hat{f}^t(\pi) - f^*(\pi)| \leq \sqrt{\frac{2 \log(2T^2 A/\delta)}{n^t(\pi)}}, \quad (3.19)$$

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

$$\bar{f}^t(\pi) = \hat{f}^t(\pi) + \sqrt{\frac{2 \log(2T^2 A/\delta)}{n^t(\pi)}}, \quad \text{and} \quad \underline{f}^t(\pi) = \hat{f}^t(\pi) - \sqrt{\frac{2 \log(2T^2 A/\delta)}{n^t(\pi)}}, \quad (3.20)$$

we obtain a valid confidence interval. With this choice—along with [Lemma 3.1](#)—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 3.1](#) 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.

Theorem 3.1 (Regret of UCB): Using the confidence bounds in (3.20), the UCB algorithm ensures that with probability at least $1 - \delta$,

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

This result is optimal up to the $\log(AT)$ factor: [Section 8.2](#) shows that every algorithm suffers $\Omega(\sqrt{AT})$ regret on some instance with Gaussian rewards. The logarithmic factor can be removed by using the same algorithm with a slightly more sophisticated confidence interval construction [20]. 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 *polynomially* 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 Theorem 3.1. Let us condition on the event in (3.19). Whenever this occurs, we have that $f^*(\pi) \in [\underline{f}^t(\pi), \bar{f}^t(\pi)]$ for all $t \in [T]$ and $\pi \in \Pi$, so the confidence intervals are valid. As a result, [Lemma 3.1](#) bounds regret in terms of the confidence width:

$$\sum_{t=1}^T f^*(\pi^*) - f^*(\pi^t) \leq \sum_{t=1}^T \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t) = \sum_{t=1}^T 2\sqrt{\frac{2 \log(2T^2 A/\delta)}{n^t(\pi^t)}} \wedge 1; \quad (3.21)$$

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

¹⁰While 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 3.2 (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 3.2. We begin by writing.

$$\sum_{t=1}^T \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 = \sum_{\pi} \sum_{t=1}^T \frac{\mathbb{I}\{\pi^t = \pi\}}{\sqrt{n^t(\pi)}} \wedge 1 = \sum_{\pi} \sum_{t=1}^{n^{T+1}(\pi)} \frac{1}{\sqrt{t-1}} \wedge 1. \quad (3.22)$$

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 3.2 and (3.21).

□

To summarize, the key steps in the proof of Theorem 3.1 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.

Subsequent sections revisit and generalize both ideas for more sophisticated settings, including contextual bandits, structured bandits, and reinforcement learning.

Remark 3.3 (Instance-dependent regret for UCB): The $\tilde{O}(\sqrt{AT})$ 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^* 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 reflects the fact that the algorithm has better regret on “nice” problem instances.

Let $\Delta(\pi) := f^*(\pi^*) - f^*(\pi)$ be the *suboptimality gap* for decision π . Then 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 $T \rightarrow \infty$, this regret bound scales only *logarithmically* in T , which improves upon the $\sqrt{T}$ -scaling of the minimax regret bound.

This concludes the main development of this section. The remaining subsections cover two extensions, the Bayesian viewpoint and adversarial or non-stochastic rewards.

3.5 Bayesian Bandits and the Posterior Sampling Algorithm*

Up to this point, we have been designed and analyzed algorithms from a *frequentist* viewpoint, in which we aim to minimize regret for a *worst-case* choice of the underlying model M^* . An alternative is to adopt a *Bayesian* viewpoint, and assume that the underlying model is drawn from a known prior $\mu \in \Delta(\mathcal{M})$.¹¹ In this case, rather than worst-case performance, we are concerned with the average regret under the prior, defined via

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

where $\mathbb{E}^{M^*}[\cdot]$ denotes the algorithm's expected regret when M^* 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 π^* 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) [198, 181].

Algorithm 3.3 Posterior Sampling

input: Prior distribution over models.

for $t = 1, \dots, T$ **do**

Set $p^t(\pi) = \mathbb{P}(\pi^* = \pi \mid \mathcal{H}^{t-1})$, where $\mathcal{H}^{t-1} = (\pi^1, r^1), \dots, (\pi^{t-1}, r^{t-1})$.

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 $\mathbb{P}(\pi^* = \cdot \mid \mathcal{H}^{t-1})$, which represents the posterior distribution over π^* given all of the data we have collected from rounds $1, \dots, t-1$. The posterior sampling algorithm simply samples the learner's action π^t from this distribution, thereby “matching” the posterior distribution of π^* .

Proposition 3.2 (Bayesian regret of posterior sampling): For any prior μ , the posterior sampling algorithm ensures that

$$\mathbf{Reg}_{\text{Bayes}}(\mu) \leq \sqrt{AT \log(A)}, \tag{3.23}$$

¹¹It is important that μ is known, otherwise this is no different from the frequentist setting.

Here we prove a simplified version of [Proposition 3.2](#); the full proof is given in [Section 3.8](#). The argument has three steps, and the reader may find it useful to keep them in mind: (i) a *decoupling* inequality bounds the per-round Bayesian regret by the estimation error of the posterior mean under the learner's own distribution p^t ; (ii) this estimation error is bounded by the *information gain* about f^* from the round's observation; (iii) the total information gain over T rounds telescopes and is at most the entropy of the prior. Step (i) is the part that generalizes furthest: [Section 5.5](#) recasts it as a bound on the Decision-Estimation Coefficient, a general complexity measure for interactive decision making.

Proof of Proposition 3.2 (simplified version). We make the following simplifying assumptions:

- ▷ We restrict to reward distributions where $M^*(\cdot \mid \pi) = \mathcal{N}(f^*(\pi), 1)$, departing, for simplicity, from the boundedness convention of this section. That is, f^* is the only part of the reward distribution that is unknown.
- ▷ f^* belongs to a known class $\mathcal{F}$, and rather than proving the regret bound in [Proposition 3.2](#), we prove a bound of the form

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

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

Since the mean reward function f^* is the only part of the reward distribution M^* 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^* rather than M^* .

Before proceeding, let us introduce some notation. The process through which we sample $f^* \sim \mu$ and run the bandit algorithm induces a joint law over $(f^*, \mathcal{H}^T)$, 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

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

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

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

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

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

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

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

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

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 3.3 (Decoupling): For any function $\bar{f} : \Pi \rightarrow \mathbb{R}$, it holds that

$$\mathbb{E}_{f^\star \sim \mu^t} [f^\star(\pi_{f^\star}) - \bar{f}(\pi_{f^\star})] \leq \sqrt{A \cdot \mathbb{E}_{f^\star \sim \mu^t} \mathbb{E}_{\pi^t \sim p^t} [(f^\star(\pi^t) - \bar{f}(\pi^t))^2]}. \quad (3.24)$$

Proof of Lemma 3.3. We show a more general result. Namely, for any $\nu \in \Delta(\mathcal{F})$ and $\bar{f} : \Pi \rightarrow \mathbb{R}$, if we define $p(\pi) = \mathbb{P}_{f \sim \nu}(\pi_f = \pi)$, then

$$\mathbb{E}_{f \sim \nu} [f(\pi_f) - \bar{f}(\pi_f)] \leq \sqrt{A \cdot \mathbb{E}_{f \sim \nu} \mathbb{E}_{\pi \sim p} [(f(\pi) - \bar{f}(\pi))^2]}. \quad (3.25)$$

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:

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

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

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

Putting these bounds together yields (3.25). $\square$

Using Lemma 3.3, we have that

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

To finish up, we show that

$$\mathbb{E} \left[\sum_{t=1}^T \mathbb{E}_{f^* \sim \mu^t} [\mathbb{E}_{\pi^t \sim p^t} [(f^*(\pi^t) - \bar{f}^t(\pi^t))^2]] \right] \leq \frac{5}{2} \log |\mathcal{F}|.$$

To do this, we need some additional information-theoretic tools.

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

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

$$\frac{2}{5} \mathbb{E} [\mathbb{E}_{f^* \sim \mu^t} \mathbb{E}_{\pi^t \sim p^t} [(f^*(\pi^t) - \bar{f}^t(\pi^t))^2]] \leq \text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^{t-1}) - \text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^t). \quad (3.26)$$

From here the result follows, because

$$\begin{aligned} \frac{2}{5} \mathbb{E} \left[\sum_{t=1}^T \mathbb{E}_{f^* \sim \mu^t} \mathbb{E}_{\pi^t \sim p^t} [(f^*(\pi^t) - \bar{f}^t(\pi^t))^2] \right] &\leq \sum_{t=1}^T [\text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^{t-1}) - \text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^t)] \\ &= \text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^0) - \text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^T) \\ &\leq \text{Ent}(\boldsymbol{f}^* \mid \boldsymbol{\mathcal{H}}^0) \\ &\leq \log |\mathcal{F}|, \end{aligned}$$

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 now prove (3.26). Conditional on the history and current action, the predictive reward is the sum of a random mean in $[0, 1]$ and independent $\mathcal{N}(0, 1)$ noise. By Hoeffding's lemma, the centered random mean is $1/4$ -sub-Gaussian, so the predictive reward is $5/4$ -sub-Gaussian. The reward conditional on f^* is also $5/4$ -sub-Gaussian. Thus, Pinsker's inequality for sub-Gaussian random variables (Lemma A.6) implies that

$$\frac{2}{5} (f^*(\pi^t) - \bar{f}^t(\pi^t))^2 \leq D_{\text{KL}}(\mathbb{P}_{\boldsymbol{r}^t | f^*, \pi^t, \mathcal{H}^{t-1}} \parallel \mathbb{P}_{\boldsymbol{r}^t | \pi^t, \mathcal{H}^{t-1}}).$$

and

$$\frac{2}{5} \mathbb{E}_{f^* \sim \mu^t} \mathbb{E}_{\pi^t \sim p^t} [(f^*(\pi^t) - \bar{f}^t(\pi^t))^2] \leq \mathbb{E}_{f^* \sim \mu^t} \mathbb{E}_{\pi^t \sim p^t} [D_{\text{KL}}(\mathbb{P}_{\boldsymbol{r}^t | f^*, \pi^t, \mathcal{H}^{t-1}} \parallel \mathbb{P}_{\boldsymbol{r}^t | \pi^t, \mathcal{H}^{t-1}})]$$

Since KL divergence satisfies $\mathbb{E}_{x \sim \mathbb{P}_X} [D_{\text{KL}}(\mathbb{P}_{Y|X=x} \parallel \mathbb{P}_Y)] = \mathbb{E}_{y \sim \mathbb{P}_Y} [D_{\text{KL}}(\mathbb{P}_{X|Y=y} \parallel \mathbb{P}_X)]$, this equals

$$\mathbb{E}_{t-1} [D_{\text{KL}}(\mathbb{P}_{\boldsymbol{f}^* | \pi^t, \boldsymbol{r}^t, \mathcal{H}^{t-1}} \parallel \mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^{t-1}})] = \mathbb{E}_{t-1} [D_{\text{KL}}(\mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^t} \parallel \mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^{t-1}})]. \quad (3.27)$$

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

$$\mathbb{E} [\mathbb{E}_{t-1} [D_{\text{KL}}(\mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^t} \parallel \mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^{t-1}})]] = \mathbb{E}_{\mathcal{H}^{t-1}} \mathbb{E}_{\mathcal{H}^t | \mathcal{H}^{t-1}} [D_{\text{KL}}(\mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^t} \parallel \mathbb{P}_{\boldsymbol{f}^* | \mathcal{H}^{t-1}})].$$

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

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

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

$$\mathbb{E}_{\mathcal{H}^{t-1}} \mathbb{E}_{\mathcal{H}^t \mid \mathcal{H}^{t-1}} [D_{\text{KL}}(\mathbb{P}_{f^* \mid \mathcal{H}^t} \parallel \mathbb{P}_{f^* \mid \mathcal{H}^{t-1}})] = \text{Ent}(f^* \mid \mathcal{H}^{t-1}) - \text{Ent}(f^* \mid \mathcal{H}^t).$$

□

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 is indeed the case for bandits, though a different analysis is required [13, 14]. However, one can show (Sections 5 and 7) 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 3.4 (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

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

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.

The entropy potential in the proof plays the same role as the pigeonhole argument of Lemma 3.2: there, each round that incurs large regret must increase the count of some action, of which there are only A ; here, each round that incurs large regret must reveal information about f^* , but only $O(\log |\mathcal{F}|)$ bits of information can be revealed in total.

3.6 Adversarial Bandits and the Exp3 Algorithm*

The stochastic model of Assumption 3.1 is restrictive: in applications such as online advertising or routing, the reward of a decision may drift over time, or may be chosen by an adversary. This subsection gives a brief introduction to the multi-armed bandit problem with non-stochastic rewards, which dispenses with Assumption 3.1 entirely. In the context of Figure 2, this moves the problem along the “data” axis, in the same way that online learning generalizes statistical learning in Section 2; the algorithm and analysis below make this analogy exact. 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 (3.4).

As one might expect, the solution for non-stochastic bandits leverages the online learning tools introduced in Section 2.2. We consider an algorithm known as Exp3 (*Exponential Weights for Exploration and Exploitation*) [23], which builds upon the exponential weights algorithm for online learning in Section 2.2. 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.

Algorithm 3.4 Exp3

input: Learning rate $\eta > 0$.

Initialize $p^1 = \text{unif}(\Pi)$.

for $t = 1, \dots, T$ **do**

 Select $\pi^t \sim p^t$ and observe loss $\ell^t(\pi^t)$.

 Form the importance-weighted loss estimate $\tilde{\ell}^t(\pi) = \ell^t(\pi) \mathbb{I}\{\pi^t = \pi\} / p^t(\pi)$ for all $\pi \in \Pi$.

 Update $p^{t+1}(\pi) \propto p^t(\pi) \exp(-\eta \tilde{\ell}^t(\pi))$.

Proposition 3.3 (Regret of Exp3): For any sequence of losses in $[0, 1]^A$ fixed in advance, the Exp3 algorithm with an appropriate choice of η ensures that

$$\mathbb{E}[\mathbf{Reg}] \lesssim \sqrt{AT \log A},$$

where the expectation is over the algorithm's randomization.

The remainder of this subsection sketches the proof.

Recall that p^t denotes the learner's randomization distribution at round t . Averaging over the sampled actions, the expected bandit regret is

$$\mathbb{E}[\mathbf{Reg}] = \mathbb{E} \left[\sum_{t=1}^T \langle p^t, \ell^t \rangle \right] - \min_{\pi \in [A]} \sum_{t=1}^T \ell^t(\pi), \quad (3.28)$$

where $\ell^t \in [0, 1]^A$ is the vector of losses at time t . The expectation also accounts for the randomness in p^t induced by past observations.

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 ℓ^t . To address this, we build an *unbiased estimate* of the vector ℓ^t from a single real-valued observation $\ell^t(\pi^t)$. At first, this might appear impossible, but it is straightforward to show that

$$\tilde{\ell}^t(\pi) = \frac{\ell^t(\pi)}{p^t(\pi)} \times \mathbb{I}\{\pi^t = \pi\} \quad (3.29)$$

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

$$\mathbb{E}_{\pi^t \sim p^t} [\tilde{\ell}^t] = \ell^t. \quad (3.30)$$

We apply exponential weights to the estimated loss vectors $\tilde{\ell}^t$. As in [Remark 2.4](#), the online learner's regret on these vectors is

$$\mathbf{Reg}_{\text{OL}} = \sum_{t=1}^T \langle p^t, \tilde{\ell}^t \rangle - \min_{\pi \in [A]} \sum_{t=1}^T \tilde{\ell}^t(\pi). \quad (3.31)$$

These estimated losses are nonnegative but can exceed one, so we use the second-order exponential weights bound. This bound holds for each realized sequence of estimated losses; taking expectations

and bounding their second moments gives $\mathbb{E}[\mathbf{Reg}_{\text{OL}}] \lesssim \sqrt{AT \log A}$. Unbiasedness then transfers the guarantee to bandit regret:

$$\mathbb{E}[\mathbf{Reg}] = \mathbb{E} \left[\sum_{t=1}^T \langle p^t, \tilde{\ell}^t \rangle \right] - \min_{\pi \in [A]} \mathbb{E} \left[\sum_{t=1}^T \tilde{\ell}^t(\pi) \right] \quad (3.32)$$

$$\leq \mathbb{E}[\mathbf{Reg}_{\text{OL}}] \lesssim \sqrt{AT \log A}. \quad (3.33)$$

The inequality follows by comparing with the best fixed action for the original loss sequence.

3.7 Bibliographic Remarks and Further Reading

Takeaways. This section introduced the multi-armed bandit as the simplest interactive decision making problem, and used it to isolate the exploration–exploitation tradeoff created by partial feedback. The contrast between ε -Greedy and UCB—uniform exploration paying $T^{2/3}$ versus adaptive exploration paying $\sqrt{T}$ —reappears in the contextual bandit setting of [Section 4](#). More broadly, three ideas recur throughout the monograph. First, optimism: by [Lemma 3.1](#), the regret of UCB is controlled by the sum of confidence widths. Second, the pigeonhole (or potential) argument of [Lemma 3.2](#), which shows that the widths cannot remain large for many rounds; it reappears in later sections as the elliptic potential lemma. Third, the information-theoretic analysis of posterior sampling, in which a decoupling inequality ([Lemma 3.3](#)) and an entropy potential replace confidence intervals altogether; this argument foreshadows the Decision-Estimation Coefficient of [Section 7](#).

Bibliographic remarks. The multi-armed bandit problem was introduced by Thompson [\[198\]](#) in the context of clinical trials and given its modern formulation by Robbins [\[177\]](#); the name refers to a slot machine, or “one-armed bandit.” Upper confidence indices originate with [\[?\]](#), who also proved the asymptotic instance-dependent lower bound; the finite-time UCB algorithm and analysis of [Theorem 3.1](#), including the instance-dependent bound in the remark that follows it, are due to Auer et al. [\[22\]](#), and the $\Omega(\sqrt{AT})$ minimax lower bound is from Auer et al. [\[23\]](#). Posterior sampling is due to Thompson [\[198\]](#); the information-theoretic analysis behind [Proposition 3.2](#) follows Russo and Van Roy [\[182\]](#), building on Russo and Van Roy [\[181\]](#), and frequentist regret bounds for posterior sampling are due to Agrawal and Goyal [\[13\]](#) and Kaufmann et al. [\[123\]](#). Exp3 and the importance-weighted loss estimator are from Auer et al. [\[23\]](#).

Further reading. Comprehensive treatments of bandits are given by Lattimore and Szepesvári [\[139\]](#), Slivkins [\[191\]](#), and Bubeck and Cesa-Bianchi [\[39\]](#). Refinements of UCB include KL-UCB [\[95\]](#), which attains the constant in the Lai–Robbins lower bound, and MOSS [\[20\]](#), which removes the logarithmic factor from the minimax bound. The Bayesian analysis of posterior sampling extends to adversarial rewards through minimax duality [\[41, 138\]](#), a perspective we return to in [Section 5.5.2](#).

3.8 Deferred Proofs

Proof of [Proposition 3.2](#) (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

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

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

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

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

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

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

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

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

where we have used that p^t was chosen to match the posterior distribution over π^* . This gives

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

We now make use of the following decoupling-type inequality, which follows from (3.25):

$$\mathbb{E}_{\pi \sim p^t}[\bar{f}_{\pi}^t(\pi) - \bar{f}^t(\pi)] \leq \sqrt{A \cdot \mathbb{E}_{\pi, \pi^* \sim p^t}[(\bar{f}_{\pi^*}^t(\pi) - \bar{f}^t(\pi))^2]}. \quad (3.34)$$

To keep notation as clear as possible going forward, let us use boldface script ($\boldsymbol{\pi}^t$, $\boldsymbol{\pi}^*$, $\boldsymbol{f}^*$, $\boldsymbol{\mathcal{H}}^t$) to refer to the abstract random variables under consideration, and use non-boldface script (π^t , π^* , f^* , $\mathcal{H}^t$) to refer to their realizations. As in the simplified proof, we show that the right-hand side in (3.34) is related to a notion of *information gain* (that is, information about π^* acquired at step t). Using Pinsker's inequality, we have

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

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

$$\mathbb{E}_{t-1}[D_{\text{KL}}(\mathbb{P}_{\boldsymbol{\pi}^* | \boldsymbol{\pi}^t, \boldsymbol{r}^t, \boldsymbol{\mathcal{H}}^{t-1}} \parallel \mathbb{P}_{\boldsymbol{\pi}^* | \boldsymbol{\mathcal{H}}^{t-1}})] = \mathbb{E}_{t-1}[D_{\text{KL}}(\mathbb{P}_{\boldsymbol{\pi}^* | \boldsymbol{\mathcal{H}}^t} \parallel \mathbb{P}_{\boldsymbol{\pi}^* | \boldsymbol{\mathcal{H}}^{t-1}})]. \quad (3.35)$$

This is quantifying how much information about π^* we gain by playing π^t and observing $\boldsymbol{r}^t$ at step t , relative to what we knew at step $t-1$. Applying (3.34) and (3.35), we have

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

We can write

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

so telescoping gives

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

□

4. CONTEXTUAL BANDITS: EXPLORATION AND GENERALIZATION

The multi-armed bandit of [Section 3](#) assumes that a single decision is best for everyone. Few applications look like this: a treatment that is right for one patient is wrong for another. This section adds *contexts*: before each decision, the learner observes side information, and the goal is to compete with the best *policy* mapping contexts to decisions. In the language of [Figure 2](#), this is the first step along the function approximation axis, and the central question is whether the statistical efficiency of supervised learning can be retained when the learner must also explore. The answer is yes, but not by optimism alone, and introduces a reduction from decision making to online regression that resolves the problem for every class. This reduction, the Estimation-to-Decisions principle, is the organizing idea of the rest of the monograph.

Organization. [Section 4.1](#) formalizes the contextual bandit problem, the regret criterion, and the role of function approximation. [Section 4.2](#) adapts the principle of optimism to contextual bandits, replacing confidence intervals by confidence sets built from least squares. [Section 4.3](#) instantiates the template for linear classes (LinUCB), and [Section 4.4](#) complements this by showing that optimistic methods can have poor performance for general classes.

[Section 4.5](#) introduces the *online regression oracle*, the main abstraction through which we reduce decision making to supervised learning. [Section 4.6](#) analyzes the simplest algorithm built on it, ϵ -Greedy, as a first attempt. This culminates in [Section 4.7](#), which gives an optimal reduction, Inverse Gap Weighting and the SquareCB algorithm.

4.1 The Contextual Bandit Problem

The simplicity of the multi-armed bandit problem comes at a cost: few real-world problems can be modeled as a multi-armed bandit directly. For example, for the problem of selecting medical treatments, the multi-armed bandit formulation presupposes 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.

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  $r^t \in \mathbb{R}$ .
```

The protocol proceeds in T rounds. At each round t , the learner observes a context $x^t \in \mathcal{X}$, selects a decision $\pi^t \in \Pi$ —possibly at random, $\pi^t \sim p^t$, where the distribution $p^t \in \Delta(\Pi)$ may depend on x^t and on the history $\mathcal{H}^{t-1} = (x^1, \pi^1, r^1), \dots, (x^{t-1}, \pi^{t-1}, r^{t-1})$ —and observes a reward r^t ; see [Figure 7](#).

As with multi-armed bandits, contextual bandits can be studied in a stochastic framework or in an adversarial framework. In this monograph, we allow the contexts $x^1, \dots, x^T$ to be generated in

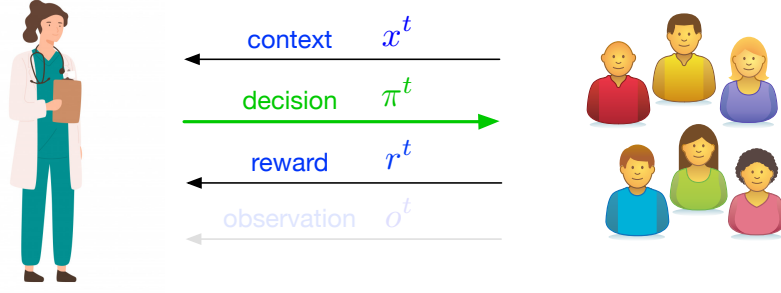


Figure 7: The contextual bandit problem. A doctor (the learner) selects a treatment (the decision) based on the patient’s context (medical history, symptoms).

an arbitrary, potentially adversarial fashion, but assume that rewards are generated from a fixed conditional distribution.

Assumption 4.1 (Stochastic Rewards): Rewards are generated independently via

$$r^t \sim M^*(\cdot \mid x^t, \pi^t), \quad (4.1)$$

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

This generalizes the stochastic multi-armed bandit framework in [Section 3](#). As in [Section 3](#), we assume throughout this section that rewards are bounded, $r^t \in [0, 1]$ almost surely, and regret bounds hold with probability at least $1 - \delta$ unless stated otherwise; all results hold for arbitrary (possibly adaptively chosen) context sequences. We define

$$f^*(x, \pi) := \mathbb{E}[r \mid x, \pi] \quad (4.2)$$

as the mean reward function under $r \sim M^*(\cdot \mid x, \pi)$, and define $\pi^*(x) := \arg \max_{\pi \in \Pi} f^*(x, \pi)$ as the optimal *policy*, which maps each context x to the optimal action for the context. Unlike the multi-armed bandit, the reward landscape now varies with the context, so no single action is best; the natural benchmark is a policy rather than an action. We measure performance via regret relative to π^* :

$$\mathbf{Reg} := \sum_{t=1}^T f^*(x^t, \pi^*(x^t)) - \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^*(x^t, \pi^t)], \quad (4.3)$$

where, as above, p^t is the learner’s action distribution at step 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 4.1 (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 reinforce-

ment learning, we aim to control the evolution of $x^1, \dots, x^T$ (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 the goal. As a first idea, if $\mathcal{X}$, the set of possible contexts, is finite, one can imagine running a separate multi-armed bandit (MAB) algorithm for each context, updating each algorithm only when its corresponding context is encountered. This leads to the following guarantee (the proof is deferred to [Section 4.9](#)).

Proposition 4.1 (Unstructured contextual bandits): Suppose $|\mathcal{X}| = K < \infty$, $|\Pi| = A < \infty$, and rewards lie in $[0, 1]$. If we run a separate UCB instance for each context, then for any (potentially adaptive) context sequence,

$$\mathbb{E}[\mathbf{Reg}] \lesssim \sqrt{AKT \log(eAKT)}.$$

We see that the regret bound scales with $|\mathcal{X}|$, 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 2](#)) 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 make the following realizability assumption.

Assumption 4.2 (Realizability): The decision-maker has access to a class $\mathcal{F} \subset \{f : \mathcal{X} \times \Pi \rightarrow \mathbb{R}\}$ such that $f^* \in \mathcal{F}$.

In keeping with the reward convention, we also assume that $f(x, \pi) \in [0, 1]$ for all $f \in \mathcal{F}$ unless otherwise specified. Using the class $\mathcal{F}$, we would like to develop algorithms that can model the underlying reward function for better decision making performance. Concretely, the goal of this section is regret that scales with $\log|\mathcal{F}|$ and has no dependence on $|\mathcal{X}|$. The best possible rate of this form, $\sqrt{AT \log|\mathcal{F}|}$, is attained in [Section 4.7](#); for linear classes, the number of actions A is replaced by the dimension d ([Section 4.3](#)). In the language of [Figure 2](#), contexts introduce generalization across $\mathcal{X}$ through function approximation, but the decision space remains unstructured.

With our goal in mind, it is natural to try leveraging the algorithms and respective guarantees we have already developed 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 greedy algorithm in [Section 3.2](#). In spite of these apparent difficulties, the remainder of this section shows that it is possible to leverage supervised learning methods to develop provable decision making methods, thereby bridging the two methodologies.

4.2 Optimism for Contextual Bandits: A 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 can be adapted to give optimal guarantees in the presence of contexts. The answer to this last question is: *it depends*. We first describe some positive results under assumptions on $\mathcal{F}$ (Sections 4.2 and 4.3), then provide a negative example (Section 4.4), and finally turn to an entirely different algorithmic principle (Section 4.7).

Optimism via confidence sets. Let us describe a general approach (or, template) for applying the principle of optimism to contextual bandits [57, 1, 180, 80]. 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 important property that $f^* \in \mathcal{F}^t$. Given such a confidence set we can define upper and lower *confidence functions* $\underline{f}^t, \bar{f}^t : \mathcal{X} \times \Pi \rightarrow \mathbb{R}$ via

$$\underline{f}^t(x, \pi) = \min_{f \in \mathcal{F}^t} f(x, \pi), \quad \bar{f}^t(x, \pi) = \max_{f \in \mathcal{F}^t} f(x, \pi). \quad (4.4)$$

These functions generalize the upper and lower confidence bounds we constructed in Section 3. Since $f^* \in \mathcal{F}^t$, they have the property that

$$\underline{f}^t(x, \pi) \leq f^*(x, \pi) \leq \bar{f}^t(x, \pi) \quad (4.5)$$

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

$$\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(x^t, \pi), \quad (4.6)$$

then as in Lemma 3.1, the optimistic action satisfies

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

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

$$\mathbf{Reg} \leq \sum_{t=1}^T \bar{f}^t(x^t, \pi^t) - \underline{f}^t(x^t, \pi^t). \quad (4.7)$$

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 (4.7) shrinks as fast as possible.

How should we construct the confidence set? Recall from (3.20) that the optimistic approach for multi-armed bandits is to select an action π^t that maximizes $\bar{f}^t(\pi) = \hat{f}^t(\pi) + \frac{\beta}{\sqrt{n^t(\pi)}}$ for some β that depends on T, A, δ . What are the corresponding quantities for our setting of contextual bandits with a class $\mathcal{F}$? Unfortunately, since in the present formulation we may never observe the same x^t more than once, the corresponding meaning of $\hat{f}^t(x, \pi)$ as an average of observed rewards makes no sense. It turns out that appealing to the variational representation of averages as minimizers of the squared loss gracefully resolves this conundrum. The idea is as follows.

Denote by $S^t(\pi)$ the set of rewards obtained when action π was chosen prior to time t , with $|S^t(\pi)| = n^t(\pi)$. As mentioned, we may write the mean reward as

$$\hat{f}^t(\pi) = \arg \min_{v^* \in \mathbb{R}} \sum_{r \in S^t(\pi)} (v^* - r)^2.$$

Similarly, the confidence set

$$\left\{ v \in \mathbb{R} : |v - \hat{f}^t(\pi)| \leq \frac{\beta}{\sqrt{n^t(\pi)}} \right\} \quad (4.8)$$

can be shown to coincide with

$$\left\{ v \in \mathbb{R} : \sum_{r \in S^t(\pi)} (v - r)^2 - \min_{v^* \in \mathbb{R}} \sum_{r \in S^t(\pi)} (v^* - r)^2 \leq \beta^2 \right\}, \quad (4.9)$$

the set of almost-minimizers of the mean squared error. To verify this when $n = n^t(\pi) > 0$, let $\bar{r} := n^{-1} \sum_{r \in S^t(\pi)} r$. Since $\sum_{r \in S^t(\pi)} (r - \bar{r}) = 0$, expansion gives

$$\sum_{r \in S^t(\pi)} (v - r)^2 = n(v - \bar{r})^2 + \sum_{r \in S^t(\pi)} (\bar{r} - r)^2.$$

The minimum is attained at $v = \bar{r}$, so the excess squared loss is $n(v - \bar{r})^2$. Its sublevel set at β^2 is exactly (4.8). When $n = 0$, the loss in (4.9) is identically zero and the confidence set is all of $\mathbb{R}$. As we show next, these re-formulations open the door to constructing the confidence sets $\mathcal{F}^t$.

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). Recall that $f(x, \pi) \in [0, 1]$ for all $f \in \mathcal{F}$ and $r^t \in [0, 1]$ almost surely. Let

$$\hat{f}^t = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} (f(x^i, \pi^i) - r^i)^2 \quad (4.10)$$

be the empirical risk minimizer at round t , and with $\beta^2 := 8 \log(|\mathcal{F}|/\delta)$ define $\mathcal{F}^1 = \mathcal{F}$ and

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} (f(x^i, \pi^i) - r^i)^2 \leq \sum_{i=1}^{t-1} (\hat{f}^t(x^i, \pi^i) - r^i)^2 + \beta^2 \right\} \quad (4.11)$$

for $t > 1$. These definitions are the bona fide generalizations of Eqs. (4.8) and (4.9), aiming to employ the structure of functions in $\mathcal{F}$ and enable generalization across the covariates and decisions. Observe that the optimization problems in Eqs. (4.10) and (4.11) pull together information across all the triples (x^i, π^i, r^i) instead of bucketing the data according to a particular context.

In summary, our confidence set $\mathcal{F}^t$ is the collection of all functions in $\mathcal{F}$ that have empirical squared error close to that of $\hat{f}^t$. We set β “just large enough” to ensure that we do not accidentally exclude f^* —a property necessary for establishing (4.5)—with the precise value for β informed by the probabilistic inequalities we explored in Section 2. 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 (4.10). 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 4.1 (Validity of least-squares confidence sets): Let $\pi^1, \dots, \pi^T$ be chosen by an arbitrary (and possibly randomized) decision-making algorithm. With probability at least $1 - \delta$, $f^\star \in \mathcal{F}^\tau$ for all $\tau \in [T]$. Moreover, with probability at least $1 - \delta$, for all $\tau \in [T]$, all $f \in \mathcal{F}^\tau$ satisfy

$$\sum_{t=1}^{\tau-1} \mathbb{E}_{\pi^t \sim p^t} [(f(x^t, \pi^t) - f^\star(x^t, \pi^t))^2] \leq 4\beta^2, \quad (4.12)$$

where $\beta^2 = 8 \log(|\mathcal{F}|/\delta)$.

Importantly, the definition of $\mathcal{F}^t$ in (4.11) does not involve the unknown $f^\star$, yet Lemma 4.1—the statistical consequence of least squares—guarantees closeness to $f^\star$.

Proof of Lemma 4.1. For $f \in \mathcal{F}$, define

$$U^t(f) = (f(x^t, \pi^t) - r^t)^2 - (f^\star(x^t, \pi^t) - r^t)^2. \quad (4.13)$$

It is straightforward to check that¹²

$$\mathbb{E}_{t-1} U^t(f) = \mathbb{E}_{t-1} (f(x^t, \pi^t) - f^\star(x^t, \pi^t))^2, \quad (4.14)$$

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. Because x^t and π^t depend on the past, Bernstein's inequality for independent variables does not apply; we use instead its martingale analogue, Freedman's inequality (Lemma A.3), which appears here for the first time and is used repeatedly in later sections. Since increments $Z^t(f)$ are bounded as $|Z^t(f)| \leq 2$ (this holds whenever $f \in [0, 1]$, $r^t \in [0, 1]$), according to Lemma A.3 with $\eta = \frac{1}{8}$, with probability at least $1 - \delta$, for all $\tau \leq T$,

$$\sum_{t=1}^{\tau} Z^t(f) \leq \frac{1}{8} \sum_{t=1}^{\tau} \mathbb{E}_{t-1} [Z^t(f)^2] + 8 \log(\delta^{-1}). \quad (4.15)$$

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

$$\mathbb{E}_{t-1} [Z^t(f)^2] \leq \mathbb{E}_{t-1} \left[((f(x^t, \pi^t) - r^t)^2 - (f^\star(x^t, \pi^t) - r^t)^2)^2 \right] \quad (4.16)$$

$$\leq 4 \mathbb{E}_{t-1} [(f(x^t, \pi^t) - f^\star(x^t, \pi^t))^2] = 4 \mathbb{E}_{t-1} U^t(f). \quad (4.17)$$

Then, after rearranging, (4.15) becomes

$$\frac{1}{2} \sum_{t=1}^{\tau} \mathbb{E}_{t-1} U^t(f) \leq \sum_{t=1}^{\tau} U^t(f) + 8 \log(\delta^{-1}). \quad (4.18)$$

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

$$\sum_{t=1}^{\tau} (f^\star(x^t, \pi^t) - r^t)^2 \leq \sum_{t=1}^{\tau} (f(x^t, \pi^t) - r^t)^2 + 8 \log(\delta^{-1}). \quad (4.19)$$

¹²We leave $\mathbb{E}_{t-1}$ on the right-hand side to include the case of randomized decisions $\pi^t \sim p^t$.

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

$$\forall f \in \mathcal{F}, \forall \tau \in [T], \quad \sum_{t=1}^{\tau} (f^*(x^t, \pi^t) - r^t)^2 \leq \sum_{t=1}^{\tau} (f(x^t, \pi^t) - r^t)^2 + 8 \log(|\mathcal{F}|/\delta), \quad (4.20)$$

and in particular, shifting index τ by 1 for brevity, and applying to the data-dependent $\hat{f}^\tau$,

$$\forall \tau \in [T + 1], \quad \sum_{t=1}^{\tau-1} (f^*(x^t, \pi^t) - r^t)^2 \leq \sum_{t=1}^{\tau-1} (\hat{f}^\tau(x^t, \pi^t) - r^t)^2 + 8 \log(|\mathcal{F}|/\delta); \quad (4.21)$$

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

$$\sum_{t=1}^{\tau-1} U^t(f) \leq \beta^2$$

since the empirical risk of f^* is never better than the empirical risk of the minimizer $\hat{f}^t$. Thus from (4.18), with probability at least $1 - \delta$, for all $\tau \leq T$,

$$\sum_{t=1}^{\tau-1} \mathbb{E}_{t-1} U^t(f) \leq 2\beta^2 + 16 \log(\delta^{-1}). \quad (4.22)$$

The second claim follows by taking union bound over $f \in \mathcal{F}^\tau \subseteq \mathcal{F}$, and by (4.14). $\square$

Lemma 4.1 is valid for any algorithm, but it is particularly useful for UCB as it establishes the validity of the confidence bounds as per (4.5); however, it 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 (4.7), but there is a mismatch: for step τ , the regret bound in (4.7) considers the width at (x^τ, π^τ) , but (4.12) only ensures closeness of functions in $\mathcal{F}^\tau$ under $(x^1, \pi^1), \dots, (x^{\tau-1}, \pi^{\tau-1})$. Section 4.3 shows that for linear models, it is possible to control this mismatch; Section 4.4 shows that this is not possible in general.

4.3 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* $\phi : \mathcal{X} \times \Pi \rightarrow \mathbb{B}_2^d(1)$, where $\mathbb{B}_2^d(1)$ 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 $\phi(x, \pi) \in \mathbb{B}_2^d(1)$. In practice, feature maps come from domain knowledge or from a representation learned offline (for example, the penultimate layer of a pretrained neural network); for the analysis, all that matters is that ϕ is fixed and known. We take $\mathcal{F}$ to be the set of linear functions given by

$$\mathcal{F} = \{(x, \pi) \mapsto \langle \theta, \phi(x, \pi) \rangle \mid \theta \in \Theta\}, \quad (4.23)$$

where $\Theta \subseteq \mathbb{B}_2^d(1)$ is the parameter set. As before, we assume $f^* \in \mathcal{F}$; we let θ^* denote the corresponding parameter vector, so that $f^*(x, \pi) = \langle \theta^*, \phi(x, \pi) \rangle$. With some abuse of notation, we associate the set of parameters Θ to the corresponding functions in $\mathcal{F}$.

The LinUCB algorithm. The following figure displays an algorithm we refer to as LinUCB [21, 57, 1], which adapts the generic template for optimistic algorithms to the case where $\mathcal{F}$ is linear in the sense of (4.23).

Algorithm 4.1 LinUCB

input: Confidence parameter $\beta > 0$.

for $t = 1, \dots, T$ **do**

 Compute the least squares solution $\hat{\theta}^t$ (over $\theta \in \Theta$) given by

$$\hat{\theta}^t = \arg \min_{\theta \in \Theta} \sum_{i=1}^{t-1} (\langle \theta, \phi(x^i, \pi^i) \rangle - r^i)^2.$$

 Define

$$\tilde{\Sigma}^t = \sum_{i=1}^{t-1} \phi(x^i, \pi^i) \phi(x^i, \pi^i)^\top + I.$$

 Given x^t , select action

$$\pi^t \in \arg \max_{\pi \in \Pi} \langle \hat{\theta}^t, \phi(x_t, \pi) \rangle + \sqrt{16\beta^2 + 4} \|\phi(x_t, \pi)\|_{(\tilde{\Sigma}^t)^{-1}}.$$

 Observe reward r^t

LinUCB instantiates the template of Section 4.2 with the least-squares confidence set (4.11): for linear classes this set is contained in an ellipsoid centered at $\hat{\theta}^t$, and maximizing a linear function over an ellipsoid gives the closed-form upper confidence bound in the last step of the algorithm; the proof below makes this precise. 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 .¹³

Theorem 4.1 (Regret of LinUCB): Let $\Theta \subseteq \mathbb{B}_2^d(1)$ be a nonempty finite set, let $\theta^* \in \Theta$, and fix $\phi : \mathcal{X} \times \Pi \rightarrow \mathbb{B}_2^d(1)$. Assume $r^t \in [-1, 1]$ almost surely, and let $T \geq 1$ and $\delta \in (0, 1)$. Break ties using a fixed rule. Let $\mathcal{F}$ be the corresponding class of linear functions in (4.23), and take $\beta^2 = 32 \log(|\mathcal{F}|/\delta)$. Then LinUCB satisfies, with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{(\beta^2 + 1)dT \log(1 + T/d)} \lesssim \sqrt{dT \log(1 + T/d)(1 + \log(|\mathcal{F}|/\delta))}$$

for any potentially adaptive sequence of contexts. More generally, for any compact, potentially infinite set $\Theta \subseteq \mathbb{B}_2^d(1)$ containing θ^* , the same first bound holds with

$$\beta^2 = 32[d \log(1 + 2T) + \log(1/\delta)] + 6.$$

In particular, with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{dT \log(1 + T/d)[d \log(1 + T) + \log(1/\delta) + 1]}.$$

¹³The theorem allows rewards in $[-1, 1]$ rather than the $[0, 1]$ convention of this section, because linear predictions with θ and ϕ in the unit ball take values in $[-1, 1]$. Lemma 4.1 applies after rescaling, with $\beta^2 = 32 \log(|\mathcal{F}|/\delta)$.

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 contexts, but across decisions. [Section 5](#) expands upon the idea of generalizing across actions.

Remark 4.2 (Connection to UCB for multi-armed bandits): We can embed the multi-armed bandit setting as a special case of linear contextual bandits by considering a single context x_0 and one-hot features $\phi(x_0, a) = e_a$, so that $d = A$. Let $N_a(t)$ count selections of action a before time t . Then

$$\Sigma^t = \text{diag}(N_1(t), \dots, N_A(t)), \quad \|e_a\|_{(\tilde{\Sigma}^t)^{-1}}^2 = \frac{1}{N_a(t) + 1}.$$

Consequently, LinUCB's score is $\hat{\theta}_a^t + \sqrt{16\beta^2 + 4}/\sqrt{N_a(t) + 1}$, and we see that in this setting, the diagonal covariance recovers the inverse-square-root count dependence of UCB's bonus.

We prove the here result for finite parameter sets; a straightforward extension to continuous parameter sets uses discretization, and is deferred to [Section 4.9](#).

Proof of [Theorem 4.1](#), finite parameter sets. The confidence set (4.11) in the generic optimistic algorithm template is

$$\mathcal{F}^t = \left\{ \theta \in \Theta : \sum_{i=1}^{t-1} (\langle \theta, \phi(x^i, \pi^i) \rangle - r^i)^2 \leq \sum_{i=1}^{t-1} (\langle \hat{\theta}^t, \phi(x^i, \pi^i) \rangle - r^i)^2 + \beta^2 \right\}, \quad (4.24)$$

where $\hat{\theta}^t$ is the least squares solution computed in LinUCB. By [Lemma 4.1](#), with the rescaling to $[-1, 1]$ described above, with probability at least $1 - \delta$, for all $t \in [T]$, all $\theta \in \mathcal{F}^t$ satisfy

$$\sum_{i=1}^{t-1} (\langle \theta - \theta^*, \phi(x^i, \pi^i) \rangle)^2 \leq 4\beta^2, \quad (4.25)$$

which means that $\mathcal{F}^t$ is a subset of¹⁴

$$\Theta' = \left\{ \theta \in \Theta : \|\theta - \theta^*\|_{\Sigma^t}^2 \leq 4\beta^2 \right\}, \quad \text{where} \quad \Sigma^t = \sum_{i=1}^{t-1} \phi(x^i, \pi^i) \phi(x^i, \pi^i)^\top. \quad (4.26)$$

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

$$\Theta'' = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}^t\|_{\tilde{\Sigma}^t}^2 \leq 16\beta^2 + 4 \right\}, \quad \text{where} \quad \tilde{\Sigma}^t = \sum_{i=1}^{t-1} \phi(x^i, \pi^i) \phi(x^i, \pi^i)^\top + I. \quad (4.27)$$

The purpose of establishing the containment $\mathcal{F}^t \subseteq \Theta''$ is purely computational, as it gives us a clean upper bound on $\bar{f}^t$ in (4.4):

$$\bar{f}^t(x, \pi) \leq \max_{\theta: \|\theta - \hat{\theta}^t\|_{\tilde{\Sigma}^t} \leq \sqrt{16\beta^2 + 4}} \langle \theta, \phi(x, \pi) \rangle = \langle \hat{\theta}^t, \phi(x, \pi) \rangle + \sqrt{16\beta^2 + 4} \|\phi(x, \pi)\|_{(\tilde{\Sigma}^t)^{-1}}, \quad (4.28)$$

¹⁴For a PSD matrix $\Sigma \succeq 0$, we define $\|x\|_\Sigma = \sqrt{\langle x, \Sigma x \rangle}$.

and similarly $\underline{f}^t(x, \pi) \geq \langle \hat{\theta}^t, \phi(x, \pi) \rangle - \sqrt{16\beta^2 + 4} \|\phi(x, \pi)\|_{(\tilde{\Sigma}^t)^{-1}}$. Since $\theta^* \in \Theta''$, the upper and lower envelopes of Θ'' bracket the true reward at every action. LinUCB maximizes this enlarged upper envelope, so the optimistic-regret argument (Lemma 3.1) bounds its regret by the width of Θ'' at the selected action. Thus,

$$\mathbf{Reg} \leq 2\sqrt{16\beta^2 + 4} \sum_{t=1}^T \|\phi(x^t, \pi^t)\|_{(\tilde{\Sigma}^t)^{-1}} \lesssim \sqrt{(\beta^2 + 1)T \sum_{t=1}^T \|\phi(x^t, \pi^t)\|_{(\tilde{\Sigma}^t)^{-1}}^2}. \quad (4.29)$$

The above upper bound has the same flavor as the one in Lemma 3.2: as we obtain more and more information in some direction v , the matrix $\tilde{\Sigma}^t$ has a larger and larger component in that direction, and for that direction v , the term $\|v\|_{(\tilde{\Sigma}^t)^{-1}}^2$ becomes smaller and smaller. Unlike the multi-armed bandit case, where each round increments the count of a single action, a single feature vector $\phi(x^t, \pi^t)$ can have components in many directions at once. To conclude, we apply a potential argument, Lemma 4.2 below, to bound

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

□

The following result is referred to as the elliptic potential lemma, and it can be viewed as a generalization of Lemma 3.2.

Lemma 4.2 (Elliptic potential lemma): Let $a_1, \dots, a_T \in \mathbb{R}^d$ satisfy $\|a_t\| \leq 1$ for all $t \in [T]$, and let $V_t = I + \sum_{s \leq t} a_s a_s^\top$. Then

$$\sum_{t=1}^T \|a_t\|_{V_{t-1}^{-1}}^2 \leq 2d \log(1 + T/d). \quad (4.30)$$

More generally, for any $\lambda > 0$, with $V_t^\lambda = \lambda I + \sum_{s < t} a_s a_s^\top$,

$$\sum_{t=1}^T \log\left(1 + \|a_t\|_{(V_t^\lambda)^{-1}}^2\right) \leq d \log(1 + \lambda^{-1}T/d). \quad (4.31)$$

Proof of Lemma 4.2. Set $u_t = a_t^\top V_{t-1}^{-1} a_t$. Since $V_{t-1} \succeq I$ and $\|a_t\|_2 \leq 1$, we have $0 \leq u_t \leq 1$. The matrix determinant lemma gives $\det V_t = \det V_{t-1}(1 + u_t)$. Also, $u \leq 2 \log(1 + u)$ on $[0, 1]$, since the derivative of $2 \log(1 + u) - u$ is $(1 - u)/(1 + u) \geq 0$. Therefore

$$\sum_{t=1}^T u_t \leq 2 \sum_{t=1}^T \log(1 + u_t) = 2 \log \frac{\det V_T}{\det V_0} = 2 \log \det V_T.$$

We further observe that if $\lambda_1, \dots, \lambda_d$ are the positive eigenvalues of V_T , the arithmetic-geometric mean inequality implies

$$\det V_T = \prod_{j=1}^d \lambda_j \leq \left(\frac{\text{tr } V_T}{d} \right)^d = \left(1 + \frac{1}{d} \sum_{t=1}^T \|a_t\|_2^2 \right)^d \leq (1 + T/d)^d.$$

Combining the two displays proves the claim. □

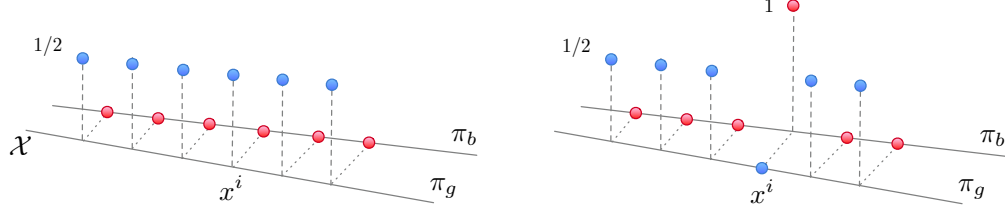


Figure 8: The construction in Example 4.1. The true function f^* (left) and a “defector” f_i (right).

To summarize, the analysis of LinUCB combined two ingredients that carry forward: confidence sets defined through excess square loss (Lemma 4.1), and a potential argument (Lemma 4.2) showing that the confidence widths at the selected actions cannot remain large. The elliptic potential lemma reappears in the analysis of LSVI-UCB in Section 9.2.

4.4 Failure of Optimism Beyond Linear Classes

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. This construction is from Foster et al. [80].

Example 4.1 (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 $\Pi = \{\pi_g, \pi_b\}$. Let $\mathcal{X} = \{x_1, \dots, x_N\}$ be a set of distinct contexts, and define a class $\mathcal{F} = \{f^*, f_1, \dots, f_N\}$ of cardinality $N + 1$ as follows. Let $f^*(x, \pi_g) = 1/2$ and $f^*(x, \pi_b) = 0$ for any $x \in \mathcal{X}$. For each $i \in [N]$, $f_i(x_j, \pi_g) = 1/2$ and $f_i(x_j, \pi_b) = 0$ for $j \neq i$, while $f_i(x_i, \pi_g) = 0$ and $f_i(x_i, \pi_b) = 1$; see Figure 8 for an illustration.

Now, consider a (realizable) problem instance in which rewards are deterministic and given by

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

which we note is a constant function with respect to the context. Since f^* is the true model, π_g is always the best action, bringing a reward of $1/2$ per round. Any time π_b is chosen, the decision-maker incurs instantaneous regret $1/2$. We now argue that if we apply the generic optimistic algorithm from Section 4.2, it chooses π_b every time a new context is encountered, incurring $\Omega(N)$ regret.

Let S^t be the set of distinct contexts encountered before round t . Clearly, the exact minimizers of empirical square loss (see (4.10)) are f^* , 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 (4.11) 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 (defined in (4.4)) than π_g , since

$$\bar{f}^t(x^t, \pi_b) = f_i(x_i, \pi_b) = 1 > \bar{f}^t(x^t, \pi_g) = f^*(x^t, \pi_g) = 1/2.$$

Hence, the cumulative regret grows by $1/2$ every time a new context is presented, and thus scales as $\Omega(N/2)$ 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 $\mathcal{F}$.

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

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

◁

This stands in contrast to supervised learning, where empirical risk minimization attains excess risk scaling with $\log|\mathcal{F}|$ for every finite class (Proposition 2.1). It is precisely the interactive nature of the problem that breaks the classical guarantee.

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 (4.7) do not shrink quickly enough. In what follows, we develop alternative methods which *do* enjoy logarithmic dependence on the size of $\mathcal{F}$, with the best of these methods (Section 4.7) 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. Section 5 introduces the eluder dimension, a structural property of the class $\mathcal{F}$ which generalizes this notion, and is sufficient for optimistic methods to achieve low regret.

4.5 Reduction to Online Regression

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 $\mathcal{F}$. This subsection develops the tool behind such principles: a reduction of decision making to *online regression*. Sections 4.6 and 4.7 then present two algorithms built on it.

For motivation, inspecting Theorem 4.1, the meaning of β^2 in (4.11), and the statistical consequence of Lemma 4.1, we observe that the regret bound of LinUCB “morally” has the form

$$\mathbf{Reg} \lesssim \left(d \cdot \sum_{\tau=1}^T \sum_{t=1}^{\tau} \mathbb{E}_{\pi^t \sim p^t} \left[(\hat{f}^{\tau+1}(x^t, \pi^t) - f^*(x^t, \pi^t))^2 \right] \right)^{1/2}$$

where $\hat{f}^{\tau+1}$ is the least squares (or, supervised learning with square loss) solution with respect to $(x^1, \pi^1, r^1), \dots, (x^\tau, \pi^\tau, r^\tau)$. However, Example 4.1 indicates that relying on past data too much is not a viable approach in general. Indeed, even if covariates x^t are i.i.d., our decisions π^t are not, and—according to Proposition 2.3—ERM may not have low regret. In hindsight, the fact that we were able to base our result for LinUCB on simple “offline” regression appears to be more of a happy accident than an instance of a deeper principle.

To make progress, we build upon *online* supervised learning as opposed to offline. To make the use of supervised learning as modular as possible, we the specific algorithm away using the notion of an *online regression oracle* [79].

Definition 4.1 (Online Regression Oracle): At each time $t \in [T]$, an *online regression oracle* returns, given

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

with $\mathbb{E}[r^i | x^i, \pi^i] = f^*(x^i, \pi^i)$ and $\pi^i \sim p^i$, a function $\hat{f}^t : \mathcal{X} \times \Pi \rightarrow \mathbb{R}$. Define the cumulative estimation error

$$\mathbf{Est}_{\text{Sq}} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left[(\hat{f}^t(x^t, \pi^t) - f^*(x^t, \pi^t))^2 \right].$$

The oracle is required to ensure that

$$\mathbf{Est}_{\text{Sq}} \leq \mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$$

with probability at least $1 - \delta$, where $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ is a deterministic bound. 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.

The quantity $\mathbf{Est}_{\text{Sq}}$ is the cumulative squared estimation error from (2.16), now averaged over the action at each round; it is a random variable that depends on the realized data. We use $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ for a deterministic upper bound on this error that holds with probability at least $1 - \delta$. Throughout, the bare symbol refers to the realized error and the argument form to the guarantee; results whose proofs are pathwise are stated in terms of the former, and the latter is substituted in when tuning parameters.

We can construct such an oracle using the online learning methods from Section 2.2. The online predictor aims to keep its square-loss regret $\mathbf{Reg}_{\text{OL}}^{\text{Sq}}$ small when predicting r^t from (x^t, π^t) . Under realizability, Lemma 2.3 converts this regret guarantee into an expected squared prediction-error bound. Freedman's inequality (Lemma A.3 in the appendix) gives the high-probability guarantee required here. For example, for finite classes of $[0, 1]$ -valued functions and bounded rewards, averaged exponential weights has $\mathbf{Reg}_{\text{OL}}^{\text{Sq}} \lesssim \log |\mathcal{F}|$ and yields $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$. Note that we make use of *online regression* oracles for the results that follow because we aim to derive regret bounds that hold for arbitrary, potentially adversarial sequences $x^1, \dots, x^T$. 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 4.7.1 for further discussion.

4.6 A First Attempt: The ε -Greedy Algorithm

With the oracle in hand, what remains is the choice of the exploration distribution p^t as a function of the estimate $\hat{f}^t$. The first general-purpose contextual bandit algorithm we study, illustrated below, is a contextual counterpart to the ε -Greedy method introduced in Section 3.

Algorithm 4.2 ε -Greedy for Contextual Bandits

input: Exploration parameter $\varepsilon \in (0, 1)$; online regression oracle.

for $t = 1, \dots, T$ **do**

Obtain $\hat{f}^t$ from online regression oracle for $(x^1, \pi^1, r^1), \dots, (x^{t-1}, \pi^{t-1}, r^{t-1})$.

Observe x^t .

With prob. ε , select $\pi^t \sim \text{unif}([A])$, and with prob. $1 - \varepsilon$, choose the greedy action

$$\hat{\pi}^t = \arg \max_{\pi \in [A]} \hat{f}^t(x^t, \pi).$$

Observe reward r^t .

At each step t , the algorithm uses an online regression oracle to compute a reward estimator $\hat{f}^t(x, \pi)$ based on the data $\mathcal{H}^{t-1}$ collected so far (the estimator is computed before x^t is revealed, and is then evaluated at x^t). 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

$$\hat{\pi}^t = \arg \max_{\pi} \hat{f}^t(x^t, \pi), \quad (4.32)$$

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

Proposition 4.2 (Regret of ε -Greedy for contextual bandits): Suppose the decision-maker has access to an online regression oracle (Definition 4.1) with a guarantee $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$. Then by choosing ε appropriately, the ε -Greedy algorithm ensures that with probability at least $1 - \delta$,

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

for any sequence $x^1, \dots, x^T$. 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. This is the first regret bound in the monograph obtained by *reducing* decision making to supervised learning: the only property of the oracle used in the proof is its estimation error, so any online regression method can be plugged in. The $T^{2/3}$ -dependence in the regret bound is suboptimal (as seen for the special case of non-contextual bandits), which Section 4.7 addresses via more deliberate exploration.

Proof of Proposition 4.2. 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 3.1, we can bound regret by

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

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

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

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

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

$$\left\{ \sum_{\pi \in \{\hat{\pi}^t, \pi^\star\}} \frac{1}{p^t(\pi)} \right\}^{1/2} \left\{ \sum_{\pi \in \{\hat{\pi}^t, \pi^\star\}} p^t(\pi) \left(f^\star(x^t, \pi) - \hat{f}^t(x^t, \pi) \right)^2 \right\}^{1/2} \quad (4.33)$$

$$\leq \sqrt{\frac{2A}{\varepsilon}} \left\{ \mathbb{E}_{\pi^t \sim p^t} \left(f^\star(x^t, \pi^t) - \hat{f}^t(x^t, \pi^t) \right)^2 \right\}^{1/2}. \quad (4.34)$$

Summing across t , this gives

$$\sum_{t=1}^T f^\star(x^t, \pi^\star(x^t)) - f^\star(x^t, \hat{\pi}^t) \leq \sqrt{\frac{2A}{\varepsilon}} \sum_{t=1}^T \left\{ \mathbb{E}_{\pi^t \sim p^t} \left(f^\star(x^t, \pi^t) - \hat{f}^t(x^t, \pi^t) \right)^2 \right\}^{1/2} \quad (4.35)$$

$$\leq \sqrt{\frac{2AT}{\varepsilon}} \left\{ \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left(f^\star(x^t, \pi^t) - \hat{f}^t(x^t, \pi^t) \right)^2 \right\}^{1/2}. \quad (4.36)$$

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

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

Whenever this occurs, we have

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

Choosing ε to balance the two terms leads to the claimed result (if the balancing value exceeds one, we take $\varepsilon = 1$; in that case $T \lesssim A \mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$, and the claimed bound holds trivially). $\square$

4.7 Inverse Gap Weighting and the Estimation-to-Decisions Principle

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 4.2](#) involved relating the instantaneous regret

$$\mathbb{E}_{\pi^t \sim p^t} [f^*(x^t, \pi^*(x^t)) - f^*(x^t, \pi^t)] \quad (4.37)$$

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

$$\mathbb{E}_{\pi^t \sim p^t} \left[(f^*(x^t, \pi^t) - \hat{f}^t(x^t, \pi^t))^2 \right] \quad (4.38)$$

between $\hat{f}^t$ and f^* 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 (4.37) in terms of (4.38). [Proposition 4.3](#) below establishes exactly that. At first glance, this distribution might appear to be somewhat arbitrary or even magical, but [Sections 5](#) and [7](#) show that it arises as a special case of a more general—and in some sense, universal—principle for designing decision making algorithms. The distribution below was introduced by Abe and Long [\[2\]](#); the form and the use we make of it here follow Foster and Rakhlin [\[79\]](#).

Definition 4.2 (Inverse Gap Weighting): Given a vector $\hat{f} = (\hat{f}(1), \dots, \hat{f}(A)) \in \mathbb{R}^A$, the Inverse Gap Weighting distribution $p = \text{IGW}_\gamma(\hat{f}(1), \dots, \hat{f}(A))$ with parameter $\gamma \geq 0$ is defined as

$$p(\pi) = \frac{1}{\lambda + 2\gamma(\hat{f}(\hat{\pi}) - \hat{f}(\pi))}, \quad (4.39)$$

where $\hat{\pi} = \arg \max_\pi \hat{f}(\pi)$ 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 (4.39). We can interpret the parameter γ as trading off exploration and exploitation. Indeed, $\gamma \rightarrow 0$ gives a uniform distribution, while $\gamma \rightarrow \infty$ amplifies the gap between the greedy action $\hat{\pi}$ and any action with $\hat{f}(\pi) < \hat{f}(\hat{\pi})$, resulting in a distribution supported only on actions that achieve the largest estimated value $\hat{f}(\hat{\pi})$.

The following fundamental technical result shows that playing the Inverse Gap Weighting distribution always suffices to link the instantaneous regret in (4.37) in to the instantaneous estimation error in (4.38).

Proposition 4.3 (Inverse gap weighting links regret to estimation error): Consider a finite decision space $\Pi = \{1, \dots, A\}$. For any vector $\hat{f} \in \mathbb{R}^A$ and $\gamma > 0$, define $p = \text{IGW}_\gamma(\hat{f}(1), \dots, \hat{f}(A))$. This strategy guarantees that for all $f^* \in \mathbb{R}^A$,

$$\mathbb{E}_{\pi \sim p} [f^*(\pi^*) - f^*(\pi)] \leq \frac{A}{\gamma} + \gamma \cdot \mathbb{E}_{\pi \sim p} [(\hat{f}(\pi) - f^*(\pi))^2]. \quad (4.40)$$

Proof of Proposition 4.3. We break the “regret” term on the left-hand side of (4.40) into three terms:

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

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

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

while the second term is at most

$$\sqrt{\mathbb{E}_{\pi \sim p}[(\hat{f}(\pi) - f^*(\pi))^2]} \leq \frac{1}{2\gamma} + \frac{\gamma}{2} \mathbb{E}_{\pi \sim p}(\hat{f}(\pi) - f^*(\pi))^2.$$

The third term can be further written as

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

The term in brackets above is equal to

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

□

The simple result we just proved is remarkable. The special IGW strategy guarantees a relation between regret and estimation error for *any* estimator $\hat{f}$ and any f^* , irrespective of the problem structure or the class $\mathcal{F}$. [Proposition 4.3](#) is the core of the development for the rest of the monograph; [Sections 5](#) and [7](#) greatly generalize it to general decision making problems and reinforcement learning.

Below, we present a contextual bandit algorithm called **SquareCB** [\[79\]](#) which makes use of the Inverse Gap Weighting distribution to achieve optimal regret.

Algorithm 4.3 SquareCB

input: Exploration parameter $\gamma > 0$; online regression oracle.

for $t = 1, \dots, T$ **do**

 Obtain $\hat{f}^t$ from online regression oracle with $(x^1, \pi^1, r^1), \dots, (x^{t-1}, \pi^{t-1}, r^{t-1})$.

 Observe x^t .

 Compute $p^t = \text{IGW}_{\gamma}(\hat{f}^t(x^t, 1), \dots, \hat{f}^t(x^t, A))$.

 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 $\hat{f}^t(x, \pi)$ based on the data $\mathcal{H}^{t-1}$ collected so far. Given this estimator, the algorithm uses Inverse Gap Weighting to compute $p^t = \text{IGW}_{\gamma}(\hat{f}^t(x^t, \cdot))$ as an exploratory distribution, then samples $\pi^t \sim p^t$.

The following result, which is a near-immediate consequence of [Proposition 4.3](#), gives a regret bound for this algorithm.

Theorem 4.2 (Regret of SquareCB): Given a class $\mathcal{F}$ with $f^* \in \mathcal{F}$, for any $\gamma > 0$ and any sequence $x^1, \dots, x^T$, SquareCB satisfies

$$\mathbf{Reg} \leq \frac{TA}{\gamma} + \gamma \cdot \mathbf{Est}_{\text{Sq}}$$

for every realization. In particular, if the decision-maker has access to an online regression oracle (Definition 4.1) with estimation error $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$, then SquareCB with $\gamma = \sqrt{TA/\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)}$ attains a regret bound of

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

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

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

Proof of Theorem 4.2. We begin with regret, then add and subtract the squared estimation error as follows:

$$\begin{aligned} \mathbf{Reg} &= \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^*(x^t, \pi^*) - f^*(x^t, \pi^t)] \\ &\leq \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left[f^*(x^t, \pi^*) - f^*(x^t, \pi^t) - \gamma \cdot (f^*(x^t, \pi^t) - \hat{f}^t(x^t, \pi^t))^2 \right] + \gamma \cdot \mathbf{Est}_{\text{Sq}}. \end{aligned}$$

By appealing to Proposition 4.3 with $\hat{f}(x^t, \cdot)$ and $f^*(x^t, \cdot)$, for each step t , we have

$$\mathbb{E}_{\pi^t \sim p^t} \left[f^*(x^t, \pi^*) - f^*(x^t, \pi^t) - \gamma \cdot (f^*(x^t, \pi^t) - \hat{f}^t(x^t, \pi^t))^2 \right] \leq \frac{A}{\gamma},$$

and thus

$$\mathbf{Reg} \leq \frac{TA}{\gamma} + \gamma \cdot \mathbf{Est}_{\text{Sq}}.$$

This holds for every realization. On the event, of probability at least $1 - \delta$, that the oracle guarantee $\mathbf{Est}_{\text{Sq}} \leq \mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ holds, choosing γ to balance the two terms yields the result. $\square$

If the online regression oracle is minimax optimal (that is, $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ 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 monograph: 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 [79] for details.

Interpretation: The Estimation-to-Decisions principle. SquareCB is the first instance of a design principle that organizes the rest of this monograph, which we call *Estimation-to-Decisions*. The principle has two ingredients: an *estimation oracle*, a supervised (online) learning algorithm that produces an estimate $\hat{f}^t$ for the unknown model with no regard for exploration, and an *exploration strategy*, a fixed rule that maps $\hat{f}^t$ to a distribution p^t over decisions. The central property is

that the exploration strategy ensures that our performance for decision making is controlled by our performance for estimation. For Inverse Gap Weighting, this is the inequality in [Proposition 4.3](#),

$$\underbrace{\mathbb{E}_{\pi \sim p}[f^*(\pi^*) - f^*(\pi)]}_{\text{regret of } p} \leq \frac{A}{\gamma} + \gamma \cdot \underbrace{\mathbb{E}_{\pi \sim p}[(\hat{f}(\pi) - f^*(\pi))^2]}_{\text{estimation error of } \hat{f} \text{ under } p} \quad \text{for all } \hat{f} \text{ and } f^*, \quad (4.41)$$

which, summed over rounds, bounds the regret of the algorithm by the estimation error of the oracle on the data the algorithm itself collects. That this is possible at all should come as a surprise: the estimate may be arbitrarily wrong on decisions the algorithm never plays, but (4.41) says that with the right choice for p , such errors cannot hurt.

The value of this reduction is that it separates the two sources of difficulty: the estimation oracle handles generalization across the class $\mathcal{F}$ (and can be any method from statistical or online learning), while the exploration strategy handles exploration. It also makes the algorithm *oracle-efficient*: each round costs one call to the regression oracle and $O(A)$ additional arithmetic. Generalizing this principle, in [Section 5](#) we ask what the best possible exploration strategy is for a *general* decision space Π ; the answer is given by a complexity measure, the Decision-Estimation Coefficient, which systematically generalizes the IGW distribution to general decision spaces. The principle is developed in even greater generality—subsuming reinforcement learning—in [Section 7](#), and [Section 8](#) further shows that it leads to a universal lower bound on regret for decision making.

This concludes the main development of this section. The remaining subsection covers an extension and can be skipped on a first reading.

4.7.1 Extension: SquareCB with Offline Regression Oracles*

When $x^1, \dots, x^T$ 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 4.3 (Offline Regression Oracle): Given

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

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

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

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 [189]. 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.

Algorithm 4.4 SquareCB with Offline Regression

input: Exploration parameters $\gamma_1, \gamma_2, \dots$ and epoch sizes $\tau_1, \tau_2, \dots$; offline regression oracle.

for $m = 1, 2, \dots$ **do**

 Obtain $\hat{f}^m$ from offline regression oracle with
 $(x^{\tau_{m-2}+1}, \pi^{\tau_{m-2}+1}, r^{\tau_{m-2}+1}), \dots, (x^{\tau_{m-1}}, \pi^{\tau_{m-1}}, r^{\tau_{m-1}})$.

for $t = \tau_{m-1} + 1, \dots, \tau_m$ **do**

 Observe x^t .

 Compute $p^t = \text{IGW}_{\gamma_m}(\hat{f}^m(x^t, 1), \dots, \hat{f}^m(x^t, A))$.

 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. The key challenge is that, while the contexts $x^1, \dots, x^T$ are i.i.d., the decisions $\pi^1, \dots, \pi^T$ evolve in a time-dependent fashion, which makes it unclear to invoke the guarantee in Definition 4.3. Nonetheless, the following remarkable result, due to Simchi-Levi and Xu [189], shows that this algorithm attains a regret bound similar to that of Theorem 4.2.

Proposition 4.4 (Regret of SquareCB with offline oracles): For $m = 1, 2, \dots$, let $\tau_m = 2^m$ and $\gamma_m = \sqrt{A\tau_{m-1}/\text{Est}_{\text{Sq}}^{\text{off}}(\mathcal{F}, \tau_{m-1}, \delta)}$. Then with probability at least $1 - \delta$, regret of SquareCB with an offline oracle is at most

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

Under mild assumptions, above bound scales as

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

For a finite class $\mathcal{F}$, we recall from Section 2 that empirical risk with the square loss (least squares) achieves $\text{Est}_{\text{Sq}}^{\text{off}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$, which gives

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

Thus, when contexts are i.i.d., the Estimation-to-Decisions reduction goes through with an offline oracle as well, at the price of an epoch schedule and a more delicate analysis.

4.8 Bibliographic Remarks and Further Reading

Takeaways. This section showed that supervised learning algorithms can be lifted into interactive decision making algorithms without sacrificing statistical efficiency: with a realizable class $\mathcal{F}$, one can achieve regret scaling with $\log|\mathcal{F}|$ and no dependence on the number of contexts. Optimism accomplishes this for linear classes (LinUCB), but fails for general classes (Example 4.1). The key idea is the Estimation-to-Decisions principle: pair an off-the-shelf regression oracle with an exploration rule, Inverse Gap Weighting, that guarantees that the regret of the chosen distribution is controlled by the estimation error of the oracle under that same distribution (Proposition 4.3). SquareCB is the simplest instance of this principle, and the rest of the monograph generalizes this to structured bandits and reinforcement learning.

Bibliographic remarks. The optimism template of [Section 4.2](#), with confidence sets defined through excess square loss ([Lemma 4.1](#)), follows Russo and Van Roy [180]. Optimism for linear contextual bandits originates with Auer [21]; the name LinUCB and the algorithm in [Algorithm 4.1](#) are from Li et al. [145], with the first regret analysis given by Chu et al. [57]. The confidence ellipsoids and elliptic potential lemma behind [Theorem 4.1](#) are due to Dani et al. [61] and Abbasi-Yadkori et al. [1]. [Example 4.1](#) is from Foster et al. [80]. The $T^{2/3}$ rate of [Proposition 4.2](#) for uniform exploration ([Algorithm 4.2](#)) with a supervised learning oracle goes back to the epoch-greedy algorithm of Langford and Zhang [133]; the online regression oracle ([Definition 4.1](#)) is the abstraction of Foster and Rakhlin [79]. The Inverse Gap Weighting distribution (4.39) was introduced by Abe and Long [2]; [Proposition 4.3](#) for linear contextual bandits. SquareCB ([Algorithm 4.3](#)), and [Theorem 4.2](#) are due to Foster and Rakhlin [79], who observed that inverse gap weighting encompasses the general contextual bandit problem and is minimax optimal. The offline variant of SquareCB, [Algorithm 4.4](#), and its regret bound are due to Simchi-Levi and Xu [189], who termed this version of the algorithm as FALCON. The perspective of *Estimation-to-Decisions* as the general principle behind these methods is from Foster et al. [84].

Further reading. Textbook treatments of linear and contextual bandits are given by Lattimore and Szepesvári [139] and Slivkins [191]. Refinements of SquareCB include instance-dependent regret bounds governed by a disagreement coefficient [85], first-order bounds [78], and adaptation to misspecification [83]. Not covered in this monograph is the *agnostic formulation* of the contextual bandit problem, where one competes with a class of policies rather than assuming realizability. This is the setting of [Exp4](#) [23] and of the oracle-efficient algorithm of Agarwal et al. [5].

4.9 Deferred Proofs

Proof of Proposition 4.1. The case $T = 1$ is immediate from $\mathbf{Reg} \leq 1$, so assume $T \geq 2$. For each pair (x, a) , consider an independent stream of rewards with distribution $M^*(\cdot | x, a)$, revealing the next entry whenever that pair is selected. This is equivalent to the stochastic multi-armed bandit protocol, even when contexts and actions depend on the past. Let $\bar{r}_{x,a,n}$ be the mean of the first n entries of this stream. Hoeffding's inequality and a union bound over KAT choices of (x, a, n) give, for $L = \log(2AKT^2)$,

$$\mathbb{P} \left(\exists x, a, 1 \leq n \leq T : |\bar{r}_{x,a,n} - f^*(x, a)| > \sqrt{\frac{L}{2n}} \right) \leq 2KATe^{-L} = \frac{1}{T}.$$

Condition on the complement of this event. Once every action has been tried for a given context, the usual optimistic comparison shows that choosing an action with $n \geq 1$ previous observations incurs regret at most $2\sqrt{L/(2n)}$. When $n = 0$, UCB chooses an untried action, whose regret is at most one.

Let $N_{x,a}$ be the number of selections of (x, a) up to time T and $N_x = \sum_a N_{x,a}$. The first selections contribute at most $\min\{A, N_x\} \leq \sqrt{AN_x}$ in context x . Since $\sum_{n=1}^{m-1} n^{-1/2} \leq 2\sqrt{m}$, the remaining selections give

$$\mathbf{Reg}_x \leq \sqrt{AN_x} + 2\sqrt{2L} \sum_{a=1}^A \sqrt{N_{x,a}} \leq (1 + 2\sqrt{2L})\sqrt{AN_x}.$$

Summing over contexts and applying Cauchy–Schwarz once more yields

$$\mathbf{Reg} \leq (1 + 2\sqrt{2L})\sqrt{A} \sum_x \sqrt{N_x} \leq (1 + 2\sqrt{2L})\sqrt{AKT}.$$

Taking expectations and using $L \lesssim \log(eAKT)$ proves the result. $\square$

LinUCB with continuous parameter sets. We now extend [Theorem 4.1](#) to continuous parameter sets. The idea is to approximate the parameter set by finitely many points, apply the concentration argument to those points, and then bound the error from the approximation. We begin with the elementary covering lemma needed for this argument.

Lemma 4.3 (Finite nets for linear predictors): For any nonempty $\Theta \subseteq \mathbb{B}_2^d(1)$ and $0 < \varepsilon \leq 1$, there is a finite $\Theta_\varepsilon \subseteq \Theta$ such that

$$|\Theta_\varepsilon| \leq (1 + 2/\varepsilon)^d, \quad \forall \theta \in \Theta \quad \min_{\theta' \in \Theta_\varepsilon} \|\theta - \theta'\|_2 \leq \varepsilon.$$

For any such approximating pair and any feature map with $\|\phi(x, \pi)\|_2 \leq 1$,

$$|f_\theta(x, \pi) - f_{\theta'}(x, \pi)| \leq \varepsilon \quad \text{for all } (x, \pi).$$

Proof of Lemma 4.3. Choose a maximal collection $\Theta_\varepsilon \subseteq \Theta$ whose distinct points are more than ε apart. The open Euclidean balls of radius $\varepsilon/2$ centered at these points are disjoint and contained in the ball of radius $1 + \varepsilon/2$. Comparing their volumes gives

$$|\Theta_\varepsilon|(\varepsilon/2)^d \leq (1 + \varepsilon/2)^d,$$

so the collection is finite with the claimed cardinality. In particular, greedily adding separated points must terminate, which establishes existence even when Θ is not closed. Maximality makes this collection an ε -net. Finally, Cauchy–Schwarz gives

$$|f_\theta(x, \pi) - f_{\theta'}(x, \pi)| = |\langle \theta - \theta', \phi(x, \pi) \rangle| \leq \|\theta - \theta'\|_2 \|\phi(x, \pi)\|_2 \leq \varepsilon.$$

$\square$

Proof of Theorem 4.1, continuous parameter sets. Recall that in this setup, LinUCB still computes least squares over the full set Θ ; we appeal to covering numbers and a finite approximation only to establish the confidence set guarantee.

Fix $0 < \varepsilon \leq 1$ and the net Θ_ε from [Lemma 4.3](#); write $N = |\Theta_\varepsilon|$. For any $\theta \in \Theta$, define

$$U_i(\theta) = (f_\theta(x^i, \pi^i) - r^i)^2 - (f^*(x^i, \pi^i) - r^i)^2, \quad D_i(\theta) = \mathbb{E}_{i-1}[U_i(\theta)].$$

Here $\mathbb{E}_{i-1}$ conditions on the past and the current context, before choosing the action and observing its reward. The stochastic reward assumption gives

$$D_i(\theta) = \mathbb{E}_{\pi^i \sim p^i}[(f_\theta(x^i, \pi^i) - f^*(x^i, \pi^i))^2].$$

All rewards and predictions lie in $[-1, 1]$. Consequently, $|D_i(\theta) - U_i(\theta)| \leq 8$ and

$$\mathbb{E}_{i-1}[(D_i(\theta) - U_i(\theta))^2] \leq \mathbb{E}_{i-1}[U_i(\theta)^2] \leq 16D_i(\theta).$$

Apply [Lemma A.3](#) with $\eta = 1/32$ to each fixed $\theta' \in \Theta_\varepsilon$, and take a union bound over the net. With probability at least $1 - \delta$, simultaneously for all net points $\theta' \in \Theta_\varepsilon$ and all $0 \leq n \leq T$,

$$\frac{1}{2} \sum_{i=1}^n D_i(\theta') \leq \sum_{i=1}^n U_i(\theta') + 32 \log(N/\delta).$$

The comparison function in this argument is the true f^* ; it need not belong to the net.

For any $\theta \in \Theta$, choose θ' with $\|\theta - \theta'\|_2 \varepsilon$. Factoring differences of squared losses, and using boundedness, gives

$$|U_i(\theta) - U_i(\theta')| \leq 4\varepsilon, \quad |D_i(\theta) - D_i(\theta')| \leq 4\varepsilon.$$

Thus the same event ensures, for every $\theta \in \Theta$ and prefix $n \leq T$,

$$\frac{1}{2} \sum_{i=1}^n D_i(\theta) \leq \sum_{i=1}^n U_i(\theta) + b, \quad b = 32 \log(N/\delta) + 6T\varepsilon. \quad (4.42)$$

Let $\beta^2 \geq b$, and consider the confidence sets (4.11) over the original class. Compactness of Θ ensures that the least-squares minimizer exists. Applying (4.42) to that minimizer, and using $D_i \geq 0$, shows that its cumulative loss is at least the cumulative loss of f^* minus b . Hence $f^* \in \mathcal{F}^t$ for every t . Every $\theta \in \mathcal{F}^t$ also has $\sum_{i < t} U_i(\theta) \leq \beta^2$, because the empirical minimizer has no greater loss than f^* . Therefore

$$\sum_{i < t} D_i(\theta) \leq 2(\beta^2 + b) \leq 4\beta^2.$$

This is exactly the confidence guarantee used in the finite-class LinUCB proof. Since LinUCB chooses its maximizing action deterministically (with fixed tie-breaking), these conditional squared errors are the quadratic forms in its empirical covariance matrix. The rest of the proof and Lemma 4.2 then give

$$\mathbf{Reg} \lesssim \sqrt{(\beta^2 + 1)dT \log(1 + T/d)}.$$

Finally, setting $\varepsilon = 1/T$, the ε -net bound gives $b \leq 32[d \log(1 + 2T) + \log(1/\delta)] + 6$, proving the stated result. $\square$

5. STRUCTURED BANDITS: EXPLORATION IN LARGE DECISION SPACES

Contextual bandits generalize across *contexts*; this section turns to the other axis of [Figure 2](#), generalizing across *decisions*. In the *structured bandit* problem, the decision space may be continuous or combinatorially large, and the learner must use structure in the reward landscape to infer the rewards of decisions it may have never played from those it has. Compared to contextual bandits, the situation is more subtle: regret must scale with $\log|\mathcal{F}|$ *and* with a measure of how hard the class $\mathcal{F}$ is to explore, and the right way to explore depends on $\mathcal{F}$. The section’s main tool is the Decision-Estimation Coefficient, which identifies the best possible exploration strategy for any class, and the resulting algorithm, Estimation-to-Decisions, which turns the principle of [Section 4](#) into a single method for that can be applied to all structured bandit problems.

Organization. [Section 5.1](#) formalizes the structured bandit problem and desiderata. [Section 5.2](#) adapts the principle of optimism to the structured setting and analyzes it based on a complexity measure called the eluder dimension, but shows that it can be exponentially suboptimal in general. [Section 5.3](#) introduces the Decision-Estimation Coefficient and the E2D algorithm, and shows that Inverse Gap Weighting is the special case for finitely many actions. [Section 5.4](#) computes the DEC for examples including linear bandits and nonparametric bandits. Lastly, [Section 5.5](#) relates the DEC back to optimism and to posterior sampling. Two optional extensions (marked $\star$) follow: [Section 5.6](#) adds contexts, and [Section 5.7](#) collects structural properties of the DEC.

5.1 The Structured Bandit Problem

The structured bandit problem generalizes the (non-contextual) multi-armed bandit problem by allowing for large, potentially infinite or continuous decision spaces. The protocol is as follows.

Structured Bandit Protocol

