

Statistical Foundations of Reinforcement Learning:
A Non-Asymptotic Perspective

Yuxin Chen
University of Pennsylvania
yuxinc@wharton.upenn.edu

Yuejie Chi
Yale University
yuejie.chi@yale.edu

Jianqing Fan
Princeton University
jqfan@princeton.edu

Gen Li
Chinese University of Hong Kong
genli@cuhk.edu.hk

Yuting Wei
University of Pennsylvania
ytwei@wharton.upenn.edu

Yuling Yan
University of Wisconsin-Madison
yuling.yan@wisc.edu

July 12, 2026

Contents

1	Introduction	2
1.1	Statistical framework and scope	3
1.2	Towards non-asymptotic statistical foundations	5
1.3	Roadmap of this monograph	5
1.4	Relation to existing RL texts	6
1.5	Notation	7
2	Preliminaries	8
2.1	Discounted infinite-horizon MDPs	8
2.2	Finite-horizon episodic MDPs	17
2.3	Model-based vs. model-free approaches	21
2.4	Concentration inequalities	22
2.5	Minimax lower bounds for multi-hypothesis testing	25
2.6	Notes	27
3	RL with a generative model: the model-based approach	28
3.1	Generative model or simulator	28
3.2	Model-based policy evaluation	29
3.3	Model-based policy learning	37
3.4	Minimax lower bounds	48
3.5	Extension: finite-horizon nonstationary MDPs	56
3.6	Appendix A: proof of auxiliary lemmas in Section 3.3.4	57
3.7	Appendix B: tie-breaking via reward perturbation	58
3.8	Notes	59
4	RL with a generative model: the model-free approach	61
4.1	Prelude: stochastic approximation	61
4.2	Temporal difference learning	62
4.3	Q-learning	72
4.4	Extension: asynchronous Q-learning	87
4.5	Extension: finite-horizon nonstationary MDPs	92

4.6	Appendix A: proof of auxiliary lemmas in Section 4.2	94
4.7	Appendix B: proof of auxiliary lemmas in Section 4.3	95
4.8	Appendix C: proof of the lower bound in Theorem 4.6	99
4.9	Notes	115
5	Online RL	117
5.1	Problem formulation	117
5.2	A detour: the UCB algorithm for multi-armed bandits	120
5.3	A model-based online RL algorithm: UCB-VI	127
5.4	An improved model-based online RL algorithm: MVP	140
5.5	Model-free online RL	151
5.6	Minimax lower bounds	156
5.7	Appendix A: proof of auxiliary lemmas in Section 5.3.5	165
5.8	Appendix B: proof of auxiliary lemmas in Section 5.4.4	175
5.9	Appendix C: proof of auxiliary lemmas in Section 5.5.1	181
5.10	Appendix D: proof of Lemma 5.20	183
5.11	Notes	185
6	Offline RL	188
6.1	Problem formulation and challenges	188
6.2	A detour: the LCB algorithm for multi-armed bandits	191
6.3	Model-based offline RL	195
6.4	Model-free offline RL	207
6.5	Minimax lower bounds	214
6.6	Extension: finite-horizon nonstationary MDPs	215
6.7	Appendix A: proof of auxiliary lemmas in Section 6.3	218
6.8	Appendix B: proof of auxiliary lemmas in Section 6.4.3	229
6.9	Appendix C: proof of the lower bound in Theorem 6.9	232
6.10	Notes	238
7	Extensions	241
7.1	Distributionally robust RL	241
7.2	Competitive multi-agent RL	250
7.3	Federated RL	257
7.4	Statistical inference in RL	262
7.5	RL with human feedback	266
7.6	Notes	272
8	Concluding remarks	275
	Acknowledgements	279
	References	280

Abstract

Reinforcement learning (RL) embodies a foundational paradigm for sequential decision-making in the face of uncertainty. The rapid growth of model complexity in emerging applications has significantly amplified the challenges of achieving data-efficient RL, particularly in sample-starved scenarios where data collection may be costly, time-consuming, or even high-stakes. Consequently, understanding and improving the data efficiency of RL algorithms has become a central problem of both fundamental theoretical interest and practical importance.

Despite decades of research, the statistical underpinnings of RL remained remarkably underdeveloped until relatively recently, particularly in the finite-sample regimes most relevant to modern practice. Over the past few years, there has been a surge of activity advancing the frontiers of non-asymptotic statistical RL theory, drawing on ideas from high-dimensional statistics and probability, information theory, and statistical learning theory. In this monograph, we present a non-asymptotic treatment for RL based on tabular Markov decision processes, focusing on three canonical sampling mechanisms — RL with a generative model, online RL, and offline RL — together with two algorithmic paradigms: model-based and model-free approaches. Centered around non-asymptotic sample complexity analysis, our treatment emphasizes minimax-optimal algorithm design, particularly methods that attain statistical optimality without significant burn-in costs, as well as sharp statistical characterizations of several widely used RL algorithms. Whenever possible, we provide largely self-contained analyses together with intuitive explanations. We also briefly discuss several extensions, such as robust RL, competitive multi-agent RL, federated RL, uncertainty quantification, and RL from human feedback. We hope this monograph equips aspiring researchers with essential tools from modern statistical RL and inspires further advances in both theory and algorithm design.

1

Introduction

Reinforcement learning (RL) (Sutton and Barto, 2018; Bertsekas, 2019), commonly formulated as sequential decision-making and learning in a Markov decision process (MDP), has garnered immense attention in recent years due to its striking empirical success across a wide range of practical domains. By enabling agents to collect data and learn through repeated, trial-and-error interaction with their environments, RL has led to major breakthroughs in applications such as competitive gaming, robotics and control, autonomous driving, e-commerce, power systems, and healthcare, often achieving state-of-the-art performance in complex decision-making tasks (Silver *et al.*, 2017; Mnih *et al.*, 2015; Tang *et al.*, 2025; Kiran *et al.*, 2021; Hu *et al.*, 2018; Bastani *et al.*, 2021; Chen *et al.*, 2022c; Yu *et al.*, 2021; Bai *et al.*, 2025). Beyond these traditional application areas, RL has recently emerged as a cornerstone for fine-tuning and aligning large language models (LLMs). In particular, post-training techniques such as Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning with Verifiable Rewards (RLVR) (Guo *et al.*, 2025; Jaech *et al.*, 2024) have proven highly effective in enhancing instruction following, safety, and sophisticated reasoning capabilities, among other things, thereby positioning RL as a key enabler in the rapid advancement of modern generative artificial intelligence (AI).

A central objective of RL is to learn, from a collection of noisy data samples, a decision-making strategy (commonly referred to as a *policy* in RL) that approximately maximizes long-term rewards in an environment (often quantified through a *value function*), without direct access to an accurate description of the environment itself. In many modern applications, the state space (i.e., the collection of possible situations an agent may encounter in the environment), the action space (i.e., the set of actions available to the agent at each decision point), and the planning horizon can all be extraordinarily large. As a result, data collection must often be carried out at a substantial scale, which may be prohibitively expensive, time-consuming, or even high-stakes, as encountered in domains such as clinical trials, online advertising, and autonomous systems. These challenges are particularly pronounced in the “sample-starved” regime, where the intrinsic complexity of the environment vastly exceeds the amount of available data, making accurate model estimation infeasible.¹ Consequently, a fundamental goal of modern RL theory is to understand how

¹Here and throughout, the “model” refers collectively to the transition kernels and the reward functions of the

reliable learning can be achieved in such data-limited scenarios. Accomplishing this goal requires developing a precise quantitative understanding of the trade-off between statistical accuracy and sample complexity, as well as designing computationally efficient algorithms that provably attain optimal or near-optimal performance.

1.1 Statistical framework and scope

To systematically address the data-efficiency challenges outlined above, this monograph develops a coherent and rigorous statistical framework that accommodates a diverse range of settings. To help orient the reader, we briefly describe three key pillars underlying the topics studied throughout this monograph: the sampling mechanisms, the core algorithmic paradigms, and the performance metrics used to evaluate learning performance.

Data collection mechanisms. The statistical challenges inherent in RL are fundamentally shaped by the mechanism through which data are collected and made available to the agent. In this monograph, we study the sample complexity of RL under several data collection mechanisms commonly adopted in the literature, with primary emphasis on the following three canonical — and perhaps most widely studied — settings.

- *RL with a generative model.* When equipped with access to a generative model or simulator, the agent can query the environment in a fully controlled manner. More specifically, the agent may sample transitions from arbitrary state-action pairs at will, effectively eliminating the need for exploration and isolating the statistical challenges inherent to policy optimization and value estimation. See Chapters 3 and 4.
- *Online RL.* In online RL, the agent interacts with the environment sequentially in real time and must adaptively determine its actions based on past observations. Since data can only be gathered through the agent’s own interactions with the environment, one must balance *exploration* (i.e., acquiring new information about the unknown environment) and *exploitation* (i.e., leveraging current knowledge of the environment to improve rewards). This fundamental tension leads to the classical exploration-exploitation trade-off, which lies at the heart of online RL. See Chapter 5.
- *Offline RL.* In offline RL, the agent is given access only to a static dataset of historical samples — typically logged by one or more (possibly unknown) behavior policies — and needs to learn without any further interaction with the environment. As a result, the agent faces fundamental challenges arising from issues such as limited data coverage and distributional mismatch between the available data and the target policy to be learned. See Chapter 6.

Algorithmic paradigms. Broadly speaking, there are at least two principal algorithmic paradigms in RL: *model-based* approaches and *model-free* approaches.

- Model-based algorithms typically decouple model estimation from policy learning. Namely, one first leverages the available data samples to estimate the unknown environment model and subsequently performs policy optimization by treating the learned model as if it were

Markov decision process describing the underlying environment, to be formalized in Chapter 2.

accurate and applying classical dynamic programming techniques dating back to Richard Bellman (Bellman, 1952; Bertsekas, 2017). A notable advantage of model-based methods is their flexibility: once a model has been learned, it can often be reused to solve new downstream tasks without requiring additional interaction with the environment or revisiting the original data samples.

- In contrast, model-free algorithms seek to learn the optimal value function or the optimal policy directly from data samples, bypassing explicit estimation of the environment’s underlying dynamics and rewards. This family of algorithms is particularly appealing in scenarios where constructing or storing an environment model is difficult, or where the environment itself may evolve over time. Furthermore, model-free methods are often more memory-efficient, as they obviate the need to store an explicit — often high-dimensional — representation of the full environment model.

Characterizing the sample efficiency of both model-based and model-free approaches has been a central focus of a vast body of recent research, and constitutes a primary theme of this monograph. As a preview, despite their higher memory and computational costs, model-based algorithms have been shown to attain statistically optimal performance across all three canonical sampling mechanisms outlined above, even in the most sample-starved regimes. In contrast, many classical model-free methods, such as Q-learning, fail to attain optimal sample complexity guarantees. More broadly, whether one can design model-free methods that achieve statistical optimality across the full spectrum of problem regimes — without sacrificing the memory efficiency that makes them attractive in practice — remains a fundamental question that has yet to be fully resolved.

Evaluation metrics. In this monograph, we evaluate the performance of RL algorithms primarily through the lens of *sample complexity*, defined within the classical *Probably Approximately Correct (PAC)* learning framework as the minimum number of samples required to guarantee, with high probability, that the statistical error falls below a prescribed threshold. Depending on the problem goal, the notion of statistical error can be tailored to the specific object of interest. For value estimation, we assess the accuracy of the estimated value function, often measured under the ℓ_∞ norm to ensure uniform precision across all initial states. For policy learning, we instead measure the performance gap between the value function induced by the learned policy and that induced by the target policy. More broadly, such sample complexity guarantees provide a precise, quantitative characterization of how statistical accuracy improves as additional data become available.

Another performance metric widely used in online settings is (cumulative) *regret*, which measures the gap between the cumulative reward that could have been achieved by following an optimal policy and the cumulative reward actually obtained by the learning algorithm over time. It is well known that regret guarantees can often be converted into PAC-style sample complexity guarantees, a connection that we will elaborate on in Chapter 5.1.

Finally, we also briefly discuss the task of uncertainty quantification (see Chapter 7.4), with the goal of assessing the uncertainty associated with estimated quantities. A common objective is to construct confidence intervals that are both statistically valid and as tight as possible.

1.2 Towards non-asymptotic statistical foundations

Despite decades of extensive research, the statistical foundations of RL in finite-sample and finite-time regimes — which carry high operational value in modern practice — remained significantly underdeveloped until the past one to two decades. Early theoretical analyses of core RL algorithms — such as temporal difference learning and Q-learning — were conducted primarily in the asymptotic regime, where the number of samples or iterations tends to infinity while all other problem parameters are held fixed. Although this line of work provided important insights into the convergence behavior of RL algorithms, it offers limited guidance on their data efficiency in modern large-scale applications. As discussed previously, many contemporary RL problems have to deal with enormous state-action spaces and long planning horizons. Treating these problem parameters as fixed constants can therefore fail to capture the statistical bottlenecks in high dimensions.

To overcome these limitations, the past several years have witnessed a flurry of activity aimed at strengthening finite-sample, finite-time theory for RL. These advances draw heavily on ideas from modern development in high-dimensional statistics, high-dimensional probability, information theory, and statistical learning theory. Yet, despite this shift toward non-asymptotic analysis, a complete understanding of the fundamental sample complexity limits remained largely elusive, even for the three canonical data collection mechanisms introduced above. As we will discuss in subsequent chapters, much of the prior literature was, at best, only able to establish sharp or optimal statistical guarantees when the available sample size exceeded an enormous threshold, leaving the sample-starved regime poorly understood until fairly recently. Indeed, even today, sharp and complete statistical characterizations of the performance of RL algorithms are available only in relatively basic settings, while many important problems and settings remain far from fully resolved.

Notably, RL in high-dimensional, sample-limited settings poses significant technical hurdles that call for both refined analysis techniques and new algorithmic ideas. In some sense, the paradigm shift from asymptotic to non-asymptotic RL theory is reminiscent of the rise of high-dimensional statistics in recent decades, where classical large-sample theory proved insufficient and ultimately motivated the development of fundamentally new mathematical techniques and algorithmic paradigms throughout modern statistics.

1.3 Roadmap of this monograph

In this monograph, we aim to present a coherent, non-asymptotic framework that covers core algorithmic principles as well as cutting-edge theoretical developments. In particular, we seek to provide sharp sample complexity characterizations for several RL algorithms, benchmarked against minimax lower bounds to assess their statistical optimality or suboptimality.

The bulk of this monograph is organized around the three classical sampling mechanisms introduced in Chapter 1.1. Specifically, we devote one or two chapters to each of these settings — namely, Chapters 3 and 4 for the generative model, Chapter 5 for online RL, and Chapter 6 for offline RL — where we present optimal or widely adopted algorithms alongside sharp statistical guarantees and minimax lower bounds. For major results, we provide largely self-contained proofs, often supplemented by intuitive intermediate analysis arguments that help deepen the reader’s theoretical understanding. Finally, we also move beyond these canonical single-agent formulations by briefly introducing several extensions of growing importance in Chapter 7, including distributionally robust RL, multi-agent RL, federated RL, as well as basic problems in statistical inference and

uncertainty quantification.

To develop these statistical foundations in the cleanest possible settings, we focus predominantly on tabular MDPs, where the state and action spaces are finite, and no low-dimensional function approximation is assumed. Notably, despite decades of study, the fundamental minimax sample complexity limits for tabular MDPs under the three canonical sampling mechanisms remained significantly unresolved for a long time and have only been settled — in several major, though not all, MDP models — very recently. We believe that obtaining a comprehensive understanding of these tabular fundamental limits constitutes an indispensable first step toward elucidating and ultimately achieving optimal statistical efficiency in RL at large, particularly in settings involving function approximation.

1.4 Relation to existing RL texts

Several existing books and monographs are particularly relevant to, or complementary with, the present monograph, and we briefly discuss a few of them here.

For introductory material, the foundational textbooks Sutton and Barto (2018), Bertsekas (2019), and Szepesvári (2010) provide excellent and comprehensive introductions to RL for readers with little or no prior background, albeit with much less emphasis on modern statistical foundations. The special cases of multi-armed bandits and contextual bandits have been covered in several books (e.g., Bubeck and Cesa-Bianchi (2012), Slivkins (2019), and Lattimore and Szepesvári (2020)), with Lattimore and Szepesvári (2020) offering perhaps the most comprehensive and up-to-date treatment of modern bandit theory. The monograph Agarwal *et al.* (2019a) has become a widely adopted resource for researchers entering RL theory, owing to its broad coverage of contemporary topics. In comparison, a substantial portion of the present monograph focuses on statistical advances in tabular RL developed after Agarwal *et al.* (2019a) first became available. For instance, fundamental questions concerning how to characterize and attain sharp statistical limits in RL with a generative model, online RL, and offline RL, as well as how to establish sharp sample complexity guarantees for Q-learning, were all resolved only subsequently.

We also note the monograph by Foster and Rakhlin (2023), which develops a statistical perspective on interactive decision-making, with a primary focus on the general framework of function approximation. In addition, Zhang (2023) contains chapters on bandits and RL that briefly discuss several modern topics, including linear mixture MDPs and the Bellman–Eluder dimension. The recent book Guo *et al.* (2026) offers a comprehensive introduction to modern development of continuous RL from the perspective of stochastic control. The emphasis of these three texts, however, differs drastically from that of the present monograph, as none of them centers on the state-of-the-art statistical development in tabular RL.

Moreover, the present monograph grew out of several tutorials and courses taught by some of the authors in recent years, which also led to lecture notes such as Chi *et al.* (2025b). In comparison, Chi *et al.* (2025b) offers a bird’s-eye view of the recent statistical and algorithmic advances in RL, including developments in policy optimization, whereas the present monograph focuses more squarely on the statistical foundations of RL, offering a thorough and in-depth treatment accompanied by largely self-contained mathematical proofs.

1.5 Notation

For clarity and consistency, we introduce below a set of notation and conventions to be used throughout this monograph. Additional notation specific to individual chapters will be introduced later as needed.

Asymptotic notation. Let $\mathcal{X} := (S, A, \frac{1}{1-\gamma}, H, \frac{1}{\varepsilon}, \frac{1}{\delta})$, where the meanings of these parameters will become clear shortly. The notation $f(\mathcal{X}) = O(g(\mathcal{X}))$ or $f(\mathcal{X}) \lesssim g(\mathcal{X})$ means there exists a universal constant $C_1 > 0$ such that $f(\mathcal{X}) \leq C_1 g(\mathcal{X})$. Similarly, $f(\mathcal{X}) = \Omega(g(\mathcal{X}))$ or $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ means that $g(\mathcal{X}) = O(f(\mathcal{X}))$, while $f(\mathcal{X}) = \Theta(g(\mathcal{X}))$ or $f(\mathcal{X}) \asymp g(\mathcal{X})$ indicates that both $f(\mathcal{X}) = O(g(\mathcal{X}))$ and $g(\mathcal{X}) = O(f(\mathcal{X}))$ hold simultaneously. Additionally, the notation $\tilde{O}(\cdot)$ (resp. $\tilde{\Omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$) is defined analogously to $O(\cdot)$ (resp. $\Omega(\cdot)$ and $\Theta(\cdot)$) but with logarithmic factors suppressed.

Vectors and matrices. For any vector $a = [a_i]_{1 \leq i \leq n} \in \mathbb{R}^n$, we overload several scalar operators in an entrywise manner. For instance, we write

$$\sqrt{a} := [\sqrt{a_i}]_{1 \leq i \leq n}, \quad |a| := [|a_i|]_{1 \leq i \leq n}, \quad \text{and} \quad a^2 := [a_i^2]_{1 \leq i \leq n}.$$

For vectors $a = [a_i]_{1 \leq i \leq n}$ and $b = [b_i]_{1 \leq i \leq n}$, the notation $a \geq b$ (resp. $a \leq b$) means that $a_i \geq b_i$ (resp. $a_i \leq b_i$) for all $1 \leq i \leq n$. We use

$$a \circ b := [a_i b_i]_{1 \leq i \leq n}$$

to represent the Hadamard product of a and b . We also let $\|A\|$ and $\text{Tr}(A)$ denote the spectral norm and the trace of a matrix A , respectively.

For any finite set $\mathcal{S} = \{1, \dots, S\}$, we let $\Delta(\mathcal{S}) = \{p \in \mathbb{R}^S \mid p \geq 0, 1^\top p = 1\}$ represent the set of probability vectors (i.e., the probability simplex) over $\mathcal{S}$. For any matrix P , we define its ℓ_1 norm (i.e., the maximum row-wise ℓ_1 norm) as

$$\|P\|_1 := \max_i \sum_j |P_{i,j}|. \quad (1.1)$$

Probability divergences. For any two probability measures P and Q defined on the same measurable space $(\Omega, \mathcal{F})$, the total-variation (TV) distance between them is defined as

$$\text{TV}(P, Q) := \sup_{A \subseteq \mathcal{F}} |P(A) - Q(A)|. \quad (1.2a)$$

When P is absolutely continuous w.r.t. Q , the Kullback-Leibler (KL) divergence and χ^2 divergence between them are defined respectively by

$$\text{KL}(P \parallel Q) := \int \left(\log \frac{dP}{dQ} \right) dP, \quad (1.2b)$$

$$\chi^2(P \parallel Q) := \int \left(\frac{dP}{dQ} - 1 \right)^2 dQ. \quad (1.2c)$$

These divergences play a central role in mathematical statistics, information theory, and probability theory as measures of the “statistical distance” between probability distributions; see, e.g., Tsybakov (2009) and Polyanskiy and Wu (2025), for in-depth discussions.

Miscellaneous. For any integer n , we denote $[n] := \{1, \dots, n\}$. We use e_i to represent the i -th standard basis vector. The notation $\mathbb{1}\{\cdot\}$ stands for the indicator function.

2

Preliminaries

We begin by reviewing the fundamentals of Markov decision processes (MDPs), the canonical framework for developing and analyzing reinforcement learning algorithms. This framework sets the stage for the discussions in the chapters that follow. We focus on two widely studied formulations: discounted infinite-horizon MDPs and finite-horizon episodic MDPs. Additionally, we present core concentration inequalities and techniques for deriving minimax lower bounds for multi-hypothesis testing, which serve as the primary analytical tools for this work.

2.1 Discounted infinite-horizon MDPs

We first introduce the basics and notation for discounted infinite-horizon MDPs. More comprehensive introductions can be found in classical references such as Puterman (2014) and Bertsekas (2017).

2.1.1 Basic components

A discounted infinite-horizon MDP can be represented by a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$, which models an agent’s sequential interaction with its environment over an unbounded time horizon (see Figure 2.1 for an illustration). Its key components are:

- (i) $\mathcal{S} = \{1, \dots, S\}$: a finite state space consisting of S distinct states. A state can be used to describe, for example, the agent’s current situation (e.g., the position of a self-driving vehicle, the health condition of a patient in a treatment model).
- (ii) $\mathcal{A} = \{1, \dots, A\}$: an action space consisting of A distinct actions available to the agent (e.g., the direction of movement in a navigation task, the choice of medication to prescribe to a patient).
- (iii) $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$: the transition kernel of the MDP, where $P(\cdot | s, a)$ specifies the probability distribution over the next state when action a is executed in state s .
- (iv) $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$: a deterministic reward function, where $r(s, a)$ indicates the immediate reward received by the agent for taking action a in state s . Rewards are normalized to lie

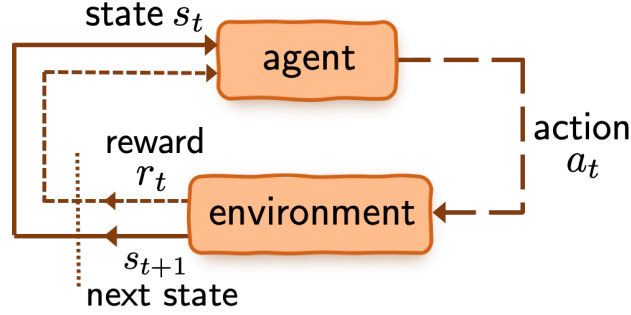


Figure 2.1: An illustration of an MDP, where an agent interacts sequentially with the environment. At time step t , the agent observes state s_t of the environment, selects action a_t , and receives an immediate reward r_t . The environment then transitions to state s_{t+1} according to the transition kernel P .

within $[0, 1]$ without loss of generality. To streamline presentation, we assume throughout this monograph that the reward function is fully known to the learner *a priori*. Note, however, that this assumption is not essential and can often be substantially relaxed.