```

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

The only change from the multi-armed bandit protocol of [Section 3](#) is that Π is no longer restricted to $\{1, \dots, A\}$. 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 be a high-dimensional vector (such as dosages for many different medications). See [Figure 9](#).
- ▷ 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.

Assumptions and regret. As in preceding sections, we assume that rewards are stochastic, and generated from a fixed model.

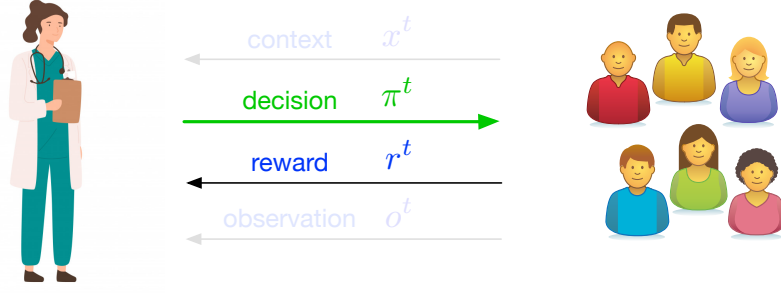


Figure 9: The structured bandit problem. A doctor (the learner) selects a continuous, high-dimensional treatment (the decision).

Assumption 5.1 (Stochastic Rewards): Rewards are generated independently via

$$r^t \sim M^*(\cdot \mid \pi^t), \quad (5.1)$$

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

We define

$$f^*(\pi) := \mathbb{E}[r \mid \pi] \quad (5.2)$$

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

$$\mathbf{Reg} := \sum_{t=1}^T f^*(\pi^*) - \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t}[f^*(\pi^t)]. \quad (5.3)$$

Here, $\pi^* := \arg \max_{\pi \in \Pi} f^*(\pi)$ as usual. We define the history as $\mathcal{H}^t = (\pi^1, r^1), \dots, (\pi^t, r^t)$. As in the preceding sections, we assume that $r^t \in [0, 1]$ almost surely, and regret bounds hold with probability at least $1 - \delta$ unless stated otherwise.

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 $\tilde{O}(\sqrt{|\Pi|T})$, which could be vacuous if Π is large relative to T . However, with no further assumptions on the underlying reward function f^* , this is unavoidable. To allow for better regret, we make assumptions on the structure of f^* that 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^* 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 *realizable* function class $\mathcal{F}$ that aims to capture our prior knowledge about f^* .

Assumption 5.2 (Realizability): The decision-maker has access to a class $\mathcal{F} \subset \{f : \Pi \rightarrow \mathbb{R}\}$ such that $f^* \in \mathcal{F}$.

As for rewards, we assume that $f(\pi) \in [0, 1]$ for all $f \in \mathcal{F}$ unless stated otherwise. Given such a class, a reasonable goal—particularly in light of the development in [Section 2](#) and [Section 4](#)—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 5.1 (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 $\mathbf{Reg} \gtrsim A \wedge T$ 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 . $\triangleleft$

What this example highlights is that generalizing across decisions is fundamentally different (and, in some sense, more challenging) than generalizing across contexts (the two can also be combined; see [Section 5.6](#)). In light of this, we 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 $\mathcal{F}$. The complexity measure we arrive at, the Decision-Estimation Coefficient, does exactly this, and characterizes the optimal regret for structured bandits up to the $\log|\mathcal{F}|$ estimation term ([Section 8](#)).

5.2 Warm-Up: Optimism and the Eluder Dimension

To see what such guarantees look like and how they can be obtained, we first investigate the optimism principle and the UCB algorithm in the structured setting. We show 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.

5.2.1 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 [\[180\]](#). Recall that $f(\pi) \in [0, 1]$ for all $f \in \mathcal{F}$ and $r^t \in [0, 1]$ almost surely. Let

$$\hat{f}^t = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} (f(\pi^i) - r^i)^2 \quad (5.4)$$

be the empirical minimizer on round t , and with $\beta^2 := 8 \log(|\mathcal{F}|/\delta)$, define confidence sets $\mathcal{F}^1 = \mathcal{F}$ and

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} (f(\pi^i) - r^i)^2 \leq \sum_{i=1}^{t-1} (\hat{f}^t(\pi^i) - r^i)^2 + \beta^2 \right\}. \quad (5.5)$$

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

$$\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi). \quad (5.6)$$

When does the confidence width shrink? Using Lemma 4.1, one can see the generalized UCB algorithm ensures that $f^* \in \mathcal{F}^t$ for all t with high probability. Whenever this happens, regret is bounded by the upper confidence width:

$$\mathbf{Reg} \leq \sum_{t=1}^T \bar{f}^t(\pi^t) - f^*(\pi^t). \quad (5.7)$$

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 (5.7) 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

$$\mathcal{F} = \left\{ \pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta \subset \mathbb{B}_2^d(1) \right\}, \quad (5.8)$$

Theorem 4.1 shows that we can bound (5.7) 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?

5.2.2 Eluder Dimension: Bounding Confidence Widths

The eluder dimension [180] 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 (5.7). It is defined for a class $\mathcal{F}$ as follows.

Definition 5.1 (Eluder Dimension): Let $\mathcal{F} \subset (\Pi \rightarrow \mathbb{R})$ and $f^* : \Pi \rightarrow \mathbb{R}$ be given, and define $\underline{\text{Edim}}_{f^*}(\mathcal{F}, \varepsilon)$ as the length of the longest sequence of decisions $\pi^1, \dots, \pi^d \in \Pi$ such that for all $t \in [d]$, there exists $f^t \in \mathcal{F}$ such that

$$|f^t(\pi^t) - f^*(\pi^t)| > \varepsilon, \quad \text{and} \quad \sum_{i < t} (f^t(\pi^i) - f^*(\pi^i))^2 \leq \varepsilon^2. \quad (5.9)$$

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

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^* well on all of the preceding points. In particular, if we form confidence sets as in (5.5) with $\beta = \varepsilon$, then the number of times the upper confidence width in (5.7) can be larger than ε is at most $\text{Edim}_{f^*}(\mathcal{F}, \varepsilon)$. We consider the definition $\text{Edim}_{f^*}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \underline{\text{Edim}}_{f^*}(\mathcal{F}, \varepsilon') \vee 1$ instead of directly working with $\underline{\text{Edim}}_{f^*}(\mathcal{F}, \varepsilon)$ to ensure monotonicity with respect to ε , which is used in the proof of Proposition 5.1.

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 $\text{Edim}(\mathcal{F}, \varepsilon)$ and $\log |\mathcal{F}|$.

Proposition 5.1 (Regret of UCB via the eluder dimension): For a finite set of functions $\mathcal{F} \subset (\Pi \rightarrow [0, 1])$, using $\beta^2 = 8 \log(|\mathcal{F}|/\delta)$, the generalized UCB algorithm guarantees that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \min_{\varepsilon > 0} \left\{ \sqrt{\text{Edim}(\mathcal{F}, \varepsilon) \cdot T \log(|\mathcal{F}|/\delta)} + \varepsilon T \right\} \lesssim \sqrt{\text{Edim}(\mathcal{F}, T^{-1/2}) \cdot T \log(|\mathcal{F}|/\delta)}. \quad (5.10)$$

For the case of linear models in (5.8), it is possible to use the elliptic potential lemma (Lemma 4.2) to show that

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

For finite classes, this gives $\mathbf{Reg} \lesssim \sqrt{dT \log(|\mathcal{F}|/\delta) \log(T)}$, which recovers the guarantee in Theorem 4.1. Another well-known example is that of *generalized linear models*. Here, we fix *link function* $\sigma : [-1, +1] \rightarrow \mathbb{R}$ and define

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

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 [180] 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

$$\text{Edim}(\mathcal{F}, \varepsilon) \lesssim \frac{L^2}{\mu^2} \cdot d \log(\varepsilon^{-1}). \quad (5.11)$$

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 $\sigma(z) = +\text{relu}(z)$ or $\sigma(z) = -\text{relu}(z)$ (either choice of sign works), where $\text{relu}(z) := \max\{z, 0\}$ is the ReLU function; this can be interpreted as a neural network with a single neuron. Here, we can have $\sigma'(z) = 0$, so (5.11) does not apply, and it turns out [144] that

$$\text{Edim}(\mathcal{F}, \varepsilon) \gtrsim e^d \quad (5.12)$$

for constant ε . That is, even for a single ReLU neuron, the eluder dimension is already exponential, which is a bit disappointing. Fortunately, the eluder dimension can be overly pessimistic: Section 5.4 shows that it is possible to do better for this class, but this requires changing the algorithm.

Proof of Proposition 5.1. Define

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

By Lemma 4.1, we have that with probability at least $1 - \delta$, for all t :

1. $f^* \in \bar{\mathcal{F}}^t$.
2. $\mathcal{F}^t \subseteq \bar{\mathcal{F}}^t$.

Let us condition on this event. As in [Lemma 3.1](#), since $f^\star \in \mathcal{F}^t$, we can upper bound

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

Now, define

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

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

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

We now appeal to the following technical lemma concerning the eluder dimension, which is Russo and Van Roy [[180](#), Lemma 3].

Lemma 5.1 (Eluder dimension controls confidence widths): Fix a function class $\mathcal{F}$, function $f^\star \in \mathcal{F}$, and parameter $\beta > 0$. For any sequence $\pi^1, \dots, \pi^T$, if we define

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

then for all $\alpha > 0$,

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

Note that for the special case where $\beta = \alpha$, the bound in [Lemma 5.1](#) 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 $\frac{\beta^2}{\alpha^2}$ that grows large when $\alpha \ll \beta$.

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

$$\sum_{t=1}^T w^t(\pi^t) \leq \sum_{t=1}^T w^t(\pi^t) \mathbb{I}\{w^t(\pi^t) > \varepsilon\} + \varepsilon T. \quad (5.13)$$

Let us order the indices $\{1, \dots, T\}$ as $\{i_1, \dots, i_T\}$, so that $w^{i_1}(\pi^{i_1}) \geq w^{i_2}(\pi^{i_2}) \geq \dots \geq w^{i_T}(\pi^{i_T})$. 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 5.1](#) (since $\alpha \leq 1 \leq \beta^2$) implies that

$$\tau \leq \sum_{t=1}^T \mathbb{I}\{w^t(\pi^t) > \alpha\} \leq \left(\frac{4\beta^2}{\alpha^2} + 1 \right) \underline{\text{Edim}}_{f^\star}(\mathcal{F}, \alpha) \leq \frac{5\beta^2}{\alpha^2} \underline{\text{Edim}}_{f^\star}(\mathcal{F}, \alpha). \quad (5.14)$$

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

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

With this, we can bound the main term in (5.13) by

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

Combining this with (5.13) gives $\mathbf{Reg} \lesssim \beta \sqrt{\text{Edim}(\mathcal{F}, \varepsilon) T} + \varepsilon T$. Since $\varepsilon > 0$ was arbitrary, we are free to minimize over it. $\square$

Proof of Lemma 5.1. Let us adopt the shorthand $d = \text{Edim}_{f^*}(\mathcal{F}, \alpha)$. We begin with a definition. We say π is α -independent of $\pi^1, \dots, \pi^t$ if there exists $f \in \mathcal{F}$ such that $|f(\pi) - f^*(\pi)| > \alpha$ and $\sum_{i=1}^t (f(\pi^i) - f^*(\pi^i))^2 \leq \alpha^2$. We say π is α -dependent on $\pi^1, \dots, \pi^t$ if for all $f \in \mathcal{F}$ with $\sum_{i=1}^t (f(\pi^i) - f^*(\pi^i))^2 \leq \alpha^2$, $|f(\pi) - f^*(\pi)| \leq \alpha$.

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

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

If there are M such disjoint sequences, we have

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

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

Next, we claim that for τ and any sequence $(\pi^1, \dots, \pi^\tau)$, there is some j such that π^j is α -dependent on at least $\lfloor \tau/d \rfloor$ disjoint subsequences of $\pi^1, \dots, \pi^{j-1}$. Let $N = \lfloor \tau/d \rfloor$, and let $B_1, \dots, B_N$ be subsequences of $\pi^1, \dots, \pi^\tau$. 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 $(\pi^{t_1}, \dots, \pi^{t_\tau})$ be the subsequence $\pi^1, \dots, \pi^\tau$ consisting of all elements for which $w^{t_i}(\pi^{t_i}) > \alpha$. Each element is dependent on at most β^2/α^2 disjoint subsequences of $(\pi^{t_1}, \dots, \pi^{t_\tau})$, and by the argument above, one element is dependent on at least $\lfloor \tau/d \rfloor$ disjoint subsequences, so we must have $\lfloor \tau/d \rfloor \leq \beta^2/\alpha^2$, which implies that $\tau \leq (\beta^2/\alpha^2 + 1)d$. $\square$

5.2.3 Failure of Optimism: The Cheating Code

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. The construction is adapted from Amin et al. [17] and Jun and Zhang [119].

Example 5.2 (Cheating Code). Let $A \in \mathbb{N}$ be a power of 2 and consider the following function class $\mathcal{F}$ (in this example, we depart from the convention of this section and allow rewards and functions to take values in $[-1, 1]$; rescaling to $[0, 1]$ changes nothing).

- ▷ The decision space is $\Pi = [A] \cup \mathcal{C}$, where $\mathcal{C} = \{c_1, \dots, c_{\log_2(A)}\}$ 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 $b(f) = (b_1(f), \dots, b_{\log_2(A)}(f)) \in \{0, 1\}^{\log_2(A)}$ be a binary encoding for the index of $\pi_f \in [A]$ (e.g., if $\pi_f = 1$, $b(f) = (0, 0, \dots, 0)$, if $\pi_f = 2$, $b(f) = (0, 0, \dots, 0, 1)$, 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 $r \sim \mathcal{N}(f^*(\pi), 1)$ under π . For each cheating action $c_i \in \mathcal{C}$, since $f^*(c_i) = -b_i(f^*) \in \{0, -1\}$, we can determine whether the value is $b_i(f^*) = 0$ or $b_i(f^*) = 1$ with high probability using $\tilde{O}(1)$ action pulls. If we do this for each $c_i \in \mathcal{C}$, which will incur $\tilde{O}(\log(A))$ regret (there are $\log(A)$ such actions and each one leads to constant regret), we can infer the binary encoding $b(f^*) = b_1(f^*), \dots, b_{\log_2(A)}(f^*)$ for the optimal action π_{f^*} with high probability. At this point, we can simply stop exploring, and commit to playing π_{f^*} for the remaining rounds, which will incur no more regret. If one is careful with the details, this gives that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \log_2(A) \cdot \log(\log_2(A)/\delta),$$

since there are $\log_2(A)$ cheating actions, and a union bound over them requires each to be pulled $O(\log(\log_2(A)/\delta))$ times. 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 , an added 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 5.1 (Suboptimality of posterior sampling): Recall the Bayesian bandit setting in [Section 3.5](#), where we showed that the posterior sampling algorithm attains regret $\tilde{O}(\sqrt{AT})$ when $\Pi = \{1, \dots, A\}$. Posterior sampling is a general-purpose algorithm, and can be applied directly to arbitrary structured bandit problems (as long as a prior is available). However, similar to UCB, the cheating code construction in [Example 5.2](#) 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.

5.3 The Decision-Estimation Coefficient: Estimation-to-Decisions for Structured Bandits

[Section 5.2](#) leaves us with two requirements: an algorithm must deliberately trade reward for information, and the tradeoff must adapt to the class $\mathcal{F}$. This section shows that both are met by a general version of the Estimation-to-Decisions principle of [Section 4](#): reduce decision making to supervised online estimation, and pair the estimation oracle with an exploration strategy that ensures that regret is controlled by estimation error.

Concretely, we keep the online regression oracle of [Section 4](#) and generalize the exploration strategy. First, we require the following oracle for supervised estimation.

Definition 5.2 (Online Regression Oracle): At each time $t \in [T]$, an *online regression oracle* returns, given

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

with $\mathbb{E}[r^i | \pi^i] = f^*(\pi^i)$ and $\pi^i \sim p^i$, a function $\hat{f}^t : \Pi \rightarrow \mathbb{R}$. Define the cumulative estimation error

$$\mathbf{Est}_{\text{Sq}} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left[(\hat{f}^t(\pi^t) - f^*(\pi^t))^2 \right].$$

The oracle is required to ensure that $\mathbf{Est}_{\text{Sq}} \leq \mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ with probability at least $1 - \delta$, where $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ is a deterministic bound. Here, $p^i(\cdot | \mathcal{H}^{i-1})$ is the randomization distribution for the decision-maker.

This is the oracle from [Definition 4.1](#) with the context omitted. As discussed there, a square-loss online regret bound on $\mathbf{Reg}_{\text{OL}}^{\text{Sq}}$ yields an estimation-error bound $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ under realizability. In particular, for finite classes of bounded functions and bounded rewards, averaged exponential weights gives $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$.

The following algorithm, which we call Estimation-to-Decisions or E2D [84, 89], is a general-purpose meta-algorithm for structured bandits.

Algorithm 5.1 Estimation-to-Decisions (E2D) for Structured Bandits

input: Exploration parameter $\gamma > 0$; online regression oracle.

for $t = 1, \dots, T$ **do**

Obtain $\hat{f}^t$ from online regression oracle with $(\pi^1, r^1), \dots, (\pi^{t-1}, r^{t-1})$.

Compute

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

Select action $\pi^t \sim p^t$ and observe reward r^t .

As in SquareCB, each round consists of an estimation step and an exploration step; the only change is the latter, which now solves a min-max problem over $p \in \Delta(\Pi)$ and $f \in \mathcal{F}$ instead of applying a closed-form rule. Why a min-max problem? Inverse Gap Weighting was tailored to finitely many unstructured actions, and the examples in Section 5.2 show that for a general class $\mathcal{F}$, the right way to explore depends on $\mathcal{F}$ in a delicate way. Rather than designing an exploration strategy by hand for each $\mathcal{F}$, E2D lets the class itself determine the strategy: given the estimate $\hat{f}^t$, it asks for the distribution over decisions that achieves the *best possible* tradeoff between regret and estimation error. The value of this tradeoff is a complexity measure (or, structural parameter) for $\mathcal{F}$ called the *Decision-Estimation Coefficient* [84, 89], given by

$$\text{dec}_\gamma(\mathcal{F}, \hat{f}) = \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p} \left[\underbrace{f(\pi_f) - f(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{(f(\pi) - \hat{f}^t(\pi))^2}_{\text{information gain for obs.}} \right]. \quad (5.15)$$

That is, the Decision-Estimation Coefficient is the value of a game in which the learner (the min player) chooses a distribution over decisions so that, against a worst-case instance f (the max player), the *regret* of the decision is controlled by the *information gain* (or, estimation error) about f relative to a reference model $\hat{f}$ —a guess for the true model. With enough information, the learner can confirm or rule out the guess, and the scale parameter γ controls how much regret they are willing to incur to do so. In general, the larger the value of $\text{dec}_\gamma(\mathcal{F}, \hat{f})$, the more difficult it is to explore.

To state a regret bound for E2D, we define

$$\text{dec}_\gamma(\mathcal{F}) = \sup_{\hat{f} \in \text{co}(\mathcal{F})} \text{dec}_\gamma(\mathcal{F}, \hat{f}). \quad (5.16)$$

Here, $\text{co}(\mathcal{F})$ denotes the set of all convex combinations of elements in $\mathcal{F}$. The reason we consider the set $\text{co}(\mathcal{F})$ is that in general, online estimation algorithms such as exponential weights will produce improper predictions with $\hat{f} \in \text{co}(\mathcal{F})$. In fact, it turns out (see Proposition 5.13) that even if we allow $\hat{f}$ to be unconstrained above, the maximizer always lies in $\text{co}(\mathcal{F})$ without loss of generality.

The main result for this section shows that the regret for E2D is controlled by the value of the DEC and the estimation error $\text{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ for the online regression oracle. The result is from Foster et al. [84].

Theorem 5.1 (Regret of E2D for structured bandits): The E2D algorithm with exploration parameter $\gamma > 0$ guarantees that for every realization,

$$\mathbf{Reg} \leq \text{dec}_\gamma(\mathcal{F}) \cdot T + \gamma \cdot \mathbf{Est}_{\text{Sq}}. \quad (5.17)$$

In particular, with an online regression oracle (Definition 5.2), with probability at least $1 - \delta$,

$$\mathbf{Reg} \leq \text{dec}_\gamma(\mathcal{F}) \cdot T + \gamma \cdot \mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta). \quad (5.18)$$

Compare (5.17) with the guarantee for SquareCB in Theorem 4.2, which has the same form with $\text{dec}_\gamma(\mathcal{F})$ replaced by A/γ : regret is controlled by the estimation error of the oracle, and the price of exploration, which was A/γ for finitely many actions, is now the DEC. We can optimize over the parameter γ in the result above, which yields

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

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

$$\mathbf{Reg} \lesssim \inf_{\gamma > 0} \max \left\{ \text{dec}_\gamma(\mathcal{F}) \cdot T, \gamma \cdot \log(|\mathcal{F}|/\delta) \right\}. \quad (5.19)$$

As desired, regret scales only with the estimation complexity $\log|\mathcal{F}|$ and the exploration complexity $\text{dec}_\gamma(\mathcal{F})$. Before interpreting the result further, we give the proof: it is the proof of Theorem 4.2 with (4.41) replaced by the definition of the DEC.

Proof of Theorem 5.1. We write

$$\begin{aligned} \mathbf{Reg} &= \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^*(\pi^*) - f^*(\pi^t)] \\ &\leq \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^*(\pi^*) - f^*(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t} \left[(f^*(\pi^t) - \hat{f}^t(\pi^t))^2 \right] + \gamma \cdot \mathbf{Est}_{\text{Sq}}. \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}_{\pi^t \sim p^t} [f^*(\pi^*) - f^*(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t} \left[(f^*(\pi^t) - \hat{f}^t(\pi^t))^2 \right] \\ &\leq \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\pi^t \sim p^t} [f(\pi_f) - f(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t} \left[(f(\pi^t) - \hat{f}^t(\pi^t))^2 \right] \right\} \\ &= \inf_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi^t) - \hat{f}^t(\pi^t))^2 \right] \\ &= \text{dec}_\gamma(\mathcal{F}, \hat{f}^t), \end{aligned} \quad (5.20)$$

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

$$\mathbf{Reg} \leq \max_{t \in [T]} \text{dec}_\gamma(\mathcal{F}, \hat{f}^t) \cdot T + \gamma \cdot \mathbf{Est}_{\text{Sq}}.$$

It remains to bound $\text{dec}_\gamma(\mathcal{F}, \hat{f}^t)$ by $\text{dec}_\gamma(\mathcal{F})$. If the oracle's estimates satisfy $\hat{f}^t \in \text{co}(\mathcal{F})$, as is the case for exponential weights, this is immediate from the definition (5.16). For arbitrary estimates $\hat{f}^t : \Pi \rightarrow \mathbb{R}$, it follows from Proposition 5.13, which shows that references outside $\text{co}(\mathcal{F})$ do not increase the DEC. The second statement follows on the event that the oracle guarantee $\mathbf{Est}_{\text{Sq}} \leq \mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ holds. $\square$

Of course, the bound is only useful if the DEC is indeed bounded, which is not immediately apparent. Section 8 shows that boundedness is not just sufficient but *necessary* for low regret, in a fairly strong quantitative sense; for now, we build intuition through examples.

5.3.1 Warmup: Computing the DEC for Multi-Armed Bandits

To understand how to apply the general DEC regret bound to specific classes of interest, we begin with the multi-armed bandit, where $\Pi = [A]$ and $\mathcal{F} = \mathbb{R}^A$. Here the Inverse Gap Weighting distribution of Section 4 turns out to be the exact minimizer.

Proposition 5.2 (IGW minimizes the DEC): For the Multi-Armed Bandit setting, where $\Pi = [A]$ and $\mathcal{F} = \mathbb{R}^A$, the Inverse Gap Weighting distribution $p = \text{IGW}_{2\gamma}(\hat{f})$ in (4.39) is the *exact* minimizer for $\text{dec}_\gamma(\mathcal{F}, \hat{f})$, and certifies that $\text{dec}_\gamma(\mathcal{F}, \hat{f}) = \frac{A-1}{4\gamma}$.

By rewriting Proposition 4.3, it is straightforward to deduce that the DEC is bounded by $\frac{A}{\gamma}$, but Proposition 5.2 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 $\text{IGW}_{2\gamma}$ as opposed IGW_γ as in Proposition 4.3; 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 5.2. We rewrite the minimax problem as

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

For any fixed p and π^* , first-order conditions for optimality imply that the choice for f that maximizes this expression is

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

This choice gives

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

and

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

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

$$\min_{p \in \Delta([A])} \max_{\pi^* \in [A]} \left\{ \mathbb{E}_{\pi \sim p} [\hat{f}(\pi^*) - \hat{f}(\pi)] + \frac{1}{4\gamma p(\pi^*)} \right\} - \frac{1}{4\gamma}. \quad (5.21)$$

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

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

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 $p = \text{IGW}_{2\gamma}(\hat{f})$, we have that for all π^* ,

$$\begin{aligned} \mathbb{E}_{\pi \sim p} [\hat{f}(\pi^*) - \hat{f}(\pi)] + \frac{1}{4\gamma p(\pi^*)} &= \mathbb{E}_{\pi \sim p} [\hat{f}(\pi^*) - \hat{f}(\pi)] + \frac{\lambda}{4\gamma} + \hat{f}(\hat{\pi}) - \hat{f}(\pi^*) \\ &= \mathbb{E}_{\pi \sim p} [\hat{f}(\hat{\pi}) - \hat{f}(\pi)] + \frac{\lambda}{4\gamma}. \end{aligned} \quad (5.22)$$

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

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

Hence, $p = \text{IGW}_{2\gamma}(\hat{f})$ achieves the optimal value.

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

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

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

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

By the KKT conditions, if we wish to show that $p \in \Delta(\Pi)$ is optimal for the objective in (5.21), it suffices to find ν such that¹⁵

$$\mathbf{0} \in \partial_p \mathcal{L}(p, \nu),$$

where ∂_p denotes the subgradient with respect to p . Recall that for a convex function $h(x) = \max_y g(x, y)$, we have $\partial_x h(x) = \text{co}(\{\nabla g(x, y) \mid g(x, y) = \max_{y'} g(x, y')\})$. As a result,

$$\partial_p \mathcal{L}(p, \nu) = \nu \mathbf{1} - \hat{f} + \text{co}(\{\nabla_p g_{\pi^*}(p) \mid g_{\pi^*}(p) = \max_{\pi'} g_{\pi'}(p)\}).$$

Now, let $p = \text{IGW}_{2\gamma}(\hat{f})$. We argue that $\mathbf{0} \in \partial_p \mathcal{L}(p, \nu)$ for an appropriate choice of ν . By (5.22), we know that $g_{\pi}(p) = g_{\pi'}(p)$ for all π, π' (p is equalizing), so the expression above simplifies to

$$\partial_p \mathcal{L}(p, \nu) = \nu \mathbf{1} - \hat{f} + \text{co}(\{\nabla_p g_{\pi^*}(p)\}_{\pi^* \in \Pi}). \quad (5.23)$$

¹⁵If $p \in \Delta(\Pi)$, the KKT condition that $\frac{d}{d\nu} \mathcal{L}(p, \nu) = 0$ is already satisfied.

Noting that $\nabla_p g_{\pi^*}(p) = -\frac{1}{4\gamma p^2(\pi^*)} e_{\pi^*}$, we compute

$$\boldsymbol{\delta} := \sum_{\pi} p(\pi) \nabla g_{\pi}(p) = \left\{ -\frac{1}{4\gamma p(\pi)} \right\}_{\pi \in \Pi} = \left\{ -\frac{\lambda}{4\gamma} - \hat{f}(\hat{\pi}) + \hat{f}(\pi) \right\}_{\pi \in \Pi},$$

which has $\boldsymbol{\delta} \in \text{co}(\{\nabla_p g_{\pi^*}(p)\}_{\pi^* \in \Pi})$. By choosing $\nu = \frac{\lambda}{4\gamma} + \hat{f}(\hat{\pi})$, we have

$$\nu \mathbf{1} - \hat{f} + \boldsymbol{\delta} = \mathbf{0},$$

so (5.23) is satisfied. □

5.3.2 Approximate Solutions and Certified Bounds on the DEC

Before proceeding, let us mention that to apply E2D, it is not necessary to exactly solve the minimax problem (5.15). Instead, let us say that a distribution $p = p(\hat{f}, \gamma)$ *certifies an upper bound* on the DEC if, given $\hat{f}$ and $\gamma > 0$, it ensures that

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right] \leq \overline{\text{dec}}_{\gamma}(\mathcal{F}, \hat{f})$$

for some known upper bound $\overline{\text{dec}}_{\gamma}(\mathcal{F}, \hat{f}) \geq \text{dec}_{\gamma}(\mathcal{F}, \hat{f})$. In this case, if we define $\overline{\text{dec}}_{\gamma}(\mathcal{F}) := \sup_{\hat{f}} \overline{\text{dec}}_{\gamma}(\mathcal{F}, \hat{f})$, then it is simple to see that using this distribution $p^t = p(\hat{f}^t, \gamma)$ within E2D gives

$$\mathbf{Reg} \leq \overline{\text{dec}}_{\gamma}(\mathcal{F}) \cdot T + \gamma \cdot \mathbf{Est}_{\text{Sq}}$$

for every realization.

5.4 Examples: Computing the Decision-Estimation Coefficient

We now compute the DEC for classes where, unlike the multi-armed bandit, the right exploration strategy is not obvious. In each case we exhibit a distribution p that certifies a bound, and read off regret via Theorem 5.1.

5.4.1 Cheating Code

We begin with the cheating code of Example 5.2, on which both UCB and posterior sampling incur $\sqrt{AT}$ regret.

Proposition 5.3 (DEC for Cheating Code): Consider the cheating code in Example 5.2. For this class $\mathcal{F}$, we have

$$\text{dec}_{\gamma}(\mathcal{F}) \lesssim \frac{\log_2(A)}{\gamma}.$$

Combining Proposition 5.3 with Theorem 5.1 gives $\mathbf{Reg} \lesssim \sqrt{\log_2(A) T \log |\mathcal{F}|}$, whether E2D plays the certifying distribution or the exact minimizer; a refinement of E2D [89] improves this to the $\log_2(A)$ bound of Example 5.2.

Proof of Proposition 5.3. To simplify exposition, we present a bound on $\text{dec}_\gamma(\mathcal{F}, \hat{f})$ for this example only for $\hat{f} \in \mathcal{F}$, not for $\hat{f} \in \text{co}(\mathcal{F})$. A similar approach (albeit with a slightly different choice for p) leads to the same bound on $\text{dec}_\gamma(\mathcal{F})$. Let $\hat{f} \in \mathcal{F}$ and $\gamma > 0$ be given, and define

$$p = (1 - \varepsilon)\pi_{\hat{f}} + \varepsilon \cdot \text{unif}(\mathcal{C}).$$

We show that if we choose $\varepsilon = 2 \frac{\log_2(A)}{\gamma}$, this strategy certifies that

$$\text{dec}_\gamma(\mathcal{F}, \hat{f}) \lesssim \frac{\log_2(A)}{\gamma}.$$

Let $f \in \mathcal{F}$ be fixed, and consider the value

$$\mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right].$$

We consider two cases. For the first, if $\pi_f = \pi_{\hat{f}}$, then we can upper bound

$$\mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right] \leq \mathbb{E}_{\pi \sim p} [f(\pi_f) - f(\pi)] = \mathbb{E}_{\pi \sim p} [f(\pi_{\hat{f}}) - f(\pi)] \leq 2\varepsilon,$$

since $f \in [-1, 1]$.

For the second case, suppose that $\pi_f \neq \pi_{\hat{f}}$. We begin by bounding

$$\mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right] \leq 2 - \gamma \cdot \mathbb{E}_{\pi \sim p} \left[(f(\pi) - \hat{f}(\pi))^2 \right],$$

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 $\pi_f \neq \pi_{\hat{f}}$, if we let $b_1, \dots, b_{\log_2(A)}$ and $b'_1, \dots, b'_{\log_2(A)}$ denote the binary representations for π_f and $\pi_{\hat{f}}$, there exists i such that $b_i \neq b'_i$. As a result, we have

$$\mathbb{E}_{\pi \sim p} \left[(f(\pi) - \hat{f}(\pi))^2 \right] \geq \frac{\varepsilon}{\log_2(A)} (f(c_i) - \hat{f}(c_i))^2 = \frac{\varepsilon}{\log_2(A)} (b_i - b'_i)^2 = \frac{\varepsilon}{\log_2(A)}.$$

We conclude that in the second case,

$$\mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right] \leq 2 - \gamma \frac{\varepsilon}{\log_2(A)}.$$

Putting the cases together, we have

$$\mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right] \leq \max \left\{ 2\varepsilon, 2 - \gamma \frac{\varepsilon}{\log_2(A)} \right\}.$$

To balance these terms, we set

$$\varepsilon = 2 \frac{\log_2(A)}{\gamma},$$

which leads to the result. $\square$

5.4.2 Linear Bandits

Linear bandits [2, 22, 61, 57, 1] are the linear contextual bandit problem of Section 4 without contexts: Π is arbitrary and $\mathcal{F} = \{\pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{B}_2^d(1)$ is a parameter set and $\phi : \Pi \rightarrow \mathbb{B}_2^d(1)$ is a fixed feature map 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*; the following is the Kiefer–Wolfowitz theorem [125].

Proposition 5.4 (G-optimal design): For any compact set $\mathcal{Z} \subseteq \mathbb{R}^d$ with $\dim \text{span}(\mathcal{Z}) = d$, there exists a distribution $p \in \Delta(\mathcal{Z})$, called the *G-optimal design*, which has

$$\sup_{z \in \mathcal{Z}} \langle \Sigma_p^{-1} z, z \rangle \leq d, \quad (5.24)$$

where $\Sigma_p := \mathbb{E}_{z \sim p} [zz^\top]$.

The G-optimal design ensures coverage in every direction of the decision space, generalizing the notion of uniform exploration for finite decision spaces. In this sense, it can be thought of as a “universal” exploratory distribution for linearly structured decision spaces. Special cases include:

- ▷ When $\mathcal{Z} = \Delta([A])$, we can take $p = \text{unif}(e_1, \dots, e_A)$ as an optimal design
- ▷ When $\mathcal{Z} = \mathbb{B}_2^d(1)$, we can again take $p = \text{unif}(e_1, \dots, e_d)$ as an optimal design.
- ▷ For any positive definite matrix $A \succ 0$, the set $\mathcal{Z} = \{z \in \mathbb{R}^d \mid \langle Az, z \rangle \leq 1\}$ is an ellipsoid. Letting $\lambda_1, \dots, \lambda_d$ and $v_1, \dots, v_d$ denote the eigenvalues and eigenvectors for A , respectively, the distribution $p = \text{unif}(\lambda_1^{-1/2} v_1, \dots, \lambda_d^{-1/2} v_d)$ 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 $\hat{f}^t$ from a supervised estimation oracle. Play $\hat{\pi}^t = \pi_{\hat{f}^t}$ with probability $1 - \varepsilon$, and sample $\pi^t \sim q$ otherwise.

It is straightforward to show that this strategy gives $\mathbf{Reg} \lesssim d^{1/3} T^{2/3} (\log |\mathcal{F}|)^{2/3}$ for linear bandits. The basic idea is to replace (4.33) in the proof of Proposition 4.2 with the optimal design property (5.24), using that the reward functions under consideration are linear.

The following result shows that by combining optimal design with inverse gap weighting, we can obtain a d/γ bound on the DEC, which leads to an improved $\sqrt{dT}$ regret bound.

Proposition 5.5 (DEC for Linear Bandits): Consider the linear bandit setting. Let a linear function $\hat{f}$ and $\gamma > 0$ be given, consider the following distribution p :

- ▷ Define $\bar{\phi}(\pi) = \phi(\pi) / \sqrt{1 + \frac{\gamma}{d}(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi))}$, where $\pi_{\hat{f}} = \arg \max_{\pi \in \Pi} \hat{f}(\pi)$.
- ▷ Let $\bar{q} \in \Delta(\Pi)$ be the G-optimal design for the set $\{\bar{\phi}(\pi)\}_{\pi \in \Pi}$, and define $q = \frac{1}{2}\bar{q} + \frac{1}{2}\mathbb{I}_{\pi_{\hat{f}}}$.
- ▷ For each $\pi \in \Pi$, set

$$p(\pi) = \frac{q(\pi)}{\lambda + \frac{\gamma}{d}(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi))},$$

where $\lambda \in [1/2, 1]$ is chosen such that $\sum_{\pi} p(\pi) = 1$.^a

This strategy certifies that

$$\text{dec}_{\gamma}(\mathcal{F}) \lesssim \frac{d}{\gamma}.$$

^aThe 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 $\text{dec}_{\gamma}(\mathcal{F}) \gtrsim \frac{d}{\gamma}$ for this setting as well, so this is the best bound we can hope for. Combining this result with [Theorem 5.1](#) and using the averaged exponential weights algorithm for estimation as in [\(5.19\)](#) gives $\mathbf{Reg} \lesssim \sqrt{dT \log(|\mathcal{F}|/\delta)}$.

Proof of Proposition 5.5. Fix $f \in \mathcal{F}$. Let us abbreviate $\eta = \frac{\gamma}{d}$. As in [Proposition 4.3](#), we break regret into three terms:

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

The first term captures the loss in exploration that we would incur if $\hat{f}$ were the true reward function, and is equal to:

$$\sum_{\pi} \frac{q(\pi)(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi))}{\lambda + \eta(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi))} \leq \sum_{\pi} \frac{q(\pi)}{\eta} \leq \frac{1}{\eta},$$

and the second term, as before, is at most

$$\sqrt{\mathbb{E}_{\pi \sim p}[(\hat{f}(\pi) - f(\pi))^2]} \leq \frac{1}{2\gamma} + \frac{\gamma}{2} \mathbb{E}_{\pi \sim p}(\hat{f}(\pi) - f(\pi))^2.$$

The third term can be written as

$$\text{(III)} = f(\pi_f) - \hat{f}(\pi_f) - (\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f)) = \langle \theta - \hat{\theta}, \phi(\pi_f) \rangle - (\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f)),$$

where $\theta, \hat{\theta} \in \Theta$ are parameters such that $f(\pi) = \langle \theta, \phi(\pi) \rangle$ and $\hat{f}(\pi) = \langle \hat{\theta}, \phi(\pi) \rangle$. Defining $\Sigma_p = \mathbb{E}_{\pi \sim p}[\phi(\pi)\phi(\pi)^{\top}]$, we can bound

$$\begin{aligned} \langle \theta - \hat{\theta}, \phi(\pi_f) \rangle &= \langle \Sigma_p^{1/2}(\theta - \hat{\theta}), \Sigma_p^{-1/2}\phi(\pi_f) \rangle \\ &\leq \|\Sigma_p^{1/2}(\theta - \hat{\theta})\|_2 \|\Sigma_p^{-1/2}\phi(\pi_f)\|_2 \leq \frac{\gamma}{2} \|\Sigma_p^{1/2}(\theta - \hat{\theta})\|_2^2 + \frac{1}{2\gamma} \|\Sigma_p^{-1/2}\phi(\pi_f)\|_2^2. \end{aligned}$$

Note that $\|\Sigma_p^{1/2}(\theta - \hat{\theta})\|_2^2 = \mathbb{E}_{\pi \sim p}[(\hat{f}(\pi) - f(\pi))^2]$ and $\|\Sigma_p^{-1/2}\phi(\pi_f)\|_2^2 = \langle \phi(\pi_f), \Sigma_p^{-1}\phi(\pi_f) \rangle$, so we have

$$(III) \leq \frac{\gamma}{2} \mathbb{E}_{\pi \sim p}[(\hat{f}(\pi) - f(\pi))^2] + \underbrace{\frac{1}{2\gamma} \langle \phi(\pi_f), \Sigma_p^{-1}\phi(\pi_f) \rangle - (\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f))}_{(IV)}.$$

To proceed, observe that

$$\begin{aligned} \Sigma_p &\succeq \frac{1}{2} \sum_{\pi} \frac{\bar{q}(\pi)}{\lambda + \eta(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi))} \phi(\pi)\phi(\pi)^\top \\ &\succeq \frac{1}{2} \sum_{\pi} \frac{\bar{q}(\pi)}{1 + \eta(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi))} \phi(\pi)\phi(\pi)^\top \succeq \frac{1}{2} \sum_{\pi} \bar{q}(\pi) \bar{\phi}(\pi) \bar{\phi}(\pi)^\top =: \frac{1}{2} \bar{\Sigma}_{\bar{q}} \end{aligned}$$

This means that we can bound

$$\begin{aligned} \langle \phi(\pi_f), \Sigma_p^{-1}\phi(\pi_f) \rangle &\leq 2 \langle \phi(\pi_f), \bar{\Sigma}_{\bar{q}}^{-1}\phi(\pi_f) \rangle \\ &= 2(1 + \eta(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f))) \langle \bar{\phi}(\pi_f), \bar{\Sigma}_{\bar{q}}^{-1}\bar{\phi}(\pi_f) \rangle \\ &\leq 2d(1 + \eta(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f))), \end{aligned}$$

where the last line uses that $\bar{q}$ is the G-optimal design for $\{\bar{\phi}(\pi)\}_{\pi \in \Pi}$. We conclude that

$$(IV) \leq \frac{2d}{2\gamma} + \frac{2d\eta}{2\gamma} (\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f)) - (\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi_f)) \leq \frac{d}{\gamma}.$$

□

Remark 5.2 (Log-determinant strategy for linear bandits): When $\Theta = \mathbb{R}^d$ and $\hat{f} \in \mathcal{F}$, an alternative strategy for linear bandits is given by

$$p \in \arg \max_{p \in \Delta(\Pi)} \left\{ \mathbb{E}_{\pi \sim p}[\hat{f}(\pi)] + \frac{1}{4\gamma} \log \det \left(\mathbb{E}_{\pi \sim p}[\phi(\pi)\phi(\pi)^\top] - \bar{\phi}_p \bar{\phi}_p^\top \right) \right\}.$$

Here $\bar{\phi}_p = \mathbb{E}_{\pi \sim p}[\phi(\pi)]$, so the matrix inside the determinant is the centered covariance of the features.^a This strategy certifies $\text{dec}_\gamma(\mathcal{F}, \hat{f}) \leq d/(4\gamma)$; see Foster et al. [83, Lemma 2].

^aThe determinant is computed on the span of the feature differences $\phi(\pi) - \phi(\pi')$, for $\pi, \pi' \in \Pi$.

5.4.3 Nonparametric Bandits

For all of the examples so far, we have shown that

$$\text{dec}_\gamma(\mathcal{F}) \lesssim \frac{\text{eff-dim}(\mathcal{F}, \Pi)}{\gamma},$$

where $\text{eff-dim}(\mathcal{F}, \Pi)$ 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* [24, 128].

We take Π to be a metric space equipped with metric ρ , and define

$$\mathcal{F} = \{f : \Pi \rightarrow [0, 1] \mid f \text{ is 1-Lipschitz w.r.t } \rho\}.$$

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

$$\forall \pi \in \Pi \quad \exists \pi' \in \Pi' \quad \text{s.t.} \quad \rho(\pi, \pi') \leq \varepsilon,$$

and let $\mathcal{N}_\rho(\Pi, \varepsilon)$ denote the size of the smallest such cover.

Proposition 5.6 (DEC for Lipschitz Bandits): Consider the Lipschitz bandit setting, and suppose that there exists $d > 0$ such that $\mathcal{N}_\rho(\Pi, \varepsilon) \leq \varepsilon^{-d}$ for all $\varepsilon > 0$. Let $\hat{f} : \Pi \rightarrow [0, 1]$ and $\gamma \geq 1$ be given and consider the following distribution:

1. Let $\Pi' \subseteq \Pi$ witness the covering number $\mathcal{N}_\rho(\Pi, \varepsilon)$ for $\varepsilon \in (0, 1]$.
2. Let p be the result of applying the inverse gap weighting strategy in (4.39) to $\hat{f}$, restricted to the (finite) decision space Π' .

By setting $\varepsilon \propto \gamma^{-\frac{1}{d+1}}$, this strategy certifies that

$$\text{dec}_\gamma(\mathcal{F}, \hat{f}) \lesssim \gamma^{-\frac{1}{d+1}}.$$

Ignoring dependence on $\mathbf{Est}_{\text{sq}}(\mathcal{F}, T, \delta)$, this result leads to regret bounds that scale as $T^{\frac{d+1}{d+2}}$ (after tuning γ in Theorem 5.1), which cannot be improved.

Proof of Proposition 5.6. 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 4.3 ensures that if we choose p using inverse gap weighting over Π' , we have

$$\mathbb{E}_{\pi \sim p}[f(\iota(\pi_f)) - f(\pi)] \leq \frac{|\Pi'|}{\gamma} + \gamma \cdot \mathbb{E}_{\pi \sim p}[(f(\pi) - \hat{f}(\pi))^2].$$

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. □

5.4.4 Further Examples: Eluder Dimension, Concave, and ReLU Bandits

We state some additional upper bounds on the DEC without proof; details can be found in [84].

Example 5.3 (Decision-Estimation Coefficient subsumes Eluder Dimension). Consider any class $\mathcal{F}$ with values in $[0, 1]$. For all $\gamma \geq e$, we have

$$\text{dec}_\gamma(\mathcal{F}) \lesssim \inf_{\varepsilon > 0} \left\{ \varepsilon + \frac{\text{Edim}(\mathcal{F} - \mathcal{F}, \varepsilon) \log^2(\gamma)}{\gamma} \right\} + \gamma^{-1}. \quad (5.25)$$

◁

As a special case, this example implies that E2D enjoys a regret bound for generalized linear bandits similar to that of UCB.

Example 5.4 (Bandits with Concave Rewards). The concave (or convex, if one considers losses rather than rewards) bandit problem [126, 77, 4, 40, 42, 135] is a generalization of the linear bandit. We take $\Pi \subseteq \mathcal{B}_2^d(1)$ and define

$$\mathcal{F} = \{f : \Pi \rightarrow [0, 1] \mid f \text{ is concave and 1-Lipschitz w.r.t } \ell_2\}.$$

For this setting, whenever $\mathcal{F} \subseteq (\Pi \rightarrow [0, 1])$, results of Lattimore [135] imply that

$$\text{dec}_\gamma(\mathcal{F}) \lesssim \frac{d^4}{\gamma} \cdot \text{polylog}(d, \gamma) \quad (5.26)$$

for all $\gamma > 0$.

◁

For the function class

$$\mathcal{F} = \left\{ f(\pi) = 1 - \text{relu}(\langle \pi, \theta \rangle) \mid \theta \in \Theta \subset \mathcal{B}_2^d(1) \right\},$$

(5.26) leads to a $\sqrt{\text{poly}(d)T}$ 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.

5.5 Relationship to Optimism and Posterior Sampling

This subsection relates the Decision-Estimation Coefficient and E2D to the two other approaches covered so far, optimism and posterior sampling.

5.5.1 E2D with Confidence Sets: Recovering 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.

Algorithm 5.2 Estimation-to-Decisions (E2D) with Confidence Sets

input: Exploration parameter $\gamma > 0$; confidence radius $\beta > 0$; online regression oracle.

for $t = 1, \dots, T$ **do**

Obtain $\hat{f}^t$ from online regression oracle with $(\pi^1, r^1), \dots, (\pi^{t-1}, r^{t-1})$.

Set

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} \mid \sum_{i < t} \mathbb{E}_{\pi^i \sim p^i} \left[(\hat{f}^i(\pi^i) - f(\pi^i))^2 \right] \leq \beta^2 \right\}.$$

Compute

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

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 $f \in \mathcal{F}^t$.¹⁶ With this change, the distribution p^t can be interpreted as the minimizer for $\text{dec}_\gamma(\mathcal{F}^t, \hat{f}^t)$.

To analyze this algorithm, we show that as long as $f^* \in \mathcal{F}^t$ for all t , the same per-step analysis as in [Theorem 5.1](#) goes through, with $\mathcal{F}$ replaced by $\mathcal{F}^t$. The argument uses nothing about the confidence sets beyond this containment, so we state the result for an arbitrary sequence of confidence sets; the construction above is one way to obtain them.

Proposition 5.7 (Regret of E2D with confidence sets): Let $\mathcal{F}^1, \dots, \mathcal{F}^T \subseteq \mathcal{F}$ be any sequence of confidence sets, where $\mathcal{F}^t$ may depend on $\mathcal{H}^{t-1}$, such that $f^* \in \mathcal{F}^t$ for all $t \in [T]$ with probability at least $1 - \delta$, and assume the conditions of [Proposition 5.13](#) hold for each $\mathcal{F}^t$. Then for any $\gamma > 0$, E2D with these confidence sets, run with an online regression oracle, ensures that with probability at least $1 - 2\delta$,

$$\text{Reg} \leq \sum_{t=1}^T \text{dec}_\gamma(\mathcal{F}^t) + \gamma \cdot \text{Est}_{\text{Sq}}(\mathcal{F}, T, \delta). \quad (5.27)$$

For the construction in [Algorithm 5.2](#) with $\beta^2 = \text{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$, the containment holds on the oracle's own event, and the bound holds with probability at least $1 - \delta$.

Proof of Proposition 5.7. Condition on the intersection of the event guaranteed by the online regression oracle, on which $\sum_{t=1}^T \mathbb{E}_{\pi \sim p^t} [(\hat{f}^t(\pi) - f^*(\pi))^2] \leq \text{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$, and the event that $f^* \in \mathcal{F}^t$ for all t ; by a union bound, this intersection has probability at least $1 - 2\delta$. For the construction in [Algorithm 5.2](#) with $\beta^2 = \text{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$, every partial sum of the oracle's error is at most β^2 on the oracle's event, so $f^* \in \mathcal{F}^t$ for every t on that event alone. By the choice of p^t and the definition

¹⁶Note that compared to the confidence sets used in UCB, a slight difference is that we compute $\mathcal{F}^t$ using the estimates $\hat{f}^1, \dots, \hat{f}^T$ 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 latter would work as well.

of the DEC, we then have

$$\begin{aligned}\mathbb{E}_{\pi \sim p^t}[f^*(\pi_{f^*}) - f^*(\pi)] &\leq \text{dec}_\gamma(\mathcal{F}^t, \hat{f}^t) + \gamma \mathbb{E}_{\pi \sim p^t}[(\hat{f}^t(\pi) - f^*(\pi))^2] \\ &\leq \text{dec}_\gamma(\mathcal{F}^t) + \gamma \mathbb{E}_{\pi \sim p^t}[(\hat{f}^t(\pi) - f^*(\pi))^2].\end{aligned}$$

The second inequality uses [Proposition 5.13](#); this step is needed because $\hat{f}^t$ need not lie in $\text{co}(\mathcal{F}^t)$. Summing over t and using the oracle bound proves the claim. $\square$

This bound is never worse than the one in [Theorem 5.1](#), but it can be smaller if the confidence sets $\mathcal{F}^1, \dots, \mathcal{F}^T$ shrink quickly.

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 $\text{dec}_\gamma(\mathcal{F}^t)$ which is never worse than usual confidence width we use in the UCB analysis.

Proposition 5.8 (UCB certifies a bound on the DEC): For any t , the UCB strategy $\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi)$ certifies that

$$\text{dec}_0(\mathcal{F}^t) \leq \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t). \quad (5.28)$$

Proof of Proposition 5.8. By choosing $\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi)$, we have that for any $\hat{f}$,

$$\begin{aligned}\text{dec}_0(\mathcal{F}^t, \hat{f}) &= \inf_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{F}^t} \mathbb{E}_{\pi \sim p} \left[\max_{\pi^*} f(\pi^*) - f(\pi) \right] \\ &\leq \sup_{f \in \mathcal{F}^t} \left[\max_{\pi^*} f(\pi^*) - f(\pi^t) \right] \\ &\leq \sup_{f \in \mathcal{F}^t} \left[\max_{\pi^*} \bar{f}^t(\pi^*) - f(\pi^t) \right] \\ &= \sup_{f \in \mathcal{F}^t} [\bar{f}^t(\pi^t) - f(\pi^t)] = \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t).\end{aligned}$$

$\square$

As we saw in the analysis of UCB for multi-armed bandits with $\Pi = \{1, \dots, A\}$ ([Section 3.4](#)), the confidence width in (5.28) might be large for a given round t , but by the pigeonhole argument ([Lemma 3.2](#)), when we sum over all rounds we have

$$\sum_{t=1}^T \text{dec}_0(\mathcal{F}^t) \leq \sum_{t=1}^T \bar{f}^t(\pi^t) - \underline{f}^t(\pi^t) \leq \tilde{O}(\sqrt{AT}).$$

Hence, although UCB does not minimize the DEC, the bound it certifies is small whenever the confidence widths shrink quickly.

Note that [Proposition 5.8](#) bounds the DEC at $\gamma = 0$; this is only meaningful for shrinking classes, since for $\mathcal{F} = \mathbb{R}^A$, [Proposition 5.2](#) gives $\text{dec}_\gamma(\mathcal{F}) \gtrsim A/\gamma \rightarrow \infty$ as $\gamma \rightarrow 0$. Allowing $\gamma > 0$ gives the following slightly stronger result, which replaces $\underline{f}^t$ by $\hat{f}^t$.

Proposition 5.9 (UCB certifies a bound on the DEC, $\gamma > 0$): For any $\gamma > 0$, the UCB strategy $\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi)$ certifies that

$$\text{dec}_\gamma(\mathcal{F}^t, \hat{f}^t) \leq \bar{f}^t(\pi^t) - \hat{f}^t(\pi^t) + \frac{1}{4\gamma}.$$

Proof of Proposition 5.9. This is a slight generalization of the proof of Proposition 5.8. By choosing $\pi^t = \arg \max_{\pi \in \Pi} \bar{f}^t(\pi)$, we have

$$\begin{aligned} \text{dec}_\gamma(\mathcal{F}^t, \hat{f}^t) &= \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}^t} \mathbb{E}_{\pi \sim p} \left[\max_{\pi^*} f(\pi^*) - f(\pi) - \gamma \cdot (\hat{f}^t(\pi) - f(\pi))^2 \right] \\ &\leq \max_{f \in \mathcal{F}^t} \left[\max_{\pi^*} f(\pi^*) - f(\pi^t) - \gamma \cdot (\hat{f}^t(\pi^t) - f(\pi^t))^2 \right] \\ &\leq \max_{f \in \mathcal{F}^t} \left[\bar{f}^t(\pi^t) - f(\pi^t) - \gamma \cdot (\hat{f}^t(\pi^t) - f(\pi^t))^2 \right] \\ &= \max_{f \in \mathcal{F}^t} \underbrace{\left[\hat{f}^t(\pi^t) - f(\pi^t) - \gamma \cdot (\hat{f}^t(\pi^t) - f(\pi^t))^2 \right]}_{\leq \frac{1}{4\gamma}} + \bar{f}^t(\pi^t) - \hat{f}^t(\pi^t). \end{aligned}$$

□

5.5.2 The Bayesian DEC: Recovering Posterior Sampling

The Decision-Estimation Coefficient (5.15) is a min-max problem. Its *dual* (max-min) analogue is

$$\underline{\text{dec}}_\gamma(\mathcal{F}, \hat{f}) = \sup_{\mu \in \Delta(\mathcal{F})} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{f \sim \mu} \mathbb{E}_{\pi \sim p} \left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \hat{f}(\pi))^2 \right]. \quad (5.29)$$

The dual has a Bayesian interpretation: the adversary selects a *prior* μ over $\mathcal{F}$, and the learner, knowing μ , trades off regret and information on average over $f \sim \mu$.

Using the minimax theorem (Lemma A.8), one can show that the Decision-Estimation Coefficient and its Bayesian counterpart coincide.

Proposition 5.10 (Equivalence of primal and dual DEC): If $\mathcal{F}$ and Π are finite, or more generally under the regularity conditions of the minimax theorem (Lemma A.8), we have

$$\text{dec}_\gamma(\mathcal{F}, \hat{f}) = \underline{\text{dec}}_\gamma(\mathcal{F}, \hat{f}). \quad (5.30)$$

Proof of Proposition 5.10. For fixed $\hat{f}$, define

$$L(p, \mu; \hat{f}) = \mathbb{E}_{f \sim \mu, \pi \sim p} [f(\pi_f) - f(\pi) - \gamma(f(\pi) - \hat{f}(\pi))^2].$$

Maximizing over $f \in \mathcal{F}$ is equivalent to taking the supremum over $\mu \in \Delta(\mathcal{F})$. The payoff L is affine in each distribution. For finite $\mathcal{F}$ and Π , both simplices are compact and L is continuous, so the minimax theorem (Lemma A.8) implies that the infimum and supremum can be exchanged. The same argument applies more generally whenever the regularity conditions of the minimax theorem hold. □

Thus, any bound on the dual DEC immediately yields a bound on the primal DEC. As an example, we can adapt the posterior sampling/probability matching strategy introduced in [Section 3](#). 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 . The following proposition makes the resulting bound explicit.

Proposition 5.11 (Posterior sampling bounds the DEC): If $|\Pi| = A < \infty$ and $\mathcal{F}$ is uniformly bounded, then for every $\gamma > 0$ and reference $\hat{f}$,

$$\underline{\text{dec}}_\gamma(\mathcal{F}, \hat{f}) \leq \frac{A}{4\gamma}.$$

Consequently, whenever the minimax identity (5.30) applies, $\text{dec}_\gamma(\mathcal{F}) \leq A/(4\gamma)$.

Proof of Proposition 5.11. Fix a prior μ and let $\bar{f}(\pi) = \mathbb{E}_{f \sim \mu} f(\pi)$. Choose the probability-matching distribution $p(\pi) = \mathbb{P}_{f \sim \mu}(\pi_f = \pi)$, using a fixed rule to choose each maximizer π_f . Write

$$V = \mathbb{E}_{f \sim \mu, \pi \sim p}[(f(\pi) - \bar{f}(\pi))^2], \quad I = \mathbb{E}_{f \sim \mu, \pi \sim p}[(f(\pi) - \hat{f}(\pi))^2].$$

Here f and π are drawn independently. Variance decomposition gives $I = V + \mathbb{E}_{\pi \sim p}[(\bar{f}(\pi) - \hat{f}(\pi))^2] \geq V$. Since the marginal distribution of π_f is p , the terms involving $\bar{f}$ cancel in the expected regret:

$$\mathbb{E}_{f \sim \mu, \pi \sim p}[f(\pi_f) - f(\pi)] = \mathbb{E}_{f \sim \mu}[f(\pi_f) - \bar{f}(\pi_f)] \leq \sqrt{AV},$$

where the inequality is [Lemma 3.3](#). Thus the objective in the dual DEC, evaluated at this p , is at most

$$\sqrt{AV} - \gamma I \leq \sqrt{AI} - \gamma I = \frac{A}{4\gamma} - \gamma \left(\sqrt{I} - \frac{\sqrt{A}}{2\gamma} \right)^2 \leq \frac{A}{4\gamma}.$$

Taking the supremum over priors proves the dual bound. The primal bound follows from [Proposition 5.10](#) and the supremum over references. $\square$

More generally, most analysis techniques for the Bayesian setting can be viewed as implicitly providing bounds on the dual Decision-Estimation Coefficient [[181](#), [41](#), [40](#), [183](#), [138](#), [135](#)]. 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 [[87](#)]. 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 [[84](#)].

5.6 Incorporating Contexts*

The Decision-Estimation Coefficient and E2D extend directly to *contextual* structured bandits, and hence to a version of SquareCB for general decision 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 4](#), 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

$$r^t \sim M^*(\cdot \mid x^t, \pi^t),$$

and define $f^*(x, \pi) = \mathbb{E}_{r \sim M^*(\cdot \mid x, \pi)}$. We assume access to a function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$, and assume access to an estimation oracle for $\mathcal{F}$ that ensures that with probability at least $1 - \delta$,

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

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 .

Algorithm 5.3 Estimation-to-Decisions (E2D) for Contextual Structured Bandits

input: Exploration parameter $\gamma > 0$; online regression oracle.

```

for  $t = 1, \dots, T$  do
  Observe  $x^t \in \mathcal{X}$ .
  Obtain  $\hat{f}^t$  from online regression oracle with  $(x^1, \pi^1, r^1), \dots, (x^{t-1}, \pi^{t-1}, r^{t-1})$ .
  Compute

```

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

Select action $\pi^t \sim p^t$ and observe reward r^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 5.12 (Regret of E2D for contextual structured bandits): The E2D algorithm with exploration parameter $\gamma > 0$ guarantees that

$$\mathbf{Reg} \leq \sup_{x \in \mathcal{X}} \text{dec}_\gamma(\mathcal{F}(x, \cdot)) \cdot T + \gamma \cdot \mathbf{Est}_{\text{Sq}} \quad (5.31)$$

for every realization, and hence, with probability at least $1 - \delta$, the same bound holds with $\mathbf{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)$ in place of $\mathbf{Est}_{\text{Sq}}$.

The proof is nearly identical to that of [Theorem 5.1](#), conditioning on x^t at each round, and is omitted.

In particular, since IGW exactly minimizes the DEC when $\mathcal{F} = \mathbb{R}^A$ ([Proposition 5.2](#)), SquareCB is the special case of Contextual E2D with $\mathcal{F} = \mathbb{R}^A$.

Going beyond the finite-action setting, it is simplest to interpret [Proposition 5.12](#) when $\mathcal{F}(x, \cdot)$ has the same structure for all contexts. One example is *contextual bandits with linearly structured decision 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 $\mathcal{G} \subset (\mathcal{X} \rightarrow \mathbb{B}_2^d(1))$ is an arbitrary function class. This setting generalizes the linear contextual bandit problem from [Section 4](#), which corresponds to the case where $\mathcal{G}$ is a set of constant functions. We can apply [Proposition 5.5](#) to conclude that $\sup_{x \in \mathcal{X}} \text{dec}_\gamma(\mathcal{F}(x, \cdot)) \lesssim \frac{d}{\gamma}$, so that [Proposition 5.12](#) gives $\text{Reg} \lesssim \sqrt{dT \cdot \text{Est}_{\text{Sq}}(\mathcal{F}, T, \delta)}$.

5.7 Additional Properties of the Decision-Estimation Coefficient*

The following proposition is what allows the supremum in [\(5.16\)](#) to be taken over $\text{co}(\mathcal{F})$, and justifies the final step in the proofs of [Theorem 5.1](#) and [Proposition 5.7](#) when the oracle's estimates do not lie in $\text{co}(\mathcal{F})$.

Proposition 5.13 (References outside the convex hull do not increase the DEC):

Suppose $\mathcal{F}$ and Π are finite. More generally, it suffices that the Bayesian DEC equivalence in [\(5.30\)](#) holds for every reference and that the pointwise mean of every prior on $\mathcal{F}$ belongs to $\text{co}(\mathcal{F})$. Then, for any $\gamma > 0$,

$$\sup_{\hat{f}: \Pi \rightarrow \mathbb{R}} \text{dec}_\gamma(\mathcal{F}, \hat{f}) = \sup_{\hat{f} \in \text{co}(\mathcal{F})} \text{dec}_\gamma(\mathcal{F}, \hat{f}).$$

Proof of [Proposition 5.13](#). The supremum over all reference functions is at least the supremum over $\text{co}(\mathcal{F})$. For the reverse inequality, fix any reference $\hat{f}$ and prior μ , and define $\bar{f}_\mu = \mathbb{E}_{f \sim \mu}[f] \in \text{co}(\mathcal{F})$. At each action π , variance decomposition gives

$$\mathbb{E}_{f \sim \mu}[(f(\pi) - \hat{f}(\pi))^2] = \mathbb{E}_{f \sim \mu}[(f(\pi) - \bar{f}_\mu(\pi))^2] + (\bar{f}_\mu(\pi) - \hat{f}(\pi))^2.$$

Let $L(p, \mu; g)$ denote the expected DEC payoff with reference g , as in the proof of [Proposition 5.10](#). The identity above implies $L(p, \mu; \hat{f}) \leq L(p, \mu; \bar{f}_\mu)$ for every p . Hence, using [\(5.30\)](#) for equality, we have that

$$\begin{aligned} \text{dec}_\gamma(\mathcal{F}, \hat{f}) &= \sup_{\mu} \inf_p L(p, \mu; \hat{f}) \\ &\leq \sup_{\mu} \inf_p L(p, \mu; \bar{f}_\mu) \\ &\leq \sup_{\mu} \sup_{\nu} \inf_p L(p, \nu; \bar{f}_\mu) = \sup_{\mu} \text{dec}_\gamma(\mathcal{F}, \bar{f}_\mu) \\ &\leq \sup_{g \in \text{co}(\mathcal{F})} \text{dec}_\gamma(\mathcal{F}, g). \end{aligned}$$

Taking the supremum over $\hat{f}$ proves the reverse inequality. □

5.8 Bibliographic Remarks and Further Reading

Takeaways. This section extended the Estimation-to-Decisions principle from finite action sets to structured decision spaces, where the learner must generalize across decisions rather than across contexts. We showed that optimism remains a viable strategy when confidence widths shrink quickly, and the eluder dimension quantifies when this happens, but the cheating code shows that this rule is not optimal for every class $\mathcal{F}$. The Decision-Estimation Coefficient resolves this by letting the class itself determine the exploration distribution: E2D solves a per-round min-max problem whose value links regret to estimation error, and recovers IGW for multi-armed bandits, optimal design for linear bandits, and covering-based exploration for Lipschitz bandits. The DEC is the quantity that [Section 7](#) generalizes beyond bandits and that [Section 8](#) shows to be necessary.

Bibliographic remarks. The generalized UCB algorithm with least-squares confidence sets, the eluder dimension, [Proposition 5.1](#), and [Lemma 5.1](#) are due to Russo and Van Roy [180]. The cheating code in [Example 5.2](#) is adapted from the constructions of Amin et al. [17] and Jun and Zhang [119]. The Decision-Estimation Coefficient, the E2D meta-algorithm ([Algorithm 5.1](#)), [Theorem 5.1](#), and the bounds in [Proposition 5.3](#), [Proposition 5.5](#), and [Proposition 5.6](#) are due to Foster et al. [84]; the $\log(T)$ refinement for the cheating code and the treatment of reference models in $\text{co}(\mathcal{F})$ appear in Foster et al. [89]. The closed-form minimizers in [Proposition 5.2](#) and for linear bandits are from Foster et al. [83, 84]. The existence of the G-optimal design is the Kiefer–Wolfowitz theorem [125]; our presentation follows Lattimore and Szepesvári [139, Chapter 21]. Lipschitz bandits in metric spaces are due to Kleinberg et al. [127, 128] and Auer et al. [24].

The results relating the DEC to optimism and posterior sampling—E2D with confidence sets ([Algorithm 5.2](#) and [Proposition 5.7](#)), the observation that UCB certifies a bound on the DEC ([Proposition 5.8](#) and [Proposition 5.9](#)), and the Bayesian DEC ([Proposition 5.10](#)) and its bound by the information ratio—are from Foster et al. [84]; the probability-matching argument in [Proposition 5.11](#) behind the Bayesian bound is the decoupling argument of Russo and Van Roy [181, 182]. E2D for contextual structured bandits ([Algorithm 5.3](#), [Proposition 5.12](#)) is from Foster et al. [84], and [Proposition 5.13](#) is implicit in Foster et al. [84, 89].

Further reading. Textbook treatments of linear and structured bandits are given by Lattimore and Szepesvári [139, Part V] and Slivkins [191, Chapter 4]. Lattimore and Szepesvári [137], complementing the minimax perspective in this section, show that optimism is asymptotically suboptimal even for finite-armed linear bandits. The information ratio originates in the Bayesian analysis of posterior sampling and information-directed sampling [181, 182, 183]; its use for minimax regret is due to Lattimore and Szepesvári [138]. Extensions of the Decision-Estimation Coefficient to adversarial decision making can be found in Foster et al. [87], Xu and Zeevi [217].

6. REINFORCEMENT LEARNING: BASIC PRINCIPLES AND FINITE STATE SPACES

The bandit problems of [Sections 3 to 5](#) feature one-shot interaction at each learning episode: a decision is made, a reward is observed, and the interaction ends. In this section, we take the next step along the interactivity axis of [Figure 2](#). In *reinforcement learning*, each decision is a multi-stage policy that acts on a state evolving in response to the learner’s actions, and the learner observes the intermediate states and rewards along the way. This extra information supports *credit assignment*—recognizing which actions made a policy good or bad so that the decision can be broken down into a sequence of single-step choices. This section treats the simplest version of the problem, *tabular* reinforcement learning, in which states and actions are finite and there is no structure to generalize across. We return to the setting, this time with function approximation and generalization, in [Section 9](#).

Organization. [Section 6.1](#) formalizes the episodic reinforcement learning problem. We first set statistical uncertainty aside: [Section 6.2](#) shows how to compute an optimal policy when the MDP is known, via dynamic programming. We then turn to the learning problem, following the structure of the preceding sections. [Section 6.3](#) shows that uniform exploration fails, now with regret exponential in the horizon. [Section 6.4](#) introduces analysis tools for reinforcement learning that are used throughout the remainder of the monograph. [Section 6.5](#) adapts the principle of optimism to unknown MDPs, and [Section 6.6](#) instantiates it with the UCB-VI algorithm. Finally, [Section 6.7](#) gives a second approach, based on Estimation-to-Decisions.

6.1 The Episodic Reinforcement Learning Problem

We consider an episodic finite-horizon reinforcement learning framework. With H denoting the *horizon*, a *Markov Decision Process* (MDP) M takes the form

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\},$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space,

$$P_h^M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$$

is the probability transition kernel at step h ,

$$R_h^M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathbb{R})$$

is the reward distribution, and $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution. We write $P_h^M(s, a) \in \Delta(\mathcal{S})$ for the transition distribution at (s, a) and $P_h^M(s' | s, a)$ for its probability mass at s' , and likewise for $R_h^M(s, a)$ and $R_h^M(\cdot | s, a)$. 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*

$$\pi = (\pi_1, \dots, \pi_H),$$

where $\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$; 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 $h = 1, \dots, H$:

- ▷ $a_h \sim \pi_h(s_h)$.
- ▷ $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(s_h, a_h)$.

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,

$$\mathbb{P}^M(s_{h+1} = s' \mid s_h, a_h) = \mathbb{P}^M(s_{h+1} = s' \mid s_h, a_h, s_{h-1}, a_{h-1}, \dots, s_1, a_1).$$

The value for a policy π under M is given by

$$f^M(\pi) := \mathbb{E}^{M, \pi} \left[\sum_{h=1}^H r_h \right], \quad (6.1)$$

where $\mathbb{E}^{M, \pi}[\cdot]$ denotes expectation under the process above. We define an optimal policy for model M as

$$\pi_M \in \arg \max_{\pi \in \Pi_{\text{rns}}} f^M(\pi). \quad (6.2)$$

Online reinforcement learning. For reinforcement learning, we focus on the *online reinforcement learning problem*, in which we interact with an unknown MDP M^* for T episodes.

Online Reinforcement Learning Protocol

for $t = 1, \dots, T$ **do**
 Select policy $\pi^t \in \Pi_{\text{rns}}$ (possibly at random, $\pi^t \sim p^t$).
 Execute π^t in M^* for one episode.
 Observe the trajectory $\tau^t = (s_1^t, a_1^t, r_1^t), \dots, (s_H^t, a_H^t, r_H^t)$.

The learner may randomize, selecting $\pi^t \sim p^t$ for a distribution $p^t \in \Delta(\Pi_{\text{rns}})$ that depends on the trajectories observed so far; note that this randomization is in addition to the randomness within the policy itself. The goal is to minimize the total regret

$$\mathbf{Reg} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^*(\pi^*) - f^*(\pi^t)] \quad (6.3)$$

against the optimal policy $\pi^* := \pi_{M^*}$ for M^* . Here and throughout, we abbreviate $f^* := f^{M^*}$ and $\pi^* := \pi_{M^*}$ whenever the true MDP M^* is meant, retaining the model subscript or superscript for a generic M . Throughout this section, rewards are bounded, $r_h \in [0, 1]$, and normalized so that $\sum_{h=1}^H r_h \in [0, 1]$ almost surely; regret bounds hold with probability at least $1 - \delta$ unless stated otherwise.

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 (6.3) coincides with the notion of regret in the respective setting. Conversely, one could view an episodic MDP as a structured bandit whose decisions are policies and whose rewards are episode returns. What distinguishes

reinforcement learning is the additional feedback: the learner observes the intermediate states and rewards, not just the return, and this is what makes credit assignment possible.

In the language of [Figure 2](#), reinforcement learning moves further along the interactivity axis; the tabular setting, like the multi-armed bandit, assumes no structure across states and actions, so their cardinalities $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ appear in the rates; function approximation is the subject of [Section 9](#).

Remark 6.1 (Alternative MDP formulations): 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. The core ideas are largely the same, but we find the formulation here to be cleanest, while capturing the essential features of the problem.

6.2 Planning via Dynamic Programming

We first set statistical uncertainty aside: suppose that the MDP M is known, so that the only task is to find a policy maximizing $f^M(\pi)$, as in [\(6.2\)](#). This is the *planning* problem. It arises on its own in games such as chess or backgammon, where the transition rules are given, and it is a prerequisite for the learning problem, in which the learner must additionally estimate the unknown M^* from interaction. We show that planning is solved by the principle of *dynamic programming*, which addresses credit assignment directly: it breaks the choice of a multi-stage policy into a sequence of single-step choices, each made with knowledge of the value of the actions that follow.

Value functions. The objects that make this decomposition possible are intermediate “reward-to-go” functions. 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 [\(6.1\)](#) reads

$$f^M(\pi) = \mathbb{E}_{s \sim d_1, a \sim \pi_1(s)} [Q_1^{M,\pi}(s, a)] = \mathbb{E}_{s \sim d_1} [V_1^{M,\pi}(s)]$$

We begin by observing that the optimal policy π_M in [\(6.2\)](#) 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. We now introduce a fundamental result, [Proposition 6.1](#), which guarantees existence of an optimal policy $\pi_M = (\pi_{M,1}, \dots, \pi_{M,H})$ that maximizes $V_1^{M,\pi}(s)$ over $\pi \in \Pi_{\text{rns}}$ for all states $s \in \mathcal{S}$ *simultaneously* (rather than just on average, as in [\(6.2\)](#)). 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 $h = 2, \dots, H$, 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 in its original form as follows [\[31\]](#):

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.*

To state the result formally, we introduce the *optimal value functions* as

$$Q_h^{M,\star}(s, a) = \max_{\pi \in \Pi_{\text{rms}}} \mathbb{E}^{M,\pi} \left[\sum_{h'=h}^H r_{h'} \mid s_h = s, a_h = a \right] \quad \text{and} \quad V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s, a) \quad (6.4)$$

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 simultaneously. Indeed, the following lemma shows that there exists π_M such that for all s, a, h ,

$$Q_h^{M,\star}(s, a) = Q_h^{M,\pi_M}(s, a), \quad \text{and} \quad V_h^{M,\star}(s) = V_h^{M,\pi_M}(s). \quad (6.5)$$

Proposition 6.1 (Bellman Optimality): The optimal value function (6.4) for MDP M can be computed via $V_{H+1}^{M,\pi_M}(s) := 0$, and for each $s \in \mathcal{S}$,

$$V_h^{M,\pi_M}(s) = \max_{a \in \mathcal{A}} \mathbb{E}^M [r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s, a_h = a]. \quad (6.6)$$

The optimal policy is given by

$$\pi_{M,h}(s) \in \arg \max_{a \in \mathcal{A}} \mathbb{E}^M [r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s, a_h = a]. \quad (6.7)$$

Equivalently, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$Q_h^{M,\pi_M}(s, a) = \mathbb{E}^M \left[r_h + \max_{a' \in \mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1}, a') \mid s_h = s, a_h = a \right], \quad (6.8)$$

and an the optimal policy is given by

$$\pi_{M,h}(s) \in \arg \max_{a \in \mathcal{A}} Q_h^{M,\pi_M}(s, a). \quad (6.9)$$

The Bellman operator. The update in (6.8) is referred to as *value iteration* (VI). It is useful to introduce a more succinct notation for this update. For an MDP M , define the *Bellman Operators* $\mathcal{T}_1^M, \dots, \mathcal{T}_H^M$ via

$$[\mathcal{T}_h^M Q](s, a) = \mathbb{E}_{r_h \sim R_h^M(s, a), s_{h+1} \sim P_h^M(s, a)} \left[r_h + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a') \right] \quad (6.10)$$

for any $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. We write the expectation above more succinctly as

$$[\mathcal{T}_h^M Q](s, a) = \mathbb{E}^M \left[r_h(s_h, a_h) + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a') \mid s_h = s, a_h = a \right] \quad (6.11)$$

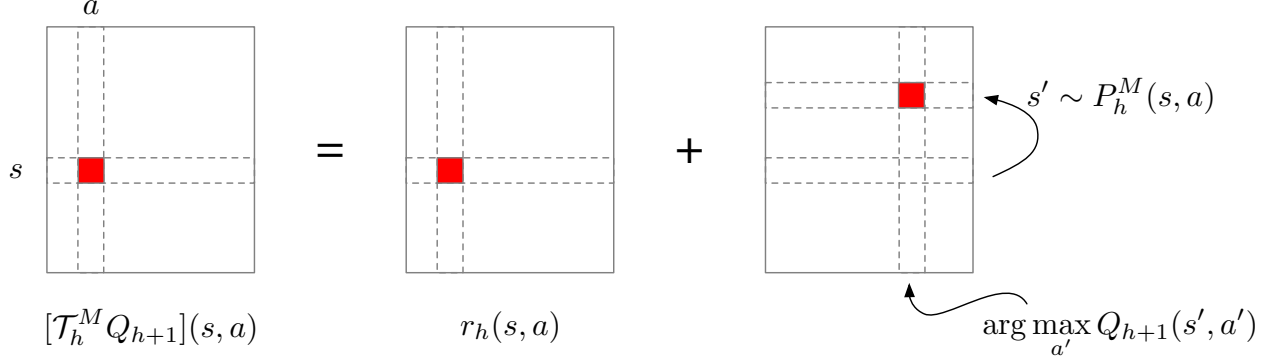


Figure 10: The Bellman operator. Applying $\mathcal{T}_h^M$ to Q_{h+1} adds the immediate reward to the expected greedy value at the next state.

Figure 10 visualizes the operator: $\mathcal{T}_h^M$ maps a function of the next layer’s state–action pairs to a function of the current layer’s, by taking one step of the dynamics and one greedy maximization. In the language of Bellman operators, (6.8) can be written as

$$Q_h^{M, \pi_M} = \mathcal{T}_h^M Q_{h+1}^{M, \pi_M}. \quad (6.12)$$

6.3 The Need for Strategic 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 monograph, 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 naive exploration with ε -Greedy, while having a suboptimal rate in T ($T^{2/3}$ as opposed to $\sqrt{T}$), gives a meaningful and useful regret guarantee for multi-armed and contextual bandits. The following classical example, a so-called “combination lock,” shows that for reinforcement learning, the performance of this strategy can be disastrous rather than merely suboptimal, as it leads to exponential regret in the horizon H .

Consider the MDP depicted in Figure 11, 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, unless by chance all of the actions sampled happen to be good; the probability that this occurs is exponentially small in H . Of course, the learner does not know in advance which action is the good one at each state, so any algorithm that does not adjust its exploration strategy adaptively must guess all H of them at once. This means that T needs to be at least $\Omega(2^H)$ to achieve nontrivial regret.

Given the failure of ε -Greedy for this example, one can ask whether other algorithmic principles also fail. As we now show, appropriate adaptations of the optimism principle and the Estimation-to-Decisions principle succeed, and yield regret bounds that are *polynomial* in the parameters $|S|$,

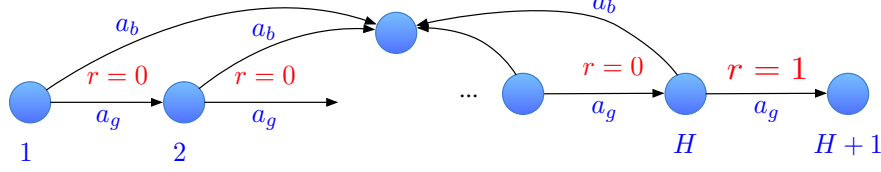


Figure 11: The combination lock MDP. The good action a_g advances along the chain and the bad action a_b leads to a terminal state; reward is received only at the end of the chain.

$|\mathcal{A}|$, and H . Before diving into the details, we present a collection of standard tools for analysis in MDPs, which are used throughout the remainder of the monograph.

6.4 Analysis Tools: Performance Difference, Bellman Residuals, and Simulation

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 actions* taken by the two policies. The simple proof, stated below, proceeds by successively changing one policy into another and keeping 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 6.1 (Performance Difference Lemma): For any $s \in \mathcal{S}$, and $\pi, \pi' \in \Pi_{\text{rs}}$,

$$V_1^{M, \pi'}(s) - V_1^{M, \pi}(s) = \sum_{h=1}^H \mathbb{E}^{M, \pi} \left[Q_h^{M, \pi'}(s_h, \pi'(s_h)) - Q_h^{M, \pi'}(s_h, a_h) \mid s_1 = s \right] \quad (6.13)$$

Proof of Lemma 6.1. Fix a pair of policies π, π' and define

$$\pi^h = (\pi_1, \dots, \pi_{h-1}, \pi'_h, \dots, \pi'_H),$$

with $\pi^1 = \pi'$ and $\pi^{H+1} = \pi$. By telescoping, we can write

$$V_1^{M, \pi'}(s) - V_1^{M, \pi}(s) = \sum_{h=1}^H V_1^{M, \pi^h}(s) - V_1^{M, \pi^{h+1}}(s). \quad (6.14)$$

Observe that for each h , we have

$$V_1^{M, \pi^h}(s) - V_1^{M, \pi^{h+1}}(s) = \mathbb{E}^{M, \pi^h} \left[\sum_{h'=1}^H r_{h'} \mid s_1 = s \right] - \mathbb{E}^{M, \pi^{h+1}} \left[\sum_{h'=1}^H r_{h'} \mid s_1 = s \right]. \quad (6.15)$$

Here, one process evolves according to (M, π^h) and the other 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, (6.15) is equal to

$$\mathbb{E}^{M, \pi} \left[Q_h^{M, \pi'}(s_h, \pi'(s_h)) - Q_h^{M, \pi'}(s_h, \pi(s_h)) \mid s_1 = s \right] \quad (6.16)$$

which can be written as

$$\mathbb{E}^{M,\pi} \left[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, a_h) \mid s_1 = s \right]. \quad (6.17)$$

□

We state the performance difference lemma for completeness. It is not needed for the main results in this monograph, but it is the starting point for many analyses in reinforcement learning, and the decomposition that follows can be viewed as a variant of it.