- (v) $\gamma \in [0, 1)$: the discount factor, which guarantees that the infinite-horizon cumulative discounted reward is always finite (as we shall see shortly). The quantity $\frac{1}{1-\gamma}$ is commonly referred to as the *effective horizon* of $\mathcal{M}$: it is proportional to the number of steps needed for a reward to be discounted to half its value, and hence reflects the time scale over which future rewards remain significant.

Here and throughout, we assume the discounted infinite-horizon MDP to be *stationary* (or *time-homogeneous*), meaning that both the transition kernel P and the reward function r are time-invariant. In addition, we introduce the following notation for convenience:

$$P_{s,a} := P(\cdot | s, a) \in \mathbb{R}^{1 \times S}. \quad (2.1)$$

It is noteworthy that this monograph focuses on regimes in which the number of states S , the number of actions A , and the effective horizon $\frac{1}{1-\gamma}$ are all large. This emphasis reflects a common challenge arising in modern RL practice, where algorithms must contend with vast state–action spaces and long horizons.

2.1.2 Policy, value function, Q-function, and occupancy distribution

At each time step, the agent needs to select actions based on the current state of the environment. An action selection rule is commonly referred to as a policy in RL. Formally speaking, a stationary policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ is a mapping that specifies, for each state $s \in \mathcal{S}$, a probability distribution over actions; namely, $\pi(a | s)$ indicates the probability of selecting action a in state s . Clearly, this formulation accommodates randomized action selection. When π is a deterministic policy, we often abuse the notation by letting $\pi(s)$ represent the action selected by policy π in state s . We further denote $P_\pi : \mathcal{S} \rightarrow \Delta(\mathcal{S})$ the equivalent transition probability when action a is drawn from $\pi(\cdot | s)$, namely,

$$\forall s \in \mathcal{S} : \quad P_\pi(\cdot | s) := \mathbb{E}_{a \sim \pi(\cdot | s)} [P(\cdot | s, a)]. \quad (2.2)$$

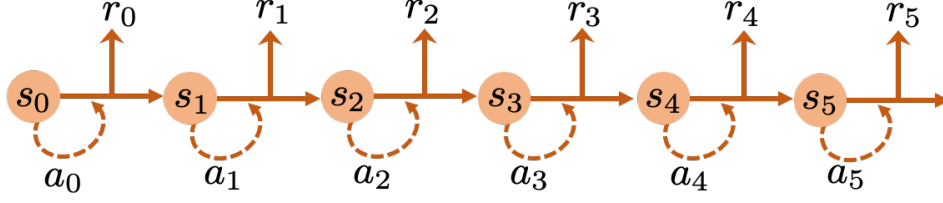


Figure 2.2: An illustration of a trajectory rolled out by an MDP, where $r_t = r(s_t, a_t)$ is the immediate reward gained at time step t .

As a remark, under the stationary assumption of the infinite-horizon MDPs considered herein, it is typically sufficient to restrict attention to time-invariant policies.

To assess the goodness of a policy, we need to introduce certain “utility functions”, which in RL are most commonly taken to be the value function and the Q-function. Consider first a sample trajectory $\{(s_t, a_t)\}_{t \geq 0}$ induced by the MDP under policy π , which governs what action a_t to take after observing the state s_t at time step t (see Figure 2.2 for an illustration). Note that the trajectory depends on the policy π as the choice of action a_t affects the next state s_{t+1} . We can then define the value function and the Q-function, formalizing how a policy’s performance is measured.

- *Value function.* The value function $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$ of policy π is defined as the expected discounted cumulative reward:

$$\forall s \in \mathcal{S} : \quad V^\pi(s) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right], \quad (2.3)$$

where the expectation is over the sample trajectory $\{(s_t, a_t)\}_{t \geq 0}$ generated under policy π , with $a_t \sim \pi(\cdot \mid s_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ for all $t \geq 0$, starting from an initial state $s_0 = s$. Intuitively, $V^\pi(s)$ represents the long-term reward an agent can expect when starting from state s and following policy π , with future rewards geometrically discounted over time. Since immediate rewards are assumed to lie in $[0, 1]$, it follows that

$$0 \leq V^\pi(s) \leq \frac{1}{1 - \gamma}$$

for every policy π . Moreover, if the initial state is randomly drawn from some state distribution $\rho \in \Delta(\mathcal{S})$, then we can also define the value function of policy π under initial state distribution ρ as:

$$V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]. \quad (2.4)$$

- *Q-function.* Analogously, the Q-function (or action-value function) $Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ of policy π can be defined such that: for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$Q^\pi(s, a) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]. \quad (2.5)$$

In contrast to the value function, the Q-function conditions on not only the initial state $s_0 = s$, but also on the initial action $a_0 = a$, with subsequent actions drawn from π . It therefore

quantifies the expected long-term return by first taking action a in state s and then following policy π thereafter. This function provides a natural way to compare the relative benefits of different actions in a given state. The intimate connection between the Q-function and the value function is reflected in the following identity:

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot | s)} [Q^\pi(s, a)]. \quad (2.6)$$

As is well known, there exists at least one deterministic policy, denoted by π^* throughout, that is optimal in the sense that it simultaneously maximizes $V^\pi(s)$ and $Q^\pi(s, a)$ for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ (see, e.g., Bertsekas (2017)). We adopt the following shorthand notation for the corresponding optimal value function and optimal Q-function:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad V^*(s) := V^{\pi^*}(s) \quad \text{and} \quad Q^*(s, a) := Q^{\pi^*}(s, a). \quad (2.7)$$

Next, we find it helpful to also introduce the *discounted occupancy distributions* associated with policy π , quantifying the frequencies with which each state (or state-action pair) is visited under policy π . Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, define the state occupancy distribution $d^\pi(s; \rho)$ and the state-action occupancy distribution $d^\pi(s, a; \rho)$ such that: for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$d^\pi(s; \rho) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0 \sim \rho), \quad (2.8a)$$

$$d^\pi(s, a; \rho) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a \mid s_0 \sim \rho), \quad (2.8b)$$

where the sample trajectory starts from $s_0 \sim \rho$ and follows policy π (i.e., $a_t \sim \pi(\cdot | s_t)$ and $s_{t+1} \sim P(\cdot | s_t, a_t)$ for all $t \geq 0$). In words, these functions describe the visitation frequencies of states and state-action pairs — with each visit discounted geometrically over time — under a policy over an infinite horizon. The prefactor $1 - \gamma$ is the reciprocal of the total weight, ensuring that these distributions are normalized.

Correspondingly, we adopt the notation of the discounted occupancy distributions associated with the optimal deterministic policy π^* , denoted by d^* , as follows:

$$d^*(s; \rho) := d^{\pi^*}(s; \rho) \quad \text{and} \quad (2.9a)$$

$$d^*(s, a; \rho) := d^{\pi^*}(s, a; \rho) = d^*(s; \rho) \mathbb{1}(a = \pi^*(s)) \quad (2.9b)$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, where the last equality in (2.9b) is valid since π^* is assumed to be deterministic. Here, we dropped the dependency on ρ for notational simplicity.

2.1.3 Bellman operators and Bellman equations

Central to the design and analysis of reinforcement learning algorithms are the Bellman operators and the Bellman equations, named after Richard Bellman (Bellman, 1952), which we now introduce.

Bellman consistency operator. Given a policy π , the Bellman consistency operator $\mathcal{T}^\pi : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \rightarrow \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ is defined such that, for any input $Q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, the (s, a) -th entry of the output $\mathcal{T}^\pi(Q)$ is given by

$$\mathcal{T}^\pi(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[\mathbb{E}_{a' \sim \pi(\cdot | s')} [Q(s', a')] \right]. \quad (2.10)$$

Intuitively, the operator $\mathcal{T}^\pi$ performs a one-step lookahead: it evaluates a state–action pair (s, a) by adding the immediate reward $r(s, a)$ to the discounted value of the next state–action pair, assuming that future actions are drawn from policy π .

It follows directly that the Q-function of policy π (cf. (2.5)) is a fixed point of this Bellman consistency operator, satisfying

$$Q^\pi = \mathcal{T}^\pi(Q^\pi). \quad (2.11)$$

The solution is also unique (see Bertsekas (2017) and also (2.15)). Combined with the relation (2.6) between V^π and Q^π , this further implies that the value function satisfies

$$V^\pi = \mathcal{T}_\pi(V^\pi), \quad (2.12a)$$

where the operator $\mathcal{T}_\pi : \mathbb{R}^S \rightarrow \mathbb{R}^S$ is defined such that for any $V \in \mathbb{R}^S$, the s -th entry of $\mathcal{T}_\pi(V)$ is given by

$$\mathcal{T}_\pi(V)(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi)} [V(s')]. \quad (2.12b)$$

Since $\mathcal{T}^\pi(Q)$ (resp. $\mathcal{T}_\pi(V)$) is a linear operator on Q (resp. V), solving the system of linear equations defined in (2.11) (resp. (2.12a)) allows for computation of the Q-function (resp. value function) for a given policy, provided that the transition kernel P and the reward function r are known *a priori*. This viewpoint is fundamental in RL for evaluating a policy's performance.

Bellman optimality operator. We now introduce another fundamental operator, the Bellman optimality operator $\mathcal{T}$, which is also a mapping from $\mathbb{R}^{S \times A}$ to itself. Concretely, given any input $Q \in \mathbb{R}^{S \times A}$, the (s, a) -th entry of the output $\mathcal{T}(Q)$ is defined as

$$\mathcal{T}(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[\max_{a' \in \mathcal{A}} Q(s', a') \right]. \quad (2.13)$$

In words, $\mathcal{T}(Q)(s, a)$ combines the immediate reward with the discounted expected return of the best possible action in the next state s' . A fundamental property of the Bellman optimality operator is that the optimal Q-function Q^* (cf. (2.7)) is its unique fixed point (Bertsekas, 2017), satisfying

$$Q^* = \mathcal{T}(Q^*). \quad (2.14)$$

As we shall see momentarily, this celebrated result underpins the design of various algorithms like Q-learning.

Contraction of Bellman operators. A remarkable property of the above Bellman operators is γ -contraction: specifically, for any $Q_1, Q_2 \in \mathbb{R}^{S \times A}$, it holds that (Bertsekas, 2017)

$$\|\mathcal{T}^\pi(Q_1) - \mathcal{T}^\pi(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty, \quad (2.15a)$$

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty. \quad (2.15b)$$

In other words, applying either the Bellman consistency operator or the Bellman optimality operator brings two Q-functions strictly closer under the ℓ_∞ norm. This contraction property is fundamental: it ensures the uniqueness of their fixed points and enables rapid convergence of many iterative algorithms in dynamic programming and RL.

Proof of γ -contraction (2.15). Regarding $\mathcal{T}^\pi$, recall the basic fact

$$|\mathbb{E}[Q_1(s, a) - Q_2(s, a)]| \leq \max_{s', a'} |Q_1(s', a') - Q_2(s', a')|$$

for any (s, a) . This combined with the definition (2.10) yields

$$\begin{aligned} |\mathcal{T}^\pi(Q_1)(s, a) - \mathcal{T}^\pi(Q_2)(s, a)| &= \gamma |\mathbb{E}[Q_1(s', a') - Q_2(s', a')]| \\ &\leq \gamma \|Q_1 - Q_2\|_\infty \end{aligned}$$

for any (s, a) pair, where the expectation is over $s' \sim P(\cdot | s, a)$ and $a' \sim \pi(\cdot | s')$. This establishes the γ -contraction property of $\mathcal{T}^\pi$.

The contraction for $\mathcal{T}$ can be proven by repeating the above arguments along with the basic fact that

$$|\max_a Q_1(s, a) - \max_a Q_2(s, a)| \leq \max_a |Q_1(s, a) - Q_2(s, a)|.$$

Details are omitted here for brevity. $\square$

2.1.4 Dynamic programming under full model specification

With the Bellman operators in place, we are now ready to introduce the classical dynamic programming algorithms for solving discounted infinite-horizon MDPs, assuming a fully known model (i.e., both the transition kernel P and the rewards r are available to the learner). The interested reader is referred to Bertsekas (2017) for in-depth discussions.

Planning. Assuming access to exact model specification (i.e., P and r), we describe two classical algorithms for computing the optimal Q-function Q^* and the optimal policy π^* , a task commonly referred to as *planning* (see Figure 2.3).

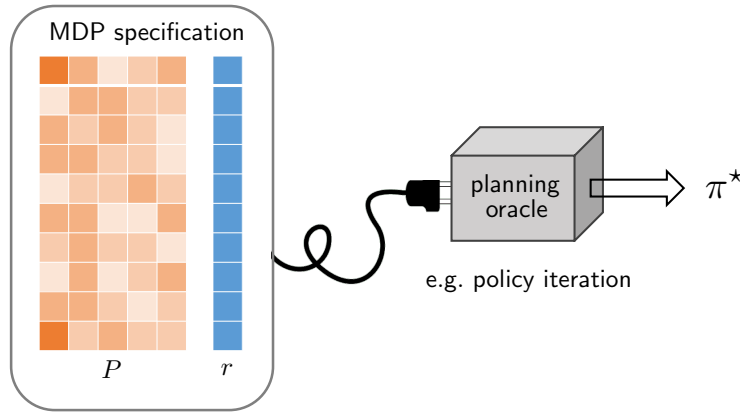


Figure 2.3: Planning in an MDP with a fully specified model. Classical dynamic algorithms, such as policy iteration and value iteration, can efficiently compute the optimal policy π^* and the associated optimal value given complete knowledge of P and r .

- *Value iteration (VI).* Initialize the Q-function estimate Q_0 . For each iteration $k = 1, 2, \dots$, update the Q-function as

$$Q_k = \mathcal{T}(Q_{k-1}), \tag{2.16}$$

with $\mathcal{T}$ the Bellman optimality operator (cf. (2.13)). In words, the Q-function estimate is refined iteratively by repeatedly applying the Bellman optimality operator to the latest Q-function estimate.

- *Policy iteration (PI)*. Initialize the policy estimate π_0 . For each iteration $k = 1, 2, \dots$, update the policy as

$$\forall s \in \mathcal{S} : \quad \pi_k(s) = \arg \max_{a \in \mathcal{A}} Q^{\pi_{k-1}}(s, a). \quad (2.17)$$

Namely, the policy is updated by acting greedily w.r.t. the Q-function of the current policy estimate.

Encouragingly, both algorithms achieve fast linear convergence to the optimal Q-function, as asserted by the following lemma.

Lemma 2.1 (Linear convergence of VI and PI). Both value iteration (cf. (2.16)) and policy iteration (cf. (2.17)) achieve

$$\|Q_k - Q^*\|_\infty \leq \gamma^k \|Q_0 - Q^*\|_\infty$$

for every $k \geq 1$, whereas in policy iteration we take $Q_k = Q^{\pi_k}$.

Proof. In light of the fixed-point property (2.14) and γ -contraction (2.15) of $\mathcal{T}$, the VI iterates obey

$$\|Q_k - Q^*\|_\infty = \|\mathcal{T}(Q_{k-1}) - \mathcal{T}(Q^*)\|_\infty \leq \gamma \|Q_{k-1} - Q^*\|_\infty.$$

Applying this inequality iteratively completes the proof for VI. The proof for PI is omitted here and can be found in Agarwal *et al.*, 2019a. $\square$

This lemma reveals that, to achieve ε -accuracy in the sense that $\|Q_k - Q^*\|_\infty \leq \varepsilon$, it takes no more than

$$\frac{1}{1-\gamma} \log \left(\frac{\|Q_0 - Q^*\|_\infty}{\varepsilon} \right)$$

iterations. Importantly, this iteration complexity is on the order of the effective horizon $\frac{1}{1-\gamma}$ (up to some logarithmic factor), irrespective of the dimension of the state-action space.

2.1.5 Vector and matrix notation

It is also convenient to introduce several pieces of vector and matrix notation, which allow our results and analysis to be presented more compactly. Here and throughout, we allow a slight abuse of notation by letting the same symbols represent both vectors/matrices and functions, whenever it is clear from context (e.g., the notation r may denote either the reward function or the corresponding reward vector).

- $r \in \mathbb{R}^{SA}$: a vector representing the reward function, so that its (s, a) -th entry is the corresponding immediate reward.
- $V^\pi, V^* \in \mathbb{R}^S$: vectors representing respectively the value function of policy π (cf. (2.3)) and the optimal value function (cf. (2.7)).
- $Q^\pi, Q^* \in \mathbb{R}^{SA}$: vectors representing respectively the Q-function Q^π of policy π (cf. (2.5)) and the optimal Q-function (cf. (2.7)).

- $P \in \mathbb{R}^{SA \times S}$: a matrix representing the transition kernel, where its (s, a) -th row is a probability vector representing $P(\cdot | s, a)$.
- $\pi(\cdot | s) \in \mathbb{R}^{1 \times A}$: a row vector $[\pi(a | s)]_{1 \leq a \leq A}$ representing the action distribution induced by policy π in state s .
- $\Pi^\pi \in \mathbb{R}^{S \times SA}$: a matrix associated with policy π taking the following form

$$\Pi^\pi = \begin{pmatrix} \pi(\cdot | 1) & & & \\ & \pi(\cdot | 2) & & \\ & & \ddots & \\ & & & \pi(\cdot | S) \end{pmatrix}. \quad (2.18)$$

- $P^\pi \in \mathbb{R}^{SA \times SA}$ and $P_\pi \in \mathbb{R}^{S \times S}$: two *square* probability transition matrices induced by policy π over the state-action pairs and the states, respectively, defined by

$$P^\pi := P \Pi^\pi \quad \text{and} \quad P_\pi := \Pi^\pi P. \quad (2.19)$$

- $r_\pi \in \mathbb{R}^S$: a reward vector associated with policy π ; namely,

$$r_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot | s)} [r(s, a)] \quad \text{for all } s \in \mathcal{S}, \quad (2.20)$$

or simply, $r_\pi = \Pi^\pi r$.

We further introduce several variance-related objects, which shall be used frequently in our statistical analysis. For any vector $V = [V_i]_{1 \leq i \leq S} \in \mathbb{R}^S$, we define the vector $\text{Var}_P(V) \in \mathbb{R}^{SA}$, whose (s, a) -th entry is given by

$$\begin{aligned} [\text{Var}_P(V)]_{(s,a)} &= \text{Var}_{P_{s,a}}(V) \\ &= \sum_{s' \in \mathcal{S}} P(s' | s, a) (V_{s'})^2 - \left(\sum_{s' \in \mathcal{S}} P(s' | s, a) V_{s'} \right)^2; \end{aligned} \quad (2.21)$$

that is, the variance of $V_{s'}$ when s' is sampled from the probability vector $P(\cdot | s, a)$. This can be expressed compactly via matrix notation:

$$\text{Var}_P(V) = P(V \circ V) - (PV) \circ (PV), \quad (2.22)$$

where we recall that $a \circ b$ denotes the Hadamard product $[a_i b_i]_{1 \leq i \leq S}$ (see Section 1.5). Similarly, we define, for any given policy π ,

$$\text{Var}_{P_\pi}(V) = P_\pi(V \circ V) - (P_\pi V) \circ (P_\pi V) \in \mathbb{R}^S. \quad (2.23)$$

The following lemma provides a perturbation bound for $\text{Var}_P(V)$.

Lemma 2.2. For any vectors $V, \hat{V} \in \mathbb{R}^S$ obeying $0 \leq V \leq \frac{1}{1-\gamma} 1_S$ and $0 \leq \hat{V} \leq \frac{1}{1-\gamma} 1_S$, one has

$$\|\text{Var}_P(\hat{V}) - \text{Var}_P(V)\|_\infty \leq \frac{4}{1-\gamma} \|\hat{V} - V\|_\infty.$$

Proof. From (2.22), we have the decomposition

$$\begin{aligned}\text{Var}_P(\hat{V}) - \text{Var}_P(V) &= P(\hat{V} \circ \hat{V} - V \circ V) - (P\hat{V}) \circ (P\hat{V}) + (PV) \circ (PV) \\ &= P((\hat{V} - V) \circ (\hat{V} + V)) - (P(\hat{V} - V)) \circ (P(\hat{V} + V)).\end{aligned}$$

Hence, it follows that

$$\begin{aligned}\|\text{Var}_P(\hat{V}) - \text{Var}_P(V)\|_\infty &\leq \|P\|_1 \|\hat{V} - V\|_\infty \|\hat{V} + V\|_\infty + \|P\|_1^2 \|\hat{V} - V\|_\infty \|\hat{V} + V\|_\infty \\ &\leq \frac{4}{1-\gamma} \|\hat{V} - V\|_\infty,\end{aligned}$$

where the last line is valid since $\|\hat{V}\|_\infty, \|V\|_\infty \leq \frac{1}{1-\gamma}$ and $\|P\|_1 = 1$. $\square$

2.1.6 Additional basic properties

Equipped with the above notation, we present a couple of basic properties that will be useful throughout.

To begin, the Bellman consistency equation (2.11) can be expressed in vector/matrix form as

$$Q^\pi = r + \gamma P V^\pi, \quad (2.24a)$$

which, when combined with (2.6), yields

$$V^\pi = \Pi^\pi Q^\pi = r_\pi + \gamma P_\pi V^\pi. \quad (2.24b)$$

Solving (2.24b) for V^π immediately gives

$$V^\pi = (I - \gamma P_\pi)^{-1} r_\pi. \quad (2.24c)$$

Substituting (2.24b) into (2.24a) and solving for Q^π yield

$$Q^\pi = r + \gamma P \Pi^\pi Q^\pi = r + \gamma P^\pi Q^\pi \implies Q^\pi = (I - \gamma P^\pi)^{-1} r. \quad (2.24d)$$

In addition, the Bellman optimality condition (2.14) admits a vector/matrix form as

$$Q^* = r + \gamma P V^*. \quad (2.25)$$

The careful reader would immediately remark that, for (2.24c) to hold, the matrix $I - \gamma P_\pi$ must be invertible. Fortunately, this requirement is automatically satisfied whenever $0 \leq \gamma < 1$. A few basic properties of this inversion are provided in the lemma below.

Lemma 2.3. For any $0 \leq \gamma < 1$ and any policy π , it holds that

$$(I - \gamma P_\pi)^{-1} = I + \sum_{i=1}^{\infty} (\gamma P_\pi)^i, \quad (2.26a)$$

$$(I - \gamma P_\pi)^{-1} \mathbf{1}_S = \frac{1}{1-\gamma} \mathbf{1}_S, \quad (2.26b)$$

$$\|(I - \gamma P_\pi)^{-1}\|_1 = \frac{1}{1-\gamma}, \quad (2.26c)$$

where $\mathbf{1}_n$ denotes the n -dimensional all-one vector. It also follows directly from (2.26a) that $(I - \gamma P_\pi)^{-1}$ is a non-negative matrix.

Additionally, the above results continue to hold if P_π is replaced with P^π and $\mathbf{1}_S$ is replaced with $\mathbf{1}_{SA}$.

Proof of Lemma 2.3. With regard to (2.26a), it suffices to note that

$$\lim_{i \rightarrow \infty} \|\gamma^i (P_\pi)^i\|_1 = \lim_{i \rightarrow \infty} \gamma^i = 0,$$

where we have used the fact that $(P_\pi)^i$ is also a probability transition matrix (so that $\|(P_\pi)^i\|_1 = 1$). This implies that the matrix geometric series in (2.26a) is convergent, thus justifying the validity of the Neumann series expansion in (2.26a). Regarding (2.26b), we can see from (2.26a) that

$$(I - \gamma P_\pi)^{-1} 1_S = \left(I + \sum_{i=1}^{\infty} (\gamma P_\pi)^i \right) 1_S = \left(1 + \sum_{i=1}^{\infty} \gamma^i \right) 1_S = \frac{1}{1 - \gamma} 1_S,$$

where the penultimate relation follows since $(P_\pi)^i$ is a probability transition matrix. This combined with the nonnegativity of $(I - \gamma P_\pi)^{-1}$ and the definition of $\|\cdot\|_1$ (cf. (1.1)) immediately justifies (2.26c). $\square$

Moreover, combining Lemma 2.3 with the Bellman equations in (2.24) reveals basic perturbation bounds. Consider two discounted infinite-horizon MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and $\tilde{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, P, \tilde{r}, \gamma)$, which differ only in their reward functions. Denote by $\tilde{Q}^\pi$ (resp. $\tilde{V}^\pi$) the Q-function (resp. value function) of policy π in $\tilde{\mathcal{M}}$, and similarly, let $\tilde{Q}^*$ (resp. $\tilde{V}^*$) represent the optimal Q-function (resp. value function) in $\tilde{\mathcal{M}}$. For any policy π , it holds that

$$\begin{aligned} \|Q^\pi - \tilde{Q}^\pi\|_\infty &= \|(I - \gamma P^\pi)^{-1} r - (I - \gamma P^\pi)^{-1} \tilde{r}\|_\infty \\ &\leq \|(I - \gamma P^\pi)^{-1}\|_1 \|r - \tilde{r}\|_\infty = \frac{1}{1 - \gamma} \|r - \tilde{r}\|_\infty, \end{aligned} \quad (2.27a)$$

and similarly,

$$\|V^\pi - \tilde{V}^\pi\|_\infty \leq \frac{1}{1 - \gamma} \|r - \tilde{r}\|_\infty. \quad (2.27b)$$

Given the validity of (2.27) for every policy π , we immediately reach

$$\begin{aligned} \|Q^* - \tilde{Q}^*\|_\infty &= \left\| \max_{\pi} Q^\pi - \max_{\pi} \tilde{Q}^\pi \right\|_\infty \leq \max_{\pi} \|Q^\pi - \tilde{Q}^\pi\|_\infty \\ &\leq \frac{1}{1 - \gamma} \|r - \tilde{r}\|_\infty, \end{aligned} \quad (2.28a)$$

and similarly,

$$\|V^* - \tilde{V}^*\|_\infty \leq \frac{1}{1 - \gamma} \|r - \tilde{r}\|_\infty. \quad (2.28b)$$

2.2 Finite-horizon episodic MDPs

The second class of MDPs studied extensively in this monograph is that of finite-horizon episodic MDPs, which we now introduce. In contrast to infinite-horizon MDPs, where the time step can continue indefinitely, finite-horizon episodic MDPs model problems with a predetermined number of steps in each episode.

2.2.1 Basic components

A finite-horizon Markov decision process, denoted by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, P, r\}$, consists of the following key components.

- (i) $\mathcal{S} = \{1, \dots, S\}$: a state space containing S different states, akin to the infinite-horizon counterpart.
- (ii) $\mathcal{A} = \{1, \dots, A\}$: an action space comprised of A distinct actions, similar to the infinite-horizon counterpart.
- (iii) H : the horizon length, which specifies the number of time steps in each episode of the MDP. This parameter dictates the planning horizon for decision-making.
- (iv) $P = \{P_h\}_{1 \leq h \leq H}$, with $P_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ denoting the transition kernel at step h (namely, $P_h(\cdot | s, a)$ stands for the transition probability of the MDP at step h when the current state-action pair is (s, a)). Conveniently, we introduce the following S -dimensional row vector

$$P_{s,a,h} := P_h(\cdot | s, a) \in \mathbb{R}^{1 \times S} \quad (2.29)$$

for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ to indicate the corresponding transition probability.

- (v) $r = \{r_h\}_{1 \leq h \leq H}$, with $r_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ denoting the reward function at step h (namely, $r_h(s, a)$ indicates the immediate reward gained at step h when taking action a in state s). Akin to the infinite-horizon setting, we assume throughout that all immediate rewards are deterministic, fully known to the learner, and lie within the interval $[0, 1]$.

Hereafter, we focus on *nonstationary* (i.e., *time-inhomogeneous*) MDPs within the finite-horizon framework, where both the transition probabilities and reward functions are allowed to vary across time steps. We are particularly interested in regimes in which the size S of the state space, the size A of the action space, and the horizon length H are all large, reflecting the challenges inherent in high-dimensional, long-horizon RL problems.

2.2.2 Policy, value function, Q-function and occupancy distribution

We now move on to introduce several important concepts tailored to finite-horizon episodic MDPs. Let us begin with the notion of a policy. A (possibly randomized) policy, denoted by $\pi = \{\pi_h\}_{1 \leq h \leq H}$ with $\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, is an action selection rule, such that $\pi_h(a | s)$ specifies the probability of choosing action a at step h when in state s . When π is a deterministic policy, we overload the notation and write $\pi_h(s)$ to denote the action selected by π in state s at step h .

Next, we define the value function and Q-function for finite-horizon episodic MDPs. To this end, consider an episode $\{(s_h, a_h)\}_{1 \leq h \leq H}$ with length H , with s_h and a_h denoting, respectively, the state and the action at step h , which serve as the basis for the definitions that follow.

- *Value function.* The value function of policy π is denoted by $V^\pi = \{V_h^\pi\}_{1 \leq h \leq H}$, where for each step $1 \leq h \leq H$, $V_h^\pi : \mathcal{S} \rightarrow \mathbb{R}$ quantifies the expected cumulative reward from a given state when starting at step h and following π until the end of the episode. More precisely, we define

$$\forall s \in \mathcal{S} : \quad V_h^\pi(s) := \mathbb{E} \left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s \right], \quad (2.30)$$

where the expectation is over the randomness of the trajectory $\{(s_t, a_t)\}_{t=h}^H$ when policy π is implemented (i.e., $a_t \sim \pi_t(\cdot | s_t)$ and $s_{t+1} \sim P_t(\cdot | s_t, a_t)$ for all $t \geq h$). Moreover, if the initial state s_1 is drawn from some state distribution $\rho \in \Delta(\mathcal{S})$, we find it convenient to define the value function of policy π under ρ as

$$V_1^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V_1^\pi(s)]. \quad (2.31)$$

- *Q-function.* Analogously, the Q-function of policy π is denoted by $Q^\pi = \{Q_h^\pi\}_{1 \leq h \leq H}$, which extends the definition of the value function by also conditioning on the first action to be taken. To be precise, for each $1 \leq h \leq H$ and every $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define

$$Q_h^\pi(s, a) := \mathbb{E} \left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s, a_h = a \right], \quad (2.32)$$

where the trajectory $\{(s_t, a_t)\}_{t=h}^H$ is sampled in a way such that $s_{t+1} \sim P_t(\cdot | s_t, a_t)$ and $a_{t+1} \sim \pi_{t+1}(\cdot | s_{t+1})$ and for all $t \geq h$. In words, $Q_h^\pi(s, a)$ quantifies the remaining expected reward of taking action a in state s at step h and then following policy π thereafter.

It is worth noting that no discount factor is needed in the above definitions of the value function and Q-function, given that they are always well-defined in finite-horizon settings. Moreover, similar to the infinite-horizon counterpart, there always exists at least one deterministic policy that simultaneously maximizes the value function and the Q-function for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ (Bertsekas, 2017). In light of this, we shall denote by $\pi^* = \{\pi_h^*\}_{1 \leq h \leq H}$ an *optimal deterministic* policy throughout this monograph, which allows us to use $\pi_h^*(s)$ to indicate the corresponding optimal action chosen in state s at step h . The resulting optimal value function and optimal Q-function are denoted respectively by $V^* = \{V_h^*\}_{1 \leq h \leq H}$ and $Q^* = \{Q_h^*\}_{1 \leq h \leq H}$:

$$\forall h \in [H]: \quad V_h^* := V_h^{\pi^*} \quad \text{and} \quad Q_h^* := Q_h^{\pi^*}. \quad (2.33)$$

Furthermore, we introduce the *occupancy distributions* associated with policy $\pi = \{\pi_h\}_{1 \leq h \leq H}$, describing how often states or state-action pairs are visited at a given step. Specifically, for any step $1 \leq h \leq H$, define

$$d_h^\pi(s; \rho) := \mathbb{P}(s_h = s \mid s_1 \sim \rho; \pi), \quad (2.34a)$$

$$d_h^\pi(s, a; \rho) := \mathbb{P}(s_h = s, a_h = a \mid s_1 \sim \rho; \pi) = d_h^\pi(s; \rho) \pi_h(a \mid s), \quad (2.34b)$$

which are conditioned on the initial state distribution $s_1 \sim \rho$ and the event that all actions are sampled according to π . As a sanity check, note that at the beginning of the episode, the occupancy state distribution simply coincides with the initial state distribution, i.e.,

$$d_1^\pi(s; \rho) = \rho(s) \quad \text{for any policy } \pi \text{ and any state } s \in \mathcal{S}. \quad (2.35)$$

For convenience, we introduce shorthand notation for the occupancy distributions induced by the optimal deterministic policy $\pi^* = \{\pi_h^*\}_{1 \leq h \leq H}$:

$$d_h^*(s) := d_h^{\pi^*}(s; \rho), \quad (2.36)$$

$$d_h^*(s, a) := d_h^{\pi^*}(s, a; \rho) = d_h^*(s) \mathbb{1}\{a = \pi_h^*(s)\} \quad (2.37)$$

for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, where the last identity holds given that policy π^* is assumed to be deterministic.

2.2.3 Bellman equations and dynamic programming

Next, we present the Bellman equations for finite-horizon MDPs, which underlie the design of classical dynamic programming algorithms.

Denote the terminal value functions as $V_{H+1}^\pi = 0$ and $Q_{H+1}^\pi = 0$ for any policy $\pi = \{\pi_h\}_{1 \leq h \leq H}$. For a given policy π , its associated value function and Q-function can be computed through backward recursion starting from the terminal states as follows: for each $h = H, \dots, 1$,

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot | s, a)} [V_{h+1}^\pi(s')], \quad (2.38a)$$

$$V_h^\pi(s) = \mathbb{E}_{a \sim \pi_h(\cdot | s)} [Q_h^\pi(s, a)] \quad (2.38b)$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. Intuitively, at each step h , the Q-value consists of two components: the immediate reward received at that step, and the expected remaining value of proceeding according to π from the next state at step $h + 1$. Analogously, given that $V_{H+1}^\star = 0$ and $Q_{H+1}^\star = 0$, the Bellman optimality equations assert that: for every $h = H, \dots, 1$,

$$Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot | s, a)} [V_{h+1}^\star(s')], \quad (2.39a)$$

$$V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a) \quad (2.39b)$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$.

In summary, given full knowledge of the model, the above recursions — often referred to as *value iteration (VI)* — allow the finite-horizon value functions and Q-functions to be computed step by step via backward induction, starting from the terminal step and proceeding backward to $h = 1$. This procedure embodies the dynamic programming principle: optimal decisions at each step can be determined by reasoning backward from the end of the episode.

2.2.4 Vector and matrix notation

Akin to the discounted infinite-horizon counterpart, it is convenient to introduce a set of vector and matrix notation that streamlines the presentation (with slight abuse of notation).

- $r_h \in \mathbb{R}^{SA}$: a vector representing the reward function at step h , so that its (s, a) -th entry is the corresponding immediate reward at step h .
- $V_h^\pi, V_h^\star \in \mathbb{R}^S$: vectors representing, respectively, the value function of policy π at step h (cf. (2.30)) and the optimal value function at step h (cf. (2.33)).
- $Q_h^\pi, Q_h^\star \in \mathbb{R}^{SA}$: vectors representing, respectively, the Q-function of policy π at step h (cf. (2.32)) and the optimal Q-function at step h (cf. (2.33)).
- $P_h \in \mathbb{R}^{SA \times S}$: a matrix representing the probability transition kernel P_h at step h , whose (s, a) -th row is the probability vector corresponding to $P_h(\cdot | s, a)$.

The Bellman equations (cf. (2.38) and (2.39)) can therefore be expressed compactly using the above vector and matrix notation as

$$Q_h^\pi = r_h + P_h V_{h+1}^\pi, \quad 1 \leq h \leq H, \quad (2.40a)$$

$$Q_h^\star = r_h + P_h V_{h+1}^\star, \quad 1 \leq h \leq H. \quad (2.40b)$$

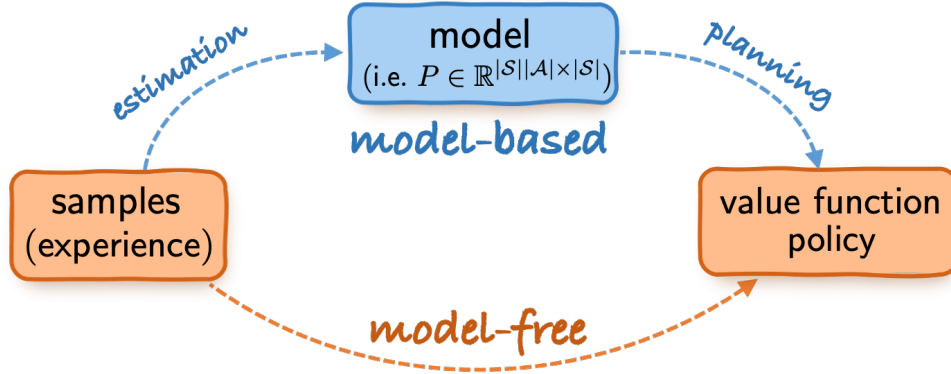


Figure 2.4: The model-based approach vs. the model-free approach. A key distinction is that model-free methods dispense with explicit model estimation and instead learn the value and policy of interest directly without the aid of an estimated model.

In addition, we introduce the notation for variances. For any vector $V = [V_i]_{1 \leq i \leq S} \in \mathbb{R}^S$, we define the scalar quantity $\text{Var}_{P_{s,a,h}}(V) \in \mathbb{R}$ and the vector $\text{Var}_{P_h}(V) \in \mathbb{R}^{S \times A}$ as

$$\text{Var}_{P_{s,a,h}}(V) = \sum_{s' \in \mathcal{S}} P_h(s' | s, a) (V_{s'})^2 - \left(\sum_{s' \in \mathcal{S}} P_h(s' | s, a) V_{s'} \right)^2, \quad (2.41a)$$

$$\begin{aligned} \text{Var}_{P_h}(V) &= [\text{Var}_{P_{s,a,h}}(V)]_{(s,a) \in \mathcal{S} \times \mathcal{A}} \\ &= P_h(V \circ V) - (P_h V) \circ (P_h V), \end{aligned} \quad (2.41b)$$

representing the variance of $V(s')$ when s' is drawn according to either a single distribution $P_{s,a,h}$ or, collectively, all rows of the transition matrix P_h .

2.3 Model-based vs. model-free approaches

Thus far, our discussion has centered on the fundamentals of MDPs and on dynamic programming algorithms given complete model knowledge. While classical dynamic programming provides powerful and efficient methods for identifying optimal policies when the model is fully specified, reinforcement learning often has to contend with more challenging scenarios in which the model is unknown *a priori* and can only be learned from data — either pre-collected or gathered in real time. In such settings, RL agents have to select their policies in the presence of uncertainty and incomplete information.

At a high level, there are at least two principal approaches in RL, which we describe below and illustrate in Figure 2.4.

- *Model-based RL*, which decouples the learning process into two stages: a model estimation stage that estimates the transition kernel and rewards using available data, and a subsequent planning stage that applies planning algorithms on the learned model to compute near-optimal policies and values. Given the use of the meta-task of model estimation, it requires maintaining and updating the learned model from collected data.

- *Model-free RL*, which circumvents the model estimation stage and seeks to learn the optimal values and policies directly. Prominent examples include Q-learning (Watkins and Dayan, 1992) and its many variants. Compared to the model-based counterpart, it holds the promise of lower space complexity, as it avoids the need to store and update a full description of the model. Following the convention in RL theory (e.g., Strehl *et al.* (2006) and Jin *et al.* (2018)), we classify an algorithm as model-free if its space complexity is substantially smaller than that required to represent the full model, irrespective of the number of samples collected.

In this work, we investigate both model-based and model-free approaches, with particular emphasis on their statistical efficiency. As the subsequent chapters will illustrate, the model-based approach is often able to achieve the best possible sample complexity, albeit at the expense of higher memory usage and computational overhead.

2.4 Concentration inequalities

Given that RL relies on learning from data, non-asymptotic concentration inequalities developed in the recent literature on high-dimensional probability and statistics (Boucheron *et al.*, 2013; Vershynin, 2018; Wainwright, 2019a) play a central role in establishing tight sample complexity bounds. For this reason, we briefly review a few commonly used concentration results.

Consider a probability transition kernel $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$. Suppose that for each state-action pair (s, a) , we have access to N independent samples drawn from $P(\cdot | s, a)$, namely,

$$s^{(i)}(s, a) \stackrel{\text{i.i.d.}}{\sim} P(\cdot | s, a), \quad i = 1, \dots, N.$$

From these samples, we can naturally form an empirical estimate $\hat{P}(\cdot | s, a)$ of the transition probability $P(\cdot | s, a)$ as follows:

$$\hat{P}(s' | s, a) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s^{(i)}(s, a) = s'\} \quad \text{for all } s' \in \mathcal{S}. \quad (2.42)$$

Taking these estimates together yields an estimate $\hat{P} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ of the probability transition matrix P , namely, the (s, a) -th row of $\hat{P}$ is given by

$$\hat{P}_{(s,a),s'} = \hat{P}(s' | s, a) \quad \text{for all } (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}. \quad (2.43)$$

The following concentration bound is a direct consequence of the classical Hoeffding inequality (Vershynin, 2018) combined with the union bound.

Lemma 2.4 (Hoeffding inequality). Consider a given state-action pair (s, a) and a finite set of vectors $\mathcal{V} \in \mathbb{R}^S$. Fix any $0 < \delta < 1$. If $\mathcal{V}$ is independent from $\hat{P}_{s,a}$, then with probability greater than $1 - \delta$,

$$|(\hat{P}_{s,a} - P_{s,a})V| \leq \sqrt{\frac{\|V\|_\infty^2 \log\left(\frac{2|\mathcal{V}|}{\delta}\right)}{2N}} \quad (2.44)$$

holds simultaneously for all $V \in \mathcal{V}$.

In particular, the Hoeffding bound (2.44) depends on V only through the maximum magnitude of its entries, without relying on any finer structural properties of V , such as its variance.

Despite its simplicity and broad applicability, Hoeffding's inequality is often conservative, especially when the underlying standard deviation is substantially smaller than the worst-case

range. In such situations, Bernstein-type inequalities can provide significantly sharper concentration bounds by explicitly incorporating variance information. To this end, we record below concentration inequalities obtained directly from the Bernstein inequality (Vershynin, 2018) in conjunction with a union bound; see also Agarwal *et al.* (2020a, Lemma 6).

Lemma 2.5 (Bernstein inequality). Consider a given state-action pair (s, a) and a finite set of vectors $\mathcal{V} \in \mathbb{R}^S$. Fix any $0 < \delta < 1$. If $\mathcal{V}$ is independent from $\hat{P}_{s,a}$, then with probability greater than $1 - \delta$,

$$|(\hat{P}_{s,a} - P_{s,a})V| \leq \sqrt{\frac{2 \log(\frac{4|\mathcal{V}|}{\delta})}{N}} \sqrt{\text{Var}_{P_{s,a}}(V)} + \frac{2\|V\|_\infty \log(\frac{4|\mathcal{V}|}{\delta})}{3N} \quad (2.45)$$

holds simultaneously for all $V \in \mathcal{V}$. Moreover, if $\mathcal{V}$ is statistically independent from $\hat{P}$, then with probability greater than $1 - \delta$, one has

$$|(\hat{P} - P)V| \leq \sqrt{\frac{2 \log(\frac{4SA|\mathcal{V}|}{\delta})}{N}} \sqrt{\text{Var}_P(V)} + \frac{2\|V\|_\infty \log(\frac{4SA|\mathcal{V}|}{\delta})}{3N} 1_{SA}. \quad (2.46)$$

Here, we recall the definitions of $\text{Var}_{P_{s,a}}(V)$ and $\text{Var}_P(V)$ in (2.21). We use 1_{SA} to denote the all-one vector, and for any non-negative vector v , $\sqrt{v}$ denotes the entrywise square root.

In words, the Bernstein-style inequalities in Lemma 2.5 reveal that the uncertainty of $\hat{P}_{s,a}V$ is governed by both its variance and the largest entrywise magnitude of V . When the sample size is large enough, the variance term typically dominates, a phenomenon that will become evident in subsequent chapters.

Furthermore, a technical challenge arises when employing the Bernstein inequality to construct confidence intervals for quantities such as $P_{s,a}V$ based on the empirical estimate $\hat{P}_{s,a}V$. Specifically, the variance term $\text{Var}_{P_{s,a}}(V)$ depends on the true transition kernel $P_{s,a}$, which is unavailable in practice. To circumvent this, it is useful to resort to the following empirical Bernstein inequality — quoted from Maurer and Pontil (2009, Theorem 4) — which substitutes the population variance with its empirical estimate, making the confidence interval computable from collected data.

Lemma 2.6 (Empirical Bernstein inequality). Consider a given state-action pair (s, a) and a finite set of vectors $\mathcal{V} \subset \mathbb{R}^S$. Fix any $0 < \delta < 1$, and suppose that $N \geq 2$. If $\mathcal{V}$ is independent from $\hat{P}_{s,a}$, then with probability greater than $1 - \delta$,

$$|(\hat{P}_{s,a} - P_{s,a})V| \leq \sqrt{\frac{2 \log(\frac{4|\mathcal{V}|}{\delta})}{N-1}} \sqrt{\text{Var}_{\hat{P}_{s,a}}(V)} + \frac{7\|V\|_\infty \log(\frac{4|\mathcal{V}|}{\delta})}{3(N-1)}$$

holds simultaneously for all $V \in \mathcal{V}$.

In addition, when coping with sums of i.i.d. Bernoulli variables, the following exponential tail bound becomes useful, which follows from standard Chernoff-type arguments.

Lemma 2.7. Let $Z_i \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$ ($1 \leq i \leq n$) be a set of independent Bernoulli random variables with mean p . Then for any $a > 0$, one has

$$\mathbb{P}\left(\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) < -an\right) \leq \exp\left(-\frac{a^2n}{2p}\right).$$

In addition, for any $\tau \geq 2pn$, one has

$$\mathbb{P}\left(\sum_{i=1}^n Z_i > \tau\right) \leq 2 \exp(-0.19\tau).$$

Proof. The first inequality can be found in Alon and Spencer (2008, Theorem A.1.13), while the second inequality is an immediate consequence of Alon and Spencer (2008, Corollary A.1.14) with $\epsilon \geq 1$ therein. $\square$

Moving beyond independent data, concentration inequalities for martingales are often needed to handle sequentially collected observations. Among the most useful are the Azuma-Hoeffding inequality (Azuma, 1967; Dubhashi and Panconesi, 2009), a martingale counterpart of Hoeffding's inequality, and Freedman's inequality (Freedman, 1975; Tropp, 2011), which serves as a martingale analogue of Bernstein's inequality. Below, we record the standard Azuma-Hoeffding inequality and a user-friendly version of Freedman's inequality (see Li *et al.* (2024a, Theorem EC.1.)).

Theorem 2.8 (Azuma-Hoeffding inequality). Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying

$$|X_k| \leq B_k \quad \text{and} \quad \mathbb{E}[X_k \mid \{X_j\}_{j:j < k}] = 0 \quad \text{for all } k \geq 1$$

for some deterministic quantities $\{B_k\}$. Then with probability exceeding $1 - \delta$, one has

$$|Y_n| \leq \sqrt{2 \sum_{k=1}^n B_k^2 \log \frac{2}{\delta}}. \quad (2.47)$$

In brief, the Azuma-Hoeffding inequality shows that the fluctuations of a martingale sum can be controlled by the sum of the squared magnitudes of its martingale difference terms. This can be viewed as a direct extension of the Hoeffding inequality (Boucheron *et al.*, 2013) from independent data to weakly dependent martingale sequences.

Theorem 2.9 (Freedman inequality). Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying

$$|X_k| \leq B \quad \text{and} \quad \mathbb{E}[X_k \mid \{X_j\}_{j:j < k}] = 0 \quad \text{for all } k \geq 1.$$

Define

$$W_n := \sum_{k=1}^n \mathbb{E}[X_k^2 \mid \{X_j\}_{j:j < k}],$$

and suppose that $W_n \leq \sigma^2$ holds deterministically. Then for any positive integer $K_0 \geq 1$, with probability at least $1 - \delta$ one has

$$|Y_n| \leq \sqrt{8 \max \left\{ W_n, \frac{\sigma^2}{2K_0} \right\} \log \frac{2K_0}{\delta}} + \frac{4}{3} B \log \frac{2K_0}{\delta}. \quad (2.48)$$

Informally, the Freedman inequality in Theorem 2.9 asserts that the deviation of a martingale sum is controlled by the cumulative conditional variance (i.e., W_n) and a uniform upper bound on the magnitude of each summand (i.e., B). The resulting concentration bound closely resembles the Bernstein inequality, with a key distinction that the total variance is now replaced by the cumulative conditional variance of the martingale sequence.