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. The quantity $Q_h - \mathcal{T}_h^M Q_{h+1}$ appearing in the second part is called the *Bellman residual* of Q at layer h under M ; it measures how far Q is from being self-consistent under one application of the Bellman operator, and appears as a central notion of estimation error in the analyses that follow.

Lemma 6.2 (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 $\pi \in \Pi_{\text{RNS}}$,

$$V_1^{M,\pi}(s) - V_1^{\widehat{M},\pi}(s) = \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi} [Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1}) \mid s_1 = s] \quad (6.18)$$

Hence, for $M, \widehat{M}$ with the same initial state distribution,

$$f^M(\pi) - f^{\widehat{M}}(\pi) = \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi} [Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1})]. \quad (6.19)$$

In addition, for any MDP M and function $Q = (Q_1, \dots, Q_H, Q_{H+1})$ with $Q_{H+1} \equiv 0$, letting $\pi_{Q,h}(s) = \arg \max_{a \in \mathcal{A}} Q_h(s, a)$, we have

$$\max_{a \in \mathcal{A}} Q_1(s, a) - V_1^{M,\pi_Q}(s) = \sum_{h=1}^H \mathbb{E}^{M,\pi_Q} [Q_h(s_h, a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h, a_h) \mid s_1 = s]. \quad (6.20)$$

and, hence,

$$\mathbb{E}_{s_1 \sim d_1} [\max_{a \in \mathcal{A}} Q_1(s_1, a)] - f^M(\pi_Q) = \sum_{h=1}^H \mathbb{E}^{M,\pi_Q} [Q_h(s_h, a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h, a_h)]. \quad (6.21)$$

Note that for the second part of [Lemma 6.2](#) $Q = (Q_1, \dots, Q_H)$ 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 6.2. We prove (6.19), and omit the proof for (6.18), which is similar but more

verbose. We have

$$\begin{aligned} \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} [Q_h^{M, \pi}(s_h, a_h) - r_h - V_{h+1}^{M, \pi}(s_{h+1})] &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} [Q_h^{M, \pi}(s_h, a_h) - V_{h+1}^{M, \pi}(s_{h+1})] - \mathbb{E}^{\widehat{M}, \pi} \left[\sum_{h=1}^H r_h \right] \\ &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} [Q_h^{M, \pi}(s_h, a_h) - V_{h+1}^{M, \pi}(s_{h+1})] - f^{\widehat{M}}(\pi). \end{aligned}$$

On the other hand, since $V_{h+1}^{M, \pi}(s) = \mathbb{E}_{a \sim \pi_h(s)}[Q_h^{M, \pi}(s, a)]$, a telescoping argument yields

$$\begin{aligned} \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} [Q_h^{M, \pi}(s_h, a_h) - V_{h+1}^{M, \pi}(s_{h+1})] &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} [V_h^{M, \pi}(s_h) - V_{h+1}^{M, \pi}(s_{h+1})] \\ &= \mathbb{E}^{\widehat{M}, \pi} [V_1^{M, \pi}(s_1)] - \mathbb{E}^{\widehat{M}, \pi} [V_{H+1}^{M, \pi}(s_{H+1})] \\ &= f^M(\pi), \end{aligned}$$

where we have used that $V_{H+1}^{M, \pi} = 0$, and that both MDPs have the same initial state distribution. We prove (6.21) (omitting the proof of (6.20)) using a similar argument. We have

$$\begin{aligned} &\sum_{h=1}^H \mathbb{E}^{M, \pi_Q} [Q_h(s_h, a_h) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)] \\ &= \sum_{h=1}^H \mathbb{E}^{M, \pi_Q} [Q_h(s_h, a_h) - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)] - \mathbb{E}^{M, \pi_Q} \left[\sum_{h=1}^H r_h \right] \\ &= \sum_{h=1}^H \mathbb{E}^{M, \pi_Q} [Q_h(s_h, a_h) - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)] - f^M(\pi_Q). \end{aligned}$$

Since $a_{h+1} = \pi_Q(s_{h+1}) = \arg \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)$, we have

$$\mathbb{E}^{M, \pi_Q} [Q_h(s_h, a_h) - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)] = \mathbb{E}^{M, \pi_Q} [Q_h(s_h, a_h) - Q_{h+1}(s_{h+1}, a_{h+1})],$$

and the result follows by telescoping. □

A third tool, the *simulation lemma*, compares the value of a single policy under two different MDPs. We state it in two parts: the first bounds the value difference by the distance between the two trajectory distributions, and the second decomposes it into per-layer errors in transition probabilities and rewards. Both are used in Section 6.7 to analyze model-based exploration.

Lemma 6.3 (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 any policy π , we have

$$|f^M(\pi) - f^{\widehat{M}}(\pi)| \leq D_{\text{TV}}(M(\pi), \widehat{M}(\pi)) \tag{6.22}$$

$$\leq D_{\text{H}}(M(\pi), \widehat{M}(\pi)) \leq \frac{1}{2\eta} + \frac{\eta}{2} D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \quad \forall \eta > 0. \tag{6.23}$$

Lemma 6.4 (Layer-wise simulation lemma): In the setting of [Lemma 6.3](#),

$$\begin{aligned} f^M(\pi) - f^{\widehat{M}}(\pi) &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[\left[(P_h^M - P_h^{\widehat{M}}) V_{h+1}^{M, \pi} \right] (s_h, a_h) \right] \end{aligned} \quad (6.24)$$

$$+ \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[\mathbb{E}_{r_h \sim R_h^M(s_h, a_h)}[r_h] - \mathbb{E}_{r_h \sim R_h^{\widehat{M}}(s_h, a_h)}[r_h] \right] \quad (6.25)$$

$$\leq \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[D_{\text{TV}} \left(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h) \right) + D_{\text{TV}} \left(R_h^M(s_h, a_h), R_h^{\widehat{M}}(s_h, a_h) \right) \right]. \quad (6.26)$$

Proof of Lemma 6.3. Let $X = \sum_{h=1}^H r_h$. Since $X \in [0, 1]$ almost surely, we have

$$\left| f^M(\pi) - f^{\widehat{M}}(\pi) \right| = \left| \mathbb{E}^{M, \pi}[X] - \mathbb{E}^{\widehat{M}, \pi}[X] \right| \leq D_{\text{TV}} \left(M(\pi), \widehat{M}(\pi) \right) \leq D_{\text{H}} \left(M(\pi), \widehat{M}(\pi) \right).$$

The final result now follows from the AM-GM inequality. $\square$

Proof of Lemma 6.4. From [Lemma 6.2](#), we have

$$\begin{aligned} f^M(\pi) - f^{\widehat{M}}(\pi) &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[Q_h^{M, \pi}(s_h, a_h) - r_h - V_{h+1}^{M, \pi}(s_{h+1}) \right] \\ &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[\left[P_h^M V_{h+1}^{M, \pi} \right] (s_h, a_h) - V_{h+1}^{M, \pi}(s_{h+1}) + \mathbb{E}_{r_h \sim R_h^M(s_h, a_h)}[r_h] - \mathbb{E}_{r_h \sim R_h^{\widehat{M}}(s_h, a_h)}[r_h] \right] \\ &= \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[\left[(P_h^M - P_h^{\widehat{M}}) V_{h+1}^{M, \pi} \right] (s_h, a_h) \right] \\ &\quad + \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[\mathbb{E}_{r_h \sim R_h^M(s_h, a_h)}[r_h] - \mathbb{E}_{r_h \sim R_h^{\widehat{M}}(s_h, a_h)}[r_h] \right] \\ &\leq \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi} \left[D_{\text{TV}} \left(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h) \right) + D_{\text{TV}} \left(R_h^M(s_h, a_h), R_h^{\widehat{M}}(s_h, a_h) \right) \right], \end{aligned}$$

where we have used that $V_{h+1}^{M, \pi}(s) \in [0, 1]$. $\square$

When rewards are known and deterministic ([Assumption 6.1](#) below), the reward distributions of any two MDPs coincide, so the reward terms in [Eqs. \(6.24\) and \(6.25\)](#) vanish.

6.5 Optimism in Reinforcement Learning

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 3.1](#) gave a way to decompose the regret of optimistic algorithms into width of confidence intervals. What is the analogue of [Lemma 3.1](#)

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* $\bar{Q}_1, \dots, \bar{Q}_H$ which are guaranteed to over-estimate the optimal value function $Q^{M^*, \star}$ for the true MDP M^* . Since the optimal value functions of M^* appear frequently, we abbreviate $Q_h^* := Q_h^{M^*, \star}$ and $V_h^* := V_h^{M^*, \star}$, retaining the model superscript only when a generic $M \in \mathcal{M}$ is intended. For multi-armed bandits, implementing optimism amounted to adding “bonuses,” constructed from past data, to estimates for the reward function. We 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 6.5 (Error decomposition for optimistic policies): Let $\{\bar{Q}_1, \dots, \bar{Q}_H\}$ be a sequence of functions $\bar{Q}_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ with the property that for all (s, a) ,

$$Q_h^{M, \star}(s, a) \leq \bar{Q}_h(s, a) \quad (6.27)$$

and set $\bar{Q}_{H+1} \equiv 0$. Let $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_H)$ be such that $\hat{\pi}_h(s) = \arg \max_a \bar{Q}_h(s, a)$. Then for all $s \in \mathcal{S}$,

$$V_1^{M, \star}(s) - V_1^{M, \hat{\pi}}(s) \leq \sum_{h=1}^H \mathbb{E}^{M, \hat{\pi}}[(\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1})(s_h, \hat{\pi}(s_h)) \mid s_1 = s]. \quad (6.28)$$

Lemma 6.5 is an analogue of Lemma 3.1 for bandits, with Bellman residuals playing the role of confidence widths. Optimism lets us pass from the unknown π_M to the algorithm’s own policy $\hat{\pi}$, and what remains is a notion of estimation error for the optimistic value functions.

Lemma 6.5 tells us that closeness of $\bar{Q}_h$ to the Bellman backup $\mathcal{T}_h^M \bar{Q}_{h+1}$ implies closeness of $\hat{\pi}$ to π_M in terms of the value. As a sanity check, if $\bar{Q}_h = Q_h^{M, \star}$, the right-hand side of (6.27) 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 $\bar{Q}$ is not optimistic, it could have been the case that small changes in $\bar{Q}_h$ exponentially degrade the quality of the policy $\hat{\pi}$.

Another important aspect of the decomposition (6.27) 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 $\hat{\pi}$ 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 6.5. Let $\bar{V}_h(s) := \max_{a \in \mathcal{A}} \bar{Q}_h(s, a)$. Just as in the proof of Lemma 3.1, the assumption that $\bar{Q}_h$ is “optimistic” implies that

$$Q_h^{M, \star}(s_h, \pi_M(s_h)) \leq \bar{Q}_h(s_h, \pi_M(s_h)) \leq \bar{Q}_h(s_h, \hat{\pi}(s_h))$$

and, hence, $V_1^{M, \star}(s) \leq \bar{V}_1(s)$. Then, (6.20) applied to $Q = \bar{Q}$ and $\pi_Q = \hat{\pi}$ states that

$$\bar{V}_1(s) - V_1^{M, \hat{\pi}}(s) = \sum_{h=1}^H \mathbb{E}^{M, \hat{\pi}}[\bar{Q}_h(s_h, a_h) - [\mathcal{T}_h^M \bar{Q}_{h+1}](s_h, a_h) \mid s_1 = s]. \quad (6.29)$$

□

Remark 6.2 (Optimism is only needed at the first step): In fact, the proof of [Lemma 6.5](#) only uses that the initial value $\bar{Q}_1$ is optimistic. However, to construct a value function with this property, the algorithms we consider proceed by backwards induction, producing optimistic estimates $\bar{Q}_1, \dots, \bar{Q}_H$ for all steps h in the process.

6.6 Optimism for Tabular MDPs: The UCB-VI Algorithm

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 actions, but require that the state and action spaces are small. The regret bounds we present 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 6.1 (Known deterministic rewards): Rewards are deterministic and known to the learner: $R_h^M(s, a) = \delta_{r_h(s, a)}$ for known $r_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, for all M .

Together with the normalization $\sum_h r_h \in [0, 1]$ of [Section 6.1](#), this implies $V_1^{M, \star}(s) \in [0, 1]$ for all s and all M .

Define, with a slight abuse of notation,

$$n_h^t(s, a) = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i) = (s, a)\}, \quad \text{and} \quad n_h^t(s, a, s') = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\},$$

as the empirical state-action and state-action-next state frequencies. We can estimate the transition probabilities via

$$\hat{P}_h^t(s' | s, a) = \frac{n_h^t(s, a, s')}{n_h^t(s, a)}. \quad (6.30)$$

Together with the known rewards, the estimated transitions define an *estimated MDP* $\hat{M}^t := \{\mathcal{S}, \mathcal{A}, \{\hat{P}_h^t\}_{h=1}^H, \{r_h\}_{h=1}^H, d_1\}$. Under [Assumption 6.1](#), an MDP is determined by its transition kernels, so estimating $M^\star$ and estimating $\{P_h^{M^\star}\}_{h=1}^H$ are the same thing, and we use the two viewpoints interchangeably. The analysis of UCB-VI below works directly with $\hat{P}_h^t$; the estimated MDP $\hat{M}^t$ is the main object in [Section 6.7](#).

The UCB-VI algorithm. The following algorithm, UCB-VI (“Upper Confidence Bound Value Iteration”) [27], combines the notion of optimism with dynamic programming.

Algorithm 6.1 UCB-VI

input: Confidence parameter $\delta \in (0, 1)$.

for $t = 1, \dots, T$ **do**

Let $\bar{V}_{H+1}^t \equiv 0$.

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 the bonus computed in (6.31).

Compute:

$$\bar{Q}_h^t(s, a) = \left\{ r_h(s, a) + \mathbb{E}_{s' \sim \hat{P}_h^t(\cdot | s, a)} \bar{V}_{h+1}^t(s') + b_{h,\delta}^t(s, a) \right\} \wedge 1. \quad (6.31)$$

Set $\bar{V}_h^t(s) = \max_{a \in \mathcal{A}} \bar{Q}_h^t(s, a)$ and $\hat{\pi}_h^t(s) = \arg \max_{a \in \mathcal{A}} \bar{Q}_h^t(s, a)$.

Collect trajectory $(s_1^t, a_1^t, r_1^t), \dots, (s_H^t, a_H^t, r_H^t)$ according to $\hat{\pi}^t$.

We analyze the UCB-VI algorithm using Lemma 6.5. In constructing functions $\bar{Q}_h$, we need to satisfy two goals: (1) ensure that with high probability (6.26) is satisfied, i.e. the value functions are optimistic; and (2) that the functions $\bar{Q}_h$ are “self-consistent,” in the sense that the Bellman residuals in (6.27) are small. The second requirement already suggests that we should define $\bar{Q}_h$ approximately as a Bellman backup $\mathcal{T}_h^{M^*} \bar{Q}_{h+1}$, going backwards for $h = H+1, \dots, 1$ as in dynamic programming, while ensuring the first requirement. In addition to these considerations, we must 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 (6.30). 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 $\mathcal{T}_h^{M^*}$ is replaced by that of the estimated MDP, $\mathcal{T}_h^{\hat{M}^t}$, which uses the estimated transition probabilities in (6.29).

The bonus functions play precisely the same role as the width of the confidence interval in (3.20): these bonuses ensure that (6.26) holds with high probability, as Lemma 6.6 shows.

The following theorem shows that with an appropriate choice of bonus, this algorithm achieves a polynomial regret bound.

Theorem 6.1 (Regret of UCB-VI): For any $\delta > 0$, UCB-VI with

$$b_{h,\delta}^t(s, a) = 2 \sqrt{\frac{\log(2SAHT/\delta)}{n_h^t(s, a)}} \quad (6.32)$$

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 6.8](#) below (using the Freedman inequality instead of the Azuma-Hoeffding inequality) yields an improved rate of $O(H\sqrt{SAT} + \text{poly}(H, S, A) \log T)$. The optimal rate is $\Theta(\sqrt{HSAT})$ (a formal lower bound is [Proposition 8.4](#) in [Section 8](#)), which is achieved using the same template with a more careful choice for the bonus $b_{h,\delta}^t$. 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 6.1](#)), as well as stationary versus non-stationary transition assumptions, can vary.

6.6.1 Optimism and Bellman Residuals Within an Episode

In what follows, we prove [Theorem 6.1](#). The analysis features two of the same ingredients as that of UCB for multi-armed bandits: bonuses to ensure optimism, and a potential argument that controls the number of times the algorithm can be surprised by bad state-action pairs. What changes is that we need [Lemma 6.5](#) to relate the algorithm's (global) suboptimality to its (local) per-state action bonuses.

To begin, we first prove several helper lemmas concerning the performance within each episode t . Throughout this subsection, we fix t and drop the superscript t . Given the estimated transitions $\{\hat{P}_h(\cdot | s, a)\}_{h,s,a}$, define the estimated MDP $\widehat{M} = \{\mathcal{S}, \mathcal{A}, \{\hat{P}_h\}_{h=1}^H, \{R_h^{M^*}\}_{h=1}^H, d_1\}$. The associated Bellman operator is

$$\mathcal{T}_h^{\widehat{M}} Q(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot | s, a)} \max_a Q(s', a) \quad (6.33)$$

for $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Consider the sequence of functions $\bar{Q}_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, $\bar{V}_h : \mathcal{S} \rightarrow [0, 1]$, for $h = 1, \dots, H+1$, with $\bar{Q}_{H+1} \equiv 0$ and

$$\bar{Q}_h(s, a) = \left\{ [\mathcal{T}_h^{\widehat{M}} \bar{Q}_{h+1}](s, a) + b_{h,\delta}(s, a) \right\} \wedge 1, \quad \text{and} \quad \bar{V}_h(s) = \max_a \bar{Q}_h(s, a) \quad (6.34)$$

for bonus functions $b_{h,\delta} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ that we leave unspecified for now. Henceforth, we follow the usual notation that for functions f, g over the same domain, $f \leq g$ 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 $\bar{Q}_1, \dots, \bar{Q}_H$ constructed above are optimistic.

Lemma 6.6 (Bonuses ensure optimism): Suppose we have estimates $\{\hat{P}_h(\cdot | s, a)\}_{h,s,a}$ and a function $b_{h,\delta} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ with the property that for all $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\left| \sum_{s'} \hat{P}_h(s' | s, a) V_h^*(s') - \sum_{s'} P_h^{M^*}(s' | s, a) V_h^*(s') \right| \leq b_{h,\delta}(s, a). \quad (6.35)$$

Then for all $h \in [H]$, we have

$$\bar{Q}_h \geq Q_h^*, \quad \text{and} \quad \bar{V}_h \geq V_h^* \quad (6.36)$$

for $\bar{Q}_h, \bar{V}_h$ defined in (6.33).

Proof of Lemma 6.6. The proof proceeds by backward induction on the statement

$$\bar{V}_h \geq V_h^*$$

with $h = H + 1$ down to $h = 1$. We start with the base case $h = H + 1$, which is trivial because $\bar{V}_{H+1} = V_{H+1}^* \equiv 0$. Now, assume $\bar{V}_{h+1} \geq V_{h+1}^*$, and let us prove the induction step. Fix $(s, a) \in \mathcal{S} \times \mathcal{A}$. If $\bar{Q}_h(s, a) = 1$, then, trivially, $\bar{Q}_h(s, a) \geq Q_h^*(s, a)$. Otherwise, $\bar{Q}_h(s, a) = \mathcal{T}_h^{\bar{M}} \bar{Q}_{h+1}(s, a) + b_{h,\delta}(s, a)$, and thus

$$\begin{aligned} \bar{Q}_h(s, a) - Q_h^*(s, a) &= b_{h,\delta}(s, a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot | s, a)} \bar{V}_{h+1}(s') - \mathbb{E}_{s' \sim P_h^{M^*}(\cdot | s, a)} V_{h+1}^*(s') \\ &\geq b_{h,\delta}(s, a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot | s, a)} V_{h+1}^*(s') - \mathbb{E}_{s' \sim P_h^{M^*}(\cdot | s, a)} V_{h+1}^*(s') \geq 0. \end{aligned}$$

This, in turn, implies that $\bar{V}_h(s) = \max_a \bar{Q}_h(s, a) \geq \max_a Q_h^*(s, a) = V_h^*(s)$, concluding the induction step. $\square$

We now analyze the effect of using an estimated model $\hat{M}$ for the Bellman operator rather than the true unknown $\mathcal{T}_h^{M^*}$.

Lemma 6.7 (Bellman residuals of the optimistic value functions): Suppose we have estimates $\{\hat{P}_h(\cdot | s, a)\}_{h,s,a}$ and $b'_{h,\delta}(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ with the property that

$$\max_{V \in \{0,1\}^S} \left| \sum_{s'} \hat{P}_h(s' | s, a) V(s') - \sum_{s'} P_h^{M^*}(s' | s, a) V(s') \right| \leq b'_{h,\delta}(s, a). \quad (6.37)$$

Then the Bellman residual satisfies

$$\bar{Q}_h - \mathcal{T}_h^{M^*} \bar{Q}_{h+1} \leq (b_{h,\delta} + b'_{h,\delta}) \wedge 1 \quad (6.38)$$

for $\bar{Q}_h, \bar{V}_h$ defined in (6.33).

Proof of Lemma 6.7. That $\bar{Q}_h - \mathcal{T}_h^{M^*} \bar{Q}_{h+1} \leq 1$ is immediate. To prove the main result, observe that

$$\bar{Q}_h - \mathcal{T}_h^{M^*} \bar{Q}_{h+1} = \left\{ \mathcal{T}_h^{\hat{M}} \bar{Q}_{h+1} + b_{h,\delta} \right\} \wedge 1 - \mathcal{T}_h^{M^*} \bar{Q}_{h+1} \leq (\mathcal{T}_h^{\hat{M}} - \mathcal{T}_h^{M^*}) \bar{Q}_{h+1} + b_{h,\delta} \quad (6.39)$$

For any $Q \in \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$,

$$(\mathcal{T}_h^{\hat{M}} - \mathcal{T}_h^{M^*})Q(s, a) = \mathbb{E}_{s' \sim \hat{P}_h(\cdot | s, a)} \max_a Q(s', a) - \mathbb{E}_{s' \sim P_h^{M^*}(\cdot | s, a)} \max_a Q(s', a) \quad (6.40)$$

$$\leq \max_{V \in \{0,1\}^S} |\mathbb{E}_{s' \sim \hat{P}_h(\cdot | s, a)} V(s') - \mathbb{E}_{s' \sim P_h^{M^*}(\cdot | s, a)} V(s')|. \quad (6.41)$$

Since the maximum is achieved at a vertex of $[0, 1]^S$, the statement follows. $\square$

6.6.2 Concentration and Regret Bound

We now bring back the time index t and show that the estimated transition probabilities in UCB-VI satisfy conditions of Lemma 6.6 and Lemma 6.7, ensuring that the functions $\bar{Q}_1^t, \dots, \bar{Q}_H^t$ are optimistic.

Lemma 6.8 (Concentration for empirical transitions): Let $\{\hat{P}_h^t\}_{h \in [H], t \in [T]}$ be defined as in (6.29). Then with probability at least $1 - \delta$, the functions

$$b_{h,\delta}^t(s, a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n_h^t(s, a)}}, \quad \text{and} \quad b'_{h,\delta}(s, a) = 8\sqrt{\frac{S \log(2SAHT/\delta)}{n_h^t(s, a)}}$$

satisfy the assumptions of Lemma 6.6 and Lemma 6.7, respectively, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$, and $t \in [T]$ simultaneously. $\square$

Proof of Lemma 6.8. We leave the proof as an exercise to the reader, as it follows similar reasoning to other concentration results throughout the monograph. $\square$

Proof of Theorem 6.1. Putting everything together, we can now prove Theorem 6.1. Under the event in Lemma 6.8, the functions $\bar{Q}_1^t, \dots, \bar{Q}_H^t$ are optimistic, which means that the conditions of Lemma 6.5 hold, and the instantaneous regret on round t (conditionally on $s_1 \sim d_1$) is at most

$$\begin{aligned} & \sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [(\bar{Q}_h^t - \mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t)(s_h^t, \hat{\pi}_h^t(s_h^t)) \mid s_1 = s] \\ & \leq \sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [(b_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t))) \wedge 1], \end{aligned}$$

where the second inequality invokes Lemma 6.7; note that the expectation is under $\hat{\pi}^t$ itself, the on-policy roll-in distribution discussed after Lemma 6.5, which is what allows the bonuses to be summed along the algorithm's own trajectories. Summing over $t = 1, \dots, T$, and applying the Azuma-Hoeffding inequality, we have that with probability at least $1 - \delta$, the regret of UCB-VI is bounded by

$$\begin{aligned} & \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [(b_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t))) \wedge 1] \\ & \lesssim \sum_{t=1}^T \sum_{h=1}^H (b_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t))) \wedge 1 + \sqrt{HT \log(1/\delta)}. \end{aligned}$$

Using the bonus definition in (6.31), the bonus term above is bounded by

$$\sum_{t=1}^T \sum_{h=1}^H \sqrt{\frac{S \log(2SAHT/\delta)}{n_h^t(s_h^t, \hat{\pi}_h^t(s_h^t))}} \wedge 1 \leq \sqrt{S \log(2SAHT/\delta)} \sum_{t=1}^T \sum_{h=1}^H \frac{1}{\sqrt{n_h^t(s_h^t, \hat{\pi}_h^t(s_h^t))}} \wedge 1 \quad (6.42)$$

The double summation can be handled in the same fashion as Lemma 3.2:

$$\begin{aligned} \sum_{t=1}^T \sum_{h=1}^H \frac{1}{\sqrt{n_h^t(s_h^t, \hat{\pi}_h^t(s_h^t))}} \wedge 1 &= \sum_{h=1}^H \sum_{(s,a)} \sum_{t=1}^T \frac{\mathbb{I}\{(s_h^t, \hat{\pi}_h^t(s_h^t)) = (s, a)\}}{\sqrt{n_h^t(s, a)}} \wedge 1 \\ &\lesssim \sum_{h=1}^H \sum_{(s,a)} \sqrt{n_h^T(s, a)} \leq H \sqrt{SAT}. \end{aligned}$$

$\square$

6.7 Estimation-to-Decisions for Tabular Reinforcement Learning

We now give a second approach to tabular reinforcement learning, based on the Estimation-to-Decisions principle of Sections 4 and 5 rather than on optimism. The estimation step is the same as in UCB-VI: from the trajectories observed so far, we form an estimated MDP $\widehat{M}^t$. However, instead of planning optimistically in $\widehat{M}^t$, we adapt the Estimation-to-Decisions principle to turn the estimate into an exploratory distribution over policies whose regret is controlled by the estimation error of $\widehat{M}^t$. Doing so requires an exploration strategy that exploits the structure of MDPs, which serves as a preview for the most general form of the Decision-Estimation Coefficient we introduce in Section 7.

Throughout this subsection, we adopt Assumption 6.1. We write $\mathcal{M}$ for the set of all tabular MDPs with state space $\mathcal{S} = [S]$, action space $\mathcal{A} = [A]$, horizon H , and the known rewards r_h ; a model $M \in \mathcal{M}$ is thus specified by its transition kernels. For $M \in \mathcal{M}$ and $\pi \in \Pi_{\text{rs}}$, $M(\pi)$ denotes the law of the trajectory $(s_1, a_1, r_1, \dots, s_H, a_H, r_H)$ under π in M .

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 .

Instantiating Estimation-to-Decisions: Policy Cover Inverse Gap Weighting. The main ingredient in our algorithm is the following subroutine which, given an estimated MDP $\widehat{M} \in \mathcal{M}$ and a parameter $\eta > 0$, computes a distribution p over policies. The strategy, *Policy Cover Inverse Gap Weighting*, combines inverse gap weighting with the notion of a *policy cover*—a collection of policies that ensures good coverage of every state [69, 152, 114].

Algorithm 6.2 Policy Cover Inverse Gap Weighting (PC-IGW)

input: Estimated model $\widehat{M} \in \mathcal{M}$; exploration parameter $\eta > 0$.

output: Distribution $p \in \Delta(\Pi_{\text{rs}})$.

Define *inverse gap weighted policy cover* $\Psi = \{\pi_{h,s,a}\}_{h \in [H], s \in [S], a \in [A]}$ via

$$\pi_{h,s,a} = \arg \max_{\pi \in \Pi_{\text{rs}}} \frac{d_h^{\widehat{M},\pi}(s, a)}{2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}. \quad (6.43)$$

For each policy $\pi \in \Psi \cup \{\pi_{\widehat{M}}\}$, define

$$p(\pi) = \frac{1}{\lambda + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}, \quad (6.44)$$

where $\lambda \in [1, 2HSA]$ is chosen such that $\sum_{\pi} p(\pi) = 1$.

return p .

The algorithm consists of two steps. First, in (6.42), 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

$$\pi_{h,s,a} := \arg \max_{\pi \in \Pi_{\text{rns}}} \frac{d_h^{\widehat{M}, \pi}(s, a)}{2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}$$

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 exploitation by trading off coverage with suboptimality. With the policy cover in hand, the second step of the algorithm computes the exploratory distribution p by simply applying inverse gap weighting to the elements of the cover and the greedy policy $\pi_{\widehat{M}}$. A naive alternative would be to apply inverse gap weighting to all of Π_{rns} at once; this would pay $|\Pi_{\text{rns}}|/\gamma$, which is exponential in S and H . The cover replaces this count by HSA , as the following proposition shows.

Proposition 6.2 (PC-IGW links regret to estimation error): Under [Assumption 6.1](#), for any $\gamma > 0$ and $\widehat{M} \in \mathcal{M}$, the PC-IGW strategy with $\eta = \frac{\gamma}{21H^2}$ ensures that for all $M \in \mathcal{M}$,

$$\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \lesssim \frac{H^3SA}{\gamma} + \gamma \cdot \mathbb{E}_{\pi \sim p}[D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))].$$

[Proposition 6.2](#) has the same form as (4.41), with H^3SA/γ in place of A/γ and Hellinger distance between trajectory distributions in place of squared error of rewards. That such an inequality holds at all is the non-trivial content of the result. The estimate $\widehat{M}$ is an entire MDP, and the Hellinger error on the right-hand side is measured only along trajectories that p actually generates; an error in the transitions at a state that p rarely visits is invisible to the right-hand side, yet could in principle be exactly what makes a policy suboptimal. The policy cover is what rules this out: every state–action pair that can be reached with non-negligible probability under *some* policy is visited by p with probability inversely proportional to that policy’s estimated suboptimality, so any misestimation that could cause regret is paid for by the oracle.

From the inequality to an algorithm. We turn the inequality in [Proposition 6.2](#) into an algorithm using the same Estimation-to-Decisions template as in [Section 4.7](#). In each episode t an online estimation oracle produces estimated transition kernels, and hence an estimated MDP $\widehat{M}^t \in \mathcal{M}$, from the trajectories observed so far, and we play $\pi^t \sim p^t := \text{PC-IGW}(\widehat{M}^t)$. The oracle now estimates distributions rather than means, so its error is measured in Hellinger distance: in direct analogy with [Definition 4.1](#), we assume an oracle whose cumulative Hellinger error

$$\mathbf{Est}_{\text{H}} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left[D_{\text{H}}^2(M^*(\pi^t), \widehat{M}^t(\pi^t)) \right]$$

satisfies, with probability at least $1 - \delta$, $\mathbf{Est}_{\text{H}} \leq \mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta)$ for some deterministic bound $\mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta)$. Informally, this guarantee states that M^* and $\widehat{M}^t$ must induce similar distributions over trajectories, on average over the policies sampled from p^t . This is the online estimation oracle of [Definition 7.1](#) in [Section 7](#), restricted to proper estimates $\widehat{M}^t \in \mathcal{M}$.

To minimize the estimation error, the following result shows that a natural choice—the empirical transition frequencies of UCB-VI, smoothed by adding one pseudo-count to each next state—has the required guarantee. This is Laplace’s rule of succession [134], applied separately for each (h, s, a) ; equivalently, this is the continuous-prior analogue of the Bayes mixture predictor of Lemma 2.5, with the uniform distribution over the simplex $\Delta(\mathcal{S})$ as the prior. The log-loss regret bound is classical [47, Chapter 9].

Proposition 6.3 (Hellinger estimation for tabular MDPs): Let $\widehat{M}^t \in \mathcal{M}$ be the MDP with transition kernels

$$\widehat{P}_h^t(s' | s, a) = \frac{n_h^t(s, a, s') + 1}{n_h^t(s, a) + S}, \quad (6.45)$$

where $n_h^t(s, a)$ and $n_h^t(s, a, s')$ are the empirical counts defined in Section 6.6. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\mathbf{Est}_H \leq HS^2A \log(e(T + S)) + 2 \log(\delta^{-1}).$$

This holds for any sequence of distributions $p^1, \dots, p^T$ used to select $\pi^1, \dots, \pi^T$.

The proof is deferred to Section 6.9. Summing Proposition 6.2 over episodes and plugging in Proposition 6.3 gives the following end-to-end regret bound.

Theorem 6.2 (Regret of Estimation-to-Decisions with PC-IGW): Under Assumption 6.1, the algorithm that plays $\pi^t \sim \text{PC-IGW}(\widehat{M}^t)$ with exploration parameter $\gamma > 0$ has the following guarantees.

1. For any estimation oracle and every realization,

$$\mathbf{Reg} \lesssim \frac{H^3SA}{\gamma} \cdot T + \gamma \cdot \mathbf{Est}_H.$$

Consequently, if the oracle has estimation error $\mathbf{Est}_H(\mathcal{M}, T, \delta)$, then with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{H^3SA \cdot T \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta)} \quad \text{for} \quad \gamma = \sqrt{H^3SA \cdot T / \mathbf{Est}_H(\mathcal{M}, T, \delta)}.$$

2. With the estimator (6.44) and $\gamma = \sqrt{H^3SA \cdot T / (HS^2A \log(e(T + S)) + 2 \log(\delta^{-1}))}$, with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{H^4S^3A^2T \log(e(T + S)) + H^3SAT \log(\delta^{-1})}.$$

Proof of Theorem 6.2. The first bound is proven identically to Theorem 4.2, with Proposition 6.2 in place of Proposition 4.3: adding and subtracting $\gamma \cdot \mathbb{E}_{\pi^t \sim p^t} \left[D_H^2 \left(M^*(\pi^t), \widehat{M}^t(\pi^t) \right) \right]$ in each episode and applying Proposition 6.2 with $M = M^*$ and $\widehat{M} = \widehat{M}^t$ bounds each episode’s regret by H^3SA/γ plus γ times that episode’s estimation error; summing over episodes gives the bound. The second bound in the first part follows on the event that $\mathbf{Est}_H \leq \mathbf{Est}_H(\mathcal{M}, T, \delta)$, by the choice of γ . $\square$

The rate in Theorem 6.2 is polynomial in all parameters, though worse than the $HS\sqrt{AT}$ rate of

UCB-VI (Theorem 6.1); the point of the argument is its generality rather than its constants. We return to this construction in Section 7 (Example 7.7), as an instance of the more general DEC theory.

Remark 6.3 (Computational efficiency): The PC-IGW strategy can be implemented in a computationally efficient fashion. Briefly, the idea is to solve (6.42) by taking a dual approach and optimizing over occupancy measures rather than policies. With this parameterization, (6.42) becomes a linear-fractional program, which can then be transformed into a standard linear program using classical techniques.

6.7.1 Preliminary Lemma: Change of Measure

The proof of Proposition 6.2 uses the simulation lemma (Lemma 6.3) from Section 6.4, together with the following “change-of-measure” lemma, which allows one to move between quantities involving an estimator $\widehat{M}$ and those involving another model M .

Lemma 6.9 (Change of measure for RL): Consider any MDPs $M, \widehat{M} \in \mathcal{M}$ (so that both satisfy Assumption 6.1). For all $p \in \Delta(\Pi)$ and $\eta > 0$ we have

$$\begin{aligned} & \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \\ & \leq \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^{\widehat{M}}(\pi)] + \eta \mathbb{E}_{\pi \sim p}[D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))] + \frac{1}{4\eta}. \end{aligned} \quad (6.46)$$

and

$$\mathbb{E}_{\pi \sim p} \mathbb{E}^{\widehat{M}, \pi} \left[\sum_{h=1}^H D_{\text{TV}}^2(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h)) \right] \leq 8H \mathbb{E}_{\pi \sim p}[D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))]. \quad (6.47)$$

Proof of Lemma 6.9. We first prove (6.45). Writing $f^M(\pi_M) - f^M(\pi) = f^M(\pi_M) - f^{\widehat{M}}(\pi) + f^{\widehat{M}}(\pi) - f^M(\pi)$ and applying (6.22) (with 2η in place of η) to the second difference, we have that for all $\eta > 0$,

$$\begin{aligned} & \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \\ & \leq \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^{\widehat{M}}(\pi)] + \eta \mathbb{E}_{\pi \sim p}[D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))] + \frac{1}{4\eta}. \end{aligned}$$

We now prove (6.46). Using Lemma A.5, we have that for all h ,

$$\mathbb{E}^{\widehat{M}, \pi} \left[D_{\text{H}}^2(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h)) \right] \leq 8D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)).$$

As a result, using that $D_{\text{TV}}^2(\cdot, \cdot) \leq D_{\text{H}}^2(\cdot, \cdot)$,

$$\mathbb{E}^{\widehat{M}, \pi} \left[\sum_{h=1}^H D_{\text{TV}}^2(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h)) \right] \leq 8H D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)).$$

Since this holds uniformly for all π , taking expectation over $\pi \sim p$ gives (6.46). □

6.7.2 Proof of Proposition 6.2

Proof of Proposition 6.2. Let $M \in \mathcal{M}$ be fixed. The main effort in the proof is to bound the quantity

$$\mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^{\widehat{M}}(\pi) \right]$$

in terms of the quantity on the right-hand side of (6.46), then apply change of measure (Lemma 6.9). We begin with the decomposition

$$\mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^{\widehat{M}}(\pi) \right] = \underbrace{\mathbb{E}_{\pi \sim p} \left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi) \right]}_{(I)} + \underbrace{f^M(\pi_M) - f^{\widehat{M}}(\pi_{\widehat{M}})}_{(II)}. \quad (6.48)$$

For the first term (I), which may be thought of as exploration bias, we have

$$\mathbb{E}_{\pi \sim p} \left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi) \right] = \sum_{\pi \in \Psi \cup \{\pi_{\widehat{M}}\}} \frac{f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)}{\lambda + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))} \leq \frac{2HSA}{\eta}, \quad (6.49)$$

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

$$f^M(\pi_M) - f^{\widehat{M}}(\pi_{\widehat{M}}) = f^M(\pi_M) - f^{\widehat{M}}(\pi_M) - (f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M)). \quad (6.50)$$

We use the simulation lemma to bound

$$f^M(\pi_M) - f^{\widehat{M}}(\pi_M) \leq \sum_{h=1}^H \mathbb{E}^{\widehat{M}, \pi_M} \left[D_{\text{TV}} \left(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h) \right) \right] = \sum_{h=1}^H \sum_{s,a} d_h^{\widehat{M}, \pi_M}(s, a) \text{err}_h^M(s, a),$$

where $\text{err}_h^M(s, a) := D_{\text{TV}}(P_h^M(s, a), P_h^{\widehat{M}}(s, a))$ (the reward terms in (6.25) vanish under Assumption 6.1). Define $\bar{d}_h(s, a) = \mathbb{E}_{\pi \sim p} [d_h^{\widehat{M}, \pi}(s, a)]$. Then, using the AM-GM inequality, we have that for any $\eta' > 0$,

$$\begin{aligned} \sum_{h=1}^H \sum_{s,a} d_h^{\widehat{M}, \pi_M}(s, a) [\text{err}_h^M(s, a)] &= \sum_{h=1}^H \sum_{s,a} d_h^{\widehat{M}, \pi_M}(s, a) \left(\frac{\bar{d}_h(s, a)}{\bar{d}_h(s, a)} \right)^{1/2} [\text{err}_h^M(s, a)] \\ &\leq \frac{1}{2\eta'} \sum_{h=1}^H \sum_{s,a} \frac{(d_h^{\widehat{M}, \pi_M}(s, a))^2}{\bar{d}_h(s, a)} + \frac{\eta'}{2} \sum_{h=1}^H \sum_{s,a} \bar{d}_h(s, a) (\text{err}_h^M(s, a))^2 \\ &= \frac{1}{2\eta'} \sum_{h=1}^H \sum_{s,a} \frac{(d_h^{\widehat{M}, \pi_M}(s, a))^2}{\bar{d}_h(s, a)} + \frac{\eta'}{2} \sum_{h=1}^H \mathbb{E}_{\pi \sim p} \mathbb{E}^{\widehat{M}, \pi} [(\text{err}_h^M(s_h, a_h))^2]. \end{aligned}$$

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

$$\frac{d_h^{\widehat{M}, \pi_M}(s, a)}{\bar{d}_h(s, a)} \leq \frac{d_h^{\widehat{M}, \pi_M}(s, a)}{d_h^{\widehat{M}, \pi_{h,s,a}}(s, a)} \cdot \frac{1}{p(\pi_{h,s,a})} \leq \frac{d_h^{\widehat{M}, \pi_M}(s, a)}{d_h^{\widehat{M}, \pi_{h,s,a}}(s, a)} \left(2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_{h,s,a})) \right),$$

where the second inequality follows from the definition of p and the fact that $\lambda \leq 2HSA$. Furthermore, since

$$\pi_{h,s,a} = \arg \max_{\pi \in \Pi_{\text{rs}}} \frac{d_h^{\widehat{M}, \pi}(s, a)}{2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))},$$

and since $\pi_M \in \Pi_{\text{rns}}$, we can upper bound by

$$\frac{d_h^{\widehat{M}, \pi_M}(s, a)}{d_h^{\widehat{M}, \pi_M}(s, a)} \left(2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M)) \right) = 2HSA + \eta(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M)). \quad (6.51)$$

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 (6.49), this establishes that

$$\begin{aligned} &f^M(\pi_M) - f^{\widehat{M}}(\pi_{\widehat{M}}) \\ &\leq \frac{H^2SA}{\eta'} + \frac{\eta'}{2} \sum_{h=1}^H \mathbb{E}_{\pi \sim p} \mathbb{E}^{\widehat{M}, \pi} [(\text{err}_h^M(s_h, a_h))^2] + \frac{\eta H}{2\eta'} (f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M)) - (f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M)). \end{aligned}$$

We set $\eta' = \frac{\eta H}{2}$ so that the latter terms cancel and we are left with

$$f^M(\pi_M) - f^{\widehat{M}}(\pi_{\widehat{M}}) \leq \frac{2HSA}{\eta} + \frac{\eta H}{4} \sum_{h=1}^H \mathbb{E}_{\pi \sim p} \mathbb{E}^{\widehat{M}, \pi} [(\text{err}_h^M(s_h, a_h))^2].$$

Combining this with (6.47) and (6.48) gives

$$\begin{aligned} &\mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^{\widehat{M}}(\pi)] \\ &\leq \frac{4HSA}{\eta} + \frac{\eta H}{4} \sum_{h=1}^H \mathbb{E}_{\pi \sim p} \mathbb{E}^{\widehat{M}, \pi} [(\text{err}_h^M(s_h, a_h))^2] \\ &= \frac{4HSA}{\eta} + \frac{\eta H}{4} \sum_{h=1}^H \mathbb{E}_{\pi \sim p} \mathbb{E}^{\widehat{M}, \pi} \left[D_{\text{TV}}^2 \left(P_h^M(s_h, a_h), P_h^{\widehat{M}}(s_h, a_h) \right) \right]. \end{aligned}$$

We conclude by applying the change-of-measure lemma (Lemma 6.9), which implies that for any $\eta' > 0$,

$$\mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)] \leq \frac{4HSA}{\eta} + (4\eta')^{-1} + (2H^2\eta + \eta') \cdot \mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right].$$

The result follows by choosing $\eta = \eta' = \frac{\gamma}{21H^2}$ (we have made no effort to optimize the constants here). $\square$

6.8 Bibliographic Remarks and Further Reading

Takeaways. This section introduced the finite-horizon episodic MDP, showed that planning in a known MDP reduces to dynamic programming, and gave two approaches to online reinforcement learning in tabular MDPs. Both build the same estimated MDP $\widehat{M}^t$ from empirical transition frequencies and differ only in what they do with it: UCB-VI plans optimistically in $\widehat{M}^t$ and is analyzed through the Bellman residuals of its optimistic value functions, while PC-IGW uses $\widehat{M}^t$ to build an exploratory distribution over a policy cover, analyzed in Hellinger distance; Section 7 shows that the latter is the DEC construction for tabular MDPs. The analysis tools in Section 6.4 recur throughout the monograph.

Bibliographic remarks. Bellman’s principle of optimality and value iteration ([Proposition 6.1](#)) go back to Bellman [31]; see Puterman [169] and Sutton and Barto [194] for textbook treatments. The performance difference lemma ([Lemma 6.1](#)) is due to Kakade [121], and the simulation lemma ([Lemma 6.3](#)) takes its name from Kearns and Singh [124]; the Bellman residual decomposition ([Lemma 6.2](#)) is a telescoping identity that appears in many forms, with Jiang and Agarwal [110] being perhaps the most notable example.

Optimism for tabular RL originates with E^3 and R-max [124, 38] and, for regret minimization, with UCRL2 [108]. UCB-VI ([Algorithm 6.1](#)) and [Theorem 6.1](#) are from Azar et al. [27]; [Lemma 6.5](#), [Lemma 6.6](#), and [Lemma 6.7](#) follow their analysis. The $\Omega(\sqrt{HSAT})$ lower bound quoted after [Theorem 6.1](#) is from Domingues et al. [66], refining Jaksch et al. [108], Osband and Van Roy [159].

The PC-IGW strategy ([Algorithm 6.2](#)), [Proposition 6.2](#), and the change-of-measure lemma ([Lemma 6.9](#)) are adapted from Foster et al. [84], where the same argument bounds the Decision-Estimation Coefficient for tabular MDPs and [Theorem 6.2](#) is a special case of the general E2D guarantee. The notion of policy cover used in the algorithm is developed in Du et al. [69], Misra et al. [152], Jin et al. [114]. The estimator in [Proposition 6.3](#) is classical [47, Chapter 9].

Further reading. The monograph of Agarwal et al. [10] covers tabular RL in depth. The bound in [Theorem 6.1](#) is loose in S and H ; Bernstein-type bonuses give the minimax rate [27, 221], and Zhang et al. [228] remove the remaining horizon dependence under the normalization of [Assumption 6.1](#); see Jiang and Agarwal [110] for why this normalization matters. Model-free counterparts of UCB-VI, which maintain optimistic Q -functions without an estimated model, begin with Jin et al. [112]. Posterior sampling for tabular RL is analyzed by Osband et al. [161], Osband and Van Roy [160].

6.9 Deferred Proofs

This proof uses somewhat standard machinery from the theory of online learning; we refer to [Lemma 7.3](#) for further background.

Proof of [Proposition 6.3](#). Under [Assumption 6.1](#), the density of a trajectory $\tau = (s_1, a_1, \dots, s_H, a_H)$ under policy π in an MDP $M \in \mathcal{M}$ is $d_1(s_1) \prod_{h=1}^H \pi_h(a_h | s_h) P_h^M(s_{h+1} | s_h, a_h)$ (rewards are a deterministic function of the trajectory, and the final transition can be ignored). Hence, for the trajectory τ^t observed at round t , the log-loss difference between the true model and the estimate satisfies

$$\log \frac{M^\star(\tau^t | \pi^t)}{\widehat{M}^t(\tau^t | \pi^t)} = \sum_{h=1}^H \log \frac{P_h^{M^\star}(s_{h+1}^t | s_h^t, a_h^t)}{\widehat{P}_h^t(s_{h+1}^t | s_h^t, a_h^t)}, \quad (6.52)$$

since the initial state and policy terms cancel. For each cell (h, s, a) , let $\mathcal{T}_{h,s,a} := \{t \in [T] : (s_h^t, a_h^t) = (s, a)\}$ be the set of rounds in which the cell is visited, and define

$$R_{h,s,a} := \sum_{t \in \mathcal{T}_{h,s,a}} \log \frac{P_h^{M^\star}(s_{h+1}^t | s, a)}{\widehat{P}_h^t(s_{h+1}^t | s, a)}, \quad (6.53)$$

the log-loss regret of the estimator (6.44) relative to the fixed distribution $P_h^{M^\star}(\cdot | s, a)$ on the subsequence of next states observed at this cell. Since every pair (t, h) belongs to exactly one cell,

summing (6.51) over t gives

$$\sum_{t=1}^T \log \frac{M^\star(\tau^t \mid \pi^t)}{\widehat{M}^t(\tau^t \mid \pi^t)} = \sum_{h=1}^H \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} R_{h,s,a}. \quad (6.54)$$

We bound $R_{h,s,a}$ for a fixed cell. Write $n = |\mathcal{T}_{h,s,a}|$ for the number of visits, $y_1, \dots, y_n \in \mathcal{S}$ for the next states observed on these visits in order, and n_j for the number of occurrences of state $j \in \mathcal{S}$ among $y_1, \dots, y_n$, so that $\sum_j n_j = n$. Let $n_j(k)$ be the number of occurrences of j among $y_1, \dots, y_k$. At the k -th visit, the counts in (6.44) are those of the first $k-1$ next states, so the estimator assigns probability

$$\widehat{P}_k(y_k) = \frac{n_{y_k}(k-1) + 1}{k-1 + S}$$

to the state y_k that is actually observed. In this notation,

$$R_{h,s,a} = \sum_{k=1}^n \log \frac{P_h^{M^\star}(y_k \mid s, a)}{\widehat{P}_k(y_k)} = \log \frac{\prod_{k=1}^n P_h^{M^\star}(y_k \mid s, a)}{\prod_{k=1}^n \widehat{P}_k(y_k)}.$$

We compute the two products separately.

The estimator's product likelihood. The denominators of $\widehat{P}_1(y_1), \dots, \widehat{P}_n(y_n)$ are $S, S+1, \dots, n+S-1$, whose product is $\frac{(n+S-1)!}{(S-1)!}$. For the numerators, fix a state j and consider the visits k at which $y_k = j$: at the first such visit $n_j(k-1) + 1 = 1$, at the second it is 2, and so on up to n_j , so these visits contribute the factor $n_j!$. Multiplying over j ,

$$\prod_{k=1}^n \widehat{P}_k(y_k) = \frac{(S-1)! \prod_{j \in \mathcal{S}} n_j!}{(n+S-1)!}. \quad (6.55)$$

(This quantity is also $\int_{\Delta(\mathcal{S})} \prod_{k=1}^n q(y_k) dq$ for the uniform distribution over the simplex, which is why (6.44) is the Bayes mixture predictor with a uniform prior [47, Chapter 9]; we do not need this fact.)

The true distribution's product likelihood. Writing $q = P_h^{M^\star}(\cdot \mid s, a)$ and $\bar{q}(j) = n_j/n$ for the empirical distribution of $y_1, \dots, y_n$, we have $\prod_{k=1}^n q(y_k) = \prod_j q(j)^{n_j}$, and

$$\log \prod_j q(j)^{n_j} = n \sum_j \bar{q}(j) \log q(j) = n \sum_j \bar{q}(j) \log \bar{q}(j) - n D_{\text{KL}}(\bar{q} \parallel q) \leq \log \prod_j \bar{q}(j)^{n_j},$$

since KL divergence is nonnegative (Lemma 7.1). In words: the empirical distribution maximizes the likelihood of the observed sequence, so the true distribution can do no better than it.

Putting everything together. Dividing the second product by (6.54),

$$R_{h,s,a} \leq \log \left(\frac{(n+S-1)!}{(S-1)! \prod_j n_j!} \prod_j \left(\frac{n_j}{n} \right)^{n_j} \right) = \log \binom{n+S-1}{S-1} + \log \left(\frac{n!}{\prod_j n_j!} \prod_j \left(\frac{n_j}{n} \right)^{n_j} \right).$$

The second logarithm is at most zero: its argument is the multinomial probability of observing exactly the counts $(n_j)_{j \in \mathcal{S}}$ in n independent draws from $\bar{q}$, which is at most 1. For the first, the

standard estimate $\binom{m}{r} \leq (em/r)^r$ with $m = n + S - 1$ and $r = S - 1$ gives (cf. Cesa-Bianchi and Lugosi [47, Chapter 9], which proves the same bound for the Laplace forecaster)

$$R_{h,s,a} \leq (S-1) \log \left(\frac{e(n+S-1)}{S-1} \right) \leq S \log(e(T+S)),$$

where the last step uses $n \leq T$ and $\frac{n+S-1}{S-1} \leq n+S-1 \leq T+S$.

Summing over the HSA cells in (6.53), the log-loss regret of the sequence $\widehat{M}^1, \dots, \widehat{M}^T$ relative to $M^\star$ is at most $HS^2A \log(e(T+S))$. Finally, Lemma 7.3 (stated in Section 7.4.1) converts log-loss regret relative to $M^\star$ into Hellinger estimation error: with probability at least $1 - \delta$, $\mathbf{Est}_H$ is at most this regret plus $2 \log(\delta^{-1})$. Since $\widehat{M}^t$ is determined by the history and τ^t is drawn from $M^\star(\pi^t)$, the lemma applies for any choice of the distributions p^t . $\square$

7. GENERAL DECISION MAKING: A UNIFIED FRAMEWORK

So far, we have covered three frameworks for interactive decision making beyond the multi-armed bandit: contextual bandits, structured bandits, and episodic reinforcement learning. For each setting, we developed algorithms as instances of the Estimation-to-Decisions principle. For structured bandits, the Decision-Estimation Coefficient (DEC) identified the best possible exploration strategy for a given class, yielding one algorithm, E2D, for all classes. For tabular reinforcement learning, the same principle was applied (with the main difference being to move from regression to estimation of entire models), but the exploration strategy and analysis were built by hand. In this section, we unify these developments by introducing a general decision making framework that captures the settings of the preceding chapters and deriving general versions of Estimation-to-Decisions and the Decision-Estimation Coefficient in this setting.