2.5 Minimax lower bounds for multi-hypothesis testing

To assess the statistical optimality of an RL algorithm, a standard approach is to establish a statistical lower bound and to demonstrate that the algorithm attains this lower limit. Notably, the development of statistical or information-theoretic lower bounds typically relies on arguments that differ substantially from those used in upper-bound analyses. While lower-bound techniques constitute a central topic in statistics and information theory (see, e.g., Tsybakov (2009) and Polyanskiy and Wu (2025)), a comprehensive treatment lies beyond the scope of this monograph. Instead, we single out a simple yet remarkably powerful and versatile tool that appears frequently in lower-bound analyses: Fano's lemma. Rooted in classical information theory and originally introduced by Robert Fano, this lemma provides a principled way to quantify the intrinsic difficulty of statistical problems through information-theoretic measures.

More specifically, suppose that there are $M + 1$ probability distributions $P_0, P_1, \dots, P_M$ on a common measurable space $(\mathcal{X}, \mathcal{A})$. Consider any test ψ that maps from $\mathcal{X}$ to $\{0, 1, \dots, M\}$, with the goal of identifying the index of the distribution generating the observed samples. The performance of such a test can be measured by its error probability under each hypothesis, namely, $P_i(\psi \neq i)$ for $i = 0, \dots, M$. The minimax probability of error is then defined as

$$p_e := \inf_{\psi} \max_{0 \leq i \leq M} P_i(\psi \neq i), \quad (2.49)$$

with the infimum taken over all possible tests. In words, p_e captures the smallest achievable worst-case error probability across all hypotheses. It is also useful to define the average probability of error as follows:

$$\bar{p}_e := \inf_{\psi} \frac{1}{M+1} \sum_{i=0}^M P_i(\psi \neq i), \quad (2.50)$$

which measures the error probability averaged over all $M + 1$ hypotheses. The Fano lemma provides a fundamental lower bound on the probability of error for testing among $M + 1$ hypotheses; the version we present below is adapted from Tsybakov (2009, Section 2.7.1).

Lemma 2.10 (Fano's lemma). Let $P_0, P_1, \dots, P_M$ be probability distributions on a common measurable space. Recall that p_e (resp. $\bar{p}_e$) is the minimax (resp. average) probability of error defined in (2.49) (resp. (2.50)).

(i) Suppose that $M = 1$. If $\text{KL}(P_1 \parallel P_0) \leq \beta < \infty$, then

$$p_e \geq \bar{p}_e \geq \max \left\{ \frac{1}{4} \exp(-\beta), \frac{1 - \sqrt{\beta/2}}{2} \right\}. \quad (2.51)$$

(ii) Suppose that $M \geq 2$. If

$$\frac{1}{M+1} \sum_{j=1}^M \text{KL}(P_j \parallel P_0) \leq \beta \log M \quad (2.52a)$$

with $0 < \beta < 1$, then

$$p_e \geq \bar{p}_e \geq \frac{\log(M+1) - \log 2}{\log M} - \beta. \quad (2.52b)$$

At a high level, Fano's lemma asserts that if the candidate distributions are close to one another in terms of KL divergence, then no test can reliably distinguish among them, and consequently, both the minimax error probability and the average error probability must remain bounded away from zero. In this way, Fano's lemma unearths an intimate connection between the performance limit of any statistical test and the information-theoretic separation between distributions (as measured by KL divergence). Guided by Fano's lemma, a common approach to establishing minimax lower bounds is to first construct a carefully chosen collection of distributions that are sufficiently close to one another and then bound their pairwise KL divergence to derive the desired lower bound.

Given the central role of KL divergence in Fano's lemma, we find it useful to record several of its key properties. We begin with the tensorization (i.e., additivity) property, which characterizes how KL divergence decomposes under independent data; see, e.g., Tsybakov (2009, Section 2.4). This simple property highlights a key advantage of KL divergence: when samples are independent, analysis of the overall KL divergence reduces to computing the KL divergence for individual samples.

Lemma 2.11. For two product distributions $P = P_1 \otimes \cdots \otimes P_n$ and $Q = Q_1 \otimes \cdots \otimes Q_n$, the KL divergence satisfies

$$\text{KL}(P \parallel Q) = \sum_{i=1}^n \text{KL}(P_i \parallel Q_i).$$

Additionally, when P and Q are both Bernoulli distributions, we adopt a slight abuse of notation and write

$$\text{KL}(p \parallel q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \quad (2.53)$$

to denote the KL divergence between $\text{Bern}(p)$ and $\text{Bern}(q)$, where $\text{Bern}(p)$ represents the Bernoulli distribution with mean p . The following lemma establishes a quadratic upper bound on this divergence, which is sometimes more convenient to work with than the original expression.

Lemma 2.12. For any $0 < p, q < 1$,

$$\text{KL}(p \parallel q) \leq \frac{(p-q)^2}{q(1-q)}.$$

Moreover, if $1/2 \leq q < p < 1$, then

$$\text{KL}(p \parallel q) \leq \text{KL}(q \parallel p) \leq \frac{(p-q)^2}{p(1-p)}.$$

Proof. We begin with the first relation. It is known that the KL divergence can be upper bounded by the χ^2 divergence (see Tsybakov (2009, Lemma 2.7)), namely,

$$\text{KL}(p \parallel q) \leq \chi^2(p \parallel q) := \frac{(p-q)^2}{q} + \frac{(1-p-(1-q))^2}{1-q} = \frac{(p-q)^2}{q(1-q)}.$$

We now move to the second relation. Let

$$a = \frac{p+q}{2} \in \left[\frac{1}{2}, 1\right] \quad \text{and} \quad b = \frac{p-q}{2} \in \left[0, \frac{1}{4}\right],$$

so that $p = a + b$ and $q = a - b$. By the definition in (2.53) and some elementary algebra, we obtain

$$\text{KL}(p \parallel q) - \text{KL}(q \parallel p) = (p+q) \log \frac{p}{q} + (2-p-q) \log \frac{1-p}{1-q}$$

$$\begin{aligned}
&= 2a \log \left(\frac{a+b}{a-b} \right) + 2(1-a) \log \frac{1-a-b}{1-a+b} \\
&=: g(a, b).
\end{aligned}$$

Taking the derivative of $g(a, b)$ with respect to b gives

$$\begin{aligned}
\frac{\partial g(a, b)}{\partial b} &= 2a \left\{ \frac{1}{a+b} + \frac{1}{a-b} \right\} - 2(1-a) \left\{ \frac{1}{1-a+b} + \frac{1}{1-a-b} \right\} \\
&= f(a) - f(1-a) \leq 0, \quad \text{where} \quad f(x) := \frac{2x}{x+b} + \frac{2x}{x-b}.
\end{aligned}$$

The last relation holds since f is decreasing on (b, ∞) and $a \geq 1-a$. Thus, for each fixed a , the function $g(a, b)$ is non-increasing in $b \geq 0$. Consequently,

$$\text{KL}(p \parallel q) - \text{KL}(q \parallel p) = g(a, b) \leq g(a, 0) = 0,$$

which proves the desired claim. $\square$

2.6 Notes

While discounted infinite-horizon stationary MDPs and finite-horizon nonstationary MDPs are arguably the two most extensively studied models, several other MDP formulations have also received considerable attention in the RL literature. Partial examples include: (i) *infinite-horizon average-reward MDPs*, where discounted cumulative rewards are replaced by long-run average rewards (Puterman, 2014; Wan *et al.*, 2021; Jin and Sidford, 2021), and (ii) *finite-horizon stationary MDPs*, in which the transition kernels and reward functions are both time-homogeneous (Ren *et al.*, 2021; Li *et al.*, 2022c; Zhang *et al.*, 2022). Each of these models bears similarities with the ones under exposition here, yet also brings its own distinctive challenges. For further details, we refer the interested reader to the aforementioned papers and the references therein.

Additionally, we direct the interested reader to Wainwright (2019a), Fan *et al.* (2020a), and Samworth and Shah (2026) for comprehensive expositions on high-dimensional statistics, and to Vershynin (2018) and Boucheron *et al.* (2013) for concentration inequalities. Furthermore, Tsybakov (2009), Polyanskiy and Wu (2025), and El Gamal and Raginsky (2026) serve as excellent references for the development of minimax lower bounds via information-theoretic techniques.

3

RL with a generative model: the model-based approach

In this chapter, we begin our study of sample complexity in RL by first examining an idealized sampling mechanism, commonly referred to as the generative model or simulator. While this sampling mechanism is a substantial simplification of practical scenarios, it suggests a clean and tractable framework that serves as a natural first step towards understanding the fundamental statistical limits of RL. Our focus in this chapter is on model-based RL with the generative model, with the discussion of model-free algorithms deferred to Chapter 4.

3.1 Generative model or simulator

Consider first the discounted infinite-horizon stationary MDP introduced in Section 2.1. Initially studied by Kearns and Singh (1999), a generative model—also referred to as a simulator—allows one to query any state-action pair and sample an independent state transition from the ground-truth transition kernel. More precisely, each query to the generative model with a pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ returns an independent sample:

$$s(s, a) \stackrel{\text{ind.}}{\sim} P(\cdot | s, a). \quad (3.1)$$

Similarly, in the case of a finite-horizon nonstationary MDP, each call to the generative model with an arbitrary state-action-step tuple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ returns an independent sample:

$$s(s, a, h) \stackrel{\text{ind.}}{\sim} P_h(\cdot | s, a). \quad (3.2)$$

Unless otherwise noted, we assume throughout that the reward function is fully known to the agent. This assumption can, in fact, be substantially relaxed without affecting the core insights underlying algorithm design and analysis.

Studying RL with a generative model offers several key advantages. To begin with, simulators have been developed in several applications to simulate complex, real-world environments, such as for autonomous driving (Shah *et al.*, 2018), robotic systems (Kumar *et al.*, 2019; Yang *et al.*, 2024a), and game playing (Bellemare *et al.*, 2013). Beyond their practical utility, generative models provide a clean abstraction: by allowing direct sampling from arbitrary state-action pairs, they

effectively decouple learning from the need for exploration. This separation simplifies the problem and makes it easier to isolate the fundamental information bottlenecks inherent to the learning part. Overall, investigating RL under this idealistic sampling mechanism helps us distill core design principles, establish theoretical benchmarks, and understand the fundamental statistical limits of RL algorithms, which in turn guide us towards more practical and robust solutions.

3.2 Model-based policy evaluation

We begin with the fundamental task of *policy evaluation*, which seeks to estimate, based on the collected samples, the value function V^π of a given policy π . Our focus throughout this section is on ℓ_∞ -accurate estimation: given a target accuracy level $\varepsilon > 0$, the aim is to find an estimate $\hat{V}^\pi$ that is ε -accurate uniformly over every state s , namely,

$$\forall s \in \mathcal{S} : \quad |\hat{V}^\pi(s) - V^\pi(s)| \leq \varepsilon.$$

Recall our assumption that all immediate rewards reside within $[0, 1]$, and hence the value function falls within the range $(0, \frac{1}{1-\gamma}]$. This implies that the range of the target accuracy level ε should be set to $\varepsilon \in [0, \frac{1}{1-\gamma}]$.

3.2.1 A plug-in approach

In view of classical Bellman equations (cf. (2.12)), the value function of policy π satisfies (see (2.24c))

$$V^\pi = (I - \gamma P_\pi)^{-1} r_\pi,$$

which depends upon the transition kernel P_π induced by π over the state space $\mathcal{S}$ (see (2.19)), the reward function r_π induced by π (see (2.20)), and the discount factor γ . Since the only unknown component of the model in our setting is the transition kernel $P_\pi \in \mathbb{R}^{S \times S}$, a natural strategy for policy evaluation is the “plug-in” approach, which starts by estimating P_π , followed by computation of the value function associated with the learned transition kernel $\hat{P}_\pi$. This two-stage paradigm, which separates model estimation from value estimation, is dubbed as the model-based approach. More formally, the model-based policy evaluation algorithm in the presence of a generative model can be described as:

- 1) *Model estimation.* For each $s \in \mathcal{S}$, draw N independent samples from the generative model:

$$\forall s \in \mathcal{S} : \quad s^{(i)}(s) \stackrel{\text{ind.}}{\sim} P_\pi(\cdot | s), \quad i = 1, \dots, N$$

with $P_\pi(\cdot | s)$ defined in (2.2), and compute the empirical transition kernel from state s as:

$$\hat{P}_\pi(s' | s) = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \{s' = s^{(i)}(s)\} \quad \text{for all } s' \in \mathcal{S}. \quad (3.3)$$

In words, $\hat{P}_\pi(s' | s)$ counts the empirical frequency of transitions from state s to state s' .

- 2) *Value estimation.* Compute the value function estimate as

$$\hat{V}^\pi = (I - \gamma \hat{P}_\pi)^{-1} r_\pi. \quad (3.4)$$

It is self-evident that the total number of samples drawn in the above model-based algorithm is $N^{\text{total}} := NS$.

Algorithm 3.1: Model-based policy evaluation with a simulator

```

1 input: sample size per state  $N$ , discount factor  $\gamma$ , reward function  $r_\pi$  induced by policy  $\pi$ .
2 for  $s \in \mathcal{S}$  do
    /* model estimation based on samples. */
3     draw  $N$  independent samples  $s^{(i)}(s) \sim P_\pi(\cdot | s)$  (cf. (2.2)).
4     compute  $\hat{P}_\pi(\cdot | s)$  based on (3.3).
    /* value estimation using the empirical model. */
5 output:  $\hat{V}^\pi = (I - \gamma \hat{P}_\pi)^{-1} r_\pi$ .
```

3.2.2 Theoretical guarantees

Despite its conceptual and algorithmic simplicity, the model-based policy evaluation algorithm described in Algorithm 3.1 turns out to achieve a form of statistical optimality in the generative model setting; namely, no other algorithms can outperform its sample complexity in a minimax sense, except possibly up to logarithmic factors. To formalize this, we first present an upper bound on the sample complexity of Algorithm 3.1 in the theorem below, and defer the discussion of a matching lower bound to Section 3.4.

Theorem 3.1 (Li et al., 2020; Li et al., 2024c). For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$ and any $0 < \delta < 1$, Algorithm 3.1 achieves

$$\forall s \in \mathcal{S} : \quad |\hat{V}^\pi(s) - V^\pi(s)| \leq \varepsilon$$

with probability at least $1 - \delta$, provided that the sample size per state exceeds

$$N \geq \frac{c_0 \log\left(\frac{S}{\delta(1-\gamma)}\right)}{(1-\gamma)^3 \varepsilon^2} \quad (3.5)$$

for some sufficiently large constant $c_0 > 0$.

In brief, Theorem 3.1 demonstrates that the sample complexity of the model-based policy evaluation algorithm in Algorithm 3.1 is no greater than

$$\tilde{O}\left(\frac{S}{(1-\gamma)^3 \varepsilon^2}\right)$$

for any meaningful target precision level ε . Notably, this requirement is often significantly lower than the sample size needed to learn the transition kernel P_π . For instance, when both $\frac{1}{1-\gamma}$ and ε are on the constant order, this sample complexity simplifies to $\tilde{O}(S)$, which can be S times smaller than what is required to learn $P_\pi \in \mathbb{R}^{S \times S}$ — an object with $O(S^2)$ degrees of freedom. Put differently, reliable value estimation becomes statistically feasible even in regimes where faithful model estimation is not. This unveils a crucial insight: while model-based algorithms conceptually involve model estimation as a meta-task, accomplishing this meta-task with high precision is not essential for policy evaluation to succeed.

3.2.3 Analysis: a first attempt

We begin with an initial attempt at bounding the value estimation error for Algorithm 3.1. This analysis strategy is straightforward yet yields a suboptimal bound.

Concretely, let us begin by deriving a simple characterization of the value estimation error. The Bellman equation (2.24b) tells us that

$$r_\pi = V^\pi - \gamma P_\pi V^\pi = (I - \gamma P_\pi) V^\pi,$$

which taken together with (3.4) implies that

$$\hat{V}^\pi = (I - \gamma \hat{P}_\pi)^{-1} r_\pi = (I - \gamma \hat{P}_\pi)^{-1} (I - \gamma P_\pi) V^\pi. \quad (3.6)$$

Invoke (3.6) to decompose the value estimation error as

$$\begin{aligned} \hat{V}^\pi - V^\pi &= (I - \gamma \hat{P}_\pi)^{-1} (I - \gamma P_\pi) V^\pi - (I - \gamma \hat{P}_\pi)^{-1} (I - \gamma \hat{P}_\pi) V^\pi \\ &= \gamma (I - \gamma \hat{P}_\pi)^{-1} (\hat{P}_\pi - P_\pi) V^\pi. \end{aligned} \quad (3.7)$$

While controlling $(I - \gamma \hat{P}_\pi)^{-1}$ can be somewhat challenging, the other term $(\hat{P}_\pi - P_\pi) V^\pi$ can be readily handled by standard concentration inequalities. Specifically, applying the Bernstein-style inequality in Lemma 2.5 with $A = 1$ and a single vector fixed V^π shows that, with probability at least $1 - \delta$,

$$|(\hat{P}_\pi - P_\pi) V^\pi| \leq C_0 \left(\sqrt{\frac{\text{Var}_{P_\pi}(V^\pi) \log(\frac{S}{\delta})}{N}} + \frac{\|V^\pi\|_\infty \log(\frac{S}{\delta})}{N} 1_S \right) \quad (3.8)$$

for some universal constant $C_0 > 0$, where $1_S \in \mathbb{R}^S$ is the all-one vector.

Regarding the term $(I - \gamma \hat{P}_\pi)^{-1}$, it follows from Lemma 2.3 that $(I - \gamma \hat{P}_\pi)^{-1}$ is a non-negative matrix and obeys

$$(I - \gamma \hat{P}_\pi)^{-1} 1_S = \frac{1}{1 - \gamma} 1_S. \quad (3.9)$$

Combining these properties with (3.7) and (3.8) shows that

$$\begin{aligned} |(I - \gamma \hat{P}_\pi)^{-1} (\hat{P}_\pi - P_\pi) V^\pi| &\leq (I - \gamma \hat{P}_\pi)^{-1} |(\hat{P}_\pi - P_\pi) V^\pi| \\ &\leq C_0 \left\{ \frac{(I - \gamma \hat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi) \log(\frac{S}{\delta})}}{\sqrt{N}} + \frac{\|V^\pi\|_\infty \log(\frac{S}{\delta})}{N} (I - \gamma \hat{P}_\pi)^{-1} 1_S \right\} \\ &= C_0 \left\{ \frac{(I - \gamma \hat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi) \log(\frac{S}{\delta})}}{\sqrt{N}} + \frac{\|V^\pi\|_\infty \log(\frac{S}{\delta})}{(1 - \gamma)N} 1_S \right\} \end{aligned} \quad (3.10)$$

occurs with probability exceeding $1 - \delta$.

We then attempt to bound $(I - \gamma \hat{P}_\pi)^{-1}$ and $\sqrt{\text{Var}_{P_\pi}(V^\pi)}$ separately:

$$\begin{aligned} \left\| (I - \gamma \hat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} \right\|_\infty &\leq \left\| (I - \gamma \hat{P}_\pi)^{-1} \right\|_1 \sqrt{\left\| \text{Var}_{P_\pi}(V^\pi) \right\|_\infty} \\ &= \frac{1}{1 - \gamma} \sqrt{\left\| \text{Var}_{P_\pi}(V^\pi) \right\|_\infty} \\ &\leq \frac{1}{(1 - \gamma)^2}, \end{aligned} \quad (3.11)$$

where the penultimate relation arises from Lemma 2.3, and the last line is valid since $\text{Var}_q(V^\pi) \leq \mathbb{E}_q[V^\pi \circ V^\pi] \leq (1 - \gamma)^{-2}$ for any probability distribution q . Substituting the above result into (3.7) and (3.10) and applying the basic inequality $\|V^\pi\|_\infty \leq (1 - \gamma)^{-1}$ give

$$\|\hat{V}^\pi - V^\pi\|_\infty = \gamma \|(I - \gamma \hat{P}_\pi)^{-1} (\hat{P}_\pi - P_\pi) V^\pi\|_\infty$$

$$\leq C_0 \left\{ \sqrt{\frac{\log(\frac{S}{\delta})}{(1-\gamma)^4 N}} + \frac{\log(\frac{S}{\delta})}{(1-\gamma)^2 N} \right\} \quad (3.12)$$

with probability exceeding $1 - \delta$.

The high-probability bound (3.12) reveals that: in order to attain ε -accurate value estimation (in an ℓ_∞ sense), it suffices to draw

$$N = \tilde{O}\left(\frac{1}{(1-\gamma)^4 \varepsilon^2}\right)$$

samples per state from the generative model. Nevertheless, this simple analysis incurs an extra multiplicative factor of $\frac{1}{1-\gamma}$ compared to the result stated in Theorem 3.1, thus falling short of tightness for problems with long effective horizons. Upon closer inspection, the separate treatment of $(I - \gamma \hat{P}_\pi)^{-1}$ and $\sqrt{\text{Var}_{P_\pi}(V^\pi)}$ overlooks their intimate connection that could allow for tighter control. Furthermore, we have not fully exploited the variance structure in (3.8); only crude upper bounds are applied. These factors together account for the suboptimality of the above analysis.

3.2.4 Analysis: a second attempt

We now refine our analysis for the value estimation error of interest, leveraging the interaction between $(I - \gamma \hat{P}_\pi)^{-1}$ and $\sqrt{\text{Var}_{P_\pi}(V^\pi)}$, as well as the underlying variance structure. To simplify the presentation and concentrate our discussion on the main term, we first note — without a formal proof — that, for N large enough,

$$(I - \gamma \hat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} \approx (I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)}.$$

This motivates us to control $(I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)}$.

Recall from Lemma 2.3 that $(I - \gamma P_\pi)^{-1}$ is a non-negative matrix with each row summing to $\frac{1}{1-\gamma}$, thus indicating that $(1-\gamma)(I - \gamma P_\pi)^{-1}$ is a valid probability transition matrix. This inspires one to push $(I - \gamma P_\pi)^{-1}$ into the square root operator, yielding

$$\begin{aligned} (I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} &= \frac{1}{1-\gamma} \left\{ (1-\gamma)(I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} \right\} \\ &\leq \frac{1}{1-\gamma} \sqrt{(1-\gamma)(I - \gamma P_\pi)^{-1} \text{Var}_{P_\pi}(V^\pi)} \\ &= \sqrt{\frac{1}{1-\gamma} (I - \gamma P_\pi)^{-1} \text{Var}_{P_\pi}(V^\pi)}. \end{aligned} \quad (3.13)$$

Here, we apply Jensen's inequality: for any probability vector q and any non-negative random variable v , it holds that $\mathbb{E}_q[\sqrt{v}] \leq \sqrt{\mathbb{E}_q[v]}$.

To better exploit the underlying variance structure, rewrite and upper-bound $\text{Var}_{P_\pi}(V^\pi)$ (cf. (2.23)) as follows:

$$\begin{aligned} \text{Var}_{P_\pi}(V^\pi) &= P_\pi(V^\pi \circ V^\pi) - (P_\pi V^\pi) \circ (P_\pi V^\pi) \\ &= P_\pi(V^\pi \circ V^\pi) - \frac{1}{\gamma^2} (V^\pi - r_\pi) \circ (V^\pi - r_\pi) \\ &= \frac{1}{\gamma^2} (\gamma^2 P_\pi - I)(V^\pi \circ V^\pi) + \frac{2}{\gamma^2} V^\pi \circ r_\pi - \frac{1}{\gamma^2} r_\pi \circ r_\pi \end{aligned}$$

$$\leq \frac{1}{\gamma^2}(\gamma P_\pi - I)(V^\pi \circ V^\pi) + \frac{2}{\gamma^2}V^\pi \circ r_\pi,$$

where the second line applies the Bellman equation $V^\pi = r_\pi + \gamma P_\pi V^\pi$ (cf. (2.24b)). This yields a simple upper bound on $(I - \gamma P_\pi)^{-1} \text{Var}_{P_\pi}(V^\pi)$:

$$\begin{aligned} 0 &\leq (I - \gamma P_\pi)^{-1} \text{Var}_{P_\pi}(V^\pi) \\ &\leq \frac{1}{\gamma^2}(I - \gamma P_\pi)^{-1}(\gamma P_\pi - I)(V^\pi \circ V^\pi) + \frac{2}{\gamma^2}(I - \gamma P_\pi)^{-1}(V^\pi \circ r_\pi) \\ &= -\frac{1}{\gamma^2}V^\pi \circ V^\pi + \frac{2}{\gamma^2}(I - \gamma P_\pi)^{-1}(V^\pi \circ r_\pi) \\ &\leq \frac{2}{\gamma^2}(I - \gamma P_\pi)^{-1}(V^\pi \circ r_\pi) \\ &\leq \frac{2\|V^\pi\|_\infty}{\gamma^2}(I - \gamma P_\pi)^{-1}r_\pi \\ &= \frac{2\|V^\pi\|_\infty}{\gamma^2}V^\pi \leq \frac{2\|V^\pi\|_\infty^2}{\gamma^2}1_S, \end{aligned} \tag{3.14}$$

where the last line uses (2.24c). Taking this together with (3.13) yields

$$0 \leq \left\| (I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} \right\|_\infty \leq \sqrt{\frac{2}{1-\gamma}} \frac{\|V^\pi\|_\infty}{\gamma}. \tag{3.15}$$

Substituting this into (3.7) and (3.10), and using the basic fact that $\|V^\pi\|_\infty \leq \frac{1}{1-\gamma}$, we arrive at

$$\begin{aligned} \|\hat{V}^\pi - V^\pi\|_\infty &\leq C_0 \left\{ \frac{\gamma}{\sqrt{N}} \left\| (I - \gamma \hat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} \right\|_\infty + \frac{\gamma}{(1-\gamma)^2 N} \right\} \\ &\approx C_0 \left\{ \frac{\gamma}{\sqrt{N}} \left\| (I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}(V^\pi)} \right\|_\infty + \frac{\gamma}{(1-\gamma)^2 N} \right\} \\ &\leq C_0 \left\{ \sqrt{\frac{2}{N(1-\gamma)^3}} + \frac{1}{(1-\gamma)^2 N} \right\} \\ &\leq 2C_0 \sqrt{\frac{2}{N(1-\gamma)^3}}, \end{aligned}$$

provided that N is sufficiently large. This reveals that ε -accurate value estimation is attainable — at least in the asymptotic regime with N (resp. ε) approaching infinity (resp. 0) — when the sample size per state exceeds the following order

$$N = \tilde{O}\left(\frac{1}{(1-\gamma)^3 \varepsilon^2}\right),$$

thereby matching the advertised result in Theorem 3.1.

However, an obvious limitation of the above analysis stems from the stringent requirement in N : the sharpness of the above analysis is only guaranteed for very large N . Additional efforts are needed to extend the claimed results to the data-limited regime, i.e., the situation where ε is not exceedingly small.

3.2.5 Sharp analysis: proof of Theorem 3.1

In order to further improve the analysis strategy outlined in Section 3.2.4, a key step lies in controlling the following term:

$$\left((I - \gamma \hat{P}_\pi)^{-1} - (I - \gamma P_\pi)^{-1} \right) \sqrt{\text{Var}_{P_\pi}(V^\pi)}, \quad (3.16)$$

as it determines the range of N within which optimality can be achieved. Interestingly, the analyses presented in Sections 3.2.3 and 3.2.4 already shed light on how to handle this term. More specifically, if we express the value estimation error as

$$\hat{V}^\pi - V^\pi = \left((I - \gamma \hat{P}_\pi)^{-1} - (I - \gamma P_\pi)^{-1} \right) r_\pi$$

according to the Bellman equation (2.24c), then the term (3.16) can be interpreted as a sort of value estimation error with the reward function taken to be $r^{(1)} = \sqrt{\text{Var}_{P_\pi}(V^\pi)}$, which can be tackled using arguments in Section 3.2.4. However, applying this approach results in a similar issue: the need to bound an even “higher-order” term

$$\left((I - \gamma \hat{P}_\pi)^{-1} - (I - \gamma P_\pi)^{-1} \right) \sqrt{\text{Var}_{P_\pi}(V^{(1)})}$$

with $V^{(1)} := (I - \gamma P_\pi)^{-1} r^{(1)}$. Motivated by this line of thought, we propose a recursive analysis approach, with the aim of controlling the residual terms in an increasingly more refined fashion.

Step 1: recursive construction of auxiliary rewards and values. To apply the above recursive arguments, construct a sequence of auxiliary “reward” and “value” functions recursively:

$$\begin{aligned} r^{(0)} &:= r_\pi, & V^{(0)} &:= (I - \gamma P_\pi)^{-1} r^{(0)}, \\ r^{(l)} &:= \sqrt{\text{Var}_{P_\pi}[V^{(l-1)}]}, & V^{(l)} &:= (I - \gamma P_\pi)^{-1} r^{(l)}, \quad l \geq 1. \end{aligned} \quad (3.17a)$$

Similarly, we find it convenient to introduce the auxiliary “value estimates” as follows:

$$\hat{V}^{(l)} := (I - \gamma \hat{P}_\pi)^{-1} r^{(l)}, \quad l \geq 0. \quad (3.17b)$$

As an immediate consequence of inequality (3.15), we can readily establish an entrywise bound on $V^{(l)}$ for any $l > 1$:

$$\begin{aligned} \|V^{(l)}\|_\infty &= \|(I - \gamma P_\pi)^{-1} r^{(l)}\|_\infty = \|(I - \gamma P_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}[V^{(l-1)}]}\|_\infty \\ &\leq \frac{\sqrt{2}}{\gamma\sqrt{1-\gamma}} \|V^{(l-1)}\|_\infty \leq \dots \leq \left(\frac{\sqrt{2}}{\gamma\sqrt{1-\gamma}} \right)^l \|V^{(0)}\|_\infty \end{aligned} \quad (3.18)$$

$$= \left(\frac{\sqrt{2}}{\gamma\sqrt{1-\gamma}} \right)^l \|V^\pi\|_\infty \leq \frac{1}{1-\gamma} \left(\frac{\sqrt{2}}{\gamma\sqrt{1-\gamma}} \right)^l, \quad (3.19)$$

where the last line holds since $\|V^\pi\|_\infty \leq \frac{1}{1-\gamma}$.

Step 2: establishing a recursion for $\hat{V}^{(l)} - V^{(l)}$. Equipped with the above notation, we observe that, for any $l \geq 0$,

$$\begin{aligned} \hat{V}^{(l)} - V^{(l)} &= (I - \gamma \hat{P}_\pi)^{-1} r^{(l)} - V^{(l)} \\ &= (I - \gamma \hat{P}_\pi)^{-1} (I - \gamma P_\pi) V^{(l)} - (I - \gamma \hat{P}_\pi)^{-1} (I - \gamma \hat{P}_\pi) V^{(l)} \end{aligned}$$

$$= \gamma(I - \gamma\hat{P}_\pi)^{-1}(\hat{P}_\pi - P_\pi)V^{(l)}, \quad (3.20)$$

where the second line follows from the identity $(I - \gamma P_\pi)V^{(l)} = r^{(l)}$ (see the construction (3.17a)).

The Bernstein inequality in Lemma 2.5 together with the union bound asserts that, with probability exceeding $1 - \delta$, one has

$$|(\hat{P}_\pi - P_\pi)V^{(l)}| \leq \sqrt{\frac{\beta_1}{N}} \sqrt{\text{Var}_{P_\pi}[V^{(l)}]} + \frac{\|V^{(l)}\|_\infty \beta_1}{N} 1_S \quad (3.21)$$

uniformly for all $0 \leq l \leq m$, where m is some large enough integer to be specified shortly, and $\beta_1 = C_0 \log \frac{Sm}{\delta}$ for some large enough constant $C_0 > 0$. Combining this result with (3.20) reveals that

$$\begin{aligned} \|\hat{V}^{(l)} - V^{(l)}\|_\infty &= \gamma \|(I - \gamma\hat{P}_\pi)^{-1}(\hat{P}_\pi - P_\pi)V^{(l)}\|_\infty \\ &\stackrel{(i)}{\leq} \gamma \|(I - \gamma\hat{P}_\pi)^{-1}|(\hat{P}_\pi - P_\pi)V^{(l)}|\|_\infty \\ &\stackrel{(ii)}{\leq} \gamma \sqrt{\frac{\beta_1}{N}} \|(I - \gamma\hat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}[V^{(l)}]}\|_\infty + \frac{\gamma \|V^{(l)}\|_\infty \beta_1}{(1 - \gamma)N} \end{aligned} \quad (3.22)$$

$$\begin{aligned} &\stackrel{(iii)}{\leq} \gamma \sqrt{\frac{\beta_1}{N}} \|\hat{V}^{(l+1)} - V^{(l+1)}\|_\infty + \gamma \sqrt{\frac{\beta_1}{N}} \|V^{(l+1)}\|_\infty + \frac{\gamma \|V^{(l)}\|_\infty \beta_1}{(1 - \gamma)N} \\ &\stackrel{(iv)}{\leq} \gamma \sqrt{\frac{\beta_1}{N}} \|\hat{V}^{(l+1)} - V^{(l+1)}\|_\infty + \left(\sqrt{\frac{2\beta_1}{(1 - \gamma)N}} + \frac{\beta_1}{(1 - \gamma)N} \right) \|V^{(l)}\|_\infty \end{aligned} \quad (3.23)$$

with probability exceeding $1 - \delta$. Here, (i) follows since $(I - \gamma\hat{P}_\pi)^{-1}$ is a non-negative matrix, (ii) comes from (3.21) and the triangle inequality, (iii) arises from the definitions of $r^{(l)}$, $V^{(l)}$ and $\hat{V}^{(l)}$, while (iv) is a consequence of (3.18). Crucially, inequality (3.23) establishes a recursive relation that enables a sharp upper bound analysis.

Step 3: upper bounding via the above recursion. Define

$$b_1 := \gamma \sqrt{\frac{\beta_1}{N}}, \quad b_2 := \sqrt{\frac{2\beta_1}{(1 - \gamma)N}} + \frac{\beta_1}{(1 - \gamma)N}, \quad b_3 := \frac{\sqrt{2}}{\gamma \sqrt{1 - \gamma}} \quad (3.24)$$

for notational simplicity. Combining inequalities (3.19) and (3.23) shows that with probability at least $1 - \delta$,

$$\begin{aligned} \|\hat{V}^{(l)} - V^{(l)}\|_\infty &\leq b_1 \|\hat{V}^{(l+1)} - V^{(l+1)}\|_\infty + b_2 \|V^{(l)}\|_\infty \\ &\leq b_1 \|\hat{V}^{(l+1)} - V^{(l+1)}\|_\infty + b_2 b_3^l \|V^{(0)}\|_\infty \end{aligned}$$

holds for all $0 \leq l < m$. Apply this relation recursively and combine terms to arrive at

$$\begin{aligned} \|\hat{V}^{(0)} - V^{(0)}\|_\infty &\leq b_1 \|\hat{V}^{(1)} - V^{(1)}\|_\infty + b_2 \|V^{(0)}\|_\infty \\ &\leq b_1 \{b_1 \|\hat{V}^{(2)} - V^{(2)}\|_\infty + b_2 b_3 \|V^{(0)}\|_\infty\} + b_2 \|V^{(0)}\|_\infty \\ &\leq \dots \\ &\leq \underbrace{b_1^m \|\hat{V}^{(m)} - V^{(m)}\|_\infty}_{=: \alpha_1} + \underbrace{b_2 \sum_{l=0}^{m-1} (b_1 b_3)^l \|V^{(0)}\|_\infty}_{=: \alpha_2}, \end{aligned} \quad (3.25)$$

leaving us with two terms to control.

Step 4: controlling quantity α_1 . To bound quantity α_1 , we invoke identity (3.22) to obtain

$$\begin{aligned}\|\widehat{V}^{(m)} - V^{(m)}\|_\infty &\leq \gamma \sqrt{\frac{\beta_1}{N}} \|(I - \gamma \widehat{P}_\pi)^{-1} \sqrt{\text{Var}_{P_\pi}[V^{(m)}]}\|_\infty + \frac{\gamma \beta_1 \|V^{(m)}\|_\infty}{(1 - \gamma)N} \\ &\leq \gamma \sqrt{\frac{\beta_1}{N}} \|(I - \gamma \widehat{P}_\pi)^{-1}\|_1 \|\sqrt{\text{Var}_{P_\pi}[V^{(m)}]}\|_\infty + \frac{\gamma \beta_1 \|V^{(m)}\|_\infty}{(1 - \gamma)N} \\ &\leq \frac{\gamma}{1 - \gamma} \sqrt{\frac{\beta_1}{N}} \|V^{(m)}\|_\infty + \frac{\gamma \beta_1 \|V^{(m)}\|_\infty}{(1 - \gamma)N},\end{aligned}$$

where the last inequality uses the basic fact $\|(I - \gamma \widehat{P}_\pi)^{-1}\|_1 = \frac{1}{1 - \gamma}$ (cf. (2.26c)) as well as the elementary relation $\text{Var}_{P_\pi}[V^{(m)}] \leq \|V^{(m)}\|_\infty^2$. This combined with inequality (3.19) yields

$$\|\widehat{V}^{(m)} - V^{(m)}\|_\infty \leq \frac{\gamma}{(1 - \gamma)^2} \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N} \right) \left(\frac{\sqrt{2}}{\gamma \sqrt{1 - \gamma}} \right)^m.$$

As a consequence, we arrive at

$$\begin{aligned}\alpha_1 &= b_1^m \|\widehat{V}^{(m)} - V^{(m)}\|_\infty \\ &\leq \left(\gamma \sqrt{\frac{\beta_1}{N}} \right)^m \frac{\gamma}{(1 - \gamma)^2} \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N} \right) \left(\frac{\sqrt{2}}{\gamma \sqrt{1 - \gamma}} \right)^m \\ &\leq \left(\sqrt{\frac{2\beta_1}{(1 - \gamma)N}} \right)^m \frac{\gamma}{(1 - \gamma)^2} \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N} \right) \\ &\leq \left(\frac{1}{e} \right)^m \frac{2}{(1 - \gamma)^2} \sqrt{\frac{\beta_1}{N}},\end{aligned}\tag{3.26}$$

where the last line holds as long as $N > \frac{2e^2}{1 - \gamma} \beta_1$.

Step 5: controlling quantity α_2 . Regarding quantity α_2 , we note that $b_1 b_3 \leq 1/2$ holds as long as $N \geq 64\beta_1/(1 - \gamma)$. Summing the terms of the geometric sequence leads to

$$b_2 \sum_{l=0}^{m-1} (b_1 b_3)^l \leq \frac{b_2}{1 - b_1 b_3} \leq \sqrt{\frac{8\beta_1}{(1 - \gamma)N}} + \frac{2\beta_1}{(1 - \gamma)N},$$

which taken collectively with the bound $\|V^{(0)}\|_\infty \leq \frac{1}{1 - \gamma}$ gives

$$\alpha_2 = \left\{ b_2 \sum_{l=0}^{m-1} (b_1 b_3)^l \right\} \|V^{(0)}\|_\infty \leq \sqrt{\frac{8\beta_1}{(1 - \gamma)^3 N}} + \frac{2\beta_1}{(1 - \gamma)^2 N},\tag{3.27}$$

provided that $N \geq \frac{64\beta_1}{1 - \gamma}$.

Step 6: putting all this together. Taking inequalities (3.25), (3.26) and (3.27) together shows that, with probability at least $1 - \delta$,

$$\|\widehat{V}^{(0)} - V^{(0)}\|_\infty \leq \left(\frac{1}{e} \right)^m \frac{2}{(1 - \gamma)^2} \sqrt{\frac{\beta_1}{N}} + \sqrt{\frac{8\beta_1}{(1 - \gamma)^3 N}} + \frac{2\beta_1}{(1 - \gamma)^2 N}$$

holds as long as $N \geq \frac{64\beta_1}{1-\gamma}$. To finish up, set $m = \lceil \log(\frac{e}{1-\gamma}) \rceil$ to derive

$$\begin{aligned} \|\widehat{V}^\pi - V^\pi\|_\infty &= \|\widehat{V}^{(0)} - V^{(0)}\|_\infty \leq 2\sqrt{\frac{8\beta_1}{(1-\gamma)^3N}} + \frac{2\beta_1}{(1-\gamma)^2N} \\ &\leq 3\sqrt{\frac{8\beta_1}{(1-\gamma)^3N}} \end{aligned} \quad (3.28)$$

with probability exceeding $1 - \delta$, provided that $N \geq \frac{\beta_1}{1-\gamma}$. This concludes the proof by recalling that $\beta_1 = C_0 \log(\frac{Sm}{\delta})$ and $m = \lceil \log(\frac{e}{1-\gamma}) \rceil$.

A slight generalization of Theorem 3.1. Before moving on, we record a slight extension of Theorem 3.1. Its proof has already been given above, but it will play a useful role in the analysis of policy learning in subsequent sections.

Theorem 3.2. Suppose that there exists some quantity $\beta_1 > 0$ such that $\{V^{(l)}\}$ (cf. (3.17)) obeys

$$\left| (\widehat{P}_\pi - P_\pi)V^{(l)} \right| \leq \sqrt{\frac{\beta_1}{N}} \sqrt{\text{Var}_{P_\pi}[V^{(l)}]} + \frac{\beta_1 \|V^{(l)}\|_\infty}{N} \mathbf{1}_S, \quad (3.29)$$

for all $0 \leq l \leq \log(\frac{e}{1-\gamma})$. Suppose that $N \geq \frac{64}{1-\gamma}\beta_1$. Then the vectors $V^\pi = (I - \gamma P_\pi)^{-1}r_\pi$ and $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1}r_\pi$ satisfy

$$\|\widehat{V}^\pi - V^\pi\|_\infty \leq \frac{3}{1-\gamma} \sqrt{\frac{8\beta_1}{N(1-\gamma)}}.$$

Importantly, Theorem 3.2 does *not* require full statistical independence between policy π and the data samples (or $\widehat{P}$), as long as the Bernstein-style condition (3.29) is satisfied. Although (3.29) holds with high probability for some reasonably small β_1 if π is statistically independent of $\widehat{P}$ (as in the proof of Theorem 3.1), this generalization also accommodates the case where π exhibits a form of “weak” statistical dependency on the data. As we shall see momentarily, this generalization underpins our subsequent analysis of policy learning.

3.3 Model-based policy learning

Having established the effectiveness of the model-based approach for policy evaluation, we now turn attention to model-based policy learning, which seeks to identify a policy that approximately maximizes the value function based on the data samples at hand. To be precise, suppose that the action space $\mathcal{A}$ contains $A \geq 2$ actions, then for any target accuracy level $\varepsilon > 0$, the aim of policy learning is to compute an ε -accurate policy π_{est} obeying

$$\forall s \in \mathcal{S} : \quad V^{\pi_{\text{est}}}(s) \geq V^*(s) - \varepsilon.$$

Akin to the policy evaluation case, the meaningful range of target precision level ε is $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Naturally, one would hope to achieve policy learning using as few samples as possible.

3.3.1 Algorithms

A model-based policy learning algorithm typically starts by constructing an empirical MDP $\widehat{\mathcal{M}}$ based on all collected data samples, mirroring what is used in model-based policy evaluation. The empirical model is then fed into the Bellman recursion — via methods such as policy iteration and value iteration introduced in Section 2.1.4 — to derive the optimal policy for the empirical model.

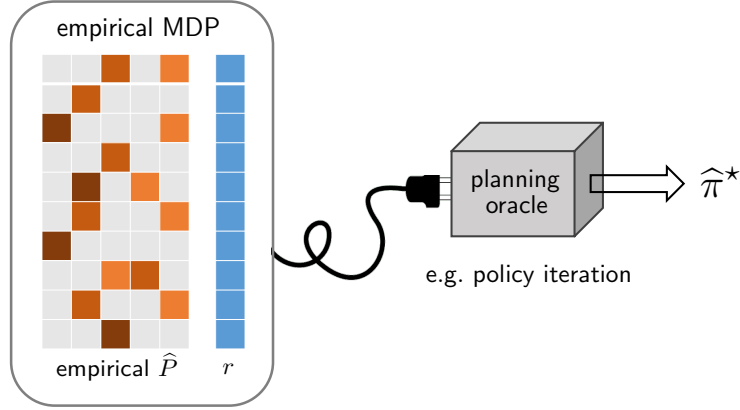


Figure 3.1: An illustration of the model-based policy learning algorithm, which first computes an empirical model and then plugs it into a planning algorithm.

A model-based approach. For concreteness, let us describe below the most natural version of the model-based policy learning algorithm.

- 1) *Model estimation.* For each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, draw N independent samples from the generative model:

$$s^{(i)}(s, a) \stackrel{\text{i.i.d.}}{\sim} P(\cdot | s, a), \quad i = 1, \dots, N,$$

and construct an empirical transition kernel $\hat{P}$ as follows

$$\forall s' \in \mathcal{S}, \quad \hat{P}(s' | s, a) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{s' = s^{(i)}(s, a)\}. \quad (3.30)$$

In words, $\hat{P}(s' | s, a)$ records the empirical frequency of transitions from (s, a) to state s' . This leads to an empirical MDP $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{P}, r, \gamma)$ constructed from the data samples.

- 2) *Planning.* Compute an optimal (deterministic) policy

$$\hat{\pi} = \arg \max_{\pi} \hat{V}^{\pi}. \quad (3.31)$$

of the empirical MDP $\widehat{\mathcal{M}}$, using, e.g., policy iteration (see Section 2.1.4). Here and below, $\hat{V}^{\pi}$ denotes the value function of policy π in the empirical MDP $\widehat{\mathcal{M}}$.

See Algorithm 3.2 for the full procedure, and Figure 3.1 for an illustration. Clearly, the total number of samples collected in the above algorithm is given by $N^{\text{total}} := NSA$.

A perturbed model-based approach. We shall also present a “randomly perturbed” variation of the above model-based algorithm, where random reward perturbation is employed to facilitate analysis. More specifically, when constructing the empirical MDP, we randomly perturb the immediate reward for each state-action pair (s, a) as follows:

$$r_p(s, a) = r(s, a) + \zeta(s, a), \quad \zeta(s, a) \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, \xi), \quad (3.32)$$

Algorithm 3.2: Model-based policy learning with a simulator

```

1 input: discount factor  $\gamma$ , reward function  $r$ , sample size per state-action pair  $N$ .
2 for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
    /* model estimation based on samples. */
3     draw  $N$  independent samples  $s^{(i)}(s, a) \stackrel{\text{i.i.d.}}{\sim} P(\cdot | s, a)$ .
4     compute  $\hat{P}(\cdot | s, a)$  based on (3.30).
    /* planning w.r.t. the empirical model. */
5 output  $\hat{\pi}$  as an optimal (deterministic) policy of the MDP  $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{P}, r, \gamma)$ .

```

Algorithm 3.3: Perturbed model-based policy learning with a simulator

```

1 input: discount factor  $\gamma$ , reward function  $r$ , sample size per state-action pair  $N$ .
2 for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
    /* obtain an empirical model and perturb it. */
3     draw  $N$  independent samples  $s^{(i)}(s, a) \stackrel{\text{i.i.d.}}{\sim} P(\cdot | s, a)$ .
4     compute  $\hat{P}(\cdot | s, a)$  based on (3.30).
5     generate a perturbed immediate reward  $r_p(s, a)$  as (3.32).
    /* planning w.r.t. the perturbed empirical model. */
6 output  $\hat{\pi}_p$  as an optimal (deterministic) policy of the MDP  $\hat{\mathcal{M}}_p = (\mathcal{S}, \mathcal{A}, \hat{P}, r_p, \gamma)$ .