Organization. Section 7.1 introduces our general decision making framework and shows how the earlier settings are special cases. Section 7.2 reviews the information-theoretic divergences needed to formally define the algorithms and complexity measures we study. Section 7.3 defines the most general form of the DEC and computes it for examples, and Section 7.4 builds on this to give the general form of the E2D algorithm. Finally, Section 7.5 introduces a second variant of the DEC, the *constrained DEC*, which leads to sharper regret bounds and sets up for Section 8.

7.1 General Decision Making Framework

We work in a simple framework we refer to as *Decision Making with Structured Observations* (DMSO), which we will show subsumes the different decision making frameworks we have encountered so far. The protocol has two ingredients beyond those of a structured bandit: a *decision space* Π , which may consist of actions, policies, or any other objects, and an *observation space* $\mathcal{O}$, which carries whatever feedback the learner receives beyond the reward.

General Decision Making Protocol

for $t = 1, \dots, T$ **do**

Learner selects a decision $\pi^t \in \Pi$ (possibly at random, $\pi^t \sim p^t$).

Nature selects a reward $r^t \in \mathcal{R} \subseteq \mathbb{R}$ and observation $o^t \in \mathcal{O}$ based on the decision.

Learner observes (r^t, o^t) .

We focus on a stochastic variant of the DMSO framework.

Assumption 7.1 (Stochastic Rewards and Observations): Rewards and observations are generated independently via

$$(r^t, o^t) \sim M^*(\cdot \mid \pi^t), \quad (7.1)$$

where $M^* : \Pi \rightarrow \Delta(\mathcal{R} \times \mathcal{O})$ 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^* . As usual, 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 7.2 (Realizability): The model class $\mathcal{M}$ contains the true model M^* .

Throughout, we assume that mean rewards are bounded, $\mathbb{E}^{M,\pi}[r] \in [0, 1]$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$; the rewards themselves may be unbounded (Gaussian, for example). Regret bounds hold with probability at least $1 - \delta$ over the draw of the rewards and observations and the learner's randomization. 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 6](#), 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

$$\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\} \tag{7.2}$$

as the induced class of mean reward functions. We evaluate the learner's performance in terms of *regret* to the optimal decision for M^* :

$$\mathbf{Reg} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^t)], \tag{7.3}$$

where $p^t \in \Delta(\Pi)$ is the learner's distribution over decisions at round t . As in [Section 6](#), we abbreviate $f^* = f^{M^*}$ and $\pi^* = \pi_{M^*}$. In the language of [Figure 2](#), this framework sits at the far corner: it is interactive, it allows arbitrary function approximation through $\mathcal{M}$, and the observation o^t can be as rich as one likes. Every setting studied so far is recovered by restricting the decision and observation spaces in different ways. Let us first see how it subsumes the structured bandit and contextual bandit problems.

Example 7.1 (Structured bandits). When there are no observations (i.e., $\mathcal{O} = \{\emptyset\}$), the DMSO framework is equivalent to *structured bandits* studied earlier in [Section 5](#). 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 (7.2). By changing the class $\mathcal{F}$, this encompasses all of the concrete examples of structured bandit problems we studied in [Section 5](#), including linear bandits, nonparametric bandits, and concave/convex bandits.

◁

Example 7.2 (Contextual bandits). The DMSO framework captures contextual bandits ([Section 4](#)) with i.i.d. contexts, a special case of the setting of [Section 4](#), which also allows adversarial contexts. To make this precise, we slightly abuse notation and think of π^t as *functions* mapping the context

x^t to an action in $[A]$. To this end, on round t , the decision-maker selects a mapping $\pi^t : \mathcal{X} \rightarrow [A]$ 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, $[A]$ be the set of actions, and $\Pi : \mathcal{X} \rightarrow [A]$ 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 . $\triangleleft$

Example 7.3 (Online reinforcement learning). The online reinforcement learning framework we introduced in Section 6 immediately falls into the DMSO framework by taking $\Pi = \Pi_{\text{rns}}$, $r^t = \sum_{h=1}^H r_h^t$, and $o^t = \tau^t$. Tabular MDPs are one choice of model class $\mathcal{M}$. Reinforcement learning with function approximation, the subject of Section 9, corresponds to other restrictions on $\mathcal{M}$ (e.g., low-rank or smooth structure in the transition dynamics); we will apply the tools of this section to each of these settings. $\triangleleft$

Other problems included in the DMSO framework which we will not cover in this monograph include partially observable MDPs [120, 113, 147, 98], where the observation is the sequence of observations and actions in an episode, and bandits with graph-structured feedback [149, 16], where the observation contains the rewards of neighboring actions. In each case the framework is the same, all that changes is the model class $\mathcal{M}$.

7.2 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 use the notion *Csiszár f -divergence*, which includes the Kullback-Leibler (KL) divergence, total variation distance, and Hellinger distance as special cases (we already made use of the latter in Section 6, but it will be helpful to review some additional properties before we continue).

Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\Omega, \mathcal{F})$. We say that $\mathbb{P}$ is *absolutely continuous* with respect to $\mathbb{Q}$ if for all events $A \in \mathcal{F}$, $\mathbb{Q}(A) = 0 \implies \mathbb{P}(A) = 0$; we denote this by $\mathbb{P} \ll \mathbb{Q}$. For a convex function $f : (0, \infty) \rightarrow \mathbb{R}$ with $f(1) = 0$, the associated f -divergence for $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$D_f(\mathbb{P} \parallel \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}} \left[f \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \right) \right] \quad (7.4)$$

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

$$D_f(\mathbb{P} \parallel \mathbb{Q}) := \int_{q>0} q f \left(\frac{p}{q} \right) d\nu + \mathbb{P}(q = 0) \cdot f'(\infty), \quad (7.5)$$

where $f'(\infty) := \lim_{x \rightarrow 0^+} x f(1/x)$. We make use of the following f -divergences, all of which have unique properties that make them useful in different contexts.¹⁷

¹⁷Another important example which appears throughout the decision making literature, but not this monograph, is the Neyman-Pearson χ^2 -divergence.

▷ Choosing $f(t) = \frac{1}{2}|t - 1|$ gives the *total variation (TV) distance*

$$D_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int \left| \frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{Q}}{d\nu} \right| d\nu,$$

which can also be written as

$$D_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

▷ Choosing $f(t) = (1 - \sqrt{t})^2$ gives *squared Hellinger distance*

$$D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}} \right)^2 d\nu.$$

▷ Choosing $f(t) = t \log(t)$ gives the *Kullback-Leibler divergence*:

$$D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) = \begin{cases} \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Note that for TV distance and Hellinger distance, we use the notation $D(\cdot, \cdot)$ rather than $D(\cdot \parallel \cdot)$ to emphasize that the divergence is symmetric. The divergences satisfy the following relationship.

Lemma 7.1 (Total variation, Hellinger, and KL): For all distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$D_{\text{TV}}^2(\mathbb{P}, \mathbb{Q}) \leq D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}) \leq D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}). \quad (7.6)$$

Proof of Lemma 7.1. Take the common dominating measure $\nu = \mathbb{P} + \mathbb{Q}$ and write $p = d\mathbb{P}/d\nu$ and $q = d\mathbb{Q}/d\nu$. By Cauchy-Schwarz,

$$\begin{aligned} D_{\text{TV}}(\mathbb{P}, \mathbb{Q}) &= \frac{1}{2} \int |\sqrt{p} - \sqrt{q}|(\sqrt{p} + \sqrt{q}) d\nu \\ &\leq \frac{1}{2} \left(\int (\sqrt{p} - \sqrt{q})^2 d\nu \right)^{1/2} \left(\int (\sqrt{p} + \sqrt{q})^2 d\nu \right)^{1/2} \leq \sqrt{D_{\text{H}}^2(\mathbb{P}, \mathbb{Q})}, \end{aligned}$$

since $\int \sqrt{pq} d\nu \leq 1$ and thus the second integral is at most 4. Squaring proves the first inequality. The second is immediate if $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) = +\infty$. Otherwise $\mathbb{P} \ll \mathbb{Q}$, and $-\log u \geq 1 - u$ gives

$$\begin{aligned} D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) &= -2 \int_{p>0} p \log \sqrt{q/p} d\nu \geq 2 \int_{p>0} p(1 - \sqrt{q/p}) d\nu \\ &= 2 - 2 \int \sqrt{pq} d\nu = D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}). \end{aligned}$$

□

It is known that $D_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = 1$ if and only if $D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}) = 2$, and $D_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}) = 0$ (more generally, $D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}) \leq 2D_{\text{TV}}(\mathbb{P}, \mathbb{Q})$). 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 7.2 (KL is bounded by Hellinger under bounded density ratios): Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\Omega, \mathcal{F})$. If $\sup_{F \in \mathcal{F}} \frac{\mathbb{P}(F)}{\mathbb{Q}(F)} \leq V$, then

$$D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) \leq (2 + \log(V)) D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}). \quad (7.7)$$

Other properties we use include:

- ▷ Boundedness of TV (by 1) and Hellinger (by 2).
- ▷ The triangle inequality for total variation distance and for (unsquared) Hellinger distance D_{H} .
- ▷ Convexity: every f -divergence is jointly convex in its arguments. In particular, for a distribution ν over distributions $\mathbb{Q}$ with mixture $\bar{\mathbb{Q}} = \mathbb{E}_{\mathbb{Q} \sim \nu}[\mathbb{Q}]$, $D_{\text{H}}^2(\mathbb{P}, \bar{\mathbb{Q}}) \leq \mathbb{E}_{\mathbb{Q} \sim \nu}[D_{\text{H}}^2(\mathbb{P}, \mathbb{Q})]$.
- ▷ The *data-processing inequality*, satisfied by every f -divergence: for any measurable map g , if $\mathbb{P}_g$ and $\mathbb{Q}_g$ denote the laws of $g(X)$ under $X \sim \mathbb{P}$ and $X \sim \mathbb{Q}$,

$$D_f(\mathbb{P}_g \parallel \mathbb{Q}_g) \leq D_f(\mathbb{P} \parallel \mathbb{Q}). \quad (7.8)$$

In words, discarding part of an observation cannot increase the divergence.

- ▷ *Chain rule* and subadditivity properties for KL and Hellinger divergence ([Lemma 8.2](#) and [Lemma 8.3](#)), stated in [Section 8](#) where they are used.
- ▷ A variational representation for TV distance:

$$D_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{g: \Omega \rightarrow [0,1]} |\mathbb{E}_{\mathbb{P}}[g] - \mathbb{E}_{\mathbb{Q}}[g]| \quad (7.9)$$

See Polyanskiy [\[165\]](#) for further background.

7.3 Decision-Estimation Coefficient for General Decision Making

To extend the Decision-Estimation Coefficient beyond bandits, the notion of information gain has to change. In a multi-armed bandit the mean rewards are a sufficient statistic, so squared error between mean reward functions is the right measure. In general decision making, the observation may carry information that the mean reward does not (a trajectory, in reinforcement learning), and so may the shape of the reward distribution itself. To address this, we measure information gain by a divergence between the distributions of rewards and observations. This handles both cases at once, and is the only change from [Section 5](#).

For a model class $\mathcal{M}$, reference model $\widehat{M}$, and scale parameter $\gamma > 0$, the Decision-Estimation Coefficient for general decision making [\[84, 89\]](#) is defined via

$$\text{dec}_{\gamma}(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[\underbrace{f^M(\pi_M) - f^M(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))}_{\text{information gain for obs.}} \right]. \quad (7.10)$$

We further define

$$\text{dec}_{\gamma}(\mathcal{M}) = \sup_{\widehat{M} \in \text{co}(\mathcal{M})} \text{dec}_{\gamma}(\mathcal{M}, \widehat{M}). \quad (7.11)$$

Compared with the structured bandit version in (5.15), two things have changed: the coefficient is defined over a class of *models* rather than mean reward functions, and the information gain $\mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right]$ measures the distance between the *distributions* of rewards and observations under M and $\widehat{M}$, not the distance between means. We have already met (7.10) once, without naming it: for tabular MDPs, Proposition 6.2 shows that PC-IGW certifies $\text{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \lesssim H^3 SA/\gamma$ for every $\widehat{M} \in \mathcal{M}$ (Example 7.7 below).

The supremum in (7.11) is over $\text{co}(\mathcal{M})$ rather than over all models because, as for structured bandits (Proposition 5.13), allowing arbitrary reference models changes the coefficient by at most a constant factor in γ .

Proposition 7.1 (References outside the convex hull): Suppose that Π and $\mathcal{M}$ are finite and rewards are bounded, so that the minimax theorem (Lemma A.8) applies; more generally, it suffices that the conclusion of Lemma A.8 holds for the payoff in (7.10), as in Proposition 5.10. Then for all $\gamma > 0$,

$$\sup_{\widehat{M}} \text{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \leq \sup_{\widehat{M} \in \text{co}(\mathcal{M})} \text{dec}_{\gamma/4}(\mathcal{M}, \widehat{M}) = \text{dec}_{\gamma/4}(\mathcal{M}), \quad (7.12)$$

where the supremum on the left is over all models $\widehat{M} : \Pi \rightarrow \Delta(\mathcal{R} \times \mathcal{O})$.

This is a slight generalization of Proposition 5.13, which gives the same statement for the squared-error coefficient with no loss in γ ; the factor of 4 is the price of replacing squared error by Hellinger distance. The proof is deferred to the end of the section.

Remark 7.1 (The choice of Hellinger distance): Any f -divergence could be substituted in (7.10), but we will see that Hellinger satisfies two properties that are essential for the analysis in this section and Section 8. First, it is dominated by KL divergence, so an estimator that minimizes the logarithmic loss controls it (Lemma 7.3), which is what we need for upper bounds. Second, it dominates total variation (Lemma 7.1), so a small information gain means the models are hard to distinguish by any test, which is what the lower bounds of Section 8 need. The latter results also use that Hellinger distance obeys a chain rule over sequences.

7.3.1 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 7.4 (Multi-armed bandit with Gaussian rewards). Let $\Pi = [A]$, $\mathcal{R} = \mathbb{R}$, $\mathcal{O} = \{\emptyset\}$. We define

$$\mathcal{M}_{\text{MAB-G}} = \{M : M(\pi) = \mathcal{N}(f(\pi), 1), f : \Pi \rightarrow [0, 1]\}.$$

We claim that

$$\text{dec}_{\gamma}(\mathcal{M}_{\text{MAB-G}}) \propto \frac{A}{\gamma}. \quad (7.13)$$

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 (5.15). Thus,

it is sufficient to argue that the squared Hellinger divergence for Gaussian distributions reduces to square difference between the means:

$$D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \propto (f^M(\pi) - f^{\widehat{M}}(\pi))^2.$$

Once this is shown, the claim follows from [Proposition 5.2](#). To prove this, first note that

$$D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \leq D_{\text{KL}}(M(\pi) \parallel \widehat{M}(\pi)) = \frac{1}{2}(f^M(\pi) - f^{\widehat{M}}(\pi))^2. \quad (7.14)$$

In the other direction, one can directly compute

$$D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) = 2 - 2 \exp\left\{-\frac{1}{8}(f^M(\pi) - f^{\widehat{M}}(\pi))^2\right\}$$

and using that $1 - \exp\{-x\} \geq (1 - e^{-1})x$ for $x \in [0, 1]$, we establish

$$D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \geq c \cdot (f^M(\pi) - f^{\widehat{M}}(\pi))^2$$

for $c = \frac{1-1/e}{8}$. $\triangleleft$

In fact, one can show that the general DEC in [\(7.10\)](#) coincides with the basic squared error version from [Section 5](#) for general structured bandit problems, not just multi-armed bandits [\[84\]](#).

Let us next consider a twist on the bandit problem that is more information-theoretic in nature, and shows why the divergence in [\(7.10\)](#) is taken between distributions rather than means.

Example 7.5 (Bandits with structured noise). Let $\Pi = [A]$, $\mathcal{R} = \mathbb{R}$, $\mathcal{O} = \{\emptyset\}$. We define

$$\mathcal{M}_{\text{MAB-SN}} = \{M_1, \dots, M_A\} \cup \{M_0\}$$

where $M_i(\pi) := \mathcal{N}(1/2, 1)$ for $\pi \neq i$ and $M_i(\pi) := \text{Ber}(3/4)$ for $\pi = i$; we further define $M_0(\pi) := \mathcal{N}(1/2, 1)$ for all $\pi \in \Pi$. We take M_0 as the reference model. Before proceeding with the calculations, observe that we can solve the general decision making problem when the underlying model is $M^* \in \mathcal{M}$ with a simple algorithm. It is sufficient to select every action in $[A]$ only once: all optimal actions have Bernoulli rewards and give $r \in \{0, 1\}$ almost surely, while the suboptimal action has Gaussian rewards, and gives $r \notin \{0, 1\}$ almost surely. Thus, if we select an action and observe a reward $r \in \{0, 1\}$, 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:

$$D_{\text{H}}^2(M_i(\pi), M_0(\pi)) = 2\mathbb{I}\{\pi = i\}.$$

To use this property to derive the upper bound on $\text{dec}_{\gamma}(\mathcal{M}_{\text{MAB-SN}}, M_0)$, first note that the maximum over M in the definition of $\text{dec}_{\gamma}(\mathcal{M}_{\text{MAB-SN}}, M_0)$ is not attained at $M = M_0$, since in that case both the divergence and regret terms are zero, irrespective of p . Now, take $p = \text{unif}[A]$. Then for any $M \in \{M_1, \dots, M_A\}$,

$$\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] = (1 - 1/A)(3/4 - 1/2),$$

and

$$\text{dec}_{\gamma}(\mathcal{M}_{\text{MAB-SN}}, M_0) \lesssim (1 - 1/A)(3/4 - 1/2) - \gamma \frac{2}{A} \lesssim \mathbb{I}\{\gamma \leq A/4\}.$$

This leads to an upper bound

$$\text{dec}_{\gamma}(\mathcal{M}_{\text{MAB-SN}}, M_0) \lesssim \mathbb{I}\{\gamma \leq A/4\} \quad (7.15)$$

which can also be shown to be tight. $\triangleleft$

Example 7.6 (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 $\pi' \in \Pi$,

$$D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \geq D_{\text{H}}^2(M_{\mathcal{O}}(\pi), \widehat{M}_{\mathcal{O}}(\pi)) \quad (7.16)$$

$$= D_{\text{H}}^2(M_{\mathcal{O}}(\pi'), \widehat{M}_{\mathcal{O}}(\pi')) \geq D_{\text{H}}^2(M_{\mathcal{R}}(\pi'), \widehat{M}_{\mathcal{R}}(\pi')). \quad (7.17)$$

Using this property, we show that for any $\widehat{M} \in \mathcal{M}$,

$$\text{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \leq \frac{1}{\gamma}. \quad (7.18)$$

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 (7.18), 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

$$\begin{aligned} f^M(\pi_M) - f^M(\pi_{\widehat{M}}) &\leq f^M(\pi_M) - f^M(\pi_{\widehat{M}}) + f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi_M) \\ &\leq 2 \cdot \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} |f^M(\pi) - f^{\widehat{M}}(\pi)| \\ &\leq 2 \cdot \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} D_{\text{H}}(M_{\mathcal{R}}(\pi), \widehat{M}_{\mathcal{R}}(\pi)). \end{aligned}$$

We then use the AM-GM inequality, which implies that for any $\gamma > 0$,

$$\begin{aligned} \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} D_{\text{H}}(M_{\mathcal{R}}(\pi), \widehat{M}_{\mathcal{R}}(\pi)) &\lesssim \gamma \cdot \max_{\pi \in \{\pi_M, \pi_{\widehat{M}}\}} D_{\text{H}}^2(M_{\mathcal{R}}(\pi), \widehat{M}_{\mathcal{R}}(\pi)) + \frac{1}{\gamma} \\ &\leq \gamma \cdot D_{\text{H}}^2(M(\pi_{\widehat{M}}), \widehat{M}(\pi_{\widehat{M}})) + \frac{1}{\gamma}, \end{aligned}$$

where the final inequality uses (7.16). This certifies that for all $M \in \mathcal{M}$, the choice for p above satisfies

$$\mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \right] \lesssim \frac{1}{\gamma},$$

so we have $\text{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \lesssim \frac{1}{\gamma}$. $\triangleleft$

Example 7.7 (Tabular reinforcement learning). Let $\mathcal{M}$ be the class of tabular MDPs with S states, A actions, horizon H , and known rewards (Assumption 6.1), viewed as a general decision making problem as in Example 7.3. By Proposition 6.2, the PC-IGW strategy certifies that

$$\text{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \lesssim \frac{H^3 S A}{\gamma} \quad \text{for all } \widehat{M} \in \mathcal{M}.$$

This is the first example in which the observation $o = \tau$ carries information beyond the reward, and the Hellinger distance between trajectory distributions is what makes the bound polynomial in S , A , and H rather than exponential. $\triangleleft$

7.4 Estimation-to-Decisions Algorithm for General Decision Making

The most general version of the Estimation-to-Decisions meta-algorithm (E2D) is given below.

Algorithm 7.1 Estimation-to-Decisions (E2D) for General Decision Making

input: Exploration parameter $\gamma > 0$; online estimation oracle.

for $t = 1, 2, \dots, T$ **do**

Obtain $\widehat{M}^t$ from online estimation oracle with $(\pi^1, r^1, o^1), \dots, (\pi^{t-1}, r^{t-1}, o^{t-1})$.

Compute // Minimizer for $\text{dec}_\gamma(\mathcal{M}, \widehat{M}^t)$.

$$p^t = \arg \min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_{\text{H}}^2(M(\pi), \widehat{M}^t(\pi)) \right]. \quad (7.19)$$

Sample decision $\pi^t \sim p^t$ and update estimation algorithm with (π^t, r^t, o^t) .

Compared to the meta-algorithm we studied for structured bandits, the main difference is that rather than trying to estimate the reward function f^* , we now estimate the underlying model M^* , as we did for tabular reinforcement learning in [Section 6.7](#). To do so, we appeal once again to the notion of an *online estimation oracle*, but this time for *model estimation*.

Definition 7.1 (Online Estimation Oracle): At each time $t \in [T]$, an *online estimation oracle* returns, given

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

with $(r^i, o^i) \sim M^*(\pi^i)$ and $\pi^i \sim p^i$, a model $\widehat{M}^t \in \widehat{\mathcal{M}}$, where $\widehat{\mathcal{M}} \supseteq \mathcal{M}$ is a fixed class of models associated with the oracle. Define the cumulative (Hellinger) estimation error

$$\mathbf{Est}_{\text{H}} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left[D_{\text{H}}^2(M^*(\pi^t), \widehat{M}^t(\pi^t)) \right]. \quad (7.20)$$

The oracle is required to ensure that $\mathbf{Est}_{\text{H}} \leq \mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta)$ with probability at least $1 - \delta$, where $\mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta)$ is a deterministic bound. Here, $p^i(\cdot \mid \mathcal{H}^{i-1})$ is the randomization distribution for the decision-maker.

This mirrors the online regression oracle of [Definition 5.2](#), with two changes: the oracle estimates the model rather than the mean reward function, and estimation error is measured by squared Hellinger distance between the distributions $M^*(\pi)$ and $\widehat{M}^t(\pi)$ rather than by squared error between means. As in the preceding sections, the bare symbol $\mathbf{Est}_{\text{H}}$ denotes the realized error and $\mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta)$ the guarantee. We use $\widehat{\mathcal{M}}$ to denote any set where the oracle's predictions are guaranteed to live. A *proper* oracle has $\widehat{\mathcal{M}} = \mathcal{M}$; the estimator for tabular MDPs in [Proposition 6.3](#) is one example. Standard online estimation algorithms such as exponential weights produce improper (mixture) estimates, for which $\widehat{\mathcal{M}} = \text{co}(\mathcal{M})$; see [Section 7.4.1](#).

E2D solves a min-max problem rather than applying an explicit rule such as inverse gap weighting, for the reasons discussed in [Section 5.3](#): the min-max problem is the definition of the best possible exploration strategy for the class, and any explicit rule that certifies an upper bound on its value can be used in its place.

Theorem 7.1 (Regret of E2D): E2D with exploration parameter $\gamma > 0$ and an online estimation oracle with estimates in $\widehat{\mathcal{M}}$ guarantees that for every realization,

$$\mathbf{Reg} \leq \sum_{t=1}^T \text{dec}_\gamma(\mathcal{M}, \widehat{M}^t) + \gamma \cdot \mathbf{Est}_H \leq \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \text{dec}_\gamma(\mathcal{M}, \widehat{M}) \cdot T + \gamma \cdot \mathbf{Est}_H. \quad (7.21)$$

Furthermore, if [Proposition 7.1](#) applies, then for any $\widehat{\mathcal{M}}$ we have $\sup_{\widehat{M} \in \widehat{\mathcal{M}}} \text{dec}_\gamma(\mathcal{M}, \widehat{M}) \leq \text{dec}_{c\gamma}(\mathcal{M})$ for an absolute constant $c \geq 1/4$, so that with probability at least $1 - \delta$,

$$\mathbf{Reg} \leq \text{dec}_{c\gamma}(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta). \quad (7.22)$$

For proper oracles, and for oracles with $\widehat{\mathcal{M}} = \text{co}(\mathcal{M})$, the first statement already gives $\text{dec}_\gamma(\mathcal{M})$ with $c = 1$. Optimizing over the parameter γ in [\(7.22\)](#) yields

$$\mathbf{Reg} \leq \inf_{\gamma > 0} \left\{ \text{dec}_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta) \right\} \leq 2 \cdot \inf_{\gamma > 0} \max \left\{ \text{dec}_\gamma(\mathcal{M}) \cdot T, \gamma \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta) \right\},$$

up to the constant c inside $\text{dec}_{c\gamma}$, which we suppress. [Section 7.4.1](#) shows that for any finite class $\mathcal{M}$, the averaged exponential weights algorithm with the logarithmic loss is an online estimation oracle with $\mathbf{Est}_H(\mathcal{M}, T, \delta) \lesssim \log(|\mathcal{M}|/\delta)$ and $\widehat{\mathcal{M}} = \text{co}(\mathcal{M})$. Putting these facts together, we see that for any finite class, it is possible to achieve

$$\mathbf{Reg} \leq \text{dec}_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \log(|\mathcal{M}|/\delta) \quad (7.23)$$

with probability at least $1 - \delta$.

Proof of [Theorem 7.1](#). We write

$$\begin{aligned} \mathbf{Reg} &= \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^t)] \\ &= \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} [f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t} [D_H^2(M^*(\pi^t), \widehat{M}^t(\pi^t))] + \gamma \cdot \mathbf{Est}_H. \end{aligned}$$

For each t , since $M^* \in \mathcal{M}$, we have

$$\begin{aligned} &\mathbb{E}_{\pi^t \sim p^t} [f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t} [D_H^2(M^*(\pi^t), \widehat{M}^t(\pi^t))] \\ &\leq \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi^t \sim p^t} [f^M(\pi_M) - f^M(\pi^t)] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t} [D_H^2(M(\pi^t), \widehat{M}^t(\pi^t))] \\ &= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_H^2(M(\pi), \widehat{M}^t(\pi))] \\ &= \text{dec}_\gamma(\mathcal{M}, \widehat{M}^t). \end{aligned} \quad (7.24)$$

Summing over all rounds t gives the first inequality in [\(7.21\)](#), and using that $\widehat{M}^t \in \widehat{\mathcal{M}}$ for all t gives the second. The final statement follows from $\sup_{\widehat{M} \in \widehat{\mathcal{M}}} \text{dec}_\gamma(\mathcal{M}, \widehat{M}) \leq \text{dec}_{c\gamma}(\mathcal{M})$, on the event that the oracle guarantee $\mathbf{Est}_H \leq \mathbf{Est}_H(\mathcal{M}, T, \delta)$ holds. $\square$

Examples for the upper bound. We now revisit the examples from [Section 7.3](#) and use E2D and [Theorem 7.1](#) to derive regret bounds for them.

Example 7.4 (Multi-armed bandit with Gaussian rewards, cont'd). For the Gaussian bandit problem from [Example 7.4](#), plugging the bound $\text{dec}_\gamma(\mathcal{M}_{\text{MAB-G}}) \lesssim A/\gamma$ into [Theorem 7.1](#) yields

$$\mathbf{Reg} \lesssim \frac{AT}{\gamma} + \gamma \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta)$$

with probability at least $1 - \delta$. Choosing $\gamma = \sqrt{AT/\mathbf{Est}_H(\mathcal{M}, T, \delta)}$ balances the terms above and gives

$$\mathbf{Reg} \lesssim \sqrt{AT \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta)}.$$

◁

Example 7.5 (Bandits with structured noise, cont'd). For the bandit-type problem with structured noise from [Example 7.5](#), the bound $\text{dec}_\gamma(\mathcal{M}_{\text{MAB-SN}}) \lesssim \mathbb{I}\{\gamma \leq A/4\}$ yields

$$\mathbf{Reg} \lesssim \mathbb{I}\{\gamma \leq A/4\} \cdot T + \gamma \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta)$$

with probability at least $1 - \delta$. We can choose $\gamma = A$, which gives

$$\mathbf{Reg} \lesssim A \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta).$$

◁

Example 7.7 (Tabular reinforcement learning, cont'd). For tabular reinforcement learning, the estimator of [Proposition 6.3](#) is proper ($\widehat{M}^t \in \mathcal{M}$), so [Example 7.7](#) bounds each round's term in (7.24) by $\text{dec}_\gamma(\mathcal{M}, \widehat{M}^t) \lesssim H^3SA/\gamma$, and the proof of [Theorem 7.1](#) gives

$$\mathbf{Reg} \lesssim \frac{H^3SA \cdot T}{\gamma} + \gamma \cdot \mathbf{Est}_H.$$

This is just a restatement of the first part of [Theorem 6.2](#). PC-IGW is an explicit strategy that certifies an upper bound on the value of (7.19) rather than solving it exactly, which suffices by the discussion of certified bounds in [Section 5](#). ◁

Further examples. The examples above are deliberately simple. Richer instances of the DEC appear throughout the monograph, and we collect pointers here. For reinforcement learning with function approximation, [Section 9.5](#) bounds the DEC for the class of MDPs with low Bellman rank ([Proposition 9.3](#)), which subsumes tabular MDPs, low-rank MDPs, and several other models; this is the most substantial application of [Theorem 7.1](#) in the monograph. For bandit problems, the calculations in [Section 5.4](#)—the cheating code ([Proposition 5.3](#)), linear bandits ([Proposition 5.5](#)), Lipschitz bandits ([Proposition 5.6](#)), and the eluder dimension, concave, and ReLU examples—carry over to the general DEC by the argument in [Example 7.4](#) (under Gaussian or bounded noise).

7.4.1 Online Estimation with Hellinger Distance

Let us now give some more detail as to how to perform the online model estimation required by [Theorem 7.1](#). Model estimation is a more challenging problem than regression, since we are estimating the underlying *conditional distribution* rather than just the conditional mean. In spite of this difficulty, estimating the model M^* with respect to Hellinger distance is a classical problem

that we can solve using the online learning tools introduced in [Section 2.2](#); in particular, online *conditional density estimation* with the log loss, which we already saw in a special case for tabular reinforcement learning in [Proposition 6.3](#).

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

$$\ell_{\log}^t(M) = \log\left(\frac{1}{m^M(r^t, o^t \mid \pi^t)}\right), \quad (7.25)$$

where we define $m^M(\cdot, \cdot \mid \pi)$ as the conditional density for (r, o) under M . As in [Section 2.2](#), we write $\mathbf{Reg}_{\text{OL}}^{\text{Log}}$ for the online predictor's regret under log loss: this compares the cumulative log loss of its model estimates with that of the best fixed model in $\mathcal{M}$. Thus,

$$\mathbf{Reg}_{\text{OL}}^{\text{Log}} = \sum_{t=1}^T \ell_{\log}^t(\widehat{M}^t) - \inf_{M \in \mathcal{M}} \sum_{t=1}^T \ell_{\log}^t(M). \quad (7.26)$$

The following result converts $\mathbf{Reg}_{\text{OL}}^{\text{Log}}$ into a bound on Hellinger estimation error $\mathbf{Est}_{\text{H}}$, which is what [Theorem 7.1](#) requires.

Lemma 7.3 (Log-loss regret controls Hellinger estimation error): For any online estimation algorithm, whenever [Assumption 7.2](#) holds, we have

$$\mathbb{E}[\mathbf{Reg}_{\text{OL}}^{\text{Log}}] \geq \mathbb{E}\left[\sum_{t=1}^T D_{\text{KL}}\left(M^*(\pi^t) \parallel \widehat{M}^t(\pi^t)\right)\right], \quad (7.27)$$

so that

$$\mathbb{E}[\mathbf{Est}_{\text{H}}] \leq \mathbb{E}[\mathbf{Reg}_{\text{OL}}^{\text{Log}}]. \quad (7.28)$$

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

$$\mathbf{Est}_{\text{H}} \leq \mathbf{Reg}_{\text{OL}}^{\text{Log}} + 2 \log(\delta^{-1}). \quad (7.29)$$

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 [[59](#), [207](#), [122](#), [103](#), [158](#), [173](#), [81](#), [148](#)], and this is complemented by theory which provides minimax rates for generic model classes [[187](#), [156](#), [45](#), [33](#)]. One example we have already seen ([Section 2](#)) is the averaged exponential weights method, which guarantees

$$\mathbf{Reg}_{\text{OL}}^{\text{Log}} \leq \log|\mathcal{M}|$$

for finite classes $\mathcal{M}$; by (7.29), it is therefore an online estimation oracle in the sense of [Definition 7.1](#) with $\mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta) = \log|\mathcal{M}| + 2 \log(\delta^{-1}) \lesssim \log(|\mathcal{M}|/\delta)$. More generally, any algorithm whose log-loss regret is bounded by a deterministic quantity $R(T)$ is an online estimation oracle with $\mathbf{Est}_{\text{H}}(\mathcal{M}, T, \delta) = R(T) + 2 \log(\delta^{-1})$. Another example is that for linear models (i.e., $m^M(r, o \mid \pi) = \langle \phi(r, o, \pi), \theta \rangle$ for a fixed feature map $\phi \in \mathbb{R}^d$), algorithms with $\mathbf{Reg}_{\text{OL}}^{\text{Log}} = O(d \log(T))$ are known [[176](#), [187](#)]. All of these algorithms output mixtures of models in $\mathcal{M}$, so $\widehat{\mathcal{M}} = \text{co}(\mathcal{M})$. The

estimator for tabular MDPs in [Proposition 6.3](#) is another instance: its proof bounds the log-loss regret of Laplace’s estimator by $HS^2A \log(e(T+S))$ and then applies [\(7.29\)](#). We refer the reader to Chapter 9 of Cesa-Bianchi and Lugosi [\[47\]](#) for further examples and discussion.

While [\(7.28\)](#) is straightforward, [\(7.29\)](#) 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 $\mathbf{Est}_H$ and $\mathbf{Reg}_{OL}^{\text{Log}}$ 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.

7.5 The Constrained Decision-Estimation Coefficient

As we have defined it so far, the DEC of [\(7.10\)](#) subtracts the information gain from the regret. To close the section, we introduce a second form of the DEC that places a *hard constraint* on the information gain instead. This *constrained* DEC yields a sharper regret bound than [Theorem 7.1](#), and is a form that we revisit for lower bounds in [Section 8](#). In what follows, we define the constrained DEC, state the regret bound it gives, and then compare the two forms.

7.5.1 Definition and Regret Upper Bound

For a model class $\mathcal{M}$, reference model $\widehat{M}$, and radius $\varepsilon > 0$, the constrained Decision-Estimation Coefficient is defined as¹⁸

$$\text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim p} \left[D_H^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] \leq \varepsilon^2 \right\}, \quad (7.30)$$

with

$$\text{dec}_\varepsilon^c(\mathcal{M}) := \sup_{\widehat{M} \in \text{co}(\mathcal{M})} \text{dec}_\varepsilon^c(\mathcal{M} \cup \{\widehat{M}\}, \widehat{M}). \quad (7.31)$$

We henceforth refer to the coefficient in [\(7.10\)](#) as the *offset* DEC to contrast it with this version. The interpretation is the same: the min player chooses a distribution over decisions, and the max player chooses a model that is hard to distinguish from the reference model yet incurs large regret; the difference is that “hard to distinguish” is now the constraint $\mathbb{E}_{\pi \sim p} \left[D_H^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] \leq \varepsilon^2$ rather than a penalty. The natural radius is $\varepsilon \approx T^{-1/2}$: after T rounds, models at Hellinger distance ε from the truth cannot be told apart, so $\text{dec}_\varepsilon^c(\mathcal{M})$ at this radius is the regret the learner cannot avoid in a single round. The union with $\{\widehat{M}\}$ is a technical convenience which makes the reference model itself an admissible alternative for the max player, so that the constrained supremum is over a nonempty set (and the value is nonnegative) even when $\widehat{M} \in \text{co}(\mathcal{M}) \setminus \mathcal{M}$.

The constrained DEC leads to upper bounds on regret through a refinement of E2D. We state the result for finite classes and omit the proof.

¹⁸We adopt the convention that the value of $\text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M})$ is zero if there exists p such that the set of $M \in \mathcal{M}$ with $\mathbb{E}_{\pi \sim p} \left[D_H^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] \leq \varepsilon^2$ is empty.

Proposition 7.2 (Regret upper bound via constrained DEC): Let $\mathcal{M}$ be finite with $M^* \in \mathcal{M}$. Suppose that for some $C_{\text{reg}} > 1$, the constrained DEC satisfies the growth condition

$$\text{dec}_\varepsilon^c(\mathcal{M}) \leq C_{\text{reg}}^2 \cdot \text{dec}_{\varepsilon/C_{\text{reg}}}^c(\mathcal{M}) \quad \text{for all } \varepsilon \in (0, 2), \quad (7.32)$$

which holds in particular whenever $\text{dec}_\varepsilon^c(\mathcal{M}) \propto \varepsilon^\rho$ for some $\rho \leq 2$. Then there is an algorithm that, for any $\delta \in (0, 1)$, ensures that with probability at least $1 - \delta$,

$$\text{Reg} \leq C \cdot \left(\text{dec}_{\bar{\varepsilon}}^c(\mathcal{M}) \cdot T \log T + \sqrt{T \log(|\mathcal{M}|/\delta)} \right), \quad \bar{\varepsilon} := C' \sqrt{\frac{\log(|\mathcal{M}|/\delta)}{T}}, \quad (7.33)$$

where $C, C' > 0$ are constants depending only on C_{reg} .

This result and the algorithm are due to Foster et al. [89]; Glasgow and Rakhlin [97] give a simpler single-pass algorithm with sharper constants. The idea behind the algorithm is to run E2D with the constrained min-max problem in place of the offset one: at round t it plays the minimizer p^t in the definition of $\text{dec}_\varepsilon^c(\mathcal{M} \cup \{\widehat{M}^t\}, \widehat{M}^t)$, where $\widehat{M}^t$ is produced by a specialized estimation oracle restricted to a *confidence set* of models consistent with the past estimates.

The regularity condition (7.32) rules out coefficients that collapse too quickly as ε decreases, and can be replaced by the same condition on any upper bound $r(\varepsilon) \geq \text{dec}_\varepsilon^c(\mathcal{M})$, in which case $r(\bar{\varepsilon})$ replaces $\text{dec}_{\bar{\varepsilon}}^c(\mathcal{M})$ in (7.33). The factor $\log T$ arises from summing the coefficient over geometrically decreasing radii and can be removed whenever $\text{dec}_\varepsilon^c(\mathcal{M}) \propto \varepsilon^\rho$ for some $\rho < 2$ [89, 97]; it is needed only in the boundary case $\rho = 2$.

7.5.2 Comparison to the Offset DEC

The offset DEC can be viewed as a Lagrangian relaxation of the constrained DEC, and always upper bounds it:

$$\begin{aligned} \text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M}) &= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] \leq \varepsilon^2 \right\} \\ &= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \inf_{\gamma \geq 0} \left\{ \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)] - \gamma \left(\mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] - \varepsilon^2 \right) \right\} \vee 0 \\ &\leq \inf_{\gamma \geq 0} \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)] - \gamma \left(\mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] - \varepsilon^2 \right) \right\} \vee 0 \\ &= \inf_{\gamma \geq 0} \left\{ \text{dec}_\gamma(\mathcal{M}, \widehat{M}) + \gamma \varepsilon^2 \right\} \vee 0. \end{aligned} \quad (7.34)$$

In the opposite direction, it is straightforward to show that

$$\text{dec}_\gamma(\mathcal{M}) \lesssim \text{dec}_{\gamma^{-1/2}}^c(\mathcal{M}).$$

This inequality is lossy, and there are indeed classes for which the constrained DEC is much smaller than the offset DEC, so that Proposition 7.2 gives a strictly better regret bound than (7.23). The following example, adapted from Foster et al. [89], exhibits the gap.

Proposition 7.3 (Improvement from constrained DEC): Fix $A \geq 4$ and $\beta \in (0, 1)$. Let $\Pi = [A] \cup \{\pi_o\}$ and $\mathcal{O} = [A] \cup \{\perp\}$, and let $\mathcal{M}_\beta = \{M_0, M_1, \dots, M_A\}$ be the following class. Under M_i for $i \in [A]$, each arm $\pi \in [A]$ gives deterministic reward $\frac{1}{2} + \frac{1}{2}\mathbb{I}\{\pi = i\}$ and observation $\perp$, while the *revealing decision* π_o gives reward 0 and observation $o = i$ with probability β and $o = \perp$ otherwise. Under M_0 , each arm gives reward $\frac{1}{2}$ and observation $\perp$, while π_o gives reward 0 and observation $o \sim \text{unif}([A])$ with probability β and $o = \perp$ otherwise. Then for all $\gamma > 0$ and all $\varepsilon^2 \leq \beta/8$,

$$\text{dec}_\gamma(\mathcal{M}_\beta) \geq \frac{1}{4 + 16\gamma\beta} - \frac{2\gamma}{A} \quad \text{and} \quad \text{dec}_\varepsilon^c(\mathcal{M}_\beta) \leq \frac{8\varepsilon^2}{\beta}. \quad (7.35)$$

Consequently, with $\beta = T^{-1/2}$ and $A = T^2$, the offset bound (7.23) is $\Omega(T^{3/4})$ for every choice of γ , whereas Proposition 7.2 gives $\mathbf{Reg} \leq \tilde{O}(\sqrt{T})$.

The proof is deferred to the end of the section, but the mechanism is easy to describe. The revealing decision is expensive, costing regret $\frac{1}{2}$ per play, but its information is cheap in aggregate: an information budget of ε^2 buys probability ε^2/β of learning the identity of the good arm outright. The constrained DEC is therefore quadratic in ε . The offset DEC misses this detail: for any γ , the offset DEC is of order $1/(\gamma\beta)$, and $\inf_\gamma \{1/(\gamma\beta) + \gamma\varepsilon^2\} = 2\varepsilon/\sqrt{\beta}$ is only linear in ε . In the language of (7.34), strong duality fails, and no single multiplier γ certifies the quadratic rate.

Despite this example, we view the constrained DEC as a conceptual tool rather than a replacement for the offset DEC. For most of the classes in this monograph, the two coefficients behave the same way, and the main use of the constrained form is for lower bounds, where it is the natural quantity; this is how Section 8 uses it. For algorithm design and upper bounds we work with the offset DEC throughout, for two reasons. It is simpler to derive algorithms from it, since E2D needs only the offset min-max problem and an estimation oracle. It is also usually simpler to find distributions p that certify bounds on it, as the examples in Section 7.3 and the calculations in Section 5 illustrate.

7.6 Bibliographic Remarks and Further Reading

Takeaways. This section replaced the separate settings studied so far by a single framework, decision making with structured observations, in which a model $M \in \mathcal{M}$ maps each decision to a distribution over rewards and observations. Two ingredients carry over from structured bandits with minimal change: the DEC, now defined with squared Hellinger distance between models in place of squared error between mean rewards, and the E2D meta-algorithm, which pairs a minimizer for the DEC with an online estimation oracle. Theorem 7.1 bounds regret in terms of the Decision-Estimation Coefficient and Hellinger estimation error, and Lemma 7.3 shows that any online density estimator with low log-loss regret is an estimation oracle in the required sense.

Bibliographic remarks. The DMSO framework of Section 7.1 is from Foster et al. [84]. The general DEC in (7.10), the online estimation oracle definition from Definition 7.1, E2D (Algorithm 7.1), and Theorem 7.1 are from Foster et al. [84], as are Example 7.4, Example 7.5, and Example 7.6; The in-expectation part of Lemma 7.3 is folklore; the high-probability part is from Foster et al. [84], and the argument through the moment generating function of the log-likelihood ratio goes back to Wong and Shen [212], van de Geer [202], Zhang [225]. Proposition 7.1 is Proposition 4.11 of Foster et al. [89]. The constrained DEC and the upper bound in Proposition 7.2 (Theorem 3.2 there) are from Foster et al. [89]; Glasgow and Rakhlin [97] give a single-pass algorithm with sharper

constants for bandit problems and extend the theory to γ -regret. [Proposition 7.3](#) is adapted from [Example 5.1](#) and [Proposition 4.4](#) of [Foster et al. \[89\]](#).

Further reading. The f -divergences in the refresher were introduced by Csiszár [\[60\]](#) and Ali and Silvey [\[15\]](#); Polyanskiy and Wu [\[166\]](#) provide a comprehensive treatment of the properties of f -divergences. The information ratio of Russo and Van Roy [\[182\]](#) is a Bayesian relative of the DEC that trades regret against information about the optimal decision; Lattimore and György [\[136\]](#) develop frequentist variants and connections to partial monitoring, and Foster et al. [\[84\]](#) connect the two quantities formally to the DEC. Adversarial rewards and observations, and bounds that scale with $\log|\Pi|$ rather than $\log|\mathcal{M}|$ through exploration by optimization, are treated by Lattimore and Szepesvári [\[140\]](#) and Foster et al. [\[87\]](#). Extensions of the framework include model-free variants whose estimation cost scales with a value function class [\[90\]](#) (see [Section 8.6.2](#)), PAC and reward-free versions [\[53\]](#), and partially observable settings [\[147, 51\]](#).

7.7 Deferred Proofs

Proof of [Proposition 7.1](#). Fix a reference model $\widehat{M}$. Since Π and $\mathcal{M}$ are finite, the payoff in [\(7.10\)](#) is bilinear in $p \in \Delta(\Pi)$ and in a prior $\mu \in \Delta(\mathcal{M})$ over models, and [Lemma A.8](#) gives

$$\text{dec}_\gamma(\mathcal{M}, \widehat{M}) = \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \right]. \quad (7.36)$$

Fix μ , and let $\overline{M}_\mu := \mathbb{E}_{M \sim \mu}[M] \in \text{co}(\mathcal{M})$ be the mixture model. For every π ,

$$\mathbb{E}_{M \sim \mu} [D_{\text{H}}^2(M(\pi), \overline{M}_\mu(\pi))] \leq \mathbb{E}_{M, M' \sim \mu} [D_{\text{H}}^2(M(\pi), M'(\pi))] \leq 4 \mathbb{E}_{M \sim \mu} [D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))],$$

where the first inequality is convexity of squared Hellinger distance in its second argument and the second is the triangle inequality for Hellinger distance with $\widehat{M}(\pi)$ as the intermediate point (both from [Section 7.2](#)). Thus the information gain relative to $\widehat{M}$ at parameter γ dominates the information gain relative to $\overline{M}_\mu$ at parameter $\gamma/4$, and substituting into [\(7.36\)](#) gives

$$\begin{aligned} \text{dec}_\gamma(\mathcal{M}, \widehat{M}) &\leq \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \frac{\gamma}{4} \cdot D_{\text{H}}^2(M(\pi), \overline{M}_\mu(\pi)) \right] \\ &\leq \sup_{\mu \in \Delta(\mathcal{M})} \text{dec}_{\gamma/4}(\mathcal{M}, \overline{M}_\mu) \leq \sup_{\overline{M} \in \text{co}(\mathcal{M})} \text{dec}_{\gamma/4}(\mathcal{M}, \overline{M}), \end{aligned}$$

where the second inequality replaces μ in the payoff by a free prior and applies [Lemma A.8](#) once more, with $\overline{M}_\mu$ as the reference. Since $\widehat{M}$ was arbitrary, this proves [\(7.12\)](#). $\square$

Proof of [Lemma 7.3](#). We first prove the in-expectation bound. By assumption, we have that

$$\sum_{t=1}^T \ell_{\log}^t(\widehat{M}^t) - \sum_{t=1}^T \ell_{\log}^t(M^\star) \leq \mathbf{Reg}_{\text{OL}}^{\text{Log}}.$$

Taking expectations, [Assumption 7.2](#) implies that

$$\sum_{t=1}^T \mathbb{E} \left[D_{\text{KL}}(M^\star(\pi^t) \parallel \widehat{M}^t(\pi^t)) \right] \leq \mathbb{E} [\mathbf{Reg}_{\text{OL}}^{\text{Log}}].$$

The bound now follows from [Lemma 7.1](#).

We now prove the high-probability bound using [Lemma A.2](#). Define $Z_t = \frac{1}{2}(\ell_{\log}^t(\widehat{M}^t) - \ell_{\log}^t(M^*))$. Applying [Lemma A.2](#) with the sequence $(-Z_t)_{t \leq T}$, we are guaranteed that with probability at least $1 - \delta$,

$$\sum_{t=1}^T -\log(\mathbb{E}_{t-1}[e^{-Z_t}]) \leq \sum_{t=1}^T Z_t + \log(\delta^{-1}) = \frac{1}{2} \sum_{t=1}^T (\ell_{\log}^t(\widehat{M}^t) - \ell_{\log}^t(M^*)) + \log(\delta^{-1}).$$

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^*} , and observe that

$$\begin{aligned} \mathbb{E}_{t-1}[e^{-Z_t} \mid \pi^t] &= \mathbb{E}_{t-1} \left[\sqrt{\frac{m^{\widehat{M}^t}(z^t \mid \pi^t)}{m^{M^*}(z^t \mid \pi^t)}} \mid \pi^t \right] \\ &= \int m^{M^*}(z \mid \pi^t) \sqrt{\frac{m^{\widehat{M}^t}(z \mid \pi^t)}{m^{M^*}(z \mid \pi^t)}} \nu(dz \mid \pi^t) \\ &= \int \sqrt{m^{M^*}(z \mid \pi^t) m^{\widehat{M}^t}(z \mid \pi^t)} \nu(dz \mid \pi^t) = 1 - \frac{1}{2} D_{\text{H}}^2(M^*(\pi^t), \widehat{M}^t(\pi^t)). \end{aligned}$$

Hence,

$$\mathbb{E}_{t-1}[e^{-Z_t}] = 1 - \frac{1}{2} \mathbb{E}_{t-1} \left[D_{\text{H}}^2(M^*(\pi^t), \widehat{M}^t(\pi^t)) \right]$$

and, since $-\log(1 - x) \geq x$ for $x \in [0, 1]$, we conclude that

$$\frac{1}{2} \sum_{t=1}^T \mathbb{E}_{t-1} \left[D_{\text{H}}^2(M^*(\pi^t), \widehat{M}^t(\pi^t)) \right] \leq \frac{1}{2} \sum_{t=1}^T (\ell_{\log}^t(\widehat{M}^t) - \ell_{\log}^t(M^*)) + \log(\delta^{-1}).$$

□

Proof of [Proposition 7.3](#). Write $\mathcal{M} = \mathcal{M}_\beta$. Under every model in $\mathcal{M}$, the revealing decision π_o has reward 0 and produces the observation $\perp$ with probability $1 - \beta$.

Lower bound for the offset DEC. Since $M_0 \in \mathcal{M} \subseteq \text{co}(\mathcal{M})$, it suffices to lower bound $\text{dec}_\gamma(\mathcal{M}, M_0)$. Fix $p \in \Delta(\Pi)$, and let $i \in [A]$ be an arm with $p(i) \leq 1/A$. We consider two choices for the max player. If $M = M_i$, the optimal decision is arm i , and

$$\mathbb{E}_{\pi \sim p}[f^{M_i}(i) - f^{M_i}(\pi)] = p(\pi_o) + \frac{1}{2}(1 - p(\pi_o) - p(i)) \geq \frac{1}{4}.$$

Since M_i and M_0 differ only at arm i , where the squared Hellinger distance is 2, and at π_o , where it is at most 2β , we also have

$$\mathbb{E}_{\pi \sim p}[D_{\text{H}}^2(M_i(\pi), M_0(\pi))] \leq \frac{2}{A} + 2\beta p(\pi_o).$$

If $M = M_0$, the regret of p is $\frac{1}{2}p(\pi_o)$ and the information gain is zero. Writing $x = p(\pi_o)$, we conclude that

$$\begin{aligned} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_{\text{H}}^2(M(\pi), M_0(\pi))] &\geq \max \left\{ \frac{1}{4} - 2\gamma\beta x - \frac{2\gamma}{A}, \frac{x}{2} \right\} \\ &\geq \frac{1}{4 + 16\gamma\beta} - \frac{2\gamma}{A}, \end{aligned}$$

where the last inequality uses that the two terms in the maximum cross at $x = (\frac{1}{4} - \frac{2\gamma}{A})/(\frac{1}{2} + 2\gamma\beta)$. Since p was arbitrary, this proves the first claim.

Upper bound for the constrained DEC. Fix a distribution $\nu \in \Delta(\{0, \dots, A\})$, let $\bar{M} = \sum_j \nu_j M_j \in \text{co}(\mathcal{M})$ be the corresponding mixture, and let $i \in \arg \max_{j \in [A]} \nu_j$. Under $\bar{M}$, arm j has mean reward $\frac{1}{2} + \frac{1}{2}\nu_j$, so $\pi_{\bar{M}} = i$, and $\nu_j \leq \frac{1}{2}$ for $j \neq i$. Set

$$p = (1 - \lambda) \cdot \mathbb{I}_i + \lambda \cdot \mathbb{I}_{\pi_o}, \quad \lambda := \frac{8\varepsilon^2}{\beta} \leq 1.$$

For $j \neq i$, $M_j(\pi_o)$ places no mass on observations in $[A] \setminus \{j\}$, while $\bar{M}(\pi_o)$ places mass $\beta(\nu_k + \nu_0/A)$ on each $k \in [A]$. Hence

$$D_{\mathbb{H}}^2(M_j(\pi_o), \bar{M}(\pi_o)) \geq \beta(1 - \nu_j - \frac{\nu_0}{A}) \geq \frac{\beta}{4},$$

using $A \geq 4$, and consequently

$$\mathbb{E}_{\pi \sim p}[D_{\mathbb{H}}^2(M_j(\pi), \bar{M}(\pi))] \geq \frac{\lambda\beta}{4} = 2\varepsilon^2 > \varepsilon^2.$$

Every M_j with $j \neq i$ is therefore excluded from the constraint set in (7.30), which leaves M_i , M_0 , and $\bar{M}$. For each of these, arm i is optimal (for M_0 , every arm is optimal) and π_o has reward 0, so the regret of p is at most λ . This proves that $\text{dec}_{\varepsilon}^c(\mathcal{M} \cup \{\bar{M}\}, \bar{M}) \leq 8\varepsilon^2/\beta$, and the second claim follows by taking the supremum over $\bar{M}$.

Rates. Take $\beta = T^{-1/2}$ and $A = T^2$, so that $\log|\mathcal{M}| \leq 3\log T$. Consider the offset bound (7.23). If $\gamma \leq T^{3/4}$, the first claim gives

$$\text{dec}_{\gamma}(\mathcal{M}) \cdot T \geq \frac{T}{4 + 16T^{1/4}} - \frac{2}{T} \geq \frac{T^{3/4}}{40}$$

for T large enough, and if $\gamma \geq T^{3/4}$, then $\gamma \log(|\mathcal{M}|/\delta) \geq T^{3/4} \log 2$. In either case (7.23) is $\Omega(T^{3/4})$. For the constrained bound, $\bar{\varepsilon}^2 \leq \beta/8$ for T large enough, so the second claim gives

$$\text{dec}_{\bar{\varepsilon}}^c(\mathcal{M}) \cdot T \log T \leq 8C'^2 \log(|\mathcal{M}|/\delta) \sqrt{T} \log T.$$

Finally, $r(\varepsilon) = 8\varepsilon^2/\beta$ satisfies the regularity condition (7.32) with $C_{\text{reg}} = 2$, so Proposition 7.2 applies with $r(\bar{\varepsilon})$ in place of $\text{dec}_{\bar{\varepsilon}}^c(\mathcal{M})$ and gives $\mathbf{Reg} \leq \tilde{O}(\sqrt{T})$. $\square$

8. LOWER BOUNDS: NECESSITY OF THE DECISION-ESTIMATION COEFFICIENT

Up to this point, we have developed algorithms and proved upper bounds on their regret. This section asks the converse question: for a given class of models $\mathcal{M}$, what is the best regret that *any* algorithm can achieve? We show that the Decision-Estimation Coefficient gives an answer to this question, lower bounding the minimax regret for any class $\mathcal{M}$ of models.

Organization. Section 8.1 formalizes algorithms and minimax regret, and Section 8.2 proves the classical $\sqrt{AT}$ lower bound for the multi-armed bandit by hand as an introduction to basic lower bound techniques. Section 8.3 proves a first lower bound based on the DEC as a warm-up, and Section 8.4 strengthens this to the main result. Finally, Section 8.5 instantiates the lower bound for bandits and reinforcement learning.

8.1 Minimax Regret and Optimality

Section 1.4 introduced the minimax perspective informally, but to state lower bounds, we need to be a bit more formal about how we define algorithms and minimax regret. For all the problems considered so far, an algorithm is fully specified¹⁹ by the collection $p = (p^1, \dots, p^T)$, where $p^t = p^t(\cdot \mid \mathcal{H}^{t-1}) \in \Delta(\Pi)$ is the algorithm's strategy for step t (a function of the history $\mathcal{H}^{t-1}$). Let $\mathbb{E}^{M,p}$ denote the expectation with respect to the observation process $(r^t, o^t) \sim M(\pi^t)$ and any randomization used by the algorithm, when $M^* = M$ is the true model. When the algorithm is fixed and clear from context, which is the case in every proof below, we drop it from the notation and write $\mathbb{E}^M$ and $\mathbb{P}^M$ for the expectation and for the law of the history $\mathcal{H}^T$ under the model M .

For a model class $\mathcal{M}$, we define the minimax regret via

$$\mathfrak{M}(\mathcal{M}, T) = \inf_{p^1, \dots, p^T} \sup_{M \in \mathcal{M}} \mathbb{E}^{M,p}[\mathbf{Reg}(T)], \quad (8.1)$$

where, throughout this section only, we write regret as $\mathbf{Reg}(T)$ to make the dependence on T explicit; the number of rounds governs how well any algorithm can distinguish between models, and thus it is useful to parametrize lower bounds by T .

8.2 Warm-Up: Lower Bound for the Multi-Armed Bandit

We begin with the multi-armed bandit, the simplest setting in which to see how a lower bound on regret is proved. Concretely, consider the multi-armed bandit problem with A actions, as studied in Section 3. To have meaningful lower bounds, we need to make sure the noise model is sufficiently confusing to the learner. To this end, we consider the Gaussian noise model, as in Example 7.4:

$$\mathcal{M}_{\text{MAB-G}} = \{M : M(\pi) = \mathcal{N}(f(\pi), 1), f : [A] \rightarrow [0, 1]\}.$$

The following result shows that no algorithm can improve on the $\sqrt{AT}$ regret of UCB and E2D by more than a constant factor.

¹⁹This includes contextual bandits if we define π as a mapping from context to action, as in the DMSO formulation.

Proposition 8.1 (Lower bound for Gaussian bandits): For any algorithm, there exists a model $M \in \mathcal{M}_{\text{MAB-G}}$ where $\mathbb{E}^M[\text{Reg}(T)] \gtrsim \sqrt{AT} \wedge T$. In particular, $\mathfrak{M}(\mathcal{M}_{\text{MAB-G}}, T) \gtrsim \sqrt{AT} \wedge T$.

Proving a lower bound of this type requires showing that *every* algorithm, including ones we have not thought of, incurs this regret on some model in the class. However, not all the models in this collection are equally difficult for a given decision-maker. For example, the model corresponding to $f = (1, 0, \dots, 0)$ is easy for an algorithm that always starts by trying the first action. The art of proving lower bounds is to find a pair or sub-family of models that (1) cannot be optimized simultaneously by any algorithm (i.e., disagree on the optimal action), yet (2) are difficult for any algorithm to distinguish. In what follows, we sketch the idea behind [Proposition 8.1](#), which is to manually build such a family of models by setting the means to be nearly, but not quite the same.

8.2.1 Anatomy of a Lower Bound

Let $\hat{f} = \Delta \cdot e_1$, where $\Delta \in (0, 1/2)$ and e_i is the i -th standard basis vector in $\mathbb{R}^A$. Further, define $f_j = \hat{f} + 2\Delta \cdot e_j$ for $j = 2, \dots, A$. Let $\widehat{M} = \mathcal{N}(\hat{f}, I_A)$ and $M_j = \mathcal{N}(f_j, I_A)$ for each $j = 2, \dots, A$. The first arm is the best decision in $\widehat{M}$, while the j th arm is the best decision in model M_j . We define $\mathcal{M}' \subseteq \mathcal{M}_{\text{MAB-G}}$ as the corresponding sub-family of A models:

$$\mathcal{M}' = \{\widehat{M}\} \cup \{M_j : j = 2, \dots, A\}. \quad (8.2)$$

We want to show that any algorithm must incur large regret on one of the models in $\mathcal{M}'$. Fix $T \in \mathbb{N}$ and an algorithm $p^1, \dots, p^T$. Each p^t is a function of the history $\mathcal{H}^{t-1}$, hence a random variable whose distribution depends on the true model. Averaging over rounds, and over the history under a model M , gives the algorithm's *average profile*,

$$p_M := \mathbb{E}^M \left[\frac{1}{T} \sum_{t=1}^T p^t \right] \in \Delta(\Pi), \quad (8.3)$$

the distribution of the decision at a uniformly random round when M is the true model. The idea behind the proof, which recurs in every lower bound of this section, is that both (1) the regret the algorithm incurs and (2) the information it collects about the model *only depend on the algorithm through its average profile*. That is, for the purpose of proving lower bounds, the profile is a sufficient statistic. The first fact we need to make precise relates regret to the average profile.

Lemma 8.1 (Regret via the average profile): For any model M and algorithm $(p^1, \dots, p^T)$, it holds that

$$\mathbb{E}^M \left[\frac{1}{T} \text{Reg}(T) \right] = \mathbb{E}_{\pi \sim p_M} [f^M(\pi_M) - f^M(\pi)], \quad (8.4)$$

The second fact we need concerns information: how well can the algorithm distinguish two models after T rounds? For this proof, we will reason about this using the KL divergence between the laws of the history under the two models. The following lemma, the *chain rule for KL divergence*, allows us to decompose the joint law of the history under the two models (a complex object to reason about) into a sum of divergences between the conditional laws over per-step observations.

Lemma 8.2 (Chain rule for KL Divergence): Let $(\mathcal{X}_1, \mathcal{F}_1), \dots, (\mathcal{X}_n, \mathcal{F}_n)$ be a sequence of measurable spaces, and let $\mathcal{X}^i = \prod_{t=1}^i \mathcal{X}_t$ and $\mathcal{F}^i = \bigotimes_{t=1}^i \mathcal{F}_t$. For each i , let $\mathbb{P}^i(\cdot | \cdot)$ and $\mathbb{Q}^i(\cdot | \cdot)$ be probability kernels from $(\mathcal{X}^{i-1}, \mathcal{F}^{i-1})$ to $(\mathcal{X}_i, \mathcal{F}_i)$. Let $\mathbb{P}$ and $\mathbb{Q}$ be the laws of $X_1, \dots, X_n$ under $X_i \sim \mathbb{P}^i(\cdot | X_{1:i-1})$ and $X_i \sim \mathbb{Q}^i(\cdot | X_{1:i-1})$ respectively. Then it holds that

$$D_{\text{KL}}(\mathbb{P} \| \mathbb{Q}) = \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^n D_{\text{KL}}(\mathbb{P}^i(\cdot | X_{1:i-1}) \| \mathbb{Q}^i(\cdot | X_{1:i-1})) \right]. \quad (8.5)$$

Applied to the joint law $\mathbb{P}^M$ over the history of an algorithm, this allows us to relate the joint KL divergence to the average profile.

Corollary 8.1 (Chain rule for interactive algorithms): Let M and $\widehat{M}$ be models, and fix an algorithm with average profile $p_{\widehat{M}}$ under $\widehat{M}$. Then the laws $\mathbb{P}^{\widehat{M}}$ and $\mathbb{P}^M$ of the history $\mathcal{H}^T$ under the two models satisfy

$$D_{\text{KL}}(\mathbb{P}^{\widehat{M}} \| \mathbb{P}^M) = \mathbb{E}^{\widehat{M}} \left[\sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \left[D_{\text{KL}}(\widehat{M}(\pi^t) \| M(\pi^t)) \right] \right] = T \cdot \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[D_{\text{KL}}(\widehat{M}(\pi) \| M(\pi)) \right].$$

Proof of Corollary 8.1. We apply Lemma 8.2 to the sequence $\pi^1, z^1, \dots, \pi^T, z^T$ 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

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}^{\widehat{M}} \| \mathbb{P}^M) &= \mathbb{E}^{\widehat{M}} \left[\sum_{t=1}^T D_{\text{KL}}(\mathbb{P}^{\widehat{M}}(\mathbf{z}^t | \mathcal{H}^{t-1}, \pi^t) \| \mathbb{P}^M(\mathbf{z}^t | \mathcal{H}^{t-1}, \pi^t)) + D_{\text{KL}}(\mathbb{P}^{\widehat{M}}(\pi^t | \mathcal{H}^{t-1}) \| \mathbb{P}^M(\pi^t | \mathcal{H}^{t-1})) \right] \\ &= \mathbb{E}^{\widehat{M}} \left[\sum_{t=1}^T D_{\text{KL}}(\widehat{M}(\pi^t) \| M(\pi^t)) \right] \end{aligned}$$

since conditionally on $\mathcal{H}^{t-1}$, the law of π^t does not depend on the model. $\square$

8.2.2 Proof of Proposition 8.1

Equipped with Lemma 8.1 and Corollary 8.1, we are now ready to prove Proposition 8.1.

Proof of Proposition 8.1. We choose $\Delta = C\sqrt{A/T}$ for an absolute constant $C \leq 1/8$, and show that any algorithm must have large regret on either $\widehat{M}$ or M_j for some $j \in \{2, \dots, A\}$. Fix an algorithm $(p^1, \dots, p^T)$. First, observe that if $p_{\widehat{M}}(1) \leq 1/2$, then the algorithm incurs regret at least $T\Delta/2 \gtrsim \sqrt{AT}$ on $\widehat{M}$ by Lemma 8.1, and we are done. Hence, we may assume that $p_{\widehat{M}}(1) > 1/2$. In this case, to choose a second model, we take any smallest coordinate j of $p_{\widehat{M}}$, i.e.,

$$j = \arg \min_{k \in [A]} p_{\widehat{M}}(k),$$

and set $M = M_j$. Equivalently,

$$M_j = \arg \max_{M \in \mathcal{M}'} \mathbb{E}_{\pi \sim p_{\widehat{M}}} [f^M(\pi_M) - f^M(\pi)]. \quad (8.6)$$

We now show that the algorithm incurs large regret on M . To this end, we first note that since $p_{\widehat{M}}(1) > 1/2$, we have $p_{\widehat{M}}(j) \leq 1/(2(A-1))$. If we can show that $p_M(j)$ is also not too large, then we are done by [Lemma 8.1](#). This step, which we may colloquially refer to as change-of-measure, is present in all the lower bounds we prove in this section.

The simplest way to do change-of-measure is to appeal to the definition of the total variation distance. Define the set $\mathcal{E} = [A] \setminus \{j\}$. We have, from [\(7.9\)](#), the data processing inequality [\(7.8\)](#), and [\(7.6\)](#),

$$p_{\widehat{M}}(\mathcal{E}) \leq p_M(\mathcal{E}) + D_{\text{TV}}(p_M, p_{\widehat{M}}) \leq p_M(\mathcal{E}) + \sqrt{D_{\text{KL}}(\mathbb{P}^{\widehat{M}} \parallel \mathbb{P}^M)}, \quad (8.7)$$

where we recall that $\mathbb{P}^{\widehat{M}}$ and $\mathbb{P}^M$ are the joint laws of the histories $\mathcal{H}^T$ under the models $\widehat{M}$ and M , respectively. Now, by [Corollary 8.1](#), we have

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}^{\widehat{M}} \parallel \mathbb{P}^M) &= \mathbb{E}^{\widehat{M}} \left[\sum_{t=1}^T D_{\text{KL}}(\widehat{M}(\pi^t) \parallel M(\pi^t)) \right] \\ &= T \cdot \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[D_{\text{KL}}(\widehat{M}(\pi) \parallel M(\pi)) \right] \\ &= T \cdot p_{\widehat{M}}(j) \cdot D_{\text{KL}}(\widehat{M}(j) \parallel M(j)) \\ &= T \cdot p_{\widehat{M}}(j) \cdot \frac{(2\Delta)^2}{2} \\ &\leq \frac{T}{2(A-1)} \cdot \frac{(2\Delta)^2}{2}. \end{aligned} \quad (8.8)$$

Here, the third equality uses that $\widehat{M}$ and M differ only on coordinate j , so only rounds in which arm j is played contribute. The fourth equality uses that the KL divergence between two Gaussian distributions with variance 1 is half the squared difference of their means, which is $(2\Delta)^2/2$. The final inequality is where the choice of j enters: the algorithm plays arm j with probability at most $1/(2(A-1))$ under $\widehat{M}$, on average over rounds, so the two models are rarely distinguished.

We now choose $\Delta = C\sqrt{A/T}$ for a constant $C \leq 1/8$. Then the right-hand side of [\(8.8\)](#) is $C^2 A/(A-1) \leq 2C^2$, which implies that the two models are statistically indistinguishable by the algorithm. Indeed, by [\(8.7\)](#),

$$p_M(\mathcal{E}) \geq p_{\widehat{M}}(\mathcal{E}) - \sqrt{D_{\text{KL}}(\mathbb{P}^{\widehat{M}} \parallel \mathbb{P}^M)} \geq 1 - \frac{1}{2(A-1)} - \sqrt{2}C \geq \frac{1}{4},$$

where we used $p_{\widehat{M}}(\mathcal{E}) = 1 - p_{\widehat{M}}(j)$ and $A \geq 2$.

Finally, every action in $\mathcal{E}$ is suboptimal under $M = M_j$, by 2Δ for arms other than 1 and by Δ for arm 1. Hence [Lemma 8.1](#) gives

$$\mathbb{E}^M[\mathbf{Reg}(T)] \geq T\Delta \cdot p_M(\mathcal{E}) \geq \frac{T\Delta}{4} = \frac{C}{4}\sqrt{AT},$$

which is the claimed bound. $\square$

The proof is specific to the Gaussian bandit only through the choice of the sub-family $\mathcal{M}'$. To isolate the general principle underlying it, and to prepare for the theorems of the next two subsections, let us summarize the main steps of the proof:

1. Pick a reference model $\widehat{M}$.
2. Fix any algorithm $(p^1, \dots, p^T)$.
3. Compute the corresponding average profile $p_{\widehat{M}}$.
4. Pick a second model M based on the average profile $p_{\widehat{M}}$ by solving (8.6) over a “difficult enough” (with respect to $\widehat{M}$) subset $\mathcal{M}'$ of the original model class $\mathcal{M}_{\text{MAB-G}}$.

What does “difficult enough” mean in the last step? The models in $\mathcal{M}'$ must be hard to distinguish from $\widehat{M}$ *by the algorithm*, that is, with respect to the profile $p_{\widehat{M}}$. In the proof, this was certified by the KL bound (8.8), which is small precisely when $\mathbb{E}_{\pi \sim p_{\widehat{M}}}[D_{\text{KL}}(\widehat{M}(\pi) \| M(\pi))] \lesssim 1/T$: the second model may differ from $\widehat{M}$ only where the algorithm rarely looks. This suggests replacing the hand-picked sub-family by the set of all models within a fixed information budget of $\widehat{M}$, which depends on the profile p under consideration,

$$\mathcal{M}'(p) = \left\{ M \in \mathcal{M} : \mathbb{E}_{\pi \sim p} \left[D_{\text{KL}}(\widehat{M}(\pi) \| M(\pi)) \right] \leq \varepsilon^2 \right\}$$

for some $\varepsilon \propto \sqrt{1/T}$. With this definition, the four steps become the optimization problem

$$\sup_{\widehat{M}} \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}'(p)} \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)]. \quad (8.9)$$

This is precisely the maximal value of $\text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M})$, the constrained DEC of (7.30), except that the DEC measures information with Hellinger distance rather than KL divergence; we switch to the former for technical reasons explained in Section 8.3. That we can define this optimization over profiles p , rather than over algorithms, is the payoff of the sufficient-statistic idea: regret and information both depend on the algorithm only through $p_{\widehat{M}}$, so it is enough to consider every profile an algorithm could produce.

Remark 8.1 (Connection to the modulus of continuity): The constrained DEC can be thought of as an interactive counterpart to the *modulus of continuity* in statistical estimation [67, 68]. In the classical setting, one observes T i.i.d. samples from an unknown distribution P belonging to a class $\mathcal{P}$, and wishes to estimate a functional $\theta(P)$ of the distribution, such as its mean, a quantile, or the value of its density at a point. The modulus of continuity of θ over $\mathcal{P}$ is

$$\omega(\varepsilon) = \sup\{|\theta(P) - \theta(Q)| : P, Q \in \mathcal{P}, D_{\text{H}}(P, Q) \leq \varepsilon\},$$

the largest change in the functional across a Hellinger ball of radius ε . It characterizes the minimax rate of estimation: the two-point method [141] shows that no estimator can achieve error better than $\omega(c/\sqrt{T})$, since two distributions at Hellinger distance $1/\sqrt{T}$ cannot be distinguished from T samples. In general the modulus gives only a lower bound, but for affine functionals over convex classes, under mild regularity conditions, this rate is also achievable [167].

One can immediately see that the constrained DEC places a similar hard constraint on Hellinger distance, and the lower bounds of the next two subsections use the same radius $\varepsilon \propto 1/\sqrt{T}$. The constrained DEC differs from the modulus of continuity only through reasons related to interaction. The functional of interest is the regret of a decision distribution p , rather than a fixed functional of the distribution, and Hellinger distance is measured under p itself, since the learner’s decisions determine what data it observes. Moreover, whereas the classical two-point method selects the pair of distributions obliviously, the pair $(\widehat{M}, M)$ in (8.9), as in the proof

of Proposition 8.1, is chosen adversarially based on the algorithm. The latter is necessary for the number of decisions to appear in the lower bound, even for multi-armed bandits.

8.3 A First Lower Bound via the Constrained Decision-Estimation Coefficient

This subsection turns the recipe of Section 8.2 into a general lower bound based on the constrained DEC. Throughout this subsection, we take $L : \mathcal{M} \times \Pi \rightarrow [0, 1]$ to be an arbitrary function, and define $\text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M})$ as the constrained DEC of (7.30) with $L(M, \pi)$ in place of $f^M(\pi_M) - f^M(\pi)$. Our motivating case of interest is

$$L(M, \pi) = f^M(\pi_M) - f^M(\pi), \quad (8.10)$$

for which $\sum_{t=1}^T L(M, \pi^t)$ is the regret and $\text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M})$ is the constrained DEC as defined in Section 7.5. The result we prove is a warm-up for the main theorem of Section 8.4 in one respect: an additional side condition ties the radius to the value of the DEC, which weakens the rate; we remove this limitation in Section 8.4

Before getting to the proof, we state a Hellinger analogue of Corollary 8.1. In Section 8.2 we measured the closeness of two models with KL divergence, which was sufficient there and has an exact, elementary chain rule. From here on we use Hellinger distance instead. Even though KL divergence is well-behaved for some classes, in general it can be much larger than Hellinger distance, and is infinite when the supports differ, so a lower bound through a KL ball would be weaker and potentially fall short of the DEC-based upper bound in Proposition 7.2. The price is that squared Hellinger distance does not satisfy an exact chain rule. It is, however, sub-additive over rounds up to a constant factor, which suffices for our purposes [92, Lemma D.2] (see also Jayram [109] and Duchi [71, Lemma 11.5.3]).

Lemma 8.3 (Sub-additivity for squared Hellinger distance): With the notation of Lemma 8.2,

$$D_H^2(\mathbb{P}, \mathbb{Q}) \leq 7 \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^n D_H^2(\mathbb{P}^i(\cdot \mid X_{1:i-1}), \mathbb{Q}^i(\cdot \mid X_{1:i-1})) \right]. \quad (8.11)$$

Applying this to the history as in the proof of Corollary 8.1 gives the following bound, in which the information collected by the algorithm again depends on it only through its profile.

Corollary 8.2 (Hellinger distance for T -round algorithms): Let M and $\widehat{M}$ be models, and fix an algorithm with average profile $p_{\widehat{M}}$ under $\widehat{M}$. Then

$$D_{\text{TV}}^2(\mathbb{P}^M, \mathbb{P}^{\widehat{M}}) \leq D_H^2(\mathbb{P}^M, \mathbb{P}^{\widehat{M}}) \leq 7T \cdot \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[D_H^2(M(\pi), \widehat{M}(\pi)) \right].$$

We can now state the first lower bound. It is expressed through the supremum of the constrained DEC $\text{dec}_\varepsilon^c(\mathcal{M}, \widehat{M})$ over *arbitrary* reference models $\widehat{M} : \Pi \rightarrow \Delta(\mathcal{R} \times \mathcal{O})$, rather than $\widehat{M} \in \text{co}(\mathcal{M})$ as in (7.31). The alternatives M in the definition still range over $\mathcal{M}$, since the true model must lie in the class; this is why the reference model is not added to the class as in (7.31). Recall that we

state the result for an arbitrary bounded function $L(M, \pi)$ of the model and the decision, not only for the regret.²⁰

Proposition 8.2 (First lower bound via the constrained DEC): Let $L : \mathcal{M} \times \Pi \rightarrow [0, 1]$ be arbitrary, and define $\mathbf{Reg} = \sum_{t=1}^T L(M, \pi^t)$. Let any $\varepsilon > 0$ satisfy

$$T\varepsilon^2 \leq \frac{1}{56} \cdot \sup_{\widehat{M}} \text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}), \quad (8.12)$$

where the supremum ranges over all models $\widehat{M} : \Pi \rightarrow \Delta(\mathcal{R} \times \mathcal{O})$. Then every algorithm satisfies

$$\sup_{M \in \mathcal{M}} \mathbb{E}^M \left[\frac{1}{T} \mathbf{Reg} \right] \geq \frac{1}{6} \sup_{\widehat{M}} \text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}).$$

Let us interpret the condition (8.12), which is the main feature of Proposition 8.2 that leaves room for improvement. For the Gaussian bandit, where $\text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}) \approx \varepsilon\sqrt{A}$ (see Section 8.5), the side condition $T\varepsilon^2 \leq \varepsilon\sqrt{A}/56$ forces $\varepsilon \lesssim \sqrt{A}/T$, and the proposition yields only $\mathbb{E}[\mathbf{Reg}(T)] \gtrsim T \cdot \varepsilon\sqrt{A} \approx A$. While this correctly shows that the lower bound must scale with the number of actions, it does not grow with T , whereas the true minimax regret is of order $\sqrt{AT}$. The main theorem of the next subsection removes the side condition and recovers the $\sqrt{AT}$ rate. To do so, it specializes L to the regret (8.10) and assumes that the reward realizations lie in $[0, 1]$; as we explain there, this is what permits a stronger change-of-measure argument.

The proof is a change-of-measure argument in the style of Section 8.2. Note that the reference model $\widehat{M}$ need not belong to $\mathcal{M}$: it serves only as the measure under which the algorithm's behavior is analyzed, and nature ultimately plays a model $M \in \mathcal{M}$. Lower bounds of this form, with arbitrary reference models, appear in Foster et al. [88].

Proof of Proposition 8.2. We follow the template of Section 8.2. Write $D = \sup_{\widehat{M}} \text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M})$, and choose $\widehat{M} : \Pi \rightarrow \Delta(\mathcal{R} \times \mathcal{O})$ with $\text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}) = D$; if the supremum is not attained, an approximate maximizer suffices, at the cost of an arbitrarily small loss in the constants below. Fix an arbitrary algorithm and let $p_{\widehat{M}}$ be its average profile under $\widehat{M}$. Choose $M \in \mathcal{M}$ as a best response to this profile in the constrained DEC, so that

$$\mathbb{E}_{\pi \sim p_{\widehat{M}}} [L(M, \pi)] \geq D, \quad \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] \leq \varepsilon^2.$$

The second condition says that M is hard to distinguish from $\widehat{M}$: by Corollary 8.2 and the side condition (8.12) on ε ,

$$D_{\text{H}}^2 \left(\mathbb{P}^M, \mathbb{P}^{\widehat{M}} \right) \leq 7T \cdot \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[D_{\text{H}}^2 \left(M(\pi), \widehat{M}(\pi) \right) \right] \leq 7T \cdot \varepsilon^2 \leq \frac{D}{8}.$$

The first condition says that the profile $p_{\widehat{M}}$ has large regret under M ; it remains to transfer this to the profile p_M , which is what the algorithm actually plays under M . Since p_M and $p_{\widehat{M}}$ are the laws of the decision at a random round under the two models, the data processing inequality gives

²⁰The proof uses nothing about the form of L beyond $L \in [0, 1]$; in particular, it places no assumption on the reward realizations, which may be unbounded (as an aside, it applies even in settings where rewards are not observed at all).

$D_{\mathbb{H}}^2(p_M, p_{\widehat{M}}) \leq D_{\mathbb{H}}^2(\mathbb{P}^M, \mathbb{P}^{\widehat{M}}) \leq D/8$, and the change-of-measure inequality for bounded functions (Lemma A.7 in Appendix A.2.2) yields

$$\mathbb{E}_{\pi \sim p_{\widehat{M}}}[\mathbf{L}(M, \pi)] \leq 3 \mathbb{E}_{\pi \sim p_M}[\mathbf{L}(M, \pi)] + 4D_{\mathbb{H}}^2(p_M, p_{\widehat{M}}) \leq 3 \mathbb{E}_{\pi \sim p_M}[\mathbf{L}(M, \pi)] + \frac{D}{2}.$$

Hence, by Lemma 8.1,

$$\mathbb{E}^M \left[\frac{1}{T} \mathbf{Reg} \right] = \mathbb{E}_{\pi \sim p_M}[\mathbf{L}(M, \pi)] \geq \frac{D}{6},$$

which proves the claim. $\square$

8.4 Main Lower Bound for the Constrained Decision-Estimation Coefficient

We now state the main result of this section: a lower bound in terms of $\text{dec}_{\varepsilon}^c(\mathcal{M})$ at radius $\varepsilon \propto T^{-1/2}$, which qualitatively matches the upper bounds in preceding sections.

Theorem 8.1 (Lower bound via the constrained DEC): Let $T \geq 1$, and let $\mathcal{M} \subseteq (\Pi \rightarrow \Delta([0, 1] \times \mathcal{O}))$ be a model class with rewards in $[0, 1]$. Define $\mathbf{L}(M, \pi) = f^M(\pi_M) - f^M(\pi)$. Then, for any T -round algorithm,

$$\sup_{M \in \mathcal{M}} \mathbb{E}^M \left[\frac{\mathbf{Reg}(T)}{T} \right] \geq \frac{1}{8} \left(\text{dec}_{\varepsilon}^c(\mathcal{M}) - 6\varepsilon - \frac{1}{T} \right), \quad \varepsilon = \frac{1}{24\sqrt{T}}.$$

It is instructive to compare Theorem 8.1 to the upper bounds of Section 7. The lower bound gives $\mathbb{E}[\mathbf{Reg}(T)] \gtrsim \text{dec}_{\varepsilon}^c(\mathcal{M}) \cdot T$ with $\varepsilon \propto T^{-1/2}$, while the constrained-DEC variant of E2D in Proposition 7.2 achieves $\mathbf{Reg}(T) \lesssim \text{dec}_{\varepsilon}^c(\mathcal{M}) \cdot T \log T + \sqrt{T \log(|\mathcal{M}|/\delta)}$ with $\varepsilon \propto \sqrt{\log(|\mathcal{M}|/\delta)/T}$. The two radii differ by a factor of $\sqrt{\log|\mathcal{M}|}$, so up to this factor and the estimation term, the upper and lower bounds coincide. In other words, the exploration strategy prescribed by the DEC is the best possible; whatever is lost relative to the minimax regret is lost in estimation, not in exploration.

Compared with Proposition 8.2, the theorem is specialized to the regret, $\mathbf{L}(M, \pi) = f^M(\pi_M) - f^M(\pi)$, and assumes that the reward realizations lie in $[0, 1]$. This specialization, together with boundedness (more generally, the condition in Remark 8.2), facilitates a stronger change-of-measure argument. The proof otherwise follows the template of Section 8.2, with one new device: the average profile is computed only up to a stopping time, the first round at which the algorithm has accumulated a prescribed amount of regret under $\widehat{M}$, so that the regret under $\widehat{M}$ and under the second model M can be compared on the same event.

Proof of Theorem 8.1. Fix any reference model $\widehat{M} \in \text{co}(\mathcal{M})$ and an algorithm. Set

$$R = \sup_{M \in \mathcal{M}} \mathbb{E}^M \left[\frac{\mathbf{Reg}(T)}{T} \right], \quad \varepsilon = \frac{1}{24\sqrt{T}}.$$

We prove

$$\text{dec}_{\varepsilon}^c(\mathcal{M}) \leq 8R + 6\varepsilon + \frac{1}{T}. \quad (8.13)$$

Since all losses are at most one, the claim is immediate when $8R + 6\varepsilon + 1/T \geq 1$. We may therefore assume the reverse inequality.

Let $\pi_{\widehat{M}}$ maximize $f^{\widehat{M}}$. For a sufficiently small $\eta > 0$, put

$$a = 4R + 2\varepsilon + \eta < 1.$$

Run the algorithm under $\widehat{M}$, and write

$$g_t = f^{\widehat{M}}(\pi_{\widehat{M}}) - \mathbb{E}_{\pi \sim p^t(\cdot | \mathcal{H}^{t-1})} [f^{\widehat{M}}(\pi)], \quad G_t = \sum_{s=1}^t g_s.$$

Let τ be the first $t \leq T$ for which $G_t \geq Ta$, with $\tau = T$ if the threshold is never reached. Define the stopped-and-padded average profile

$$p_{\widehat{M}} = \frac{1}{T} \mathbb{E}^{\widehat{M}} \left[\sum_{t=1}^{\tau} p^t(\cdot | \mathcal{H}^{t-1}) + (T - \tau) \delta_{\pi_{\widehat{M}}} \right]. \quad (8.14)$$

Because $0 \leq g_t \leq 1$, this profile has reference loss

$$\mathbb{E}_{\pi \sim p_{\widehat{M}}} [f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)] = \frac{1}{T} \mathbb{E}^{\widehat{M}} [G_{\tau}] < a + \frac{1}{T}. \quad (8.15)$$

For this particular profile, let

$$B = \sup_{\substack{N \in \mathcal{M} \cup \{\widehat{M}\} \\ \mathbb{E}_{\pi \sim p_{\widehat{M}}} [D_{\text{H}}^2(N(\pi), \widehat{M}(\pi))] \leq \varepsilon^2}} \mathbb{E}_{\pi \sim p_{\widehat{M}}} [f^N(\pi_N) - f^N(\pi)]. \quad (8.16)$$

Since $p_{\widehat{M}}$ is one admissible choice in the DEC infimum,

$$\text{dec}_{\varepsilon}^c(\mathcal{M} \cup \{\widehat{M}\}, \widehat{M}) \leq B. \quad (8.17)$$

We claim that

$$B \leq 2a + \sqrt{2}\varepsilon + \frac{1}{T}. \quad (8.18)$$

Suppose otherwise. Choose a model M attaining the supremum in (8.16); if it is not attained, choose a sufficiently accurate approximate maximizer. Denote its objective value by b , so that

$$b > 2a + \sqrt{2}\varepsilon + \frac{1}{T}. \quad (8.19)$$

The choice $M = \widehat{M}$ has objective value less than $a + 1/T$ by (8.15); hence the selected model belongs to $\mathcal{M}$. It also satisfies

$$\mathbb{E}_{\pi \sim p_{\widehat{M}}} [D_{\text{H}}^2(M(\pi), \widehat{M}(\pi))] \leq \varepsilon^2. \quad (8.20)$$

Since the reward is observed and belongs to $[0, 1]$,

$$|f^M(\pi) - f^{\widehat{M}}(\pi)|^2 \leq 2D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)).$$

Thus, (8.20) and Cauchy–Schwarz imply

$$\mathbb{E}_{\pi \sim p_{\widehat{M}}} [|f^M(\pi) - f^{\widehat{M}}(\pi)|] \leq \sqrt{2}\varepsilon.$$

Together with (8.15) and (8.19), this gives

$$\begin{aligned} f^M(\pi_M) - f^{\widehat{M}}(\pi_{\widehat{M}}) &= b + \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[f^M(\pi) - f^{\widehat{M}}(\pi) \right] \\ &\quad - \mathbb{E}_{\pi \sim p_{\widehat{M}}} \left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi) \right] > a. \end{aligned} \quad (8.21)$$

Now run the original algorithm under M , defining G_t and τ from its realized history by the same formulas. Let

$$X = \sum_{t=1}^{\tau} \mathbb{E}_{\pi \sim p^t(\cdot|\mathcal{H}^{t-1})} \left[|f^M(\pi) - f^{\widehat{M}}(\pi)| \right].$$

For a realized history, write the conditional pseudo-regret through time s as

$$\overline{\mathbf{Reg}}_M(s) = \sum_{t=1}^s [f^M(\pi_M) - \mathbb{E}_{\pi \sim p^t(\cdot|\mathcal{H}^{t-1})} [f^M(\pi)]] .$$

For every stopped history, (8.21) implies

$$\overline{\mathbf{Reg}}_M(\tau) \geq \tau a + G_\tau - X \geq T a - X. \quad (8.22)$$

Indeed, if $\tau = T$, the first term is at least Ta ; if $\tau < T$, the threshold term G_τ is at least Ta .

Under $\widehat{M}$, feasibility and the definition of $p_{\widehat{M}}$ give

$$\begin{aligned} \mathbb{E}^{\widehat{M}}[X^2] &\leq T \mathbb{E}^{\widehat{M}} \left[\sum_{t=1}^{\tau} \mathbb{E}_{\pi \sim p^t(\cdot|\mathcal{H}^{t-1})} \left[|f^M(\pi) - f^{\widehat{M}}(\pi)|^2 \right] \right] \\ &\leq 2T^2 \varepsilon^2. \end{aligned} \quad (8.23)$$

Consequently, $X \leq 2T\varepsilon$ with probability at least $1/2$ under $\widehat{M}$. Padding each stopped transcript with a common cemetery symbol and applying the Hellinger chain rule (Corollary 8.2) gives

$$D_{\text{H}}^2(\mathbb{P}_{\tau}^{\widehat{M}}, \mathbb{P}_{\tau}^M) \leq 7T\varepsilon^2, \quad D_{\text{TV}}(\mathbb{P}_{\tau}^{\widehat{M}}, \mathbb{P}_{\tau}^M) \leq \sqrt{14T\varepsilon^2} < \frac{1}{4}.$$

Thus, $X \leq 2T\varepsilon$ with probability at least $1/4$ under M . Since conditional instantaneous regret is nonnegative,

$$\begin{aligned} \mathbb{E}^M \left[\frac{\mathbf{Reg}(T)}{T} \right] &= \mathbb{E}^M \left[\frac{\overline{\mathbf{Reg}}_M(T)}{T} \right] \\ &\geq \mathbb{E}^M \left[\frac{\overline{\mathbf{Reg}}_M(\tau)}{T} \right] \geq \frac{1}{4}(a - 2\varepsilon) = R + \frac{\eta}{4}, \end{aligned}$$

contradicting the definition of R . This proves (8.18).

Combining (8.17) and (8.18), and then letting $\eta \downarrow 0$, gives

$$\text{dec}_{\varepsilon}^c(\mathcal{M} \cup \{\widehat{M}\}, \widehat{M}) \leq 8R + (4 + \sqrt{2})\varepsilon + \frac{1}{T} \leq 8R + 6\varepsilon + \frac{1}{T}.$$

Taking the supremum over $\widehat{M} \in \text{co}(\mathcal{M})$ proves (8.13). Rearranging completes the proof. $\square$

The theorem is stated for rewards in $[0, 1]$; the following remark extends general reward distributions with bounded means, which covers the Gaussian examples below.

Remark 8.2 (Unbounded rewards): The conclusion of [Theorem 8.1](#) extends, up to absolute constants, to unbounded rewards provided that the mean rewards lie in $[0, 1]$ and, for some absolute constant c , all $M \in \mathcal{M}$, $\bar{M} \in \text{co}(\mathcal{M})$, and $\pi \in \Pi$ satisfy

$$|f^M(\pi) - f^{\bar{M}}(\pi)|^2 \leq c \cdot D_{\text{H}}^2(M(\pi), \bar{M}(\pi)).$$

Indeed, these are the only consequences of bounded rewards used in the proof. The numerical coefficient of ε may depend on c . This includes unit-variance Gaussian rewards with means in $[0, 1]$, including their mixtures as reference models. Indeed, a standard Cauchy–Schwarz argument gives

$$|\mathbb{E}_P[r] - \mathbb{E}_Q[r]|^2 \leq 2(\mathbb{E}_P[r^2] + \mathbb{E}_Q[r^2]) \cdot D_{\text{H}}^2(P, Q),$$

and the second moments are at most two, so $c = 8$ suffices.

8.5 Examples

We now instantiate the lower bound in [Theorem 8.1](#) for concrete model classes. We begin by revisiting examples from [Section 7.3](#). Each example lower bounds $\text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M})$ for a single, conveniently chosen reference model $\widehat{M} \in \text{co}(\mathcal{M})$, which suffices since $\text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}) \leq \text{dec}_{\varepsilon}^c(\mathcal{M})$ by [\(7.31\)](#).

Example 7.4 (Multi-armed bandit with Gaussian rewards, cont'd). Let us lower bound the constrained DEC for the Gaussian bandit problem from [Example 7.4](#). Set $\widehat{M}(\pi) = \mathcal{N}(1/2, 1)$, and let $\{M_1, \dots, M_A\} \subseteq \mathcal{M}$ 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 to be chosen below. Observe that for all i , $\mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2(M_i(\pi), \widehat{M}(\pi)) \right] \leq \frac{1}{2} \Delta^2 p(i)$ by [\(7.14\)](#), and $\mathbb{E}_{\pi \sim p} [f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi)] = (1 - p(i))\Delta$, so we can lower bound

$$\begin{aligned} \text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}) &= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p} [f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim p} \left[D_{\text{H}}^2(M(\pi), \widehat{M}(\pi)) \right] \leq \varepsilon^2 \right\} \\ &\geq \inf_{p \in \Delta(\Pi)} \max_i \left\{ (1 - p(i))\Delta \mid p(i) \frac{\Delta^2}{2} \leq \varepsilon^2 \right\} \end{aligned}$$

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 satisfies the constraint $p(i) \frac{\Delta^2}{2} \leq \varepsilon^2$, and we are left with

$$\text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}) \geq (1 - p(i))\Delta \geq \varepsilon \sqrt{A/2},$$

since $1 - p(i) \geq 1/2$.

Plugging this lower bound on the constrained Decision-Estimation Coefficient into [Theorem 8.1](#), with the extension in [Remark 8.2](#), yields

$$\mathbb{E}[\text{Reg}] \geq \Omega(\sqrt{AT}).$$

This recovers [Proposition 8.1](#), with the DEC automatically selecting the hard alternative model. $\triangleleft$

Generalizing the argument above, the following proposition gives a useful recipe for lower bounding the Decision-Estimation Coefficient for any model class that “embeds” a family of models with properties similar to the multi-armed bandit problem.

Proposition 8.3 (DEC lower bound for hard families): Let a reference model $\widehat{M}$ be given, and suppose that a class $\mathcal{M}$ contains a sub-class $\{M_1, \dots, M_N\}$ and collection of decisions $\pi_1, \dots, \pi_N$ with the property that for all i :

1. $D_{\text{H}}^2(M_i(\pi), \widehat{M}(\pi)) \leq \beta^2 \cdot \mathbb{I}\{\pi = \pi_i\}.$
2. $f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \alpha \cdot \mathbb{I}\{\pi \neq \pi_i\}.$

Then

$$\text{dec}_{\varepsilon}^c(\mathcal{M}, \widehat{M}) \gtrsim \alpha \cdot \mathbb{I}\left\{\varepsilon \geq \beta/\sqrt{N}\right\}.$$

The examples that follow can be obtained by applying this result with an appropriate sub-family; see Foster et al. [84, 89] for detailed proofs.

Example 7.5 (Bandits with structured noise, cont'd). Recall the bandit-type problem with structured noise from [Example 7.5](#), where we have $\mathcal{M}_{\text{MAB-SN}} = \{M_1, \dots, M_A\} \cup \{M_0\}$, with $M_i(\pi) = \mathcal{N}(1/2, 1)\mathbb{I}\{\pi \neq i\} + \text{Ber}(3/4)\mathbb{I}\{\pi = i\}$ and $M_0(\pi) = \mathcal{N}(1/2, 1)$ for every π . With M_0 as the reference model, the sub-family $\{M_1, \dots, M_A\}$ satisfies the conditions of [Proposition 8.3](#) with $\alpha = 1/4$ and $\beta^2 = 2$. As a result, we have $\text{dec}_{\varepsilon}^c(\mathcal{M}_{\text{MAB-SN}}) \gtrsim \mathbb{I}\left\{\varepsilon \geq \sqrt{2/A}\right\}$, which yields

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \widetilde{\Omega}(A)$$

if we apply [Theorem 8.1](#), with the extension in [Remark 8.2](#).

◁

Example 7.6 (Bandits with full information, cont'd). Consider the full-information variant of the bandit setting in [Example 7.6](#). By adapting the argument in [Example 7.4](#), one can show that

$$\text{dec}_{\varepsilon}^c(\mathcal{M}) \gtrsim \varepsilon,$$

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 5](#). For these, it is convenient to have a version of the lower bound stated directly in terms of the reward function class $\mathcal{F}$, in the language of [Section 5](#). For $\widehat{f} : \Pi \rightarrow \mathbb{R}$, define the square-loss constrained DEC

$$\text{dec}_{\varepsilon}^{c, \text{sq}}(\mathcal{F}, \widehat{f}) = \inf_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\pi \sim p}[f(\pi_f) - f(\pi)] \mid \mathbb{E}_{\pi \sim p}[(f(\pi) - \widehat{f}(\pi))^2] \leq \varepsilon^2 \right\},$$

the analogue of [\(7.30\)](#) with squared error in place of Hellinger distance.

Corollary 8.3 (Lower bound for structured bandits): Let $\mathcal{F} \subseteq (\Pi \rightarrow [0, 1])$, and let $\mathcal{M}_{\mathcal{F}} = \{M_f : f \in \mathcal{F}\}$ with $M_f(\pi) = \mathcal{N}(f(\pi), 1)$ be the induced structured bandit class with unit-variance Gaussian rewards. Then for all $\varepsilon > 0$,

$$\text{dec}_{\varepsilon}^c(\mathcal{M}_{\mathcal{F}}) \geq \sup_{\hat{f} \in \mathcal{F}} \text{dec}_{\sqrt{2}\varepsilon}^{c, \text{sq}}(\mathcal{F}, \hat{f}),$$

and consequently every algorithm has $\mathbb{E}[\mathbf{Reg}(T)] \gtrsim T \cdot \sup_{\hat{f} \in \mathcal{F}} \text{dec}_{c/\sqrt{T}}^{c, \text{sq}}(\mathcal{F}, \hat{f}) - \sqrt{T}$ for an absolute constant $c > 0$.

We omit a formal proof, but briefly, the first claim holds because, by (7.14), $D_{\text{H}}^2(M_f(\pi), M_{\hat{f}}(\pi)) \leq \frac{1}{2}(f(\pi) - \hat{f}(\pi))^2$, so the square-loss ball of radius $\sqrt{2}\varepsilon$ around $\hat{f}$ is contained in the Hellinger ball of radius ε around $M_{\hat{f}} \in \mathcal{M}_{\mathcal{F}}$, and enlarging the max player's feasible set can only increase the value.

The examples that follow are instances of this corollary.

Example 8.1 (Linear bandits). Consider the linear bandit setting in Section 5.4.2, with $\mathcal{F} = \{\pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{B}_2^d(1)$ is a parameter set and $\phi : \Pi \rightarrow \mathbb{R}^d$ 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

$$\text{dec}_{\varepsilon}^c(\mathcal{M}) \gtrsim \varepsilon \sqrt{d},$$

which gives

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{dT}.$$

◁

Example 8.2 (Nonparametric bandits). Consider the Lipschitz bandit setting in Section 5.4.3, where Π is a metric space with metric ρ , and

$$\mathcal{F} = \{f : \Pi \rightarrow [0, 1] \mid f \text{ is 1-Lipschitz w.r.t } \rho\}.$$

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

$$\text{dec}_{\varepsilon}^c(\mathcal{M}) \gtrsim \varepsilon^{\frac{2}{d+2}},$$

which leads to $\mathbb{E}[\mathbf{Reg}] \gtrsim T^{\frac{d+1}{d+2}}$.

◁

Each of these lower bounds is matched, up to logarithmic factors, by an upper bound proved earlier (Proposition 5.5 for linear bandits and Proposition 5.6 for Lipschitz bandits), both via E2D with the corresponding offset DEC. In both cases the offset and constrained coefficients have the same order, which is why the lower bounds are proven through the latter and the upper bounds are proven through the former without loss.

Tabular reinforcement learning. Our final example is the class of tabular MDPs from [Section 6](#), for which [Section 6.7](#) gave the upper bound $\mathbf{Reg} \lesssim \sqrt{\text{poly}(H, S, A) \cdot T}$ via PC-IGW. We state, but do not prove, a complementary lower bound on the constrained DEC [\[84, 89\]](#).

Proposition 8.4 (DEC lower bound for tabular RL): 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

$$\text{dec}_\varepsilon^c(\mathcal{M}) \gtrsim \varepsilon \sqrt{HSA}.$$

Using [Theorem 8.1](#), this gives $\mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{HSA T}$, matching the upper bound of [Section 6.7](#) up to polynomial factors in H .

8.6 Tighter Regret Bounds for the Decision-Estimation Coefficient*

The main gap between the most general upper and lower bounds based on the Decision-Estimation Coefficient concerns estimation. All of the upper bounds we have given pay an estimation cost, represented by $\log|\mathcal{M}|$ for a finite model class, that is absent from the DEC lower bound. Each of the bounds we have presented can be tight, depending on the model class, but sharper guarantees are possible in many special cases of interest. In this optional epilogue, we present some general techniques for obtaining such refinements by tailoring estimation to the decision-making task. Obtaining a complete understanding of when and how the estimation gap can be closed remains an active area of research; see, for example, [Chen et al. \[52\]](#).

8.6.1 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 $D_H^2(M(\pi), \widehat{M}(\pi))$, which is fundamental in the sense that it leads to lower bounds on the optimal regret ([Theorem 8.1](#)). 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”) $D^\pi(\cdot \parallel \cdot)$, we define

$$\text{dec}_\gamma^D(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^\pi(\widehat{M} \parallel M) \right]. \quad (8.24)$$

This variant of the DEC naturally leads to regret bounds in terms of estimation error under $D^\pi(\cdot \parallel \cdot)$. Note that we use notation $D^\pi(\widehat{M} \parallel M)$ instead of say, $D(\widehat{M}(\pi), M(\pi))$, 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 E2D assumes that at each round, the online estimation oracle provides a point estimate $\widehat{M}^t$. In some settings, it is 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

$$\overline{\text{dec}}_\gamma^D(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M} \sim \nu} \left[D^\pi(\widehat{M} \parallel M) \right] \right] \quad (8.25)$$

for distributions $\nu \in \Delta(\mathcal{M})$. We additionally define $\overline{\text{dec}}_\gamma^D(\mathcal{M}) = \sup_{\nu \in \Delta(\mathcal{M})} \overline{\text{dec}}_\gamma^D(\mathcal{M}, \nu)$.

Algorithm 8.1 E2D for General Divergences and Randomized Estimators

input: Exploration parameter $\gamma > 0$; divergence $D(\cdot \parallel \cdot)$; randomized online estimation oracle.

for $t = 1, 2, \dots, T$ **do**

Obtain randomized estimate $\nu^t \in \Delta(\mathcal{M})$ from estimation oracle with $\{(\pi^i, r^i, o^i)\}_{i < t}$.

Compute // Eq. (8.25).

$$p^t = \arg \min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M} \sim \nu^t} \left[D^\pi(\widehat{M} \parallel M) \right] \right].$$

Sample decision $\pi^t \sim p^t$ and update estimation algorithm with (π^t, r^t, o^t) .

Algorithm. A generalization of E2D that incorporates general divergences and randomized estimators is given at the start of this subsection. The algorithm is identical to E2D (Section 7.4), with the only differences being that i) we play the distribution that solves the minimax problem (8.25) with the user-specified divergence $D^\pi(\cdot \parallel \cdot)$ 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 $\nu^1, \dots, \nu^T \in \Delta(\mathcal{M})$ with respect to the given divergence $D^\pi(\cdot \parallel \cdot)$, which we define as

$$\mathbf{Est}_D := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}_{\widehat{M}^t \sim \nu^t} \left[D^{\pi^t}(\widehat{M}^t \parallel M^\star) \right]. \quad (8.26)$$

We have the following guarantee.

Proposition 8.5 (Regret of E2D with general divergences): The algorithm E2D for General Divergences and Randomized Estimators with exploration parameter $\gamma > 0$ guarantees that

$$\mathbf{Reg} \leq \overline{\text{dec}}_\gamma^D(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_D \quad (8.27)$$

for every realization.

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* $\psi : \mathcal{M} \rightarrow \Psi$ with the property that we can write (overloading notation)

$$D^\pi(M \parallel M') = D^\pi(\psi(M) \parallel \psi(M')), \quad f^M(\pi) = f^{\psi(M)}(\pi), \quad \text{and} \quad \pi_M = \pi_{\psi(M)}$$

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

$$\mathbf{Est}_D := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}_{\widehat{\psi}^t \sim \nu^t} \left[D^{\pi^t}(\widehat{\psi}^t \parallel M^\star) \right]. \quad (8.28)$$

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 **Est** that scale with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.

For example, in structured bandit problems, one can work with the divergence

$$D_{\text{Sq}}(\widehat{M}(\pi), M(\pi)) := (f^M(\pi) - f^{\widehat{M}}(\pi))^2$$

which uses the mean reward function as a sufficient statistic, i.e. $\psi(M) = f^M$. Here, it is clear that one can achieve $\mathbf{Est}_D \lesssim \log|\mathcal{F}|$, which improves upon the rate $\mathbf{Est}_H \lesssim \log|\mathcal{M}|$ for Hellinger distance, and recovers the specialized version of the E2D algorithm we considered in [Section 5](#). 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 9](#) for details.

Does randomized estimation help? Note that whenever D is convex in the first argument, we have $\overline{\text{dec}}_\gamma^D(\mathcal{M}) \leq \sup_{\widehat{M} \in \text{co}(\mathcal{M})} \text{dec}_\gamma^D(\mathcal{M}, \widehat{M}) = \text{dec}_\gamma^D(\mathcal{M})$ (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 [\(8.25\)](#) 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 8.6 (Randomized estimators do not improve the DEC): Let D be any bounded divergence with the property that for all models $M, M', \widehat{M}$ and $\pi \in \Pi$,

$$D^\pi(M \parallel M') \leq C \left(D^\pi(\widehat{M} \parallel M) + D^\pi(\widehat{M} \parallel M') \right). \quad (8.29)$$

Then for all $\gamma > 0$,

$$\sup_{\widehat{M}} \text{dec}_\gamma^D(\mathcal{M}, \widehat{M}) \leq \overline{\text{dec}}_{\gamma/(2C)}^D(\mathcal{M}). \quad (8.30)$$

Squared Hellinger distance is symmetric and satisfies Condition [\(8.29\)](#) with $C = 2$. Hence, writing $\overline{\text{dec}}_\gamma^H(\mathcal{M})$ as shorthand for $\overline{\text{dec}}_\gamma^D(\mathcal{M})$ with $D = D_H^2(\cdot, \cdot)$, we obtain the following corollary.

Proposition 8.7 (Randomized DEC for Hellinger distance): Suppose that $\mathcal{R} \subseteq [0, 1]$. Then for all $\gamma > 0$,

$$\overline{\text{dec}}_\gamma^H(\mathcal{M}) \leq \sup_{\widehat{M} \in \text{co}(\mathcal{M})} \text{dec}_\gamma^H(\mathcal{M}, \widehat{M}) \leq \sup_{\widehat{M}} \text{dec}_\gamma^H(\mathcal{M}, \widehat{M}) \leq \overline{\text{dec}}_{\gamma/4}^H(\mathcal{M}).$$

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 $\overline{\text{dec}}_\gamma^H(\mathcal{M}, \nu)$ can be simpler to compute than strategies that minimize $\text{dec}_\gamma^H(\mathcal{M}, \widehat{M})$ for $\widehat{M} \in \text{co}(\mathcal{M})$.