```

where $\text{Unif}(0, \xi)$ denotes the uniform distribution between 0 and some small quantity $\xi > 0$ (to be specified momentarily).¹ The perturbed empirical MDP is then taken to be $\hat{\mathcal{M}}_p = (\mathcal{S}, \mathcal{A}, \hat{P}, r_p, \gamma)$, using the probability transition kernel $\hat{P}$ (cf. (3.30)) and the perturbed reward function $r_p = [r_p(s, a)]_{(s,a) \in \mathcal{S} \times \mathcal{A}}$. The perturbed model-based approach then returns the optimal policy

$$\hat{\pi}_p^* := \arg \max_{\pi} \hat{V}_p^{\pi} \quad (3.33)$$

of the perturbed empirical MDP $\hat{\mathcal{M}}_p$, where $\hat{\pi}_p^*$ is taken to be a deterministic policy. Here and throughout, we shall employ $\hat{V}_p^{\pi}$ to denote the value function of policy π in the perturbed empirical MDP $\hat{\mathcal{M}}_p$. The whole procedure is summarized in Algorithm 3.3. Clearly, the total sample size of this variant is also $N^{\text{total}} := N\mathcal{S}\mathcal{A}$. As we shall see momentarily, random perturbation of the reward function helps break ties and ensure the uniqueness of the optimal policy, thereby simplifying analysis.

3.3.2 Theoretical guarantees

We now state the sample complexity guarantee for the perturbed model-based method in Algorithm 3.3.

Theorem 3.3 (Li *et al.*, 2020; Li *et al.*, 2024c). There exist some universal constants $c_0, c_1 > 0$ such that: for any $\delta > 0$ and any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the policy $\hat{\pi}_p^*$ returned by the perturbed model-based

¹Note that perturbation is only invoked when running the planning algorithms and does not require collecting new samples.

algorithm in Algorithm 3.3 achieves

$$\forall s \in \mathcal{S}, \quad V^{\hat{\pi}_p^*}(s) \geq V^*(s) - \varepsilon \quad (3.34a)$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad Q^{\hat{\pi}_p^*}(s, a) \geq Q^*(s, a) - \gamma\varepsilon \quad (3.34b)$$

with probability at least $1 - \delta$, provided that the perturbation size is $\xi = \frac{c_1(1-\gamma)\varepsilon}{S^5 A^5}$ and that the sample size per state-action pair exceeds

$$N \geq \frac{c_0 \log\left(\frac{SA}{(1-\gamma)\varepsilon\delta}\right)}{(1-\gamma)^3 \varepsilon^2}. \quad (3.35)$$

In a nutshell, the above theorem demonstrates that the model-based approach succeeds in finding an ε -optimal policy as soon as the total sample complexity exceeds the order of $\frac{SA}{(1-\gamma)^3 \varepsilon^2}$ (modulo some log factor). This sample complexity is provably minimax optimal, as we will elaborate in Section 3.4. Importantly, this result imposes no restriction on the range of ε and, in particular, the accuracy level ε is permitted to be as large as $\frac{1}{1-\gamma}$. As a consequence, this theorem is applicable even in regimes with small-to-moderate sample sizes, since its validity is guaranteed as long as

$$N \geq \tilde{\Omega}\left(\frac{1}{1-\gamma}\right). \quad (3.36)$$

Noteworthy, Condition (3.36) is necessary for learning a policy that performs strictly better than a random guess, and hence characterizes the full range of “meaningful” sample sizes.

From a computational standpoint, both the empirical VI and PI algorithms w.r.t. $\hat{\mathcal{M}}_p$ are able to compute $\hat{\pi}_p^*$ within $O\left(\frac{1}{1-\gamma} \log\left(\frac{SA}{(1-\gamma)\varepsilon\delta}\right)\right)$ iterations, thereby yielding computationally lightweight solutions.

3.3.3 Analysis: a first attempt

We now proceed to establish performance guarantees for policy learning. Note that it suffices to analyze the suboptimality (3.34a) of $\hat{\pi}_p^*$ w.r.t. value functions. Clearly, if (3.34a) can be established, then the Bellman equation gives

$$\begin{aligned} Q^{\hat{\pi}_p^*}(s, a) - Q^*(s, a) &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[V^{\hat{\pi}_p^*}(s')] - (r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[V^*(s')]) \\ &= \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[V^{\hat{\pi}_p^*}(s') - V^*(s')] \geq -\gamma\varepsilon, \end{aligned}$$

as claimed in (3.34b).

In this subsection, we begin by examining the performance of the model-based policy learning algorithm in Algorithm 3.2 without reward perturbation; we will return shortly to elucidate the role of reward perturbation. We first present a simple approach to establishing asymptotically optimal statistical guarantees (for large enough N) for Algorithm 3.2, albeit at the cost of a significant burn-in requirement. For simplicity, in this subsection we assume (without providing a formal proof) that

$$\hat{P} \approx P, \quad (3.37)$$

which holds in the asymptotic regime as $N \rightarrow \infty$, but still leaves substantial room for improvement from a non-asymptotic perspective. In Section 3.3.4, we will relax this assumption (and the large- N requirement) and extend the analysis all the way to the most sample-limited regime. Before proceeding, let us introduce some useful notation:

- $\widehat{V}^\pi$ and $\widehat{Q}^\pi$: the value function and Q-function of policy π in the empirical MDP $\widehat{\mathcal{M}}$, respectively;
- $\widehat{\pi}^*$: an optimal *deterministic* policy of $\widehat{\mathcal{M}}$ (returned by Algorithm 3.2);
- $\widehat{V}^* := \widehat{V}^{\widehat{\pi}^*}$ and $\widehat{Q}^* := \widehat{Q}^{\widehat{\pi}^*}$: the optimal value function and optimal Q-function in $\widehat{\mathcal{M}}$.

To bound the suboptimality of $\widehat{\pi}^*$ in terms of the value function, we start with the following elementary decomposition:

$$\begin{aligned}
0 \leq V^* - V^{\widehat{\pi}^*} &= V^{\pi^*} - V^{\widehat{\pi}^*} \\
&= (V^{\pi^*} - \widehat{V}^{\pi^*}) + (\widehat{V}^{\pi^*} - \widehat{V}^{\widehat{\pi}^*}) + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \\
&\leq (V^{\pi^*} - \widehat{V}^{\pi^*}) + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}),
\end{aligned} \tag{3.38}$$

where the last line follows since $\widehat{\pi}^* = \arg \max_{\pi} \widehat{V}^\pi$ and hence $\widehat{V}^{\pi^*} \leq \widehat{V}^{\widehat{\pi}^*}$. We are thus left with two terms to control (see (3.38)), both of which are concerned with certain policy evaluation errors (i.e., the policy evaluation errors w.r.t. π^* and $\widehat{\pi}^*$).

Step 1: controlling $V^{\pi^*} - \widehat{V}^{\pi^*}$. Given that π^* is the optimal policy of the true MDP and hence independent of the data, we can readily accomplish this step by invoking the policy evaluation result presented in Theorem 3.1. Specifically, taking $\pi = \pi^*$ in the analysis of Theorem 3.1 (see (3.28)) shows that, with probability at least $1 - \delta$,

$$\|\widehat{V}^{\pi^*} - V^{\pi^*}\|_\infty \leq c_1 \sqrt{\frac{\log\left(\frac{S}{\delta(1-\gamma)}\right)}{N(1-\gamma)^3}} \tag{3.39}$$

as long as $N \geq \frac{c_2 \log\left(\frac{S}{\delta(1-\gamma)}\right)}{1-\gamma}$, where $c_1, c_2 > 0$ are universal constants.

Step 2: controlling $\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}$. Bounding the policy evaluation error $\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}$ is, however, substantially more complex than bounding $V^{\pi^*} - \widehat{V}^{\pi^*}$. The complexity stems from the intricate statistical dependency between $\widehat{\pi}^*$ and the data samples, rendering Theorem 3.1 inapplicable. We present here an argument valid in the asymptotic regime as $N \rightarrow \infty$. Since this approach is relatively loose in the non-asymptotic regime — and will be sharpened in Section 3.3.4 — we omit the control of second-order (or higher-order) terms here in order to streamline the presentation and better highlight the main ideas.

Note that the Bellman equation (2.24c) implies that

$$\begin{aligned}
\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*} &= (I - \gamma \widehat{P}_{\widehat{\pi}^*})^{-1} r_{\widehat{\pi}^*} - V^{\widehat{\pi}^*} \\
&= (I - \gamma \widehat{P}_{\widehat{\pi}^*})^{-1} \left\{ (I - \gamma P_{\widehat{\pi}^*}) V^{\widehat{\pi}^*} - (I - \gamma \widehat{P}_{\widehat{\pi}^*}) V^{\widehat{\pi}^*} \right\} \\
&= \gamma (I - \gamma \widehat{P}_{\widehat{\pi}^*})^{-1} (\widehat{P}_{\widehat{\pi}^*} - P_{\widehat{\pi}^*}) V^{\widehat{\pi}^*}
\end{aligned} \tag{3.40}$$

$$= \gamma (I - \gamma \widehat{P}_{\widehat{\pi}^*})^{-1} (\widehat{P}_{\widehat{\pi}^*} - P_{\widehat{\pi}^*}) V^{\pi^*} + \gamma (I - \gamma \widehat{P}_{\widehat{\pi}^*})^{-1} (\widehat{P}_{\widehat{\pi}^*} - P_{\widehat{\pi}^*}) (V^{\widehat{\pi}^*} - V^{\pi^*}) \tag{3.41}$$

$$\approx \gamma (I - \gamma \widehat{P}_{\widehat{\pi}^*})^{-1} (\widehat{P}_{\widehat{\pi}^*} - P_{\widehat{\pi}^*}) V^{\pi^*}, \tag{3.42}$$

where we remind the readers of the matrix notation P_π and r_π in (2.19) and (2.20), respectively. Informally, the first term in (3.41) is easier to cope than its counterpart in (3.40), since V^{π^*} is a fixed

vector independent from $\hat{P}$; meanwhile, the last term in (3.41) can be regarded as a second-order term (as it involves both $\hat{P} - P$ and $V^{\hat{\pi}^*} - V^{\pi^*}$), and is therefore expected to be asymptotically negligible as $N \rightarrow \infty$, a key observation that yields (3.42).

To bound (3.42), it is first seen from (2.19) and the Bernstein inequality in Lemma 2.5 that, with probability greater than $1 - \delta$,

$$\begin{aligned}
|(\hat{P}_{\hat{\pi}^*} - P_{\hat{\pi}^*})V^{\pi^*}| &= \Pi_{\hat{\pi}^*}|(\hat{P} - P)V^{\pi^*}| \\
&\leq \Pi_{\hat{\pi}^*} \left(\sqrt{\frac{2 \log(\frac{4SA}{\delta})}{N}} \sqrt{\text{Var}_P(V^{\pi^*})} + \frac{2\|V^{\pi^*}\|_\infty \log(\frac{4SA}{\delta})}{3N} 1_{SA} \right) \\
&\leq \sqrt{\frac{2 \log(\frac{4SA}{\delta})}{N}} \sqrt{\text{Var}_{P_{\hat{\pi}^*}}(V^{\pi^*})} + \frac{2 \log(\frac{4SA}{\delta})}{3N(1-\gamma)} 1_S \\
&\approx \sqrt{\frac{2 \log(\frac{4SA}{\delta})}{N}} \sqrt{\text{Var}_{P_{\hat{\pi}^*}}(V^{\pi^*})}, \tag{3.43}
\end{aligned}$$

where the penultimate line holds since $\|V^{\pi^*}\|_\infty \leq \frac{1}{1-\gamma}$, and in the last line we omit the higher-order term proportional to $1/N$ (as the other term scales with $1/\sqrt{N}$). In view of the non-negativity of $(I - \gamma\hat{P}_{\hat{\pi}^*})^{-1}$ (cf. Lemma 2.3), it follows from (3.43) that

$$\begin{aligned}
\gamma|(I - \gamma\hat{P}_{\hat{\pi}^*})^{-1}(\hat{P}_{\hat{\pi}^*} - P_{\hat{\pi}^*})V^{\pi^*}| &\leq \gamma(I - \gamma\hat{P}_{\hat{\pi}^*})^{-1}|(\hat{P}_{\hat{\pi}^*} - P_{\hat{\pi}^*})V^{\pi^*}| \\
&\approx \gamma(I - \gamma P_{\hat{\pi}^*})^{-1}|(\hat{P}_{\hat{\pi}^*} - P_{\hat{\pi}^*})V^{\pi^*}| \\
&\lesssim \gamma \sqrt{\frac{2 \log(\frac{4SA}{\delta})}{N}} (I - \gamma P_{\hat{\pi}^*})^{-1} \sqrt{\text{Var}_{P_{\hat{\pi}^*}}(V^{\pi^*})} \\
&\approx \gamma \sqrt{\frac{2 \log(\frac{4SA}{\delta})}{N}} (I - \gamma P_{\hat{\pi}^*})^{-1} \sqrt{\text{Var}_{P_{\hat{\pi}^*}}(V^{\hat{\pi}^*})} \tag{3.44}
\end{aligned}$$

$$\leq \sqrt{\frac{4 \log(\frac{4SA}{\delta})}{(1-\gamma)N}} \|V^{\hat{\pi}^*}\|_\infty 1_S \leq \sqrt{\frac{4 \log(\frac{4SA}{\delta})}{(1-\gamma)^3 N}} 1_S \tag{3.45}$$

with probability exceeding $1 - \delta$, where the approximation in the second and the fourth lines relies on the assumption (3.37) (and hence $V^{\hat{\pi}^*} \approx V^{\pi^*}$), and the last line invokes (3.15) and the basic property $\|V^{\hat{\pi}^*}\|_\infty \leq \frac{1}{1-\gamma}$. We remark that the approximation in (3.44) serves to align the policy associated with the value function and the one associated with the transition kernel, thereby enabling the application of (3.15). It is worth emphasizing that the use of property (3.15) here is crucial for obtaining the asymptotically sharp bound, as it carefully leverages the underlying variance structure to sharpen the horizon dependency.

Step 3: putting all this together. To finish up, substituting (3.39), (3.42) and (3.45) into (3.38), we can conclude that

$$\|V^* - V^{\hat{\pi}^*}\|_\infty \lesssim \sqrt{\frac{\log(\frac{SA}{\delta(1-\gamma)})}{(1-\gamma)^3 N}}$$

holds with high probability when $N \rightarrow \infty$.

Nevertheless, while the arguments above — similar in spirit to those used in Azar *et al.* (2013) — suffice to establish the desirable statistical bound for the model-based algorithm asymptotically, they rely on an excessively large sample size requirement on N ; for instance, working out the full

details may require N to be at least polynomial in S . To overcome this limitation, a substantially more refined analysis strategy is needed, which we introduce next.

3.3.4 Sharp analysis for Theorem 3.3: leave-one-out analysis

We now present a refined analysis for establishing sharp statistical bounds for $\|V^* - \widehat{V}^{\widehat{\pi}^*}\|_\infty$. In light of the decomposition (3.38) and the bound (3.39), it suffices to derive a tight bound on $\|\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}\|_\infty$ that remains valid even when N is small or moderately sized.

Leave-one-out analysis: key ideas. As alluded to previously, the most challenging part for bounding $\|\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}\|_\infty$ lies in coping with the intricate statistical dependency between $\widehat{\pi}^*$ and $\widehat{P}$. To better illustrate this issue, recall our decomposition (3.40) of the value estimation error as follows:

$$\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*} = \gamma(I - \gamma\widehat{P}_{\widehat{\pi}^*})^{-1}(\widehat{P}_{\widehat{\pi}^*} - P_{\widehat{\pi}^*})V^{\widehat{\pi}^*}. \quad (3.46)$$

Given that $\widehat{\pi}^*$ — and hence $V^{\widehat{\pi}^*}$ — is statistically dependent on $\widehat{P}$, one cannot directly invoke the Bernstein inequality (or other standard concentration inequalities) to control the term $(\widehat{P}_{\widehat{\pi}^*} - P_{\widehat{\pi}^*})V^{\widehat{\pi}^*}$.

To remedy this issue, we resort to a powerful decoupling technique known as the “*leave-one-out analysis*”. Originally developed in the high-dimensional probability and statistics literature (e.g., El Karoui (2018), Ma *et al.* (2020), and Chen *et al.* (2021c)), the leave-one-out analysis was first introduced to RL theory by Agarwal *et al.* (2019b), and later on refined by Li *et al.* (2020) and Li *et al.* (2024c) to sharpen sample complexity analysis. At a high level, this approach consists of the following key ingredients:

- *Construction of leave-one-out auxiliary MDPs.* For each state-action pair (s, a) , construct a small collection of auxiliary MDPs whose transition kernels are nearly identical to that of the original empirical MDP, except that all data samples directly associated with (s, a) are discarded. We refer to these MDPs as “leave-one-out MDPs”, as each of them effectively drops one row of data samples when estimating the transition kernel. The construction is carefully designed so that at least one of these leave-one-out MDPs nearly preserves the optimal Q-function of the original MDP.
- *Stability in model-based policy learning.* Since leave-one-out auxiliary MDPs represent a slight perturbation of the original empirical MDP, the resulting optimal policies are often expected to be exceedingly close to the optimal policy of the empirical MDP. This reflects a form of “stability” in leave-one-out perturbations, a property typically required in leave-one-out decoupling arguments. Moreover, under suitable separation conditions, one can identify, for each state-action pair (s, a) , a leave-one-out MDP from the constructed set that shares exactly the same optimal policy with the original empirical MDP.
- *Harnessing local statistical independence in leave-one-out MDPs.* Since a leave-one-out MDP is built without samples collected for, say, the (s, a) pair, it is independent of the empirical transition vector $\widehat{P}_{s,a}$. As a result, the (s, a) -th entry of the term $(\widehat{P} - P)V^{\widehat{\pi}^*}$ can be bounded using the following observation

$$(\widehat{P}_{s,a} - P_{s,a})V^{\widehat{\pi}^*} = (\widehat{P}_{s,a} - P_{s,a})V^{\widehat{\pi}^*_{\text{loo}}}, \quad (3.47)$$

with $\hat{\pi}_{\text{loo}}^*$ the optimal policy of a leave-one-out MDP that shares the same optimal policy as $\hat{\pi}^*$ while being independent from data tied to (s, a) . Such independence enables the application of standard concentration inequalities to control (3.47), which is crucial in analyzing (3.46) despite complicated statistical dependency.

In what follows, we present a proof that formalizes the above high-level ideas under a certain separation condition that rules out ties in the optimal policy. Specifically, given any $0 < \omega < 1$, we define the following event:

$$\mathcal{B}_\omega := \left\{ \hat{Q}^*(s, \hat{\pi}^*(s)) - \max_{a: a \neq \hat{\pi}^*(s)} \hat{Q}^*(s, a) \geq \omega \text{ for all } s \in \mathcal{S} \right\}. \quad (3.48)$$

where we recall that $\hat{Q}^*$ represents the optimal Q-function of the empirical MDP $\widehat{\mathcal{M}}$. On the event $\mathcal{B}_\omega$, the optimal policy $\hat{\pi}^*$ of $\widehat{\mathcal{M}}$ is clearly unique, since for each state s , the action $\hat{\pi}^*(s)$ yields a strictly higher Q-value compared to any alternative action. Finally, we will show in Section 3.7 how the separation condition in (3.48) can be relaxed with the aid of random reward perturbations.

Preliminaries: absorbing MDPs with identical optimal Q-functions. Before diving into the main analysis, we first explain how, given any MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, one can construct an absorbing MDP — an MDP that contains absorbing state-action pairs (i.e., states that, once entered, cannot be exited by taking the associated action) — while preserving the same optimal Q-functions.

Concretely, consider any state-action pair (s, a) and any given scalar u obeying $|u| \leq 1/(1 - \gamma)$. We construct an auxiliary MDP with the same state space $\mathcal{S}$, action space $\mathcal{A}$ and discount factor γ — denoted by $\mathcal{M}_{s,a,u}$ — that mirrors the original MDP $\mathcal{M}$ except that state s becomes absorbing if action a is taken in this state. More precisely, the transition kernel of $\mathcal{M}_{s,a,u}$ — denoted by $P_{\mathcal{M}_{s,a,u}}$ — is taken to be

$$\begin{aligned} P_{\mathcal{M}_{s,a,u}}(s \mid s, a) &= 1, \\ P_{\mathcal{M}_{s,a,u}}(s' \mid s, a) &= 0, & \text{for all } s' \neq s, \\ P_{\mathcal{M}_{s,a,u}}(\cdot \mid s', a') &= P(\cdot \mid s', a'), & \text{for all } (s', a') \neq (s, a), \end{aligned} \quad (3.49)$$

where the first two lines in (3.49) ensure that state s is absorbing if action a is taken in state s . Meanwhile, the reward function of $\mathcal{M}_{s,a,u}$ — denoted by $r_{\mathcal{M}_{s,a,u}}$ — is taken to be

$$\begin{aligned} r_{\mathcal{M}_{s,a,u}}(s, a) &= u, \\ r_{\mathcal{M}_{s,a,u}}(s', a') &= r(s', a'), & \text{for all } (s', a') \neq (s, a), \end{aligned} \quad (3.50)$$

so that the immediate rewards coincide with those of $\mathcal{M}$ everywhere except at (s, a) . For notational convenience, let $Q_{s,a,u}^\pi$ (resp. $Q_{s,a,u}^*$) represent the Q-function of policy π (resp. optimal Q-function) in $\mathcal{M}_{s,a,u}$, let $V_{s,a,u}^\pi$ (resp. $V_{s,a,u}^*$) represent the value function of policy π (resp. optimal value function) in $\mathcal{M}_{s,a,u}$, and denote by $\pi_{s,a,u}^*$ the optimal policy of $\mathcal{M}_{s,a,u}$.

As it turns out, by properly choosing the immediate reward u , we can guarantee that the absorbing MDP $\mathcal{M}_{s,a,u}$ and the original MDP $\mathcal{M}$ share the same optimal Q-function and value function, as formalized in the following lemma.

Lemma 3.4. Setting $u^* = r(s, a) + \gamma(PV^*)_{(s,a)} - \gamma V^*(s)$, one has

$$Q_{s,a,u^*}^* = Q^* \quad \text{and} \quad V_{s,a,u^*}^* = V^*. \quad (3.51)$$

The proof of this lemma consists of comparing the Bellman optimality equations of $\mathcal{M}$ and $\mathcal{M}_{s,a,u^*}$, and is deferred to Section 3.6.1.

Remark 3.1. It is easily seen that u^* falls within the range $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$.

In summary, for each (s, a) pair, we can form an (s, a) -absorbing MDP $\mathcal{M}_{s,a,u^*}$ that shares the same optimal Q-function and optimal value function as the original MDP $\mathcal{M}$. Moreover, it decouples the randomness arising from the (s, a) pair, as we will soon demonstrate.

Step 1: constructing leave-one-out auxiliary MDPs. As alluded to previously, the leave-one-out analysis employed here requires constructing a collection of leave-one-out auxiliary MDPs. Concretely, for each (s, a) pair and each scalar $u \in [-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$, we can apply the method described above (see (3.49) and (3.50)) to construct an (s, a) -absorbing empirical MDP — denoted by $\widehat{\mathcal{M}}_{s,a,u}$ — associated with the empirical MDP $\widehat{\mathcal{M}}$. In other words, the transition kernel of $\widehat{\mathcal{M}}_{s,a,u}$ coincides with $\widehat{P}$ except that its (s, a) -th row is made absorbing, while the reward function of $\widehat{\mathcal{M}}_{s,a,u}$ agrees with r except that its (s, a) -th entry is replaced by u .

Recall that the (s, a) -th row of $\widehat{P}$ is derived from data samples associated with (s, a) , whereas in $\widehat{\mathcal{M}}_{s,a,u}$ this row is replaced by a standard basis vector with 1 in the s -th entry and 0 elsewhere. As a result, we call each $\widehat{\mathcal{M}}_{s,a,u}$ (for any fixed u) a leave-one-out auxiliary MDP, as it discards one row of data samples. This leave-one-out construction offers a key advantage: for any fixed u , $\widehat{\mathcal{M}}_{s,a,u}$ is statistically independent of all data samples associated with (s, a) , or equivalently, of $\widehat{P}_{s,a}$.

Throughout the rest of the proof, we use the following notation. Let $\widehat{Q}_{s,a,u}^\pi$ (resp. $\widehat{Q}_{s,a,u}^*$) denote the Q-function of policy π (resp. optimal Q-function) in $\widehat{\mathcal{M}}_{s,a,u}$; similarly, use $\widehat{V}_{s,a,u}^\pi$ (resp. $\widehat{V}_{s,a,u}^*$) to represent the value function of policy π (resp. optimal value function) in $\widehat{\mathcal{M}}_{s,a,u}$. Also, $\widehat{\pi}_{s,a,u}^*$ stands for an optimal *deterministic* policy in $\widehat{\mathcal{M}}_{s,a,u}$.