8.6.2 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 [226] (see also Dann et al. [62], Agarwal and Zhang [3], Zhong et al. [229]), which we refer to as *optimistic estimation*. The results we present here are based on Foster et al. [90].

Let a divergence $D^\pi(\cdot \parallel \cdot)$ be fixed. An *optimistic estimation oracle* $\mathbf{Alg}_{\text{Est}}$ is an algorithm which, at each step t , given $\mathcal{H}^{t-1} = (\pi^1, r^1, o^1), \dots, (\pi^{t-1}, r^{t-1}, o^{t-1})$, 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

$$\mathbf{OptEst}_\gamma^D := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}_{\widehat{M}^t \sim \nu^t} \left[D^{\pi^t}(\widehat{M}^t \parallel M^*) + \gamma^{-1}(f^{M^*}(\pi_{M^*}) - f^{\widehat{M}^t}(\pi_{\widehat{M}^t})) \right]. \quad (8.31)$$

This quantity is similar to (8.26), but incorporates a bonus term

$$\gamma^{-1}(f^{M^*}(\pi_{M^*}) - f^{\widehat{M}^t}(\pi_{\widehat{M}^t})),$$

which encourages the estimation algorithm to *over-estimate* the optimal value $f^{M^*}(\pi_{M^*})$ for the underlying model, leading to a form of optimism. The online estimation algorithm must control this combined criterion to obtain the regret bounds below.

Example 8.3 (Structured bandits). Consider any structured bandit problem with decision space Π , function class $\mathcal{F} \subseteq (\Pi \rightarrow [0, 1])$, and $\mathcal{O} = \{\emptyset\}$. 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 $\mathbb{E}[\mathbf{OptEst}_\gamma^D] \lesssim \log|\mathcal{F}| + \sqrt{T \log|\mathcal{F}|}/\gamma$ for $D = D_{\text{Sq}}(\cdot, \cdot)$ [226]. $\triangleleft$

Optimistic Estimation-to-Decisions. We consider 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 $(\pi^1, r^1, o^1), \dots, (\pi^{t-1}, r^{t-1}, o^{t-1})$ collected so far. The algorithm then uses the estimator to compute a distribution $p^t \in \Delta(\Pi)$ and samples π^t from this distribution.

Algorithm 8.2 Optimistic E2D (E2D.Opt)

input: Exploration parameter $\gamma > 0$; divergence $D(\cdot \| \cdot)$; optimistic estimation oracle.

for $t = 1, 2, \dots, T$ **do**

Obtain randomized estimate $\nu^t \in \Delta(\mathcal{M})$ from optimistic estimation oracle with

$\{(\pi^i, r^i, o^i)\}_{i < t}$.

Compute

// Eq. (8.32).

$$p^t = \arg \min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \mathbb{E}_{\widehat{M} \sim \nu^t} \left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^M(\pi) - \gamma \cdot D^\pi(\widehat{M} \| M) \right].$$

Sample decision $\pi^t \sim p^t$ and update estimation algorithm with (π^t, r^t, o^t) .

The main change relative to the version of E2D for general divergences in Section 8.6.1 is that the minimax problem in E2D.Opt is derived from an “optimistic” variant of the DEC tailored to the optimistic estimation error in (8.31). This quantity, which we refer to as the *Optimistic Decision-Estimation Coefficient*, is defined for $\nu \in \Delta(\mathcal{M})$ as

$$\text{o-dec}_\gamma^D(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \mathbb{E}_{\widehat{M} \sim \nu} \left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f^M(\pi) - \gamma \cdot D^\pi(\widehat{M} \| M) \right]. \quad (8.32)$$

and

$$\text{o-dec}_\gamma^D(\mathcal{M}) = \sup_{\nu \in \Delta(\mathcal{M})} \text{o-dec}_\gamma^D(\mathcal{M}, \nu). \quad (8.33)$$

The Optimistic DEC is the same as the generalized DEC in (8.25), except that the optimal value $f^M(\pi_M)$ in (8.25) 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 (8.25) would otherwise be unbounded.

Remark 8.3 (Sufficient statistics): When the divergence D admits a sufficient statistic $\psi : \mathcal{M} \rightarrow \Psi$, for any distribution $\nu \in \Delta(\mathcal{M})$, if we define $\nu \in \Delta(\Psi)$ via $\nu(\psi) = \nu(\{M \in \mathcal{M} : \psi(M) = \psi\})$, we have

$$\text{o-dec}_\gamma^D(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \mathbb{E}_{\psi \sim \nu} \left[f^\psi(\pi_\psi) - f^M(\pi) - \gamma \cdot D^\pi(\psi \| M) \right].$$

In this case, by overloading notation slightly, we may simplify the definition in (8.33) to

$$\text{o-dec}_\gamma^D(\mathcal{M}) = \sup_{\nu \in \Delta(\Psi)} \text{o-dec}_\gamma^D(\mathcal{M}, \nu).$$

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 8.8 (Regret of optimistic E2D): E2D.Opt ensures that

$$\mathbf{Reg} \leq \mathbf{o-dec}_\gamma^D(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{OptEst}_\gamma^D \quad (8.34)$$

for every realization.

This regret bound has the same structure as that of [Proposition 8.5](#), but the DEC and estimation error are replaced by their optimistic counterparts.

When does optimistic estimation help? When does the regret bound in [Proposition 8.8](#) improve upon its non-optimistic counterpart in [Proposition 8.5](#)? It turns out that for *asymmetric* divergences such as those found in the context of reinforcement learning, the regret bound in (8.34) can be much smaller than the corresponding bound in (8.27); see [Section 9.5](#) for an example. However, for symmetric divergences such as Hellinger distance, we now show that the result never improves upon [Proposition 8.5](#).

Given a divergence D , we define the *flipped divergence*, which swaps the first and second arguments, by

$$\check{D}^\pi(\widehat{M} \parallel M) := D^\pi(M \parallel \widehat{M}).$$

Proposition 8.9 (Equivalence of optimistic DEC and randomized DEC): Assume that For all pairs of models $M, \widehat{M} \in \text{co}(\mathcal{M})$, we have $(f^{\widehat{M}}(\pi) - f^M(\pi))^2 \leq L_{\text{lip}}^2 \cdot D^\pi(\widehat{M} \parallel M)$ for a constant $L_{\text{lip}} > 0$. Then for all $\gamma > 0$,

$$\overline{\text{dec}}_{3\gamma/2}^{\check{D}}(\mathcal{M}) - \frac{L_{\text{lip}}^2}{2\gamma} \leq \mathbf{o-dec}_\gamma^D(\mathcal{M}) \leq \overline{\text{dec}}_{\gamma/2}^{\check{D}}(\mathcal{M}) + \frac{L_{\text{lip}}^2}{2\gamma}. \quad (8.35)$$

This result shows that the optimistic DEC with divergence D is equivalent to the generalized DEC in (8.25), but with the arguments to the divergence flipped. Thus, for symmetric divergences, the quantities are equivalent. In particular, we can combine [Proposition 8.9](#) with [Proposition 8.6](#) to derive the following corollary for Hellinger distance.

Proposition 8.10 (Optimistic DEC for Hellinger distance): Suppose that rewards are bounded in $[0, 1]$. Then for all $\gamma > 0$,

$$\mathbf{o-dec}_{2\gamma}^H(\mathcal{M}) - \frac{1}{\gamma} \leq \sup_{\overline{M}} \text{dec}_\gamma^H(\mathcal{M}, \overline{M}) \leq \mathbf{o-dec}_{\gamma/6}^H(\mathcal{M}) + \frac{3}{\gamma}.$$

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 8.9](#) shows that to match the guarantee in [Proposition 8.8](#), optimism is not required, and it suffices to run the non-optimistic algorithm of [Section 8.6.1](#). However, as discussed in [Section 9.5](#), 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 and E2D.Opt leads to meaningful guarantees.

8.7 Bibliographic Remarks and Further Reading

Takeaways. This section closes the loop on the DEC: [Theorem 8.1](#) shows that the constrained DEC is not only sufficient for low regret, as E2D established in [Proposition 7.2](#), but necessary, so that whatever separates E2D from the minimax regret is a cost of estimation. Every lower bound in the section rests on one mechanism: a change of measure between two models that are statistically close (with closeness aggregated over the interaction by the chain rule for KL or Hellinger divergence) yet whose near-optimal decisions are far apart. Here, the potentially surprising takeaway is that the DEC—dual to its role in the Estimation-to-Decisions upper bound—can encompass this lower bound mechanism into a single optimization problem.

Bibliographic remarks. The two-point method is classical [\[141\]](#); see Tsybakov [\[200, Chapter 2\]](#) or Donoho and Liu [\[67\]](#). The $\Omega(\sqrt{AT})$ bandit lower bound in [Proposition 8.1](#) is due to Auer et al. [\[23\]](#); the presentation through [Lemma 8.1](#) and [Corollary 8.1](#) follows Lattimore and Szepesvári [\[139, Chapter 15\]](#), and [Lemma 8.3](#) is from Jayram [\[109\]](#), Foster et al. [\[92\]](#).

Lower bounds in terms of the DEC were introduced by Foster et al. [\[84\]](#); the constrained form and [Theorem 8.1](#) are due to Foster et al. [\[89\]](#), simplified for this text. The general-loss lower bound of [Proposition 8.2](#), with arbitrary reference models, is from Foster et al. [\[88\]](#), where it arises from decision making with hidden rewards. The hard family of [Proposition 8.3](#) and the examples in [Section 8.5](#) follow Foster et al. [\[84, 89\]](#), recovering classical rates: Dani et al. [\[61\]](#) for [Example 8.1](#), Kleinberg et al. [\[127\]](#) for [Example 8.2](#), and Jaksch et al. [\[108\]](#), Osband and Van Roy [\[159\]](#), Domingues et al. [\[66\]](#) for [Proposition 8.4](#). In the starred material, [Proposition 8.5](#), [Proposition 8.6](#), [Proposition 8.7](#), [Proposition 8.9](#), and [Proposition 8.8](#) are from Foster et al. [\[90\]](#); optimistic estimation itself is due to Zhang [\[226\]](#).

Further reading. The estimation term separating the upper and lower bounds cannot be removed in general [\[89\]](#); Chen et al. [\[52\]](#), building on Foster et al. [\[87\]](#), close this gap up to polynomial factors for the special case of convex classes, and highlight barriers to further improvement. For more background on the modulus of continuity in statistical estimation, which can be viewed as a non-interactive predecessor to the Decision-Estimation Coefficient, see Donoho and Liu [\[67, 68\]](#). Chen and Rakhlin [\[50\]](#) extend the upper and lower bounds of this section to environments that change over time, within a hybrid stochastic/adversarial framework that covers local differential privacy, statistical-query learning, and robust decision making.

9. REINFORCEMENT LEARNING: FUNCTION APPROXIMATION AND LARGE STATE SPACES

This section returns to reinforcement learning, now without the assumption that the state space is small. The protocol is the same as [Section 6](#), but the goal is regret that does not depend on the number of states. The main tool we use is function approximation: a class of models for the transition dynamics, or a class of value functions, that lets the learner generalize across states as it explores. Compared with the contextual and structured bandit problems, where function approximation already appeared, reinforcement learning adds credit assignment on top of generalization, and the combination is what makes this problem challenging. There are two questions this section is organized around: what structure, beyond realizability, makes sample-efficient learning possible, and which algorithms exploit it. In the language of [Figure 2](#), this section covers the far corner of the figure, covering the frontier of theoretical understanding.

Organization. [Section 9.1](#) recalls the setting and shows that realizability alone, of the model, the value function, or the policy, is not enough for sample-efficient learning. [Section 9.2](#) studies the first structural assumption, low-rank transitions, and the LSVI-UCB algorithm. [Section 9.3](#) introduces the more general notion of Bellman rank and the BiLinUCB algorithm, and [Section 9.4](#) works through MDP families that have low Bellman rank. Finally, [Section 9.5](#) instantiates the Estimation-to-Decisions principle by bounding the Decision-Estimation Coefficient for classes with low Bellman rank, and [Section 9.6](#) closes with a survey of the landscape of tractability and open directions.

9.1 Setting and Realizability

Recall ([Section 6](#)) that for reinforcement learning, an MDP M takes the form

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\},$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P_h^M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is the probability transition kernel at step h , $R_h^M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathbb{R})$ is the reward distribution, and $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution. All of the results in this section take $\Pi = \Pi_{\text{rs}}$, and we assume that rewards are nonnegative with $\sum_{h=1}^H r_h \in [0, 1]$ almost surely, unless otherwise specified. As in [Section 6](#), $Q_h^* := Q_h^{M^*,*}$ and $V_h^* := V_h^{M^*,*}$ denote the optimal value functions of the true MDP M^* . Note that the online reinforcement learning framework is a special case of the general decision making setting in [Section 7](#), but the algorithms we develop in this section are tailored to the MDP structure.

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^* .
- ▷ *Value function realizability*: We have a class $\mathcal{Q}$ of state-action value functions (Q -functions) that contains the optimal function Q^* for the underlying MDP.
- ▷ *Policy realizability*: We have a class Π of policies that contains an optimal policy π^* .

To relate these assumptions, define the *induced classes*

$$\mathcal{Q}_{\mathcal{M}} := \{Q^{M,\star} : M \in \mathcal{M}\}, \quad \Pi_{\mathcal{Q}} := \{\pi_Q : Q \in \mathcal{Q}\}, \quad (9.1)$$

where π_Q is the greedy policy for Q , with a fixed rule for breaking ties. We see that model realizability for $\mathcal{M}$ implies value function realizability for $\mathcal{Q}_{\mathcal{M}}$. Likewise, value function realizability for $\mathcal{Q}$ implies policy realizability for $\Pi_{\mathcal{Q}}$, since a greedy policy for $Q^\star$ is optimal by [Proposition 6.1](#). In general however, the converse implications do not hold.

Ideally, we would hope that whenever one of (model, value function, policy) realizability holds, we can obtain regret bounds that scale with the complexity of the corresponding 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 $|S|$ or other properties of the underlying MDP, analogous to the situation for statistical learning. Unfortunately, the following result, due to Krishnamurthy et al. [\[130\]](#), shows that this is too much to ask for.

Proposition 9.1 (Realizability is insufficient): For any $S \in \mathbb{N}$ and $H \in \mathbb{N}$, there exists a class of horizon- H MDPs $\mathcal{M}$ with $|S| = S$, $|\mathcal{A}| = 2$, and $\log|\mathcal{M}| = \log(S)$, yet any algorithm must have

$$\mathbb{E}[\text{Reg}] \gtrsim \sqrt{\min\{S, 2^H\} \cdot T \wedge T}.$$

The interpretation of this result is that even if model realizability holds, any algorithm needs regret that scales with $\min\{|S|, |\mathcal{M}|, 2^H\}$. This means additional structural assumptions on the underlying MDP $M^\star$ —beyond realizability—are required if we want to obtain sample-efficient learning guarantees. The same construction also rules out such guarantees under value function or policy realizability: the induced classes in [\(9.1\)](#) satisfy

$$|\Pi_{\mathcal{Q}_{\mathcal{M}}}| \leq |\mathcal{Q}_{\mathcal{M}}| \leq |\mathcal{M}|,$$

so passing to these classes does not increase the class-size complexity.

Remark 9.1 (Comparison to structured bandits): [Proposition 9.1](#) is analogous to the impossibility result we proved for structured bandits ([Example 5.1](#)), which is subsumed by the RL framework. That result required a large number of actions, while [Proposition 9.1](#) holds even when $|\mathcal{A}| = 2$.

Remark 9.2 (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 [\[222, 116\]](#), or assume that the class $\mathcal{Q}$ obeys certain notions of consistency with respect to the Bellman operator for $M^\star$ (*Bellman completeness*; see Munos and Szepesvári [\[155\]](#), Chen and Jiang [\[55\]](#)).

The remainder of this section explores different structural assumptions that facilitate low regret for reinforcement learning with function approximation. Briefly, the idea is 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 begin with reinforcement learning with linear models ([Section 9.2](#)), then explore a general structural property known as Bellman rank ([Section 9.3](#)).

9.2 Low-Rank MDPs and LSVI-UCB

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 corresponding to the true model M^* , generalizing the linear bandit setting in [Section 5](#):

$$Q_h^*(s, a) = \langle \phi(s, a), \theta_h^{M^*} \rangle, \quad \forall h \in [H] \quad (9.2)$$

where $\phi(s, a) \in \mathbb{B}_2^d(1)$ is a feature map that is known to the learner and $\theta_h^{M^*} \in \mathbb{B}_2^d(1)$ is an unknown parameter vector. Equivalently, we can define

$$\mathcal{Q} = \left\{ Q = (Q_1, \dots, Q_H) \text{ with } Q_h(s, a) = \langle \phi(s, a), \theta_h \rangle \mid \theta_h \in \mathbb{B}_2^d(1) \forall h \right\}, \quad (9.3)$$

and assume that $Q^* \in \mathcal{Q}$. This is called the *Linear- Q^** 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^* 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; the following result is due to Weisz et al. [\[210\]](#) (see also Wang et al. [\[209\]](#)).

Proposition 9.2 (Linear Q^* is insufficient): For any $d \in \mathbb{N}$ and $H \in \mathbb{N}$ sufficiently large, any algorithm for the Linear- Q^* model must have

$$\mathbb{E}[\mathbf{Reg}] \gtrsim \min\left\{2^{\Omega(d)}, 2^{\Omega(H)}\right\} \wedge T.$$

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^* 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 begin with.

We mention in passing that [Proposition 9.2](#) can be proven by lower bounding the Decision-Estimation Coefficient [\[84\]](#).

The Low-Rank MDP model. We therefore 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

$$P_h^M(s' \mid s, a) = \langle \phi(s, a), \mu_h^M(s') \rangle, \quad \text{and} \quad \mathbb{E}[r_h \mid s, a] = \langle \phi(s, a), w_h^M \rangle. \quad (9.4)$$

Here, $\phi(s, a) \in \mathbb{B}_2^d(1)$ 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 $w_h^M \in \mathbb{B}_2^d(\sqrt{d})$ is an unknown parameter vector. Additionally, for simplicity, we assume that

$$\left\| \sum_{s' \in \mathcal{S}} |\mu_h^M(s')| \right\| \leq \sqrt{d} \quad (9.5)$$

which in particular holds in the tabular example below. 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 (9.4) implies that the transition matrix has rank at most d , thus facilitating information sharing and generalization across states, even when the cardinality of $\mathcal{S}$ and $\mathcal{A}$ is large or infinite. For this reason, we refer to MDPs with this structure as *low-rank MDPs* [175, 220, 115, 8].

Just as linear bandits generalize unstructured multi-armed bandits, the low-rank MDP model (9.4) generalizes tabular RL, which corresponds to the special case in which $d = |\mathcal{S}| \cdot |\mathcal{A}|$, $\phi(s, a) = \mathbf{e}_{s,a}$, and $(\mu_h(s'))_{s,a} = P_h^M(s' | s, a)$.

Properties of low-rank MDPs. The linear structure of the transition probabilities and mean rewards is a significantly more stringent assumption than linearity of $Q_h^*(s, a)$ in (9.2). Notably, it implies that Bellman backups of *arbitrary functions* are linear.

Lemma 9.1 (Bellman backups for low-rank MDPs): For any low-rank MDP $M \in \mathcal{M}$ and any $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and any $h \in [H]$, the Bellman operator is linear in ϕ :

$$[\mathcal{T}_h^M Q](s, a) = \langle \phi(s, a), \theta_Q^M \rangle$$

for some $\theta_Q^M \in \mathbb{R}^d$. Furthermore, for $Q : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, under the scaling assumption (9.5), it holds that $\|\theta_Q^M\| \leq 2\sqrt{d}$. Finally, for any policy $\pi = (\pi_1, \dots, \pi_H)$, functions $Q_h^{M, \pi}$ are linear in ϕ for every h .

Proof of Lemma 9.1. We have

$$[\mathcal{T}_h^M Q](s, a) = \langle \phi(s, a), w_h^M \rangle + \sum_{s'} P_h^M(s' | s, a) \max_{a'} Q(s', a') \quad (9.6)$$

$$= \langle \phi(s, a), w_h^M \rangle + \sum_{s'} \langle \phi(s, a), \mu_h^M(s') \rangle \max_{a'} Q(s', a') \quad (9.7)$$

$$= \left\langle \phi(s, a), w_h^M + \sum_{s'} \mu_h^M(s') \max_{a'} Q(s', a') \right\rangle. \quad (9.8)$$

and, under the aforementioned assumptions,

$$\|\theta_Q^M\| \leq \|w_h^M\| + \left\| \sum_{s'} \mu_h^M(s') \max_{a'} Q(s', a') \right\| \leq 2\sqrt{d}. \quad (9.9)$$

□

9.2.1 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 an influential paper of Jin et al. [115]. 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

$$\bar{Q}_h^t(s, a) \geq Q_h^*(s, a)$$

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.

Algorithm 9.1 LSVI-UCB

input: Radius $R > 0$.

for $t = 1, \dots, T$ **do**

Let $\bar{Q}_{H+1}^t \equiv 0$.

for $h = H, \dots, 1$ **do**

Compute least-squares estimator

$$\hat{\theta}_h^t = \arg \min_{\theta \in \mathcal{B}_2^d(2\sqrt{d})} \sum_{i < t} \left(\langle \phi(s_h^i, a_h^i), \theta \rangle - r_h^i - \max_a \bar{Q}_{h+1}^t(s_{h+1}^i, a) \right)^2,$$

and let $\hat{Q}_h^t(s, a) := \langle \phi(s, a), \hat{\theta}_h^t \rangle$.

Define

$$\bar{\Sigma}_h^t = \sum_{i < t} \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top + I.$$

Compute bonus:

$$b_{h,\delta}^t(s, a) = \sqrt{R} \|\phi(s, a)\|_{(\bar{\Sigma}_h^t)^{-1}}.$$

Compute optimistic value function:

$$\bar{Q}_h^t(s, a) = \max \left\{ 0, \min \left\{ 1, \hat{Q}_h^t(s, a) + b_{h,\delta}^t(s, a) \right\} \right\}.$$

Set $\bar{V}_h^t(s) = \max_{a \in \mathcal{A}} \bar{Q}_h^t(s, a)$ and $\hat{\pi}_h^t(s) = \arg \max_{a \in \mathcal{A}} \bar{Q}_h^t(s, a)$.

Collect trajectory $(s_1^t, a_1^t, r_1^t), \dots, (s_H^t, a_H^t, r_H^t)$ according to $\hat{\pi}^t$.

Two features of the algorithm deserve comment. First, the regression target at layer h is $r_h + \max_a \bar{Q}_{h+1}^t(s_{h+1}, a)$, the empirical Bellman backup of the optimistic value function computed one layer later; [Lemma 9.1](#) shows that this can always be approximated by a linear function. Second, the bonus $b_{h,\delta}^t$ is the elliptic bonus from LinUCB ([Section 4.3](#)), but now applied layer by layer.

The LSVI-UCB algorithm enjoys the following regret bound.

Theorem 9.1 (Regret of LSVI-UCB): Under the scaling assumption [\(9.5\)](#), let $\delta \in (0, 1)$ and set $R = c \cdot d^2 \log(2dHT/\delta)$ for a sufficiently large numerical constant c . Then, with probability at least $1 - \delta$, LSVI-UCB satisfies

$$\mathbf{Reg} \lesssim H \sqrt{d^3 T} \log(2dHT/\delta). \quad (9.10)$$

The bound is polynomial in H , d , and $\log(1/\delta)$, and independent of the number of states. The reason the horizon enters only polynomially, in contrast with [Proposition 9.2](#), is that low-rank structure controls how regression errors propagate: the error at layer h is a linear function of ϕ , so

it is confined to the d directions the data have not yet covered. The dependence on d and H in (9.10) is not optimal, but can be sharpened with similar techniques [11].

9.2.2 Proof of Theorem 9.1

The starting point of our analysis for UCB-VI was Lemma 6.5, which states that it is sufficient to construct optimistic estimates $\{\bar{Q}_1, \dots, \bar{Q}_H\}$ (i.e. $Q_h^* \leq \bar{Q}_h$) such that the Bellman residuals $\mathbb{E}^{M^*, \hat{\pi}}[(\bar{Q}_h - \mathcal{T}_h^{M^*} \bar{Q}_{h+1})(s_h, a_h)]$ are small under the greedy (with respect to $\bar{Q}$'s) policy $\hat{\pi}$. 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 $\bar{Q}_h$ to be the empirical Bellman backup $\mathcal{T}_h^{\widehat{M}} \bar{Q}_{h+1}$, 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. Subtleties arise in the analysis of this regression step due to lack of independence.

Technical lemmas for regression. Recall from the definition of the Bellman operator (Section 6.2) that for any fixed $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$,

$$\mathbb{E}^{M^*} \left[r_h^i + \max_a Q(s_{h+1}^i, a) \mid s_h^i, a_h^i \right] = [\mathcal{T}_h^{M^*} Q](s_h^i, a_h^i). \quad (9.11)$$

However, for layer h , the regression problem within LSVI-UCB concerns a *data-dependent* function $Q = \bar{Q}_{h+1}^t$ (with $i < t$), which is chosen as a function of all the trajectories $\tau^1, \dots, \tau^{t-1}$. This dependence implies that the regression problem solved by LSVI-UCB is not of the type studied, say, in Proposition 2.1. Instead, in the language of Section 2.1, 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 allows us to appeal to uniform convergence arguments. In particular, for every h and t , $\bar{Q}_h^t$ belongs to the class

$$\bar{\mathcal{Q}} := \left\{ (s, a) \mapsto \max \left\{ 0, \min \left\{ 1, \langle \theta, \phi(s, a) \rangle + \sqrt{R} \|\phi(s, a)\|_{(\bar{\Sigma})^{-1}} \right\} \right\} : \begin{array}{l} \|\theta\| \leq 2\sqrt{d}, \quad \sigma_{\min}(\bar{\Sigma}) \geq 1 \end{array} \right\}. \quad (9.12)$$

To make use of this fact, we first state a result on least squares with adaptively collected covariates and data-dependent targets. It comes in two parts: a finite-class version, which we prove in full, and the version for the classes at hand, whose proof is a covering argument that we only sketch.

Lemma 9.2 (Least squares with data-dependent targets): Let $(\mathcal{F}_i)_{i \geq 0}$ be a filtration, and let $x_1, \dots, x_T \in \mathcal{X}$ be such that x_i is $\mathcal{F}_{i-1}$ -measurable. Let $\mathcal{G}$ be an index set, and for each $g \in \mathcal{G}$ let $y_1(g), \dots, y_T(g) \in \mathbb{R}$ be outcomes such that $y_i(g)$ is $\mathcal{F}_i$ -measurable and, for a function $f_g \in \mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R}\}$ and a constant $R \geq 1$,

$$\mathbb{E}[y_i(g) \mid \mathcal{F}_{i-1}] = f_g(x_i) \quad \text{and} \quad |y_i(g) - f_g(x_i)| \leq R \quad \text{almost surely.} \quad (9.13)$$

The random variables $\{y_i(g)\}_{g \in \mathcal{G}}$ may be correlated. For each $g \in \mathcal{G}$, define the least-squares solution

$$\hat{f}_g \in \arg \min_{f \in \mathcal{F}} \sum_{i=1}^T (y_i(g) - f(x_i))^2.$$

1. If $\mathcal{F}$ and $\mathcal{G}$ are finite, then with probability at least $1 - \delta$, simultaneously for all $g \in \mathcal{G}$,

$$\sum_{i=1}^T (\hat{f}_g(x_i) - f_g(x_i))^2 \leq 8R^2 \log(|\mathcal{F}||\mathcal{G}|/\delta). \quad (9.14)$$

2. Suppose instead that $\mathcal{X} = \mathbb{B}_2^d(1)$, that $\mathcal{F} = \{x \mapsto \langle \theta, x \rangle : \|\theta\| \leq 2\sqrt{d}\}$, and that $\mathcal{G} = \bar{\mathcal{Q}}$ is the class in (9.12), with the maps $g \mapsto y_i(g)$ and $g \mapsto f_g$ both 1-Lipschitz with respect to the sup-norm on $\bar{\mathcal{Q}}$. Then with probability at least $1 - \delta$, simultaneously for all $g \in \mathcal{G}$,

$$\sum_{i=1}^T (\hat{f}_g(x_i) - f_g(x_i))^2 \lesssim R^2 d^2 \log(dT/\delta). \quad (9.15)$$

Proof of Lemma 9.2. We prove Part 1 in full. Fix $g \in \mathcal{G}$ and $f \in \mathcal{F}$, and write $\epsilon_i := y_i(g) - f_g(x_i)$ for the noise and $v_i := f(x_i) - f_g(x_i)$ for the difference of the two functions at x_i . By (9.13), ϵ_i is conditionally mean zero and bounded by R given $\mathcal{F}_{i-1}$, while v_i is $\mathcal{F}_{i-1}$ -measurable. Hoeffding's lemma ((2.27)) therefore gives, for every $\lambda > 0$,

$$\log \mathbb{E}_{i-1} \left[e^{\lambda \epsilon_i v_i} \right] \leq \frac{\lambda^2 R^2 v_i^2}{2}.$$

Applying Lemma A.2 to the sequence $\lambda \epsilon_i v_i$, we conclude that with probability at least $1 - \delta'$,

$$\sum_{i=1}^T \epsilon_i v_i \leq \frac{\lambda R^2}{2} \sum_{i=1}^T v_i^2 + \frac{\log(1/\delta')}{\lambda}. \quad (9.16)$$

Taking $\delta' = \delta/(|\mathcal{F}||\mathcal{G}|)$ and a union bound, (9.16) holds simultaneously for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$ with probability at least $1 - \delta$. On this event, fix g and apply (9.16) with $f = \hat{f}_g$, which is admissible because the bound holds for every $f \in \mathcal{F}$ at once. Optimality of $\hat{f}_g$ gives

$$\sum_{i=1}^T (y_i(g) - \hat{f}_g(x_i))^2 \leq \sum_{i=1}^T (y_i(g) - f_g(x_i))^2,$$

and expanding the squares with $y_i(g) = f_g(x_i) + \epsilon_i$ yields the basic inequality

$$\sum_{i=1}^T (\hat{f}_g(x_i) - f_g(x_i))^2 \leq 2 \sum_{i=1}^T \epsilon_i (\hat{f}_g(x_i) - f_g(x_i)).$$

Combining the two displays with $\lambda = 1/(2R^2)$,

$$\sum_{i=1}^T (\hat{f}_g(x_i) - f_g(x_i))^2 \leq \frac{1}{2} \sum_{i=1}^T (\hat{f}_g(x_i) - f_g(x_i))^2 + 4R^2 \log(|\mathcal{F}||\mathcal{G}|/\delta),$$

and rearranging proves (9.14).

For Part 2 we give only a sketch. Let $\mathcal{F}'$ be a cover of $\mathcal{F}$ at scale $1/T$ in the sup-norm on $\mathcal{B}_2^d(1)$ and $\mathcal{G}'$ a cover of $\bar{\mathcal{Q}}$ at scale $1/T$ in the sup-norm; standard volumetric arguments give $\log|\mathcal{F}'| \lesssim d \log(dT)$, since $\mathcal{F}$ is parameterized by $\theta \in \mathbb{R}^d$, and $\log|\mathcal{G}'| \lesssim d^2 \log(dT)$, since $\bar{\mathcal{Q}}$ is parameterized by $\theta \in \mathbb{R}^d$ and a matrix $\bar{\Sigma} \in \mathbb{R}^{d \times d}$ with $\sigma_{\min}(\bar{\Sigma}) \geq 1$. Clipping to $[0, 1]$ does not increase the sup-norm distance between functions, so the same covering bound applies to the clipped class. Part 1 applied to the covers gives (9.14) with $\log(|\mathcal{F}'||\mathcal{G}'|/\delta) \lesssim d^2 \log(dT/\delta)$. Passing from the covers back to $\mathcal{F}$ and $\bar{\mathcal{Q}}$ costs an additive constant: by the Lipschitz assumptions, replacing g by its nearest cover point changes every outcome $y_i(g)$ and every value $f_g(x_i)$ by at most $1/T$, and replacing $\hat{f}_g$ by its nearest cover point changes each value by at most $1/T$, so that all sums of squares over T terms change by $O(1)$. We omit the details. $\square$

We may now apply Lemma 9.2 to analyze the regression step of LSVI-UCB.

Lemma 9.3 (Confidence bound for LSVI-UCB regression): With probability at least $1 - \delta$, we have that for all t and h ,

$$\sum_{i < t} \left(\hat{Q}_h^t(s_h^i, a_h^i) - [\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s_h^i, a_h^i) \right)^2 \lesssim d^2 \log(2dHT/\delta). \quad (9.17)$$

Proof of Lemma 9.3. Fix $t \in [T]$ and $h \in [H]$. We apply Part 2 of Lemma 9.2 with T replaced by $t - 1$, then take a union bound over t and h .

Let $\mathcal{F}_i$ be the sigma-field generated by the first i trajectories together with (s_h^{i+1}, a_h^{i+1}) , and for $Q \in \bar{\mathcal{Q}}$ set

$$x_i = \phi(s_h^i, a_h^i), \quad y_i(Q) = r_h^i + \max_a Q(s_{h+1}^i, a).$$

Then x_i is $\mathcal{F}_{i-1}$ -measurable and $y_i(Q)$ is $\mathcal{F}_i$ -measurable, and by the Markov property,

$$\mathbb{E}[y_i(Q) \mid \mathcal{F}_{i-1}] = [\mathcal{T}_h^{M^*} Q](s_h^i, a_h^i) = \langle x_i, \theta_Q^{M^*} \rangle,$$

where $\|\theta_Q^{M^*}\| \leq 2\sqrt{d}$ by Lemma 9.1, since Q takes values in $[0, 1]$. Thus $f_Q := \langle \cdot, \theta_Q^{M^*} \rangle$ lies in the linear class $\mathcal{F}$ of Part 2.

We now verify the remaining hypotheses of Lemma 9.2. Rewards and value functions lie in $[0, 1]$, so $y_i(Q)$ and $f_Q(x_i)$ lie in $[0, 2]$ and the noise is bounded by $R = 2$. Both $Q \mapsto y_i(Q)$ and $Q \mapsto \mathcal{T}_h^{M^*} Q$ are 1-Lipschitz in the sup-norm.

Conclusion. The least-squares solution in LSVI-UCB is $\hat{Q}_h^t = \hat{f}_Q$ for $Q = \bar{Q}_{h+1}^t \in \bar{\mathcal{Q}}$. This choice of Q depends on the data, but Part 2 holds simultaneously for all $Q \in \bar{\mathcal{Q}}$, so it applies. Taking a union bound over t and h gives (9.17). $\square$

Establishing optimism. The next lemma shows that closeness of the regression estimate to the Bellman backup on the data $\{(s_h^i, a_h^i)\}_{i < t}$ 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 $\bar{Q}_1^t, \dots, \bar{Q}_H^t$ are optimistic.

Lemma 9.4 (Optimism for LSVI-UCB): With R chosen as in [Theorem 9.1](#), whenever the event in [Lemma 9.3](#) occurs, we have that for all (s, a, h) and $t \in [T]$,

$$\left| \hat{Q}_h^t(s, a) - [\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s, a) \right| \leq \sqrt{R} \cdot \|\phi(s, a)\|_{(\bar{\Sigma}_h^t)^{-1}} = b_{h,\delta}^t(s, a). \quad (9.18)$$

and

$$\bar{Q}_h^t(s, a) \geq Q_h^*(s, a). \quad (9.19)$$

Proof of Lemma 9.4. Writing the Bellman backup, via [Lemma 9.1](#), as

$$[\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s, a) = \langle \phi(s, a), \theta_h^t \rangle$$

for some $\theta_h^t \in \mathbb{R}^d$ with $\|\theta_h^t\|_2 \leq 2\sqrt{d}$, we have that

$$\begin{aligned} \left| \hat{Q}_h^t(s, a) - [\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s, a) \right| &= \left| \langle \phi(s, a), \hat{\theta}_h^t - \theta_h^t \rangle \right| \\ &= \left| \langle (\bar{\Sigma}_h^t)^{-1/2} \phi(s, a), (\bar{\Sigma}_h^t)^{1/2} (\hat{\theta}_h^t - \theta_h^t) \rangle \right| \\ &\leq \|\phi(s, a)\|_{(\bar{\Sigma}_h^t)^{-1}} \cdot \|\hat{\theta}_h^t - \theta_h^t\|_{\bar{\Sigma}_h^t}. \end{aligned}$$

[Lemma 9.3](#) then implies (9.18), since

$$\|\hat{\theta}_h^t - \theta_h^t\|_{\bar{\Sigma}_h^t}^2 = (\hat{\theta}_h^t - \theta_h^t)^\top \left(\sum_{i < t} \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top + I \right) (\hat{\theta}_h^t - \theta_h^t) \quad (9.20)$$

$$= \sum_{i < t} \left(\hat{Q}_h^t(s_h^i, a_h^i) - [\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s_h^i, a_h^i) \right)^2 + \|\hat{\theta}_h^t - \theta_h^t\|^2 \quad (9.21)$$

and $\|\hat{\theta}_h^t - \theta_h^t\|^2 \lesssim d$ by (9.9).

To show (9.19), we proceed by induction on $\bar{V}_h^t \geq V_h^*$, as in the proof of [Lemma 6.6](#). We start with the base case $h = H + 1$, which has $\bar{V}_{H+1}^t = V_{H+1}^* \equiv 0$. Assuming $\bar{V}_{h+1}^t \geq V_{h+1}^*$, we first observe that $\mathcal{T}_h^{M^*}$ is monotone and $\mathcal{T}_h^{M^*} \bar{V}_{h+1}^t \geq \mathcal{T}_h^{M^*} V_{h+1}^* = Q_h^*$. Hence,

$$\hat{Q}_h^t = \hat{Q}_h^t - \mathcal{T}_h^{M^*} \bar{V}_{h+1}^t + \mathcal{T}_h^{M^*} \bar{V}_{h+1}^t \quad (9.22)$$

$$\geq \hat{Q}_h^t - \mathcal{T}_h^{M^*} \bar{V}_{h+1}^t + Q_h^* \quad (9.23)$$

$$\geq -b_{h,\delta}^t + Q_h^* \quad (9.24)$$

and thus $\hat{Q}_h^t + b_{h,\delta}^t \geq Q_h^*$. Since $Q_h^* \in [0, 1]$, clipping to $[0, 1]$ preserves this inequality, so $\bar{Q}_h^t \geq Q_h^*$. In particular, the lower clipping is inactive on this event. This, in turn, implies $\bar{V}_h^t \geq V_h^*$. $\square$

Finishing the proof. With the technical results above established, the proof of [Theorem 9.1](#) follows fairly quickly.

Proof of Theorem 9.1. Apply [Lemma 9.3](#) with failure probability $\delta/2$; the choice of R in [Theorem 9.1](#) accommodates this change by increasing c . On this event, since $\bar{Q}$ is optimistic by [Lemma 9.4](#), we have that for each timestep t ,

$$\begin{aligned} f^*(\pi^*) - f^*(\hat{\pi}^t) &\leq \mathbb{E}_{s_1 \sim d_1} [\bar{V}_1^t(s_1)] - f^*(\hat{\pi}^t) \\ &= \sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [\bar{Q}_h^t(s_h, a_h) - [\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s_h, a_h)] \end{aligned}$$

by [Lemma 6.2](#). Using the definition of $b_{h,\delta}^t$ and [Lemma 9.4](#), we have

$$\sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [\bar{Q}_h^t(s_h, a_h) - [\mathcal{T}_h^{M^*} \bar{Q}_{h+1}^t](s_h, a_h)] \lesssim \sqrt{R} \sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [\|\phi(s_h, a_h)\|_{(\bar{\Sigma}_h^t)^{-1}}].$$

Summing over all timesteps t gives

$$\mathbf{Reg} \lesssim \sqrt{R} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M^*, \hat{\pi}^t} [\|\phi(s_h, a_h)\|_{(\bar{\Sigma}_h^t)^{-1}}].$$

Each feature norm is at most one, so their sum over an episode lies in $[0, H]$. Azuma's inequality (the martingale form of [Lemma 2.7](#)) implies that, with probability at least $1 - \delta/2$, the preceding bound is at most a constant times

$$\sqrt{R} \sum_{t=1}^T \sum_{h=1}^H \|\phi(s_h^t, a_h^t)\|_{(\bar{\Sigma}_h^t)^{-1}} + H \sqrt{RT \log(2/\delta)}.$$

The elliptic potential lemma ([Lemma 4.2](#)) now allows us to bound

$$\sum_{t=1}^T \|\phi(s_h^t, a_h^t)\|_{(\bar{\Sigma}_h^t)^{-1}} \lesssim \sqrt{dT \log(1 + T/d)}$$

for each h . Substituting $R = c \cdot d^2 \log(2dHT/\delta)$ and bounding both $\log(1 + T/d)$ and $\log(2/\delta)$ by $\log(2dHT/\delta)$ gives [\(9.10\)](#). A union bound over the two events gives the claimed probability. $\square$

9.3 Bellman Rank and BiLinUCB

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 $Q^* \in \mathcal{Q}$. In [Section 9.2](#), 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 introduce a structural property, *Bellman rank* [[111](#), [70](#), [116](#)], which allows for sample-efficient learning with general classes $\mathcal{Q}$, and subsumes a number of well-studied MDP families, including:

- ▷ Low-Rank MDPs [218, 115, 8]
- ▷ Block MDPs and reactive POMDPs [69].
- ▷ MDPs with Linear Q^* and V^* [70].
- ▷ MDPs with low occupancy complexity [70].
- ▷ Linear mixture MDPs [154, 26].
- ▷ Linear dynamical systems (LQR) [63].

Section 9.4 defines and works through several of these examples.

Building intuition. Bellman rank is a property of the underlying MDP M^* 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 $Q_h(s, a) = \langle \phi(s, a), \theta_h^Q \rangle$ be an arbitrary linear value function. Observe that since M is a Low-Rank MDP, we have $[\mathcal{T}_h^M Q](s, a) = \langle \phi(s, a), \tilde{\theta}_h^{M,Q} \rangle$, 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

$$\mathbb{E}^{M,\pi} [Q_h(s_h, a_h) - r_h - \max_a Q_{h+1}(s_{h+1}, a)] = \langle \mathbb{E}^{M,\pi} [\phi(s_h, a_h)], \theta_h^Q - \tilde{\theta}_h^{M,Q} \rangle \quad (9.25)$$

$$= \langle X_h^M(\pi), W_h^M(Q) \rangle, \quad (9.26)$$

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 $W_h^M(Q) := \theta_h^Q - \tilde{\theta}_h^{M,Q} \in \mathbb{R}^d$ 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 $\mathcal{E}_h(\cdot, \cdot) \in \mathbb{R}^{\Pi \times \mathcal{Q}}$ (we write $\mathcal{E}_h^M$ when the dependence on the MDP M needs to be explicit) with

$$\mathcal{E}_h(\pi, Q) := \mathbb{E}^{M,\pi} [Q_h(s_h, a_h) - (r_h + \max_a Q_{h+1}(s_{h+1}, a))], \quad (9.27)$$

then the property (9.26) implies that $\text{rank}(\mathcal{E}_h(\cdot, \cdot)) \leq d$. Bellman rank is an abstraction of this property.²¹

Definition 9.1 (Bellman rank): For an MDP M with value function class $\mathcal{Q}$ and policy class Π , the Bellman rank is defined as

$$d_B(M) = \max_{h \in [H]} \text{rank}(\{\mathcal{E}_h(\pi, Q)\}_{\pi \in \Pi, Q \in \mathcal{Q}}). \quad (9.28)$$

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

$$\mathcal{E}_h(\pi, Q) = \langle X_h^M(\pi), W_h^M(Q) \rangle \quad (9.29)$$

for all $\pi \in \Pi$ and $Q \in \mathcal{Q}$.

The factorization in (9.29) is what controls distribution shift. Informally, it says that there are only d directions in which the state distribution of a new policy can surprise the learner, as far as the class $\mathcal{Q}$ is concerned. This property was used implicitly in the analysis of LSVI-UCB, and holds in many settings beyond the low-rank MDP model as we will soon see.

²¹Bellman rank was originally introduced in the pioneering work of Jiang et al. [111]. The definition of Bellman rank we present, which is slightly different from the original definition, is taken from the later work of Du et al. [70], Jin et al. [116], and is often referred to as *Q-type* Bellman rank.

9.3.1 The BiLinUCB Algorithm

We now present an algorithm, BiLinUCB [70], which attains low regret for MDPs with low Bellman rank under the realizability assumption that

$$Q^* \in \mathcal{Q}.$$

Like many of the algorithms we have covered, BiLinUCB is based on confidence sets and optimism, though the way we construct the confidence sets and implement optimism is new.

PAC versus regret. For technical reasons, we do not directly give a regret bound for BiLinUCB. Instead, we prove a *PAC* (“probably approximately correct”) guarantee. For PAC, the algorithm plays for T episodes, then outputs a final policy $\hat{\pi}$, and its performance is measured via

$$f^*(\pi^*) - f^*(\hat{\pi}). \quad (9.30)$$

That is, instead of considering cumulative performance as with regret, we are only concerned with final performance. For PAC, we want to ensure that $f^*(\pi^*) - f^*(\hat{\pi}) \leq \varepsilon$ 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 $O(\frac{C}{\varepsilon^2})$ episodes via online-to-batch conversion. In the other direction, if we have an algorithm that achieves PAC error ε using $O(\frac{C}{\varepsilon^2})$ 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 5), with the property that $Q^* \in \mathcal{Q}$ 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 $\{\hat{\mathcal{E}}_h^k(Q)\}_{h \in [H]}$ which approximate the Bellman residual $\mathcal{E}_h(\pi^k, Q)$ for all $Q \in \mathcal{Q}$. Then, in (9.31), 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 $\pi^1, \dots, \pi^k$. Eliminating value functions with large Bellman residual is a natural idea, because we know from the Bellman equation that Q^* has zero Bellman residual.

Algorithm 9.2 BiLinUCB

input: $\beta > 0$; iteration count $K \in \mathbb{N}$; batch size $n \in \mathbb{N}$.

output: Policy $\hat{\pi}$.

$\mathcal{Q}^1 \leftarrow \mathcal{Q}$.

for iteration $k = 1, \dots, K$ **do**

 Compute optimistic value function:

$$Q^k = \arg \max_{Q \in \mathcal{Q}^k} \mathbb{E}_{s_1 \sim d_1} [Q_1(s_1, \pi_Q(s_1))].$$

 and let $\pi^k := \pi_{Q^k}$.

for $l = 1, \dots, n$ **do**

 Execute π^k for an episode and observe trajectory $(s_1^{k,l}, a_1^{k,l}, r_1^{k,l}), \dots, (s_H^{k,l}, a_H^{k,l}, r_H^{k,l})$.

 Compute confidence set

$$\mathcal{Q}^{k+1} = \left\{ Q \in \mathcal{Q} \mid \sum_{i \leq k} (\hat{\mathcal{E}}_h^i(Q))^2 \leq \beta^2 \quad \forall h \in [H] \right\}, \quad (9.31)$$

 where

$$\hat{\mathcal{E}}_h^i(Q) := \frac{1}{n} \sum_{l=1}^n \left(Q_h(s_h^{i,l}, a_h^{i,l}) - r_h^{i,l} - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}^{i,l}, a) \right).$$

Let $\hat{k} = \arg \max_{k \in [K]} \hat{V}^k$, where $\hat{V}^k := \frac{1}{n} \sum_{l=1}^n \sum_{h=1}^H r_h^{k,l}$.

return $\hat{\pi} = \pi^{\hat{k}}$.

Main guarantee. The main result for this section is the following PAC guarantee for BiLinUCB.

Theorem 9.2 (Sample complexity of BiLinUCB): Suppose rewards are nonnegative with $\sum_{h=1}^H r_h \leq 1$ almost surely, and let $\mathcal{Q}$ be a finite class with $Q^* \in \mathcal{Q}$, $Q_h(s, a) \in [0, 1]$, and $Q_{H+1} \equiv 0$ for every $Q \in \mathcal{Q}$. Suppose M^* has Bellman rank $d_B(M^*) \leq d$, with the corresponding factors bounded as $\|X_h^{M^*}(\pi)\|_2 \leq B_X$ and $\|W_h^{M^*}(Q)\|_2 \leq B_W$ for all $h \in [H]$, $\pi \in \Pi_{\mathcal{Q}}$, and $Q \in \mathcal{Q}$, where $B_X, B_W \geq 1$.

Fix $\varepsilon, \delta \in (0, 1)$. For sufficiently large absolute constants $c_1, c_2, c_3 > 0$, set

$$K = \lceil c_1 H d \log \left(2 + \frac{B_X^2 B_W^2 n}{d} \right) \rceil, \quad \beta^2 = c_2 K \frac{\log(2|\mathcal{Q}|HK/\delta)}{n},$$

and choose $n \geq c_3 H^2 K \log(2|\mathcal{Q}|HK/\delta)/\varepsilon^2$. Then BiLinUCB returns a policy $\hat{\pi}$ satisfying $f^*(\pi^*) - f^*(\hat{\pi}) \leq \varepsilon$ with probability at least $1 - \delta$. In particular, choosing n appropriately gives this guarantee using

$$\tilde{O} \left(\frac{H^4 d^2 \log(2|\mathcal{Q}|/\delta)}{\varepsilon^2} \right)$$

episodes, where $\tilde{O}$ also suppresses logarithmic dependence on $B_X B_W$.

That is, low Bellman rank suffices to learn a near-optimal policy, with sample complexity that

depends polynomially on d , H , and ε^{-1} , and logarithmically on $|\mathcal{Q}|$. This reflects that BiLinUCB 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*, that is, under the state distribution of the policy the algorithm actually executes, similar to what we saw for UCB-VI and LSVI-UCB; as noted above, the optimism here is only on average over the initial state distribution.
- ▷ The confidence set construction (9.31) 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 $\tilde{O}(d)$ “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. The Bellman rank property is only used for analysis, and the algorithm does not explicitly compute or estimate the factorization.

For the low-rank MDP example above, take bounded linear Q -functions with $\|\theta_h^Q\|_2 \leq 2\sqrt{d}$. Then $X_h^M(\pi) = \mathbb{E}^{M,\pi}[\phi(s_h, a_h)]$ has norm at most one, while $W_h^M(Q) = \theta_h^Q - \tilde{\theta}_h^{M,Q}$ has norm at most $4\sqrt{d}$ by Lemma 9.1. Thus we can take $B_X = 1$ and $B_W = 4\sqrt{d}$, which affects only logarithmic factors in the guarantee.

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

$$\mathbf{Reg} \leq \tilde{O}\left((H^4 d^2 \log(|\mathcal{Q}|/\delta))^{1/3} \cdot T^{2/3}\right).$$

Here too, $\tilde{O}$ suppresses logarithmic dependence on $B_X B_W$. Under an additional assumption known as *Bellman completeness*, which asks that Bellman backups of functions in $\mathcal{Q}$ remain in $\mathcal{Q}$, it is possible to attain $\sqrt{T}$ with a variant of this algorithm that uses a slightly different confidence set construction [116]. Low-rank MDPs are Bellman complete by Lemma 9.1, which is why LSVI-UCB attains $\sqrt{T}$ regret there.

One may ask why optimism is used at all when the goal is only a PAC guarantee. At its core, optimism is a device for choosing informative policies: the optimistic value function is either nearly correct, in which case its greedy policy is nearly optimal, or wrong, in which case executing its policy exposes the error. For general reinforcement learning (compared to say, bandits, where crude exploration suffices) a PAC learning algorithm needs this property as much as a regret minimization algorithm does. The only difference is that a PAC learner may enjoy better quantitative guarantees, as it need not pay for exploration in every round.

9.3.2 Proof of Theorem 9.2

Recall the factorization in Theorem 9.2:

$$\mathcal{E}_h(\pi, Q) = \langle X_h^{M^*}(\pi), W_h^{M^*}(Q) \rangle \quad \forall \pi \in \Pi_{\mathcal{Q}}, Q \in \mathcal{Q}.$$

For simplicity, we give the proof for $B_X = B_W = 1$.²²

²²For general bounds, the regularization term $\sqrt{\lambda} + \beta$ in (9.40) becomes $\sqrt{\lambda}B_W + \beta$, and the potential logarithm becomes $\log(1 + B_X^2 K/(\lambda d))$. Choosing $\lambda = \beta^2/B_W^2$ gives the stated dependence on $B_X B_W$.

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 9.5 (Validity of the BiLinUCB confidence sets): For any $\delta > 0$, if we set $\beta^2 = c \cdot K \frac{\log|\mathcal{Q}| + \log(HK/\delta)}{n}$, 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

$$\sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2 \lesssim \beta^2 \quad \forall h \in [H]. \quad (9.32)$$

2. $Q^* \in \mathcal{Q}^k$.

Proof of Lemma 9.5. Using Hoeffding's inequality and a union bound (Lemma 2.7), we have that with probability at least $1 - \delta$, for all $k \in [K]$, $h \in [H]$, and $Q \in \mathcal{Q}$,

$$\left| \hat{\mathcal{E}}_h^k(Q) - \mathcal{E}_h(\pi^k, Q) \right| \leq C \cdot \sqrt{\frac{\log(|\mathcal{Q}|HK/\delta)}{n}}, \quad (9.33)$$

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 $Q \in \mathcal{Q}$,

$$\sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2 \leq 2 \sum_{i < k} (\hat{\mathcal{E}}_h^i(Q))^2 + 2 \sum_{i < k} (\mathcal{E}_h(\pi^i, Q) - \hat{\mathcal{E}}_h^i(Q))^2.$$

For $Q \in \mathcal{Q}^k$, the definition of $\mathcal{Q}^k$ implies that $\sum_{i < k} (\hat{\mathcal{E}}_h^i(Q))^2 \leq \beta^2$, while (9.33) gives

$$\sum_{i < k} (\mathcal{E}_h(\pi^i, Q) - \hat{\mathcal{E}}_h^i(Q))^2 \leq C^2 K \frac{\log(|\mathcal{Q}|HK/\delta)}{n} \leq \beta^2,$$

provided the constant c in the definition of β^2 is at least C^2 . Combining these bounds proves Part 1.

For Part 2, Bellman optimality gives $\mathcal{E}_h(\pi, Q^*) = 0$ for every π . Thus, for all k and h ,

$$\begin{aligned} \sum_{i < k} (\hat{\mathcal{E}}_h^i(Q^*))^2 &= \sum_{i < k} (\hat{\mathcal{E}}_h^i(Q^*) - \mathcal{E}_h(\pi^i, Q^*))^2 \\ &\leq C^2(k-1) \frac{\log(|\mathcal{Q}|HK/\delta)}{n} \\ &\leq C^2 K \frac{\log(|\mathcal{Q}|HK/\delta)}{n} \leq \beta^2, \end{aligned}$$

where we used (9.33) and the same choice of c . Consequently, $Q^* \in \mathcal{Q}^k$ for all k , as required. $\square$

The next result shows that whenever the event in the previous lemma holds, the value functions constructed by BiLinUCB are optimistic.

Lemma 9.6 (Consequences of the confidence event): Whenever the event in [Lemma 9.5](#) occurs, the following properties hold:

1. Define

$$\Sigma_h^k = \sum_{i < k} X_h^{M^*}(\pi^i) X_h^{M^*}(\pi^i)^\top. \quad (9.34)$$

For all $k \in [K]$, all $Q \in \mathcal{Q}^k$ satisfy

$$\|W_h^{M^*}(Q)\|_{\Sigma_h^k}^2 \lesssim \beta^2. \quad (9.35)$$

2. For all k , Q^k is optimistic in the sense that

$$\mathbb{E}_{s_1 \sim d_1}[Q_1^k(s_1, \pi_Q(s_1))] \geq \mathbb{E}_{s_1 \sim d_1}[Q_1^*(s_1, \pi^*(s_1))] = f^*(\pi^*). \quad (9.36)$$

Proof of Lemma 9.6. For Part 1, recall that by the Bilinear class property, we can write $\mathcal{E}_h(\pi^k, Q) = \langle X^{M^*}(\pi^k), W^{M^*}(Q) \rangle$, so that [\(9.32\)](#) implies that

$$\|W^{M^*}(Q)\|_{\Sigma_h^k}^2 = \sum_{i < k} \langle X^{M^*}(\pi^i), W^{M^*}(Q) \rangle^2 = \sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2 \lesssim \beta^2.$$

For Part 2, we observe that for all k , since $Q^* \in \mathcal{Q}^k$, we have

$$\begin{aligned} \mathbb{E}_{s_1 \sim d_1}[Q_1^k(s_1, \pi_Q(s_1))] &= \sup_{Q \in \mathcal{Q}^k} \mathbb{E}_{s_1 \sim d_1}[Q_1(s_1, \pi_Q(s_1))] \\ &\geq \mathbb{E}_{s_1 \sim d_1}[Q_1^*(s_1, \pi^*(s_1))] \\ &= f^*(\pi^*). \end{aligned}$$

□

Proving the main result. Equipped with the lemmas above, we prove [Theorem 9.2](#).

Proof of Theorem 9.2. 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 9.5](#), applied with failure probability $\delta/2$. Whenever this event occurs, [Lemma 9.6](#) implies that Q^k is optimistic, so we can bound

$$f^*(\pi^*) - f^*(\pi^k) \leq \mathbb{E}_{s_1 \sim d_1}[Q_1^k(s_1, \pi_{Q^k}(s_1))] - f^*(\pi^k) \quad (9.37)$$

$$= \sum_{h=1}^H \mathbb{E}^{M^*, \pi^k} \left[Q_h^k(s_h, a_h) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}, a) \right] \quad (9.38)$$

$$= \sum_{h=1}^H \langle X_h^{M^*}(\pi^k), W_h^{M^*}(Q^k) \rangle, \quad (9.39)$$

where the first equality uses the Bellman residual decomposition ([Lemma 6.2](#)), and the second equality uses the Bellman rank assumption. For any $\lambda \geq 0$, using Cauchy-Schwarz, we can bound

$$\sum_{h=1}^H \langle X_h^{M^*}(\pi^k), W_h^{M^*}(Q^k) \rangle \leq \sum_{h=1}^H \|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}} \|W_h^{M^*}(Q^k)\|_{\lambda I + \Sigma_h^k}$$

For each $h \in [H]$, applying the bound in (9.35) gives

$$\|W_h^{M^*}(Q^k)\|_{\lambda I + \Sigma_h^k} \leq \sqrt{\lambda \|W_h^{M^*}(Q^k)\|_2^2 + \beta^2} \leq \lambda^{1/2} + \beta,$$

where we have used that $\|W_h^{M^*}(Q^k)\|_2 \leq 1$ by assumption. This allows us to bound

$$\sum_{h=1}^H \langle X_h^{M^*}(\pi^k), W_h^{M^*}(Q^k) \rangle \lesssim (\lambda^{1/2} + \beta) \cdot \sum_{h=1}^H \|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}. \quad (9.40)$$

If we can find a policy π^k for which the right-hand side of (9.40) is small, this policy will be guaranteed to have low regret. The following lemma shows that such a policy is guaranteed to exist.

Lemma 9.7 (Existence of a well-covered iteration): For any $\lambda > 0$, as long as $K \geq Hd \log(1 + \lambda^{-1}K/d)$, there exists $k \in [K]$ such that

$$\|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}^2 \lesssim \frac{Hd \log(1 + \lambda^{-1}K/d)}{K} \quad \forall h \in [H]. \quad (9.41)$$

We choose $\lambda = \beta^2$, which implies that it suffices to take $K \gtrsim Hd \log(1 + n/d)$ to satisfy the condition in Lemma 9.7. By choosing k to satisfy (9.41) and plugging this bound into (9.40), we conclude that the policy π^k has

$$f^*(\pi^*) - f^*(\pi^k) \lesssim H\beta \sqrt{\frac{Hd \log(1 + \beta^{-2}K/d)}{K}} \lesssim \tilde{O}\left(H^{3/2} \sqrt{\frac{d \log(|\mathcal{Q}|/\delta)}{n}}\right) \lesssim \varepsilon \quad (9.42)$$

as desired.

Finally, we need to argue that the policy $\hat{\pi}$ 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 $1 - \delta/2$, for all k ,

$$|f^*(\pi^k) - \hat{V}^k| \lesssim \sqrt{\frac{\log(2K/\delta)}{n}},$$

which implies that $f^*(\hat{\pi}) \gtrsim f^*(\pi^k) - \sqrt{\frac{\log(2K/\delta)}{n}}$. The error term here is of lower order than (9.42). A union bound gives overall success probability at least $1 - \delta$. □

Proof of Lemma 9.7.

Proof of Lemma 9.7. The proof uses the logarithmic form of the elliptic potential lemma, (4.31) in Lemma 4.2. For any $\lambda > 0$, applying (4.31) for each $h \in [H]$ and summing gives

$$\sum_{k=1}^K \sum_{h=1}^H \log(1 + \|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}^2) \leq Hd \log(1 + \lambda^{-1}K/d).$$

This implies that there exists k such that

$$\sum_{h=1}^H \log(1 + \|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}^2) \leq \frac{Hd \log(1 + \lambda^{-1}K/d)}{K},$$

which means that for all $h \in [H]$, $\log(1 + \|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}^2) \leq \frac{Hd \log(1 + \lambda^{-1}K/d)}{K}$, or equivalently:

$$\|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}^2 \leq \exp\left(\frac{Hd \log(1 + \lambda^{-1}K/d)}{K}\right) - 1.$$

As long as $K \geq Hd \log(1 + \lambda^{-1}K/d)$, using that $e^x \leq 1 + 2x$ for $0 \leq x \leq 1$, we have

$$\|X_h^{M^*}(\pi^k)\|_{(\lambda I + \Sigma_h^k)^{-1}}^2 \leq 2 \frac{Hd \log(1 + \lambda^{-1}K/d)}{K}.$$

□

9.4 Examples and Generalizations of Bellman Rank

This subsection works through some well known examples of classes with low Bellman rank, computing the factorization and rank d for each. Crucially, we see that many of these examples support low rank even when the value function class under consideration is nonlinear.

9.4.1 Examples

We start with familiar examples, then introduce new models that allow for nonlinear function approximation.

Example 9.1 (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

$$X_h^M(\pi) = \{d_h^{M, \pi}(s, a)\}_{s \in \mathcal{S}, a \in \mathcal{A}} \in \mathbb{R}^{SA}$$

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 $d_B(M) \leq SA$.

◁

Example 9.2 (Low-Rank MDPs). The calculation in (9.26) shows that by choosing $X_h^M(\pi) := \mathbb{E}^{M,\pi}[\phi(s_h, a_h)] \in \mathbb{R}^d$ and $W_h^M(Q) := \theta_h^Q - \tilde{\theta}_h^{M,Q} \in \mathbb{R}^d$, any Low-Rank MDP M has $d_B(M) \leq d$. For this class, BiLinUCB gives a weaker guarantee than LSVI-UCB, because it does not exploit Bellman completeness; see the discussion of regret bounds in Section 9.3.

◁

We now explore Bellman rank for some MDP families that have not already been covered.

Example 9.3 (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

$$d_h^{M,\pi}(s, a) = \langle \phi^M(s, a), \theta_h^{M,\pi} \rangle. \quad (9.43)$$

Note that neither ϕ^M nor $\theta_h^{M,\pi}$ 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

$$\begin{aligned} \mathcal{E}_h(\pi, Q) &= \mathbb{E}^{M,\pi} \left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a) \right) \right] \\ &= \sum_{s,a} d_h^{M,\pi}(s, a) \cdot \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \\ &= \sum_{s,a} \langle \phi^M(s, a), \theta_h^{M,\pi} \rangle \cdot \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \\ &= \left\langle \theta_h^{M,\pi}, \sum_{s,a} \phi^M(s, a) \cdot \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \right\rangle. \end{aligned}$$

It follows that if we define

$$X_h^M(\pi) = \theta_h^{M,\pi}$$

and

$$W_h^M(Q) = \sum_{s,a} \phi^M(s, a) \cdot \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right],$$

then $\mathcal{E}_h(\pi, Q) = \langle X_h^M(\pi), W_h^M(Q) \rangle$, which shows that $d_B(M) \leq d$.

This setting includes tabular MDPs²³ and allows for *nonlinear* function approximation: whenever the occupancies satisfy (9.43), the Bellman rank is at most d for *any class* $\mathcal{Q}$, including classes of neural networks or other nonlinear models.

◁

We close with two more new examples.

Example 9.4 (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 form²⁴

$$r_h = -s_h^\top \Lambda^M s_h - a_h^\top \Gamma^M a_h$$

²³Low-rank MDPs satisfy a related condition on state occupancies. This condition can also be handled within the bilinear framework, with a modified exploration rule; see Du et al. [70].

²⁴LQR is typically stated in terms of losses; we negate because we consider rewards. Because the noise is Gaussian, the rewards in the LQR setting are technically unbounded, and fall outside this section's $[0, 1]$ -bounded reward assumption; this requires a straightforward extension of the results in Section 9.2.

for matrices $\Lambda^M, \Gamma^M \succeq 0$. 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 $d_B(M) \leq d^2 + 1$. The basic idea is to choose $X_h^M(\pi) = (\text{vec}(\mathbb{E}^{M,\pi}[s_h s_h^\top]), 1)$ and use the quadratic structure of the value functions. $\triangleleft$

Example 9.5 (Linear $Q^\star/V^\star$). Linearity of $Q^\star$ alone is not enough ([Proposition 9.2](#)), but linearity of both $Q^\star$ and $V^\star$ is.

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

$$\mathcal{Q} = \left\{ Q \mid Q_h(s, a) = \langle \phi(s, a), \theta_h \rangle : \theta_h \in \mathbb{R}^d, \exists w \max_{a \in \mathcal{A}} \langle \phi(s, a), \theta_h \rangle = \langle \psi(s), w \rangle \forall s \right\}.$$

Then $d_B(M) \leq 2d$. We do not prove this result, but the basic idea is to choose $X_h^M(\pi) = \mathbb{E}^{M,\pi}[(\phi(s_h, a_h), \psi(s_{h+1}))]$. $\triangleleft$

See Jiang et al. [\[111\]](#), Du et al. [\[70\]](#) for further examples.

9.4.2 Generalizations: V-Type Bellman Rank

While we gave [\(9.29\)](#) 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 [\[111\]](#), which asserts that for all $\pi \in \Pi$ and $Q \in \mathcal{Q}$,

$$\mathbb{E}^{M,\pi} \mathbb{E}_{s_{h+1}|s_h, a_h \sim \pi_Q(s_h)}^M \left[Q_h(s_h, a_h) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a) \right] = \langle X_h^M(\pi), W_h^M(Q) \rangle. \quad (9.44)$$

This is the same as the definition [\(9.29\)](#) (which is typically referred to as *Q-type* Bellman rank), except that we take $a_h = \pi_Q(s_h)$ instead of $a_h = \pi(s_h)$.²⁵ With an appropriate modification, 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 [\(9.29\)](#), with a canonical example being *Block MDPs*.

Example 9.6 (Block MDP). The Block MDP [\[111, 69, 152\]](#) 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 $M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}, \{R_h^M\}, d_1\}$ 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

$$s_h \sim q_h^M(z_h),$$

²⁵The name “V-type” refers to the fact that [\(9.44\)](#) only depends on Q through the induced V -function $s \mapsto Q_h(s, \pi_Q(s))$, while [\(9.29\)](#) depends on the full Q -function, hence “Q-type”.

where $q_h^M : \mathcal{Z} \rightarrow \Delta(\mathcal{S})$ is an *emission distribution* with the property that $\text{supp}(q_h(z)) \cap \text{supp}(q_h(z')) = \emptyset$ if $z \neq z'$. This property (*decodability*) ensures that there exists a unique mapping $\phi_h^M : \mathcal{S} \rightarrow \mathcal{Z}$ that maps the observed state s_h to the corresponding latent state z_h . We assume that $R_h^M(s, a) = R_h^M(\phi_h^M(s), a)$, which implies that optimal policy π_M depends only on the endogenous latent state, i.e. $\pi_{M,h}(s) = \pi_{M,h}(\phi_h^M(s))$.

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 $\text{poly}(H, |\mathcal{Z}|, |\mathcal{A}|) \cdot \sqrt{T}$ 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 $\text{poly}(|\mathcal{Z}|, |\mathcal{A}|, H, \log|\mathcal{F}|)$, where $\mathcal{F}$ is an appropriate class of function approximators (typically either a value function class $\mathcal{Q}$ containing $Q^{M,*}$ 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

$$\begin{aligned} & \mathbb{E}^{M,\pi} \mathbb{E}_{s_{h+1}|a_h, s_h \sim \pi_Q(s_h)}^M \left[Q_h(s_h, \pi_Q(s_h)) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, \pi_Q(s_{h+1})) \right] \\ &= \sum_{z \in \mathcal{Z}} d_h^{M,\pi}(z) \mathbb{E}_{s \sim q_h^M(z)} \mathbb{E}_{s_{h+1}|s_h, a_h \sim \pi_Q(s_h)}^M \left[Q_h(s_h, \pi_Q(s_h)) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, \pi_Q(s_{h+1})) \right]. \end{aligned}$$

where $d_h^{M,\pi}(z)$ denotes the probability that the latent state at layer h is z under π . This implies that we can take

$$X_h^M(\pi) = \{d_h^{M,\pi}(z)\}_{z \in \mathcal{Z}}$$

and

$$W_h^M(Q) = \left\{ \mathbb{E}_{s \sim q_h^M(z)} \mathbb{E}_{s_{h+1}|s_h, a_h \sim \pi_Q(s_h)}^M \left[Q_h(s_h, \pi_Q(s_h)) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, \pi_Q(s_{h+1})) \right] \right\}_{z \in \mathcal{Z}}$$

so that the V-type Bellman rank is at most $|\mathcal{Z}|$. This means that as long as $\mathcal{Q}$ contains $Q^{M,*}$, we can obtain sample complexity guarantees that scale with $|\mathcal{Z}|$ rather than $|\mathcal{S}|$, as desired. $\triangleleft$

There are a number of variants of Bellman rank, including Bilinear rank [70] and Bellman-Eluder dimension [116], which subsume and slightly generalize both Bellman rank definitions.

9.5 Estimation-to-Decisions for Low Bellman Rank

An alternative to BiLinUCB is to run E2D (Algorithm 7.1) over a realizable model class $\mathcal{M}$ with low Bellman rank. By Theorem 7.1, all that is required is a bound on the DEC for the class at hand. The following result [84] shows that the Decision-Estimation Coefficient is always bounded for classes with low Bellman rank.

Proposition 9.3 (DEC for Bellman rank): Let $\mathcal{M}$ be a class of MDPs with nonnegative rewards and $\sum_{h=1}^H r_h \leq 1$ almost surely under every model and policy. Suppose every $Q \in \mathcal{Q}_{\mathcal{M}}$ satisfies $Q_h(s, a) \in [0, 1]$ for all h, s, a , and every $M \in \mathcal{M}$ has Bellman rank at most d relative to the induced classes $\mathcal{Q}_{\mathcal{M}}$ and $\Pi_{\mathcal{Q}_{\mathcal{M}}}$ from (9.1). Then, for every $\gamma > 0$, we have

$$\text{dec}_{\gamma}(\mathcal{M}) \lesssim \frac{H^2 d}{\gamma}. \quad (9.45)$$

By (7.22), E2D with $\gamma \propto H\sqrt{dT/\mathbf{Est}_H(\mathcal{M}, T, \delta)}$ has $\mathbf{Reg} \lesssim H\sqrt{dT \cdot \mathbf{Est}_H(\mathcal{M}, T, \delta)}$ with probability at least $1 - \delta$ whenever we have access to a realizable model class satisfying the assumptions of Proposition 9.3. As a special case, for any finite class $\mathcal{M}$, averaged exponential weights has online log-loss regret $\mathbf{Reg}_{\text{OL}}^{\text{Log}} \leq \log|\mathcal{M}|$. The conversion in Lemma 7.3 controls $\mathbf{Est}_H$, which in turn gives the RL regret bound

$$\mathbb{E}[\mathbf{Reg}] \lesssim H\sqrt{dT \log|\mathcal{M}|} \quad (9.46)$$

in expectation, via the first inequality in (7.21). For tabular MDPs, $d = SA$, and this recovers the guarantee of Theorem 6.2 with an improved dependence on the horizon. We do not prove Proposition 9.3, but interested readers can refer to Foster et al. [84, Theorem 7.1]. The proof uses a more general version of the PC-IGW algorithm from Section 6.7, incorporating optimal design in the space of policies; see Foster et al. [84, Theorem 7.2].

Value-based guarantees via optimistic estimation. Estimation-to-Decisions as described above is a model-based method, and in general, the model estimation complexity $\log|\mathcal{M}|$ in (9.46) 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 Theorem 9.2, 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 8.6.1 and 8.6.2. One such choice is the divergence

$$D_{\text{sbr}}^{\pi}(Q \parallel M) = \sum_{h=1}^H \left(\mathbb{E}^{M, \pi} \left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a) \right) \right] \right)^2,$$

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 8.6.2. One can show that the optimistic DEC for this divergence is bounded as

$$\text{o-dec}_{\gamma}^{D_{\text{sbr}}}(\mathcal{M}) \lesssim \frac{H \cdot d}{\gamma}.$$

This implies that E2D.Opt, with an appropriate choice of estimation algorithm tailored to $D_{\text{sbr}}^{\pi}(\cdot \parallel \cdot)$, 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 $D_{\text{sbr}}^{\pi}(\cdot \parallel \cdot)$, it is critical to appeal to optimistic estimation to derive this result. Indeed, the non-optimistic generalized DEC $\text{dec}_{\gamma}^{D_{\text{sbr}}}$ does not enjoy a polynomial bound. See Foster et al. [90] for details.

9.6 Concluding Remarks and Broader Outlook

This section showed that for reinforcement learning with function approximation, realizability alone, whether of the model, the optimal value function, or the optimal policy, does not permit sample-efficient learning, and that additional structure controlling distribution shift is required. We developed two such structures: low-rank transitions, exploited by LSVI-UCB through least-squares value iteration with elliptic bonuses (Theorem 9.1), and the more general notion of low Bellman rank, exploited by BiLinUCB through optimism on average and elimination of value functions with large Bellman residual (Theorem 9.2), and, through the bound on the DEC in Proposition 9.3, by E2D. We close with a survey of the broader landscape of tractability, and with open directions for reinforcement learning and the theory developed in this monograph as a whole.

The landscape of tractability. Within the Bellman-rank family, other algorithms include OLIVE [111], which eliminates value functions using a single adversarially chosen policy per round, and GOLF [116], which attains $\sqrt{T}$ regret under Bellman completeness and is analyzed through the Bellman-eluder dimension, a combination of Bellman rank with the eluder dimension of Russo and Van Roy [180]. Bilinear classes [70] and witness rank [193] extend these ideas to model-based function approximation, and Chen et al. [53] and Zhong et al. [229] unify the resulting guarantees through the DEC and related measures. Representation learning for low-rank and block MDPs is treated by Agarwal et al. [8], Uehara et al. [201], Zhang et al. [227], and Mhammedi et al. [150], and inherent Bellman error offers an alternative to Bellman completeness [222]. For an expanded textbook treatment of theoretical guarantees in reinforcement learning with function approximation, see Agarwal et al. [10].

Bellman rank is one of several structural conditions under which sample-efficient reinforcement learning with function approximation is possible, and it is worth knowing the others. The first is *coverability* [215, 18, 151], which asks that a single (unknown) distribution over states cover the occupancy of every policy, and is satisfied by tabular, low-rank, and block MDPs; this facilitates a different type of control over distribution shift which subsumes some settings beyond the scope of Bellman rank.

The second is *exogenous* structure, in which part of the state evolves independently of the agent’s actions [65]; for exogenous block MDPs, the endogenous latent state can be recovered with sample complexity independent of the exogenous part [76, 74, 132].

The third is *partial observability*: while learning in general POMDPs is intractable, structural conditions such as undercompleteness [113], observability and short-term memory [98, 75], and weak revealing or *B*-stability [147, 51] permit sample-efficient learning, as does the predictive state representation model [224].

These conditions are incomparable in general. Coverability captures classes outside the scope of Bellman rank, and the converse also holds; and no single condition is known to capture every class for which sample-efficient learning is possible. Characterizing such classes is an ongoing research program.

Topics beyond the scope of the monograph. Two large parts of the reinforcement learning literature fall outside the scope of this monograph, as discussed in Section 1.3. *Offline reinforcement learning* removes exploration altogether: the learner receives a fixed dataset and must find a good policy without further interaction, so that the central issues become coverage of the data and the realizability and completeness of the function class [155, 55, 213, 208, 86], with pessimism playing the role that optimism plays here [117, 174, 214, 223]; see Levine et al. [143] for a survey.

Policy gradient methods [211, 195, 184, 185], which optimize a parameterized policy directly and are the workhorse of practical deep reinforcement learning, are analyzed by Agarwal et al. [9]; their guarantees require the data distribution to already cover the relevant states, and comparatively little is known about exploration for policy gradient methods beyond the tabular and linear settings [44, 7], even though they succeed in practice in domains where strong exploration is not required.

Open directions and broader outlook. Stepping back, we close with a broader view of where the theory developed in this monograph stands and where it is headed.

Closing the statistical complexity gap. The most immediate question is a purely technical one.

The upper and lower bounds of [Sections 7](#) and [8](#) match up to the cost of estimation: the upper bounds pay $\log|\mathcal{M}|$ for a finite class and the lower bounds do not, and the radii at which the two are evaluated differ by the same factor. For specific classes the gap can be closed by hand, and [Section 8.6](#) and [Chen et al. \[52\]](#) describe some more general techniques, but a unified account of the statistical complexity of interactive decision making, with matching upper and lower bounds for every class, does not yet exist. The gap is most pronounced in reinforcement learning, where model-based guarantees scale with the complexity of the model class and value-based guarantees with that of the value function class; value-based guarantees obtainable through the DEC ([Section 9.5](#)) are not yet sharp.

Computational foundations. This monograph has taken a purely statistical perspective. Our upper bounds are attained by algorithms that involve primitives like solving a min-max problem at every round, or maintaining confidence sets over value functions, and we have not asked whether these steps can be carried out efficiently. For contextual bandits the answer is yes, which is what makes the reduction to regression oracles practical. For reinforcement learning with function approximation the picture is different: there are settings in which sample-efficient learning is possible but computationally efficient learning is not, under standard complexity-theoretic assumptions [[99](#), [178](#)]. A computational theory of reinforcement learning, one that says which structural conditions permit computationally efficient exploration and which do not, is in its infancy, and we regard it as among the most interesting technical problems in the area.

Rethinking the foundations. One can read the state of affairs in this section more pessimistically. Two facts stand out. First, unlike contextual bandits, where exploration algorithms are now deployed at scale ([Section 1.2](#)), provably efficient exploration algorithms for reinforcement learning have not made their way into practice in any robust form. Second, every positive result in this section rests on some kind of structural assumption. These assumptions are delicate, and designing algorithms around them has been a slow process, proceeding one assumption and one paper at a time. Taken together, these may be a sign that the problem as we have posed it, learning to act from scratch in an unknown environment with only a function class as prior knowledge, is simply too hard, and that progress requires relaxing the setting rather than lengthening the list of assumptions.

Pre-trained models. Pre-trained models offer one such relaxation, and the most consequential one to date. In the language model paradigm, reinforcement learning is not used to learn from scratch but to adapt a model that already encodes a great deal about the task, through post-training with verifiable or human feedback [[56](#), [192](#), [163](#), [28](#), [170](#)]. This has been enormously successful, and a science is beginning to emerge around questions like what the pre-trained model must provide for post-training to succeed [[107](#), [54](#), [93](#)]. From the viewpoint of this monograph, the pre-trained model plays the role that the class $\mathcal{M}$ played for us, that of prior knowledge about the environment, but in a far richer and less explicit form, and the statistical and computational complexity of decision making relative to such a prior is largely uncharted.

Language models are, from the standpoint of this monograph, a simple decision making problem: a single decision per prompt, or at most a short horizon. The ideas developed here around exploration, credit assignment, and distribution shift matter most for longer-horizon problems, including robotics, embodied agents, and scientific discovery. We hope to see successful approaches in these settings to combine the exploration principles of this monograph with pre-training and imitation learning [[168](#), [179](#), [91](#)], the latter supplying the coverage that exploration from scratch cannot, and understanding such combinations is wide open.

Beyond the black box. Finally, throughout this monograph we have treated estimation as a black

box: an oracle that returns a model or a value function together with a guarantee, whose internals we never examine. This was deliberate, and is what allows us to state one principle covering every function class. But the models now in use are deep neural networks, and a black-box view discards most of what is known about them: the representations they learn, the inductive biases of their architectures and of the optimizers that train them, and the way their behavior scales with data and compute. Each of these bears on decision making. How learning and exploration interact inside a neural network, rather than across an oracle interface, may be where the next theory (or science) of decision making begins.

9.7 Bibliographic Remarks

[Proposition 9.1](#) is adapted from Krishnamurthy et al. [130], who introduced the rich-observation setting that later became the Block MDP. [Proposition 9.2](#) is due to Weisz et al. [210], who proved the lower bound for planning with a generative model; Wang et al. [209] provided a simpler construction and showed that the hardness persists in online RL, even under a constant action suboptimality gap. The low-rank MDP model, in the form used here, is from Yang and Wang [218], Jin et al. [115], who use the name *linear MDP*. LSVI-UCB ([Algorithm 9.1](#)) and its analysis—[Lemma 9.1](#), [Theorem 9.1](#), [Lemma 9.3](#), and [Lemma 9.4](#)—follow Jin et al. [115].

Bellman rank was introduced by Jiang et al. [111] in the V-type form of [\(9.44\)](#); the Q-type definition in [\(9.29\)](#) is from Du et al. [70], Jin et al. [116]. BiLinUCB ([Algorithm 9.2](#)), [Theorem 9.2](#), and [Lemma 9.5](#), [Lemma 9.6](#), and [Lemma 9.7](#) follow Du et al. [70], specialized from bilinear classes to Bellman rank and to finite $\mathcal{Q}$; the logarithmic form of the elliptic potential lemma, [\(4.31\)](#), is Lattimore and Szepesvári [139, Lemma 19.4]. The examples in [Section 9.4](#) are drawn from Jiang et al. [111] and Du et al. [70]. [Proposition 9.3](#) is Foster et al. [84, Theorem 7.1], and the value-based guarantee via optimistic estimation is from Foster et al. [90].

ACKNOWLEDGMENTS

We thank Fan Chen, Margalit Glasgow, Noah Golowich, Yanjun Han, Sham Kakade, Jian Qian, Yunbei Xu, and Yunzong Xu for many helpful discussions on the topics in the manuscript over the years, and for being amazing research collaborators and friends. We also thank the students of MIT’s 9.522 course, particularly the Fall 2022 and Fall 2023 iterations, for invaluable feedback on the content and presentation of the material in the monograph.

Statement on AI use. The authors used AI tools in the final stages of this project for copy-editing, restructuring, and fixing technical errors. All technical content, proofs, and final wording were written and checked by the authors, who take full responsibility for the manuscript.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 35284–35297, 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, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.
- [6] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, S. Sen, and A. Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016.
- [7] A. Agarwal, M. Henaff, S. Kakade, and W. Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, 2020.
- [8] A. Agarwal, S. Kakade, A. Krishnamurthy, and W. Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [9] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- [10] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. *Reinforcement Learning: Theory and Algorithms*. 2022. Monograph draft, Version 3 (January 31, 2022), <https://rltheorybook.github.io>.
- [11] A. Agarwal, Y. Jin, and T. Zhang. Vo q l: Towards optimal regret in model-free rl with non-linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR, 2023.
- [12] R. Agrawal. Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.
- [13] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, volume 23, pages 39.1–39.26. PMLR, 2012.

- [14] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.
- [15] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B*, 28(1):131–142, 1966.
- [16] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, 2015.
- [17] K. Amin, M. Kearns, and U. Syed. Bandits, query learning, and the haystack dimension. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 87–106. JMLR Workshop and Conference Proceedings, 2011.
- [18] P. Amortila, D. J. Foster, N. Jiang, A. Sekhari, and T. Xie. Harnessing density ratios for online reinforcement learning. In *International Conference on Learning Representations*, 2024.
- [19] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.
- [20] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 217–226, 2009.
- [21] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- [22] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.
- [23] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- [24] 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.
- [25] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 21, pages 89–96, 2008. Journal version: Jaksch, Ortner, and Auer, *Journal of Machine Learning Research* 11:1563–1600, 2010.
- [26] 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.
- [27] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.
- [28] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022.

- [29] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.
- [30] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 2016.
- [31] R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.
- [32] A. Bietti, A. Agarwal, and J. Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.
- [33] 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.
- [34] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- [35] W. Black, E. Ilhan, A. Marchini, and V. Markeviciute. AdaptEx: A self-service contextual bandit platform. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023. doi: 10.1145/3604915.3608870.
- [36] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.
- [37] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.
- [38] R. I. Brafman and M. Tennenholtz. R-max—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.
- [39] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- [40] S. Bubeck and R. Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589, 2016.
- [41] 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.
- [42] 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.
- [43] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.
- [44] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.
- [45] N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pages 12–18. ACM, 1999.

- [46] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.
- [47] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- [48] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.
- [49] O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, 2011.
- [50] F. Chen and A. Rakhlin. Decision making in changing environments: Robustness, query-based learning, and differential privacy. *arXiv preprint arXiv:2501.14928*, 2025.
- [51] F. Chen, Y. Bai, and S. Mei. Partially observable RL with B-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022.
- [52] F. Chen, D. J. Foster, Y. Han, J. Qian, A. Rakhlin, and Y. Xu. Assouad, Fano, and Le Cam with Interaction: A Unifying Lower Bound Framework and Characterization for Bandit Learnability. In *Advances in Neural Information Processing Systems*, volume 37, pages 75585–75641, 2024.
- [53] F. Chen, S. Mei, and Y. Bai. Unified algorithms for RL with decision-estimation coefficients: PAC, reward-free, preference-based learning, and beyond. *The Annals of Statistics*, 53(1):426–456, 2025.
- [54] F. Chen, A. Huang, N. Golowich, S. Malladi, A. Block, J. T. Ash, A. Krishnamurthy, and D. J. Foster. The coverage principle: How pre-training enables post-training. In *International Conference on Learning Representations (ICLR)*, 2026.
- [55] J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.
- [56] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.
- [57] 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.
- [58] T. M. Cover. Behavior of sequential predictors of binary sequences. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 263–272, Prague, 1967. Publishing House of the Czechoslovak Academy of Sciences.
- [59] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.
- [60] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.
- [61] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, 2008.

- [62] 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.
- [63] 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.
- [64] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- [65] T. G. Dietterich, G. Trimponias, and Z. Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In *International Conference on Machine Learning*, pages 1262–1270, 2018.
- [66] O. D. Domingues, P. Ménard, E. Kaufmann, and M. Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.
- [67] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, I. Technical Report 137a, Department of Statistics, University of California, Berkeley, 1987.
- [68] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, 19(2):633–667, 1991.
- [69] 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.
- [70] 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.
- [71] J. C. Duchi. Lecture notes on statistics and information theory. 2023.
- [72] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178. AUAI Press, 2011.
- [73] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.
- [74] Y. Efroni, D. J. Foster, D. Misra, A. Krishnamurthy, and J. Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, 2022.
- [75] Y. Efroni, C. Jin, A. Krishnamurthy, and S. Miryoosefi. Provable reinforcement learning with a short-term memory. In *International Conference on Machine Learning*, 2022.
- [76] Y. Efroni, D. Misra, A. Krishnamurthy, A. Agarwal, and J. Langford. Provable RL with exogenous distractors via multistep inverse dynamics. In *International Conference on Learning Representations (ICLR)*, 2022.

- [77] 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.
- [78] D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Neural Information Processing Systems (NeurIPS)*, 2021.
- [79] D. J. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.
- [80] D. J. Foster, A. Agarwal, M. Dudík, H. Luo, and R. E. Schapire. Practical contextual bandits with regression oracles. *International Conference on Machine Learning*, 2018.
- [81] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. *Conference on Learning Theory*, 2018.
- [82] D. J. Foster, A. Rakhlin, and K. Sridharan. Online learning: Sufficient statistics and the burkholder method. *Conference on Learning Theory*, 2018.
- [83] D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020.
- [84] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.
- [85] D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, page 2059, 2021.
- [86] D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, 2022.
- [87] D. J. Foster, A. Rakhlin, A. Sekhari, and K. Sridharan. On the complexity of adversarial decision making. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- [88] D. J. Foster, D. P. Foster, N. Golowich, and A. Rakhlin. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *Conference on Learning Theory*, 2023.
- [89] D. J. Foster, N. Golowich, and Y. Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 3969–4043, 2023.
- [90] D. J. Foster, N. Golowich, J. Qian, A. Rakhlin, and A. Sekhari. Model-free reinforcement learning with the decision-estimation coefficient. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [91] D. J. Foster, A. Block, and D. Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666, 2024.

- [92] D. J. Foster, Y. Han, J. Qian, and A. Rakhlin. Online estimation via offline estimation: An information-theoretic framework. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [93] D. J. Foster, Z. Mhammedi, and D. Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *Conference on Learning Theory (COLT)*, 2025.
- [94] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
- [95] A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.
- [96] J. Gerriets. VowpalWabbit 8.9.0 release notes. <https://vowpalwabbit.org/blog/vowpalwabbit-8.9.0.html>, 2020. Accessed 2026-09-16.
- [97] M. Glasgow and A. Rakhlin. Tight bounds for γ -regret via the decision-estimation coefficient. *arXiv preprint arXiv:2303.03327*, 2023.
- [98] N. Golowich, A. Moitra, and D. Rohatgi. Learning in observable POMDPs, without computationally intractable oracles. In *Advances in Neural Information Processing Systems*, 2022.
- [99] N. Golowich, A. Moitra, and D. Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1953–1967. IEEE, 2024.
- [100] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.
- [101] P. Harsha, C. Subramanian, N. Abe, S. Subramanian, A. Ba, K. A. Fernández Román, M. Longinos Garrido, and C. Narayanaswami. Practical contextual bandits for large-scale structured discrete constrained optimization problems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 850–859, 2025. doi: 10.1145/3711896.3737086.
- [102] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- [103] E. Hazan and S. Kale. An online portfolio selection algorithm with regret logarithmic in price variation. *Mathematical Finance*, 25(2):288–310, 2015.
- [104] L. He, Y. Wu, and G. Jindal. Enhancing personalized CRM communication with contextual bandit strategies. Uber Engineering Blog, <https://www.uber.com/blog/enhancing-personalized-crm/>, 2025. Accessed 2026-09-16.
- [105] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.
- [106] A. Huang, A. Block, D. J. Foster, D. Rohatgi, C. Zhang, M. Simchowitz, J. T. Ash, and A. Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *International Conference on Learning Representations*, 2025.

- [107] A. Huang, A. Block, Q. Liu, N. Jiang, A. Krishnamurthy, and D. J. Foster. Is best-of-N the best of them? coverage, scaling, and optimality in inference-time alignment. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 25075–25126, 2025.
- [108] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.
- [109] T. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 562–573. Springer, 2009.
- [110] N. Jiang and A. Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.
- [111] 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.
- [112] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.
- [113] C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. *Advances in Neural Information Processing Systems*, 33, 2020.
- [114] 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.
- [115] 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.
- [116] C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34: 13406–13418, 2021.
- [117] Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, 2021.
- [118] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.
- [119] K.-S. Jun and C. Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 33, 2020.
- [120] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.
- [121] S. M. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
- [122] A. Kalai and S. Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 3:423–440, 2002.

- [123] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213, 2012.
- [124] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002.
- [125] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.
- [126] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17:697–704, 2004.
- [127] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.
- [128] R. Kleinberg, A. Slivkins, and E. Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.
- [129] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.
- [130] A. Krishnamurthy, A. Agarwal, and J. Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.
- [131] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.
- [132] A. Lamb, R. Islam, Y. Efroni, A. Didolkar, D. Misra, D. J. Foster, L. Molu, R. Chari, A. Krishnamurthy, and J. Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023.
- [133] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.
- [134] P.-S. Laplace. *Essai philosophique sur les probabilités*. Courcier, Paris, 1814. English translation: *A Philosophical Essay on Probabilities*, Dover, 1951.
- [135] T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3):311–334, 2020.
- [136] T. Lattimore and A. György. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.
- [137] T. Lattimore and C. Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 728–737, 2017.
- [138] 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.
- [139] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

- [140] T. Lattimore and C. Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pages 2488–2515. PMLR, 2020.
- [141] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.
- [142] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3):591–613, 2009.
- [143] S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
- [144] G. Li, P. Kamath, D. J. Foster, and N. Srebro. Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.
- [145] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM, 2010.
- [146] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.
- [147] Q. Liu, A. Chung, C. Szepesvári, and C. Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220, 2022.
- [148] H. Luo, C.-Y. Wei, and K. Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, 2018.
- [149] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, 2011.
- [150] Z. Mhammedi, A. Block, D. J. Foster, and A. Rakhlin. Efficient model-free exploration in low-rank MDPs. In *Advances in Neural Information Processing Systems*, 2023.
- [151] Z. Mhammedi, D. J. Foster, and A. Rakhlin. The power of resets in online reinforcement learning. *Advances in Neural Information Processing Systems*, 37:12334–12407, 2024.
- [152] 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.
- [153] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- [154] 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.
- [155] R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

- [156] M. Opper and D. Haussler. Worst case prediction over sequences under log loss. In G. Cybenko, D. P. O’Leary, and J. Rissanen, editors, *The Mathematics of Information Coding, Extraction and Distribution*, volume 107 of *The IMA Volumes in Mathematics and its Applications*, pages 81–90. Springer, New York, 1999.
- [157] F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
- [158] L. Orseau, T. Lattimore, and S. Legg. Soft-Bayes: Prod for mixtures of experts with log-loss. In *International Conference on Algorithmic Learning Theory*, pages 372–399. PMLR, 2017.
- [159] I. Osband and B. Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.
- [160] I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.
- [161] I. Osband, D. Russo, and B. Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26:3003–3011, 2013.
- [162] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034, 2016.
- [163] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.
- [164] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.
- [165] Y. Polyanskiy. Information theoretic methods in statistics and computer science. Class notes, EPFL 2019 / ENSAE-CREST 2020. https://people.lids.mit.edu/yp/homepage/sdpi_course.shtml, 2020.
- [166] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. Lecture notes for MIT 6.441 / UIUC ECE 563 / Yale S&DS 664, 2014. Superseded by *Information Theory: From Coding to Learning*, Cambridge University Press, 2025.
- [167] Y. Polyanskiy and Y. Wu. Dualizing Le Cam’s method for functional estimation, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*, 2019.
- [168] D. A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, 1988.
- [169] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.
- [170] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.
- [171] A. Rakhlin and K. Sridharan. Statistical learning and sequential prediction, 2012. Available at <http://www.mit.edu/~rakhlin/courses/stat928/stat928-notes.pdf>.

- [172] A. Rakhlin and K. Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pages 1704–1722, 2017.
- [173] 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.
- [174] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, 2021.
- [175] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 811–820. ACM, 2010.
- [176] J. Rissanen. Complexity of strings in the class of markov sources. *IEEE Transactions on Information Theory*, 32(4):526–532, 1986.
- [177] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
- [178] D. Rohatgi and D. J. Foster. Necessary and sufficient oracles: Toward a computational taxonomy for reinforcement learning. *Conference on Learning Theory (COLT)*, 2025.
- [179] S. Ross, G. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.
- [180] 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.
- [181] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.
- [182] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.
- [183] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.
- [184] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.
- [185] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [186] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [187] Y. M. Shtar’kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987. English translation: Problems of Information Transmission, 23(3):175–186, 1987.

- [188] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.
- [189] D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.
- [190] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958.
- [191] A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.
- [192] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- [193] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.
- [194] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [195] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999.
- [196] H. Tang, R. Houthoofd, D. Foote, A. Stooke, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. #Exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.
- [197] A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 2017.
- [198] 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.
- [199] A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.
- [200] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.
- [201] M. Uehara, X. Zhang, and W. Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.
- [202] S. A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.
- [203] A. Vangara and A. Egg. Contextual bandits in payment processing: Non-uniform exploration and supervised learning. *arXiv preprint arXiv:2412.00569*, 2024.
- [204] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

- [205] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.
- [206] V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pages 371–383. Morgan Kaufmann, 1990.
- [207] 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.
- [208] R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations (ICLR)*, 2021.
- [209] Y. Wang, R. Wang, and S. M. Kakade. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- [210] 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.
- [211] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.
- [212] W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.
- [213] T. Xie and N. Jiang. Q^* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.
- [214] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.
- [215] T. Xie, D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade. The role of coverage in online reinforcement learning. In *ICLR*, 2023.
- [216] T. Xie, D. J. Foster, A. Krishnamurthy, C. Rosset, A. Awadallah, and A. Rakhlin. Exploratory preference optimization: Harnessing implicit Q^* -approximation for sample-efficient RLHF. In *International Conference on Learning Representations (ICLR)*, 2025.
- [217] Y. Xu and A. Zeevi. Bayesian design principles for frequentist sequential learning. In *International Conference on Machine Learning*, 2023.
- [218] 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.
- [219] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.
- [220] H. Yao, C. Szepesvári, B. Á. Pires, and X. Zhang. Pseudo-mdps and factored linear action models. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, ADPRL 2014, Orlando, FL, USA, December 9-12, 2014*, pages 1–9. IEEE, 2014.

- [221] A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.
- [222] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*, 2020.
- [223] W. Zhan, B. Huang, A. Huang, N. Jiang, and J. D. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775, 2022.
- [224] W. Zhan, M. Uehara, W. Sun, and J. D. Lee. PAC reinforcement learning for predictive state representations. In *International Conference on Learning Representations*, 2023.
- [225] T. Zhang. From ϵ -entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.
- [226] T. Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.
- [227] X. Zhang, Y. Song, M. Uehara, M. Wang, A. Agarwal, and W. Sun. Efficient reinforcement learning in block MDPs: A model-free representation learning approach. In *International Conference on Machine Learning*, 2022.
- [228] Z. Zhang, X. Ji, and S. Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021.
- [229] H. Zhong, W. Xiong, S. Zheng, L. Wang, Z. Wang, Z. Yang, and T. Zhang. Generalized eluder coefficient: A unified framework for interactive decision making in Markov decision processes, partially observable Markov decision processes, and beyond. *Mathematics of Operations Research*, 2026. Articles in advance.
- [230] Z. Zhu, R. de Salvo Braz, J. Bhandari, D. Jiang, Y. Wan, Y. Efroni, L. Wang, R. Xu, H. Guo, A. Nikulkov, D. Korenkevych, U. Dogan, F. Cheng, Z. Wu, and W. Xu. Pearl: A production-ready reinforcement learning agent. *Journal of Machine Learning Research*, 25(273):1–30, 2024.

A. TECHNICAL TOOLS

A.1 Probabilistic Inequalities

A.1.1 Proof of Bernstein's Inequality

Proof of Lemma 2.8. Set $X_i = Z_i - \mathbb{E} Z_i$. If $v = 0$, then $X_i = 0$ almost surely and the result is immediate. Otherwise $B > 0$. For $0 \leq \lambda < 3/B$, boundedness permits expansion of the exponential in expectation. Since $\mathbb{E} X_i = 0$, $\mathbb{E} X_i^2 = v^2$, and $|\mathbb{E} X_i^k| \leq B^{k-2} v^2$ for $k \geq 2$,

$$\mathbb{E} e^{\lambda X_i} \leq 1 + v^2 \sum_{k=2}^{\infty} \frac{\lambda^k B^{k-2}}{k!} \leq 1 + \frac{\lambda^2 v^2}{2(1 - \lambda B/3)} \leq \exp \left\{ \frac{\lambda^2 v^2}{2(1 - \lambda B/3)} \right\}.$$

Here $k! \geq 2 \cdot 3^{k-2}$ for $k \geq 2$. Independence and Markov's inequality imply, for any $s > 0$ and $0 < \lambda < 3/B$,

$$\mathbb{P} \left(\sum_{i=1}^T X_i \geq s \right) \leq \exp \left\{ -\lambda s + \frac{T \lambda^2 v^2}{2(1 - \lambda B/3)} \right\}.$$

Put $u = \log(1/\delta)$, $a = \sqrt{2u/(Tv^2)}$, and choose $\lambda = a/(1 + Ba/3)$. For $s = \sqrt{2Tv^2u} + Bu/3$, substitution makes the exponent exactly $-u$. Dividing the resulting bound by T proves the claim. $\square$

A.1.2 Tail Bounds with Stopping Times

Lemma A.1 (Hoeffding's inequality with adaptive stopping time): For i.i.d. random variables $Z_1, \dots, Z_T$ taking values in $[a, b]$ almost surely, with probability at least $1 - \delta$,

$$\frac{1}{T'} \sum_{i=1}^{T'} Z_i - \mathbb{E}[Z] \leq (b - a) \sqrt{\frac{\log(T/\delta)}{2T'}} \quad \forall 1 \leq T' \leq T. \quad (\text{A.1})$$

As a consequence, for any random time $\tau \in [T]$, with probability at least $1 - \delta$,

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

The choice of τ may depend arbitrarily on the data. └

Proof of Lemma A.1. Lemma 2.7 states that for any fixed $T' \in [T]$, with probability at least $1 - \delta$,

$$\frac{1}{T'} \sum_{i=1}^{T'} Z_i - \mathbb{E}[Z] \leq (b - a) \sqrt{\frac{\log(1/\delta)}{2T'}}.$$

(A.1) follows by applying this result with $\delta' = \delta/T$ and taking a union bound over all T choices for $T' \in [T]$. On this event, the bound holds at every $T' \in [T]$, so substituting $T' = \tau$ proves (A.2). $\square$

A.1.3 Tail Bounds for Martingales

Lemma A.2 (Martingale Chernoff bound): For any sequence of real-valued random variables $(X_t)_{t \leq T}$ adapted to a filtration $(\mathcal{F}_t)_{t \leq T}$, it holds that with probability at least $1 - \delta$, for all $T' \leq T$,

$$\sum_{t=1}^{T'} X_t \leq \sum_{t=1}^{T'} \log(\mathbb{E}_{t-1}[e^{X_t}]) + \log(\delta^{-1}), \quad (\text{A.3})$$

where $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ denotes conditional expectation.

Proof of Lemma A.2. Write $\psi_t = \log(\mathbb{E}_{t-1}[e^{X_t}])$ and define

$$Z_0 = 1, \quad Z_\tau = \exp\left(\sum_{t=1}^{\tau} (X_t - \psi_t)\right) = \prod_{t=1}^{\tau} e^{X_t - \psi_t}.$$

If $\psi_t = +\infty$, interpret the corresponding factor Z_t as zero. In this case, the right-hand side of (A.3) is infinite for all $T' \geq t$, so the desired bound is automatic.

To proceed, we observe that the process $(Z_\tau)_{\tau=0}^T$ is a nonnegative supermartingale. Indeed, for each $1 \leq \tau \leq T$, on the event $\psi_\tau < \infty$,

$$\mathbb{E}_{\tau-1}[Z_\tau] = Z_{\tau-1} e^{-\psi_\tau} \mathbb{E}_{\tau-1}[e^{X_\tau}] = Z_{\tau-1},$$

while on the event $\psi_\tau = +\infty$ the conditional expectation is zero. Hence $\mathbb{E}_{\tau-1}[Z_\tau] \leq Z_{\tau-1}$ in either case. Since $Z_0 = 1$, Ville's inequality (e.g., Howard et al. [105]) gives

$$\mathbb{P}\left(\max_{1 \leq \tau \leq T} Z_\tau \geq \delta^{-1}\right) \leq \delta.$$

With probability at least $1 - \delta$, therefore, $Z_{T'} < \delta^{-1}$ for all $1 \leq T' \leq T$. Taking logarithms whenever $\psi_1, \dots, \psi_{T'}$ are finite proves (A.3). $\square$

The next result is a martingale counterpart to Bernstein's inequality (Lemma 2.8).

Lemma A.3 (Freedman's inequality (Bernstein for martingales)): Let $(X_t)_{t \leq T}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathcal{F}_t)_{t \leq T}$. If $|X_t| \leq R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1 - \delta$, for all $T' \leq T$,

$$\sum_{t=1}^{T'} X_t \leq \eta \sum_{t=1}^{T'} \mathbb{E}_{t-1}[X_t^2] + \frac{\log(\delta^{-1})}{\eta}.$$

Proof of Lemma A.3. Without loss of generality, let $R = 1$, and fix $\eta \in (0, 1)$. The result follows by invoking Lemma A.2 with ηX_t in place of X_t , and by the facts that $e^a \leq 1 + a + (e - 2)a^2$ for $a \leq 1$ and $1 + b \leq e^b$ for all $b \in \mathbb{R}$. $\square$

The following result follows by applying Lemma A.3 to both tails.

Lemma A.4 (Multiplicative Freedman inequality): Let $(X_t)_{t \leq T}$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_t)_{t \leq T}$. If $0 \leq X_t \leq R$ almost surely, then with probability at least $1 - \delta$,

$$\sum_{t=1}^T X_t \leq \frac{3}{2} \sum_{t=1}^T \mathbb{E}_{t-1}[X_t] + 2R \log(2\delta^{-1}),$$

and

$$\sum_{t=1}^T \mathbb{E}_{t-1}[X_t] \leq 2 \sum_{t=1}^T X_t + 4R \log(2\delta^{-1}).$$

Proof of Lemma A.4. Suppose $R > 0$, and write $\mu_t = \mathbb{E}_{t-1}[X_t]$ and $Y_t = X_t - \mu_t$. Then $\mathbb{E}_{t-1}[Y_t] = 0$, $|Y_t| \leq R$, and

$$\mathbb{E}_{t-1}[Y_t^2] = \mathbb{E}_{t-1}[X_t^2] - \mu_t^2 \leq R\mu_t.$$

Apply Lemma A.3 to (Y_t) and $(-Y_t)$ with $\eta = 1/(2R) \in (0, 1/R)$ and failure probability $\delta/2$ for each application. Set $S = \sum_{t=1}^T X_t$, $M = \sum_{t=1}^T \mu_t$, and $L = \log(2/\delta)$. A union bound gives, with probability at least $1 - \delta$,

$$S - M \leq \frac{1}{2}M + 2RL, \quad M - S \leq \frac{1}{2}M + 2RL.$$

Rearranging yields

$$S \leq \frac{3}{2}M + 2RL, \quad M \leq 2S + 4RL.$$

□

A.2 Information Theory

A.2.1 Properties of Hellinger Distance

Lemma A.5 (Hellinger distance for conditional distributions): For any distributions $\mathbb{P}$ and $\mathbb{Q}$ over a pair of random variables (X, Y) ,

$$\mathbb{E}_{X \sim \mathbb{P}_X} [D_{\mathbf{H}}^2(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X})] \leq 4D_{\mathbf{H}}^2(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}).$$

Proof of Lemma A.5. Since squared Hellinger distance is an f -divergence, we have

$$\mathbb{E}_{X \sim \mathbb{P}_X} [D_{\mathbf{H}}^2(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X})] = D_{\mathbf{H}}^2(\mathbb{P}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{P}_X).$$

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,

$$\begin{aligned} \mathbb{E}_{X \sim \mathbb{P}_X} [D_{\mathbf{H}}^2(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X})] &\leq 2D_{\mathbf{H}}^2(\mathbb{P}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{Q}_X) + 2D_{\mathbf{H}}^2(\mathbb{Q}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{Q}_X) \\ &= 2D_{\mathbf{H}}^2(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}) + 2D_{\mathbf{H}}^2(\mathbb{P}_X, \mathbb{Q}_X) \\ &\leq 4D_{\mathbf{H}}^2(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}), \end{aligned}$$

where the final line follows from the data processing inequality. □

A.2.2 Change-of-Measure Inequalities

The first inequality below is a sub-Gaussian variant of Pinsker's inequality; see Russo and Van Roy [182].

Lemma A.6 (Pinsker for sub-Gaussian random variables): Suppose that $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$ are both σ^2 -sub-Gaussian. Then

$$|\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[Y]| \leq \sqrt{2\sigma^2 \cdot D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q})}.$$

The next inequality is from Foster et al. [84].

Lemma A.7 (Multiplicative Pinsker-type inequality for Hellinger distance): Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathcal{F})$. For all $h : \mathcal{X} \rightarrow \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) \cdot D_{\text{H}}^2(\mathbb{P}, \mathbb{Q})}. \quad (\text{A.4})$$

In particular,

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leq 3\mathbb{E}_{\mathbb{Q}}[h(X)] + 4R \cdot D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}). \quad (\text{A.5})$$

Proof of Lemma A.7. Let a measurable event A be fixed. Let $p = \mathbb{P}(A)$ and $q = \mathbb{Q}(A)$. Then we have

$$\frac{(p - q)^2}{2(p + q)} \leq (\sqrt{p} - \sqrt{q})^2 \leq D_{\text{H}}^2((p, 1 - p), (q, 1 - q)) \leq D_{\text{H}}^2(\mathbb{P}, \mathbb{Q}),$$

where the third inequality is the data-processing inequality. It follows that

$$|p - q| \leq \sqrt{2(p + q)D_{\text{H}}^2(\mathbb{P}, \mathbb{Q})}.$$

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. $\square$

A.3 Minimax Theorem

We recall Sion's minimax theorem [190].

Lemma A.8 (Sion's Minimax Theorem): Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets in linear topological spaces, and assume $\mathcal{X}$ is compact. Let $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y). \quad (\text{A.6})$$