Step 2: linking $\widehat{\pi}^*$ with a small set of policies independent of $\widehat{P}_{s,a}$. With the above leave-one-out construction and Lemma 3.4 in place, one might be tempted to use $\widehat{\mathcal{M}}_{s,a,\widehat{u}^*}$, with

$$\widehat{u}^* := r(s, a) + \gamma(\widehat{P}\widehat{V}^*)_{(s,a)} - \gamma\widehat{V}^*(s),$$

as a surrogate for the empirical MDP $\widehat{\mathcal{M}}$, since the two share identical optimal Q-functions. However, a subtle issue arises: $\widehat{\mathcal{M}}_{s,a,\widehat{u}^*}$ is no longer statistically independent of $\widehat{P}_{s,a}$, because the choice of $\widehat{u}^*$ explicitly depends on $\widehat{P}_{s,a}$. Unfortunately, this dependence undermines the key advantage of the leave-one-out construction described earlier.

Fortunately, the entire statistical dependency of $\widehat{\mathcal{M}}_{s,a,\widehat{u}^*}$ on $\widehat{P}_{s,a}$ is encapsulated in a single scalar parameter $\widehat{u}^*$. Inspired by techniques from high-dimensional probability (Vershynin, 2018), a plausible strategy is to invoke a standard covering argument: construct a finite, yet sufficiently dense, cover of the feasible range of $\widehat{u}^*$ (i.e., $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$), and then approximate $\widehat{\mathcal{M}}_{s,a,\widehat{u}^*}$ using a group of auxiliary MDPs $\{\widehat{\mathcal{M}}_{s,a,u}\}$ with u selected from this cover.

To formalize this idea, form an epsilon-net of the interval $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$ with precision ϵ , defined as

$$\mathcal{N}_\epsilon := \{-n_\epsilon\epsilon, \dots, -\epsilon, 0, \epsilon, \dots, n_\epsilon\epsilon\}, \quad \text{with } n_\epsilon = \left\lfloor \frac{1}{(1-\gamma)\epsilon} \right\rfloor, \quad (3.52)$$

which has cardinality at most $\lceil \frac{2}{(1-\gamma)\epsilon} \rceil + 1$. For each $u \in \mathcal{N}_\epsilon$, we construct an (s, a) -absorbing MDP $\widehat{\mathcal{M}}_{s,a,u}$ as described previously. Since $\mathcal{N}_\epsilon$ is statistically independent of $\widehat{P}_{s,a}$ for any fixed ϵ , each $\widehat{\mathcal{M}}_{s,a,u}$ with $u \in \mathcal{N}_\epsilon$ remains statistically independent of $\widehat{P}_{s,a}$. Notably, under the separation condition described in $\mathcal{B}_\omega$ (cf. (3.48)), one can find a point $u \in \mathcal{N}_\epsilon$ (for some suitably chosen ϵ) such that its associated absorbing MDP shares the same optimal policy as the original empirical MDP.

Lemma 3.5. Consider any $\omega > 0$, and suppose the event $\mathcal{B}_\omega$ (cf. (3.48)) holds. Then for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$ (cf. (3.52)), such that

$$\widehat{\pi}^* = \widehat{\pi}_{s,a,u_0}^*. \quad (3.53)$$

The proof of this lemma is postponed to Section 3.6.2. Importantly, this lemma unveils that, under the separation condition, analyzing $\widehat{\pi}^*$ reduces to studying a small collection of leave-one-out MDPs that are statistically independent of $\widehat{P}_{s,a}$. The entire statistical dependence of $\widehat{\pi}^*$ on $\widehat{P}_{s,a}$ is thus encapsulated in the choice of an auxiliary MDP from this small collection. This insight paves the way for effective decoupling of statistical dependencies.

Step 3: establishing optimal value estimation accuracy. Armed with these leave-one-out auxiliary MDPs, we are ready to derive the desired statistical bound under the separation condition. More concretely, we intend to establish the following lemma, which in turn proves Theorem 3.3 for the cases where the separation condition $\mathcal{B}_\omega$ holds.

Lemma 3.6. Given $0 < \omega < 1$ and $\delta > 0$, suppose that $\mathcal{B}_\omega$ (cf. (3.48)) holds. Then, with probability at least $1 - 2\delta$, it holds that

$$\|\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}\|_\infty \leq c_1 \sqrt{\frac{\log(\frac{SA}{(1-\gamma)\omega\delta})}{N(1-\gamma)^3}}, \quad (3.54a)$$

$$0 \leq V^* - V^{\widehat{\pi}^*} \leq 2c_1 \sqrt{\frac{\log(\frac{SA}{(1-\gamma)\omega\delta})}{N(1-\gamma)^3}} 1_S, \quad (3.54b)$$

provided that $N \geq \frac{c_0 \log(\frac{SA}{(1-\gamma)\delta\omega})}{1-\gamma}$. Here, $c_0, c_1 > 0$ are some large enough universal constants.

Proof of Lemma 3.6. To begin, once the first bound (3.54a) is proven, then the second bound (3.54b) follows immediately by substituting (3.54a) and (3.39) into (3.38). Thus, the main task reduces to establishing the value estimation bound (3.54a) for $\widehat{\pi}^*$. Throughout this proof, we would like to make use of the sequence $\{V^{(l)}\}$ introduced in (3.17) for any policy π ; for clarity, we shall write $V_\pi^{(l)}$ instead to make explicit the dependency on policy π .

A key ingredient of this proof is Theorem 3.2, which provides value estimation error bounds in settings where the policy of interest and $\widehat{P}$ exhibit some sort of weak statistical dependency. By “weak statistical dependency”, we refer to the Bernstein-style condition (3.29) for small enough β_1 , a property that can be verified with the aid of leave-one-out auxiliary MDPs.

Specifically, for each state-action pair (s, a) , construct the epsilon-net $\mathcal{N}_{(1-\gamma)\omega/4}$ as in the expression (3.52). By construction, the set of policies $\{\widehat{\pi}_{s,a,u}^*\}_{u \in \mathcal{N}_{(1-\gamma)\omega/4}}$ is independent of $\widehat{P}_{s,a}$. The Bernstein inequality in Lemma 2.5 taken together with the union bound, thus guarantees that with

probability at least $1 - \delta$,

$$\left| (\hat{P}_{s,a} - P_{s,a}) V_{\hat{\pi}_{s,a,u}^*}^{(l)} \right| \leq \sqrt{\frac{\beta_1 (\text{Var}_P[V_{\hat{\pi}_{s,a,u}^*}^{(l)}])_{(s,a)}}{N}} + \frac{\|V_{\hat{\pi}_{s,a,u}^*}^{(l)}\|_\infty \beta_1}{N} \quad (3.55)$$

holds uniformly over all $0 \leq l \leq \log \frac{e}{1-\gamma}$, all $u \in \mathcal{N}_{(1-\gamma)\omega/4}$, and every $(s, a) \in \mathcal{S} \times \mathcal{A}$, where β_1 is given by

$$\beta_1 = 2 \log \left(\frac{4 \log \left(\frac{e}{1-\gamma} \right) |\mathcal{N}_{(1-\gamma)\omega/4}| SA}{\delta} \right) \leq 2 \log \left(\frac{32 SA \log \left(\frac{e}{1-\gamma} \right)}{(1-\gamma)^2 \omega \delta} \right),$$

where we have used the fact $|\mathcal{N}_{(1-\gamma)\omega/4}| \leq \frac{8}{(1-\gamma)^2 \omega}$. In addition, for any $0 < \omega < 1$, Lemma 3.5 guarantees that for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$ such that $\hat{\pi}^* = \hat{\pi}_{s,a,u_0}^*$. Invoking this important fact, we obtain

$$\begin{aligned} \left| (\hat{P}_{s,a} - P_{s,a}) V_{\hat{\pi}^*}^{(l)} \right| &= \left| (\hat{P}_{s,a} - P_{s,a}) V_{\hat{\pi}_{s,a,u_0}^*}^{(l)} \right| \\ &\leq \sqrt{\frac{\beta_1}{N}} \sqrt{(\text{Var}_P[V_{\hat{\pi}_{s,a,u_0}^*}^{(l)}])_{(s,a)}} + \frac{\|V_{\hat{\pi}_{s,a,u_0}^*}^{(l)}\|_\infty \beta_1}{N} \\ &= \sqrt{\frac{\beta_1}{N}} \sqrt{(\text{Var}_P[V_{\hat{\pi}^*}^{(l)}])_{(s,a)}} + \frac{\|V_{\hat{\pi}^*}^{(l)}\|_\infty \beta_1}{N}. \end{aligned}$$

Taking together the above results for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ demonstrates that, with probability exceeding $1 - \delta$,

$$\begin{aligned} \left| (\hat{P}_{\hat{\pi}^*} - P_{\hat{\pi}^*}) V_{\hat{\pi}^*}^{(l)} \right| &= \left| \Pi^{\hat{\pi}^*} (\hat{P} - P) V_{\hat{\pi}^*}^{(l)} \right| \\ &\leq \sqrt{\frac{\beta_1}{N}} \sqrt{\Pi^{\hat{\pi}^*} \text{Var}_P[V_{\hat{\pi}^*}^{(l)}]} + \frac{\|V_{\hat{\pi}^*}^{(l)}\|_\infty \beta_1}{N} 1_S \\ &= \sqrt{\frac{\beta_1}{N}} \sqrt{\text{Var}_{P_{\hat{\pi}^*}}[V_{\hat{\pi}^*}^{(l)}]} + \frac{\|V_{\hat{\pi}^*}^{(l)}\|_\infty \beta_1}{N} 1_S, \end{aligned}$$

where we have also used the assumption that $\hat{\pi}^*$ is a deterministic policy. The above derivation validates the assumption required for Theorem 3.2. As a result, if $N > \frac{1}{1-\gamma} \beta_1$ and $0 < \omega < 1$, then Theorem 3.2 yields

$$\|\hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*}\|_\infty \leq \frac{3}{1-\gamma} \sqrt{\frac{8\beta_1}{N(1-\gamma)}} \lesssim \sqrt{\frac{\log \left(\frac{SA}{(1-\gamma)\omega\delta} \right)}{N(1-\gamma)^3}} \quad (3.56)$$

as claimed. $\square$

Step 4: relaxing the separation condition via reward perturbation. Thus far, our analysis has relied crucially on the separation condition described in $\mathcal{B}_\omega$ (cf. (3.48)); in particular, the proof argument may break down if $\hat{\pi}^*$ is non-unique or insufficiently distinguishable from competing policies. It is therefore crucial to ensure that the optimal policy $\hat{\pi}^*$ is clearly separated from other alternatives. As it turns out, this separation condition can be guaranteed with high probability by introducing small random perturbations to the reward function. Such perturbations have minimal impacts on the optimal policy and the corresponding optimal values, yet they suffice to ensure uniqueness of the optimal policy. The details are postponed to Section 3.7.

3.4 Minimax lower bounds

To confirm the optimality of model-based algorithms, we present a matching minimax lower bound under the generative model, originally established in the seminal work of Azar *et al.*, 2013. This result demonstrates that the performance guarantees achieved by the model-based approach are, in fact, statistically optimal (up to logarithmic factors).

Theorem 3.7. Consider any $0 < \varepsilon < \frac{1}{32(1-\gamma)}$ and $\gamma \in [2/3, 1)$, and suppose $SA \geq 16$. Assume that N independent samples are drawn for each state-action pair from the generative model, where

$$N \leq \frac{c_N \log(SA)}{(1-\gamma)^3 \varepsilon^2}$$

for some small universal constant $c_N > 0$.

- (i) For any (deterministic) policy estimator $\hat{\pi}$, there is a γ -discounted infinite-horizon MDP $\mathcal{M}$ with S states and A actions such that

$$\mathbb{P}_{\mathcal{M}}(\|Q^{\hat{\pi}} - Q^*\|_{\infty} > \varepsilon) \geq 1/4.$$

- (ii) For any Q-function estimator $\hat{Q}$, there exists a γ -discounted infinite-horizon MDP $\mathcal{M}$ with S states and A actions such that

$$\mathbb{P}_{\mathcal{M}}(\|\hat{Q} - Q^*\|_{\infty} > \varepsilon) \geq 1/4.$$

Here, $\mathbb{P}_{\mathcal{M}}(\cdot)$ denotes the probability distribution under $\mathcal{M}$.

As asserted by Theorem 3.7, no algorithm can reliably learn an ε -optimal policy, or an ε -accurate optimal Q-function, for every γ -discounted infinite-horizon MDP, unless the total sample size exceeds $\tilde{\Omega}(\frac{SA}{(1-\gamma)^3 \varepsilon^2})$. By comparing this minimax lower bound with the performance guarantees of Algorithm 3.1 (see Theorem 3.1, corresponding the policy evaluation setting with $A = 1$), as well as Algorithm 3.3 (see Theorem 3.3), we conclude that the model-based approach is information-theoretically optimal (up to logarithmic factors) in sample complexity for both policy evaluation and policy learning, across the entire range of precision levels ε .

Finally, we remark that Azar *et al.* (2013) established the first tight minimax lower bound on the sample complexity, applicable to both generative model and online sampling settings. Although the statement of Azar *et al.* (2013, Theorem 3) focuses on estimating optimal Q-functions and restricts the precision level to $\varepsilon \leq O(1)$, the lower bound developed therein actually extends to policy learning and holds for the broader range $\varepsilon \leq O(\frac{1}{1-\gamma})$. The original proof in Azar *et al.* (2013) relies on a change-of-measure argument by analyzing the likelihood ratio between two carefully constructed MDPs. In contrast, the proof to be presented shortly in this monograph adopts an information-theoretic approach based on Fano's inequality (see Lemma 2.10).

3.4.1 Analysis: lower bound for a simpler small-scale case

Before presenting the proof of Theorem 3.7, we first consider a related problem of Q-function estimation in a family of small-scale MDPs containing two states and one action. As it turns out, such a simplified setting suffices to illustrate the key ideas underlying the application of Fano's lemma for lower bound analyses, while also revealing the fundamental bottleneck associated with the effective horizon that arises in more general policy learning and value estimation tasks.

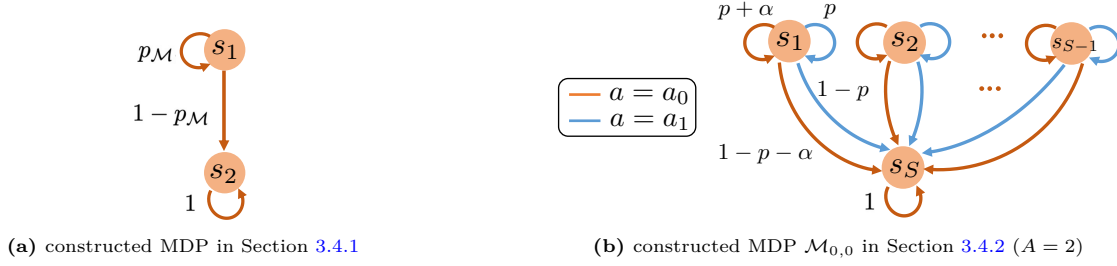


Figure 3.2: Illustration of MDP construction for the minimax lower bounds.

Consider any $0 < \varepsilon < \frac{1}{32(1-\gamma)}$ and any $\gamma \in [2/3, 1)$. Our goal in this subsection is to show that one can construct two MDPs $\mathcal{M}_0$ and $\mathcal{M}_1$ — each with two states and a single action — such that if N independent samples are drawn from the generative model for each state-action pair and

$$N \leq \frac{0.003}{(1-\gamma)^3 \varepsilon^2}, \quad (3.57)$$

then any estimator $\hat{Q}$ must satisfy

$$\mathbb{P}_{\mathcal{M}_i} \left(|\hat{Q}(s_1, a) - Q_{\mathcal{M}_i}^*(s_1, a)| > \varepsilon \right) > 0.1 \quad \text{for some } i \in \{0, 1\}. \quad (3.58)$$

Here, we denote by $\mathbb{P}_{\mathcal{M}_0}(\cdot)$ (resp. $\mathbb{P}_{\mathcal{M}_1}(\cdot)$) the probability measure under $\mathcal{M}_0$ (resp. $\mathcal{M}_1$). In other words, no algorithm can simultaneously estimate the optimal Q-function for both $\mathcal{M}_0$ and $\mathcal{M}_1$ to the desired level of precision ε , unless the sample size per state-action pair exceeds the order of $\frac{1}{(1-\gamma)^3 \varepsilon^2}$. The main idea of the proof is as follows: we identify two simple MDPs that are difficult to distinguish, formulate a binary hypothesis testing problem to determine which of these two MDPs generates the observed samples, and then relate the resulting minimax testing error (obtained via Fano's lemma) to the estimation risk.

Step 1: constructing hard MDP instances. The first step is to construct two MDP instances, deliberately designed to be nearly indistinguishable unless sufficiently many samples are available. We begin by specifying the common structure and properties shared by the two MDPs under consideration. Specifically, each MDP $\mathcal{M}$ of interest is described as follows; see Figure 3.2(a) for an illustration.

- *State space.* The state space $\mathcal{S}$ contains 2 states, given by

$$\mathcal{S} = \{s_1, s_2\}. \quad (3.59)$$

- *Action space.* In each state $s \in \mathcal{S}$, the action space is a singleton set $\mathcal{A} = \{a\}$.
- *Probability transition kernel.* We now specify the transition kernel of the MDP $\mathcal{M}$, denoted by $P_{\mathcal{M}}$.

- For state s_1 , set

$$P_{\mathcal{M}}(s | s_1, a) = \begin{cases} p_{\mathcal{M}}, & \text{if } s = s_1 \\ 1 - p_{\mathcal{M}}, & \text{if } s = s_2 \end{cases} \quad (3.60a)$$

for some parameter $0 < p_{\mathcal{M}} < 1$. That is, when the current state is s_1 , the process either remains in the same state or transitions to s_2 .

- State s_2 is set to be an absorbing state, i.e.,

$$P_{\mathcal{M}}(s_2 | s_2, a) = 1. \quad (3.60b)$$

- *Immediate rewards.* The reward function is chosen to be

$$r(s, a) = \begin{cases} 1, & \text{if } s = s_1, \\ 0, & \text{else.} \end{cases} \quad (3.61)$$

Clearly, a key parameter governing the optimal Q-function in the above construction of $\mathcal{M}$ is $p_{\mathcal{M}}$. One can readily calculate the optimal Q-function in $\mathcal{M}$, denoted by $Q_{\mathcal{M}}^*$, through the Bellman equation:

$$Q_{\mathcal{M}}^*(s_1, a) = \frac{1}{1 - \gamma p_{\mathcal{M}}}. \quad (3.62)$$

Proof of (3.62). First, since s_2 is an absorbing state with 0 immediate reward, it follows that $V_{\mathcal{M}}^*(s_2) = 0$. Second, the Bellman equation gives

$$\begin{aligned} V_{\mathcal{M}}^*(s_1) &= Q_{\mathcal{M}}^*(s_1, a) = r(s_1, a) + \gamma \{p_{\mathcal{M}} V_{\mathcal{M}}^*(s_1) + (1 - p_{\mathcal{M}}) V_{\mathcal{M}}^*(s_2)\} \\ &= 1 + \gamma p_{\mathcal{M}} V_{\mathcal{M}}^*(s_1), \end{aligned}$$

and hence $Q_{\mathcal{M}}^*(s_1, a) = V_{\mathcal{M}}^*(s_1) = \frac{1}{1 - \gamma p_{\mathcal{M}}}$. $\square$

In the sequel, we focus on two MDPs constructed as above, denoted by $\mathcal{M}_0$ and $\mathcal{M}_1$, which differ only in their choice of $p_{\mathcal{M}}$ (cf. (3.60a)):

$$p_{\mathcal{M}} = \begin{cases} p, & \text{if } \mathcal{M} = \mathcal{M}_0, \\ p + \alpha, & \text{if } \mathcal{M} = \mathcal{M}_1, \end{cases} \quad (3.63)$$

where p and α are parameters to be specified shortly, satisfying $1/2 \leq p < p + \alpha < 1$. In what follows, we denote by $\mathbb{E}_{\mathcal{M}_0}[\cdot]$ (resp. $\mathbb{E}_{\mathcal{M}_1}[\cdot]$) the expectation under $\mathcal{M}_0$ (resp. $\mathcal{M}_1$). The main step of the proof is to analyze the (in)distinguishability of these two MDPs given samples drawn from the generative model.

Step 2: applying Fano's lemma. Suppose that for each state-pair, N independent samples are drawn from the generative model. In this setting, all useful information is contained in whether the process transitions to s_2 after executing action a in state s_1 .

Let $(s_1, a, s_{\text{next}}^{(i)})$ ($1 \leq i \leq N$) denote the N independent samples drawn w.r.t. the state-action pair (s_1, a) . We introduce a set of random variables $\iota_1, \dots, \iota_N$, where $\iota_i = \mathbb{1}\{s_{\text{next}}^{(i)} = s_1\}$ is an indicator of staying in state s_1 in the i -th sample, and is hence an independent Bernoulli random variable. Let P_0 and P_1 denote the distributions of all observed samples under $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively, and denote by P_{0, ι_i} and P_{1, ι_i} the distributions of ι_i under $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. In view of the tensorization property of KL divergence (cf. Lemma 2.11), we have

$$\text{KL}(P_1 \parallel P_0) = \sum_{i=1}^N \text{KL}(P_{1, \iota_i} \parallel P_{0, \iota_i}) = N \text{KL}(p + \alpha \parallel p)$$

$$\leq \frac{N\alpha^2}{p(1-p)} \leq \frac{2N\alpha^2}{1-p} =: \beta, \quad (3.64)$$

where $\text{KL}(p + \alpha \| p)$ denotes the KL divergence between $\text{Bern}(p + \alpha)$ and $\text{Bern}(p)$, and the last line arises from Lemma 2.12 and the property $p \geq 1/2$. As a consequence, Fano's lemma (see Lemma 2.10) implies that the minimax probability of error necessarily exceeds

$$p_e := \inf_{\psi} \max_{0 \leq i \leq 1} \mathbb{P}_{\mathcal{M}_i}(\psi \neq i) \geq \max \left\{ \frac{1}{4} \exp(-\beta), \frac{1 - \sqrt{\beta/2}}{2} \right\}, \quad (3.65)$$

where the infimum is taken over all possible tests ψ . In particular,

$$p_e \geq \max \left\{ \frac{1}{4} \exp(-\beta), \frac{1 - \sqrt{\beta/2}}{2} \right\} \geq 0.1 \quad \text{if } \beta \leq 0.9. \quad (3.66)$$

Step 3: choosing parameters p and α . In light of (3.66), to guarantee that p_e is lower bounded by 0.1, it suffices to choose the parameters p and α to ensure that $\beta \leq 0.9$. Recall that N needs to be no larger than the order of $\frac{1}{(1-\gamma)^3 \varepsilon^2}$ (see (3.57)). Motivated by this, we select p, α as

$$p = \frac{4\gamma - 1}{3\gamma} \quad \text{and} \quad \alpha = \frac{2(1 - \gamma p)^2 \varepsilon}{\gamma^2} = \frac{32(1 - \gamma)^2 \varepsilon}{9\gamma^2}. \quad (3.67)$$

It is readily seen that as $\gamma \rightarrow 1$, we have $p \rightarrow 1$ and $\alpha \rightarrow 0$, so that $p_{\mathcal{M}_0}$ and $p_{\mathcal{M}_1}$ become increasingly similar and harder to differentiate. Moreover, one can verify that $1 - p \asymp 1 - \gamma$, $\alpha \asymp (1 - \gamma)^2 \varepsilon$, and hence $\alpha^2/(1 - p) \asymp (1 - \gamma)^3 \varepsilon^2$, which matches the inverse of the upper bound (3.57) on N .

A few simple yet useful properties are summarized below.

Lemma 3.8. Consider the above MDPs with parameters (3.67). Suppose that $0 < \varepsilon < \frac{1}{32(1-\gamma)}$ and $\gamma \in [2/3, 1)$. Then one has

- (i) $1/2 \leq p < p + \alpha < 1$ and $1 - p \geq 2\alpha$;
- (ii) $Q_{\mathcal{M}_1}^*(s_1, a) - Q_{\mathcal{M}_0}^*(s_1, a) > 2\varepsilon$;
- (iii) $\frac{\alpha^2}{1-p} = \frac{32^2(1-\gamma)^3 \varepsilon^2}{27\gamma^3} \leq 128(1-\gamma)^3 \varepsilon^2$.

Proof of Lemma 3.8. Regarding (i), it suffices to see that

$$1 - p = \frac{1 - \gamma}{3\gamma} > \frac{64(1 - \gamma)^2 \varepsilon}{9\gamma^2} = 2\alpha \quad \text{and} \quad p = \frac{4\gamma - 1}{3\gamma} > \frac{1}{2}$$

under our conditions on ε and γ . Regarding (ii), (3.62) tells us that

$$\begin{aligned} Q_{\mathcal{M}_1}^*(s_1, a) - Q_{\mathcal{M}_0}^*(s_1, a) &= \frac{1}{1 - \gamma(p + \alpha)} - \frac{1}{1 - \gamma p} \\ &= \frac{\gamma\alpha}{(1 - \gamma p)(1 - \gamma(p + \alpha))} > 2\varepsilon, \end{aligned} \quad (3.68)$$

where the last inequality uses the choice (3.67) of α . Property (iii) also follows from direct calculation under our conditions on γ . $\square$

Combine Lemma 3.8 with (3.66) and the definition (3.64) of β to demonstrate that the minimax probability of testing error obeys

$$p_e \geq 0.1 \quad \text{as long as } N \leq \frac{0.003}{(1 - \gamma)^3 \varepsilon^2}. \quad (3.69)$$

Step 4: connecting testing error with estimation risk. Finally, we show how the minimax probability of testing error can be leveraged to bound the minimax estimation risk. Recall from Lemma 3.8(ii) that $Q_{\mathcal{M}_1}^*(s_1, a) - Q_{\mathcal{M}_0}^*(s_1, a) > 2\varepsilon$. Suppose instead that there exists an estimator $\hat{Q}$ such that

$$\mathbb{P}_{\mathcal{M}_i}(|\hat{Q}(s_1, a) - Q_{\mathcal{M}_i}^*(s_1, a)| \leq \varepsilon) > 0.9, \quad \forall i \in \{0, 1\}. \quad (3.70)$$

Based on this estimator, consider the test

$$\psi = \begin{cases} 0, & \text{if } \hat{Q}(s_1, a) \leq \frac{1}{1-\gamma p} + \varepsilon \\ 1, & \text{else.} \end{cases}$$

By construction, under $\mathcal{M}_0$ we have

$$\begin{aligned} \mathbb{P}_{\mathcal{M}_0}(\psi = 0) &= \mathbb{P}_{\mathcal{M}_0}\left(\hat{Q}(s_1, a) \leq \frac{1}{1-\gamma p} + \varepsilon\right) \\ &= \mathbb{P}_{\mathcal{M}_0}(\hat{Q}(s_1, a) \leq Q_{\mathcal{M}_0}^*(s_1, a) + \varepsilon) > 0.9, \end{aligned}$$

where the last inequality arises from the assumption (3.70). An analogous argument shows that $\mathbb{P}_{\mathcal{M}_1}(\psi = 1) > 0.9$. As a result, the testing error of ψ satisfies

$$p_e(\psi) := \max\{\mathbb{P}_{\mathcal{M}_0}(\psi \neq 0), \mathbb{P}_{\mathcal{M}_1}(\psi \neq 1)\} < 0.1.$$

This contradicts the lower bound (3.69) on the minimax error probability, thereby establishing the claimed result (3.58).

3.4.2 Proof of Theorem 3.7

We now turn to establishing the lower bound in Theorem 3.7, covering both policy learning and estimation of optimal Q-functions. Our construction and analysis largely build upon those developed in Section 3.4.1, with modifications to expand the state-action space to include S states and A actions. At a high level, we again construct a family of similar MDPs and formulate a multi-hypothesis testing problem aimed at identifying the underlying data-generating MDP; we then connect the minimax error probability of this testing problem with the performance of any RL algorithm.

Step 1: constructing hard MDP instances. Let us construct a family of γ -discounted infinite-horizon MDPs with S states and A actions. We begin by presenting their common structure before describing their differences in the transition kernels.

- *State space.* The state space $\mathcal{S}$ contains S states, given by

$$\mathcal{S} = \{s_1, \dots, s_S\}.$$

The state space $\mathcal{S}$ is partitioned into the following subsets:

$$\mathcal{S}_1 = \{s_1, \dots, s_{S-1}\}, \quad \mathcal{S}_2 = \{s_S\},$$

where each subset is associated with distinct transition probabilities, to be specified shortly.

- *Action space.* For ease of exposition, we allow the action space to vary across states. For any state in $\mathcal{S}_1$, the action space contains A actions, i.e.,

$$\mathcal{A} := \{a_0, a_1, \dots, a_{A-1}\}.$$

In contrast, the action space associated with s_S is the singleton set $\{a_0\}$.

- *Immediate rewards.* The reward function is chosen to be

$$r(s, a) = \begin{cases} 1, & \text{if } s \in \mathcal{S}_1, \\ 0, & \text{else.} \end{cases}$$

Since only states in $\mathcal{S}_1$ yield positive rewards, this structure favors policies that induce a higher probability of remaining within $\mathcal{S}_1$.

Having described the common structure shared by this family of MDPs, we now specify the optimal policy and transition kernel for each instance. In particular, we construct a collection of

$$M = 1 + (S - 1)(A - 1) \quad (3.71)$$

MDP instances:

$$\mathcal{C}_{\text{mdp}} = \{\mathcal{M}_{0,0}\} \cup \{\mathcal{M}_{j,k} \mid 1 \leq j < S, 1 \leq k < A\}. \quad (3.72)$$

As in Section 3.4.1, the construction is parameterized by two scalars p and α satisfying $1/2 \leq p < p + \alpha < 1$ and $1 - p \geq 2\alpha$. See Figure 3.2(b) for an illustration of $\mathcal{M}_{0,0}$ with $A = 2$.

- *Optimal policy.* The optimal policy of $\mathcal{M}_{j,k}$, denoted by $\pi_{j,k}^*$, is chosen such that

$$\pi_{0,0}^*(s) = a_0 \quad \forall s \in \mathcal{S}_1 \quad (3.73a)$$

and for any $1 \leq j < S$ and $1 \leq k < A$,

$$\pi_{j,k}^*(s) = \begin{cases} a_k, & \text{if } s = s_j, \\ a_0, & \text{if } s \in \mathcal{S}_1 \text{ and } s \neq s_j. \end{cases} \quad (3.73b)$$

The transition probability kernels will be constructed to ensure consistency with these optimal policies.

- *Probability transition kernel.* The probability transition kernel of $\mathcal{M}_{j,k}$, denoted by $P_{\mathcal{M}_{j,k}}$, is defined below.

– In any state s_i with $1 \leq i < S$ (so that $s_i \in \mathcal{S}_1$), set

$$P_{\mathcal{M}_{j,k}}(s \mid s_i, a) = \begin{cases} p + \alpha, & \text{if } a = \pi_{j,k}^*(s_i) \text{ and } s = s_i, \\ 1 - p - \alpha, & \text{if } a = \pi_{j,k}^*(s_i) \text{ and } s = s_S, \\ p, & \text{if } a \neq \pi_{j,k}^*(s_i) \text{ and } s = s_i, \\ 1 - p, & \text{if } a \neq \pi_{j,k}^*(s_i) \text{ and } s = s_S, \\ 0 & \text{else.} \end{cases}$$

In words, executing the action $\pi_{j,k}^*(s_i)$ in state $s_i \in \mathcal{S}_1$ has a higher probability of remaining in the same state. Given that states in $\mathcal{S}_1$ are associated with higher immediate rewards, this design ensures that taking action $\pi_{j,k}^*(s_i)$ in state s_i yields a larger Q-value.

- The last state s_S is set to be an absorbing state, i.e.,

$$P_{\mathcal{M}_{j,k}}(s_S | s_S, a_0) = 1.$$

Evidently, the transition probabilities of these MDP instances differ only in those corresponding to the optimal actions; all non-optimal actions share the same transition probabilities.

Repeating the same arguments as for (3.62) immediately reveals that: for each $s \in \mathcal{S}_1$,

$$Q_{\mathcal{M}_{j,k}}^*(s, a) = \begin{cases} \frac{1}{1-\gamma(p+\alpha)} & \text{if } a = \pi_{j,k}^*(s), \\ \frac{1}{1-\gamma p} & \text{if } a \neq \pi_{j,k}^*(s). \end{cases} \quad (3.74)$$

This characterization is clearly consistent with the optimal policy defined in (3.73).

Step 2: applying Fano's lemma. To invoke Fano's lemma stated in Lemma 2.10, we need to examine the pairwise KL divergences among the constructed MDP instances. Recall that N independent samples are drawn from the generative model for each state-action pair. Let $P_{\mathcal{M}}$ and $\pi_{\mathcal{M}}^*$ denote the distribution of the observed samples and the optimal policy under $\mathcal{M}$, respectively. By construction, the optimal policies of $\mathcal{M}_{j,k}$ (for $j \neq 0$) and $\mathcal{M}_{0,0}$ differ in exactly one state, leading to two state-action pairs with different transition probabilities in $P_{\mathcal{M}_{j,k}}$ and $P_{\mathcal{M}_{0,0}}$. Consequently, the corresponding sample distributions differ only in the $2N$ samples associated with these two state-action pairs.

Among these $2N$ samples, N of them contribute a KL divergence of $\text{KL}(p + \alpha \| p)$ each, while the remaining N contribute $\text{KL}(p \| p + \alpha)$ each. Therefore, the overall KL divergence between $P_{\mathcal{M}_{j,k}}$ and $P_{\mathcal{M}_{0,0}}$ can be expressed as

$$\text{KL}(P_{\mathcal{M}_{j,k}} \| P_{\mathcal{M}_{0,0}}) = N \text{KL}(p + \alpha \| p) + N \text{KL}(p \| p + \alpha)$$

for $j \neq 0$. Repeating the arguments for (3.64), we further arrive at

$$\begin{aligned} \text{KL}(P_{\mathcal{M}_{j,k}} \| P_{\mathcal{M}_{0,0}}) &\leq \frac{N\alpha^2}{p(1-p)} + \frac{N\alpha^2}{(p+\alpha)(1-p-\alpha)} \\ &\leq \frac{2N\alpha^2}{(p+\alpha)(1-p-\alpha)} \leq \frac{8N\alpha^2}{1-p}, \end{aligned} \quad (3.75)$$

with the proviso that $1/2 \leq p < p + \alpha < 1$ and $1 - p - \alpha \geq (1 - p)/2$ (a consequence of $1 - p \geq 2\alpha$).

Define the minimax probability of testing error as

$$p_e := \inf_{\psi} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{P}_{\mathcal{M}}(\psi \neq \mathcal{M}), \quad (3.76)$$

where the infimum is taken over all tests that map the observed data to an MDP in $\mathcal{C}_{\text{mdp}}$ (cf. (3.72)). Fano's lemma (cf. Lemma 2.10) gives

$$p_e \geq \frac{\log(M+1) - \log 2}{\log M} - \frac{1}{4} \geq \frac{1}{4}, \quad (3.77)$$

provided that $M \geq 4$ and

$$\frac{1}{M+1} \sum_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \text{KL}(P_{\mathcal{M}} \| P_{\mathcal{M}_{0,0}}) \leq \frac{1}{4} \log M. \quad (3.78)$$

With (3.75) in place, this KL condition (3.78) — and hence (3.77) — can be guaranteed if

$$\frac{8N\alpha^2}{1-p} \leq \frac{1}{4} \log \frac{SA}{4} \leq \frac{1}{4} \log M. \quad (3.79)$$

Step 3: choosing parameters p and α . It then comes down to choosing the parameters p and α so as to satisfy (3.79). Akin to Section 3.4.1, we set

$$p = \frac{4\gamma - 1}{3\gamma} \quad \text{and} \quad \alpha = \frac{2(1 - \gamma p)^2 \varepsilon}{\gamma^2} = \frac{32(1 - \gamma)^2 \varepsilon}{9\gamma^2},$$

From Lemma 3.8(iii), it follows that $\frac{\alpha^2}{1-p} \leq 128(1 - \gamma)^3 \varepsilon^2$, and as a result, Condition (3.79) is satisfied provided that

$$N \leq \frac{1}{4096} \frac{1}{(1 - \gamma)^3 \varepsilon^2} \log \frac{SA}{4}.$$

Step 4: connecting testing error with performance of policy learning. We now relate the minimax probability of testing error p_e to the statistical risk of policy learning. As shown in (3.74) and Lemma 3.8(ii), for any $\mathcal{M} \in \mathcal{C}_{\text{mdp}}$ and any state $s \in \mathcal{S}_1$, the optimal action is separated from all suboptimal actions by the following gap:

$$Q_{\mathcal{M}}^*(s, \pi_{\mathcal{M}}^*(s)) - Q_{\mathcal{M}}^*(s, a) > 2\varepsilon, \quad \forall a \neq \pi_{\mathcal{M}}^*(s). \quad (3.80)$$

Now suppose that there exists a (deterministic) policy estimate $\hat{\pi}$ achieving

$$\mathbb{P}_{\mathcal{M}}(\|Q^{\hat{\pi}} - Q_{\mathcal{M}}^*\|_{\infty} > \varepsilon) < 1/4, \quad \forall \mathcal{M} \in \mathcal{C}_{\text{mdp}}. \quad (3.81)$$

In view of (3.80), the only policy that can satisfy $\|Q^{\hat{\pi}} - Q_{\mathcal{M}}^*\|_{\infty} \leq \varepsilon$ is $\pi_{\mathcal{M}}^*$. Hence, (3.81) is equivalent to

$$\mathbb{P}_{\mathcal{M}}(\hat{\pi} \neq \pi_{\mathcal{M}}^*) < 1/4, \quad \forall \mathcal{M} \in \mathcal{C}_{\text{mdp}}. \quad (3.82)$$

By construction, the optimal policy $\pi_{\mathcal{M}}^*$ uniquely identifies $\mathcal{M}$ within the collection $\mathcal{C}_{\text{mdp}}$. Define the test

$$\psi_{\text{policy}} = \mathcal{M}_{\hat{\pi}},$$

with $\mathcal{M}_{\hat{\pi}}$ indicating the MDP in $\mathcal{C}_{\text{mdp}}$ whose optimal policy coincides with $\hat{\pi}$. Then it follows that

$$\mathbb{P}_{\mathcal{M}}(\psi_{\text{policy}} \neq \mathcal{M}) = \mathbb{P}_{\mathcal{M}}(\hat{\pi} \neq \pi_{\mathcal{M}}^*) < 1/4, \quad \forall \mathcal{M} \in \mathcal{C}_{\text{mdp}},$$

which contradicts the lower bound (3.77). We have thus established that for any policy estimator $\hat{\pi}$, there exists some $\mathcal{M} \in \mathcal{C}_{\text{mdp}}$ such that

$$\mathbb{P}_{\mathcal{M}}(\|Q^{\hat{\pi}} - Q_{\mathcal{M}}^*\|_{\infty} > \varepsilon) \geq 1/4.$$

Step 5: connecting testing error with Q-function estimation risk. It remains to link the minimax probability of testing error p_e with the estimation risk for learning optimal Q-functions. Suppose for the moment that there exists an estimator $\hat{Q}$ such that

$$\mathbb{P}_{\mathcal{M}}(\|\hat{Q} - Q_{\mathcal{M}}^*\|_{\infty} > \varepsilon) < 1/4, \quad \forall \mathcal{M} \in \mathcal{C}_{\text{mdp}}. \quad (3.83)$$

Combined with (3.80), this reveals that: under every $\mathcal{M} \in \mathcal{C}_{\text{mdp}}$, the following event occurs with probability at least 3/4:

$$\begin{aligned} \hat{Q}(s, \pi_{\mathcal{M}}^*(s)) - \hat{Q}(s, a) &\geq Q_{\mathcal{M}}^*(s, \pi_{\mathcal{M}}^*(s)) - Q_{\mathcal{M}}^*(s, a) - 2\|\hat{Q} - Q_{\mathcal{M}}^*\|_{\infty} \\ &> 2\varepsilon - 2\varepsilon = 0 \end{aligned} \quad (3.84)$$

simultaneously for all $s \in \mathcal{S}$ and all $a \neq \pi_{\mathcal{M}}^*(s)$. As an immediate consequence, if we define π_{est} to be the greedy policy w.r.t. $\hat{Q}$ (namely, $\pi_{\text{est}}(s) = \arg \max_a \hat{Q}(s, a)$ for all $s \in \mathcal{S}$), then under $\mathcal{M}$ one must have $\pi_{\text{est}} = \pi_{\mathcal{M}}^*$.

In light of this observation, consider the following test:

$$\psi_{\text{est}} = \mathcal{M}_{\pi_{\text{est}}},$$

where $\mathcal{M}_{\pi_{\text{est}}}$ denotes the MDP in $\mathcal{C}_{\text{mdp}}$ whose optimal policy coincides with π_{est} . It is then readily seen that

$$\mathbb{P}_{\mathcal{M}}(\psi_{\text{est}} \neq \mathcal{M}) = \mathbb{P}_{\mathcal{M}}(\pi_{\text{est}} \neq \pi_{\mathcal{M}}^*) < 1/4, \quad \forall \mathcal{M} \in \mathcal{C}_{\text{mdp}},$$

resulting in a contradiction with the minimax lower bound (3.77). Therefore, we conclude that for any estimator $\hat{Q}$, there exists some $\mathcal{M} \in \mathcal{C}_{\text{mdp}}$ such that

$$\mathbb{P}_{\mathcal{M}}(\|\hat{Q} - Q_{\mathcal{M}}^*\|_{\infty} > \varepsilon) \geq 1/4.$$

3.5 Extension: finite-horizon nonstationary MDPs

Moving beyond discounted infinite-horizon stationary MDPs, we now briefly discuss model-based policy learning for finite-horizon nonstationary MDPs. The proof techniques are similar in spirit to — and in fact simpler than — those for the discounted infinite-horizon setting, owing in part to the availability of independent samples across different steps h over the entire horizon. For brevity, we refer interested readers to Li *et al.* (2024c) for the full proof details.

A model-based approach. We begin by describing what is arguably the most natural version of the model-based policy learning algorithm under the generative model. In essence, the algorithm performs model-based planning using the empirical model; unlike its discounted infinite-horizon counterpart in Algorithm 3.3, it does not involve reward perturbation.

- 1) *Model estimation.* For each state-action-step tuple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, draw N independent samples from the generative model:

$$s^{(i)}(s, a, h) \stackrel{\text{i.i.d.}}{\sim} P_h(\cdot | s, a), \quad i = 1, \dots, N,$$

and construct the empirical transition kernels $\{\hat{P}_h\}_{h=1}^H$ via

$$\forall s' \in \mathcal{S}, \quad \hat{P}_h(s' | s, a) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s' = s^{(i)}(s, a, h)\}. \quad (3.85)$$

Namely, $\hat{P}_h(s' | s, a)$ stands for the empirical frequency of transitions from (s, a) to state s' at step h .

- 2) *Planning.* Compute an optimal policy $\hat{\pi}^* = \{\hat{\pi}_h^*\}_{h=1}^H$ for the empirical MDP $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{P}, r, H)$.

Note that the optimal policy $\hat{\pi}_h^*$ of the empirical MDP $\widehat{\mathcal{M}}$ at step h is computed via backward induction starting from $h = H$. This backward recursion ensures that $\hat{\pi}_h^*$ depends solely on the empirical transition kernels at or after step h , and as a result, $\hat{\pi}_h^*$ is statistically independent of the empirical transitions $\{\hat{P}_j\}_{1 \leq j < h}$. Consequently, the statistical dependencies across steps are naturally decoupled. In contrast to the discounted infinite-horizon setting, this obviates the need for random reward perturbations, which were introduced there to facilitate the analysis. As the following theorem shows, this simple model-based approach already achieves the desirable sample complexity.

Theorem 3.9 (Li *et al.*, 2024c). There exists some universal constant $c_0 > 0$ such that: for any $\delta > 0$ and $0 < \varepsilon \leq H$, the policy $\hat{\pi}^*$ returned by the aforementioned model-based algorithm obeys

$$\begin{aligned} \forall (s, h) \in \mathcal{S} \times [H], \quad & V_h^{\hat{\pi}^*}(s) \geq V_h^*(s) - \varepsilon \\ \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H], \quad & Q_h^{\hat{\pi}^*}(s, a) \geq Q_h^*(s, a) - \varepsilon \end{aligned}$$

with probability at least $1 - \delta$, provided that the sample size for every tuple (s, a, h) exceeds

$$N \geq \frac{c_0 H^3 \log\left(\frac{HSA}{\delta}\right)}{\varepsilon^2}. \quad (3.86)$$

Akin to the discounted infinite-horizon setting, the model-based approach achieves ε -accuracy as long as the sample size per state-action-step tuple (s, a, h) exceeds the order

$$\frac{H^3}{\varepsilon^2}$$

up to logarithmic factors. This guarantee, which holds for the full range $\varepsilon \in (0, H]$, mirrors the bound (3.35) for the discounted infinite-horizon setting, with the effective horizon $\frac{1}{1-\gamma}$ now replaced by the finite horizon length H . Since there are SAH distinct (s, a, h) -tuples in total, the overall sample complexity scales as

$$\tilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right). \quad (3.87)$$

The quadruple horizon dependence H^4 in this bound — as opposed to the cubic scaling $1/(1-\gamma)^3$ seen in the discounted infinite-horizon case — stems from time inhomogeneity. Because the transition kernels P_h may vary across steps, one incurs an additional factor of H in the sample complexity. Encouragingly, the sample complexity (3.87) is provably minimax optimal in view of the matching lower bound (Domingues *et al.*, 2021).

3.6 Appendix A: proof of auxiliary lemmas in Section 3.3.4

3.6.1 Proof of Lemma 3.4

To prove this lemma, it suffices to verify that V^* and Q^* satisfy the Bellman optimality equations (2.14) w.r.t. $\mathcal{M}_{s,a,u^*}$. For notational simplicity, let P^{abs} and r^{abs} denote the probability transition matrix and the reward function associated with $\mathcal{M}_{s,a,u^*}$, respectively.

For any pair (s, a) , by construction we have

$$r^{\text{abs}}(s, a) + \gamma(P^{\text{abs}}V^*)_{(s,a)} = u^* + \gamma V^*(s).$$

Note that V^* obeys the Bellman equation w.r.t. $\mathcal{M}$, so that $Q^*(s, a) = r(s, a) + \gamma(PV^*)_{(s,a)}$. This together with our choice of u^* gives

$$u^* + \gamma V^*(s) = Q^*(s, a) - \gamma V^*(s) + \gamma V^*(s) = Q^*(s, a).$$

Taking the above identities together leads to

$$Q^*(s, a) = r^{\text{abs}}(s, a) + \gamma(P^{\text{abs}}V^*)_{(s,a)}. \quad (3.88)$$

Next, consider any state-action pair $(s', a') \neq (s, a)$. Recalling again the construction of $\mathcal{M}_{s,a,u^*}$, we see that

$$r^{\text{abs}}(s', a') + \gamma(P^{\text{abs}}V^*)_{(s',a')} = r(s', a') + \gamma(PV^*)_{(s',a')} = Q^*(s', a'),$$

where the last identity follows from the Bellman equation w.r.t. $\mathcal{M}$.

Combining the above two cases implies that V^* and Q^* satisfy the Bellman optimality equation w.r.t. $\mathcal{M}_{s,a,u^*}$, thus concluding the proof.

3.6.2 Proof of Lemma 3.5

In view of Lemma 3.4, if we set $\hat{u}^* := r(s, a) + \gamma(\hat{P}\hat{V}^*)_{s,a} - \gamma\hat{V}^*(s)$, then it holds that

$$\hat{Q}_{s,a,\hat{u}^*}^* = \hat{Q}^* \quad \text{and} \quad \hat{V}_{s,a,\hat{u}^*}^* = \hat{V}^*.$$

In addition, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$ such that $|\hat{u}^* - u_0| \leq (1-\gamma)\omega/4$, which combined with the basic property (2.28) gives

$$\|\hat{Q}^* - \hat{Q}_{s,a,u_0}^*\|_\infty = \|\hat{Q}_{s,a,\hat{u}^*}^* - \hat{Q}_{s,a,u_0}^*\|_\infty \leq \frac{|\hat{u}^* - u_0|}{1-\gamma} \leq \frac{\omega}{4}. \quad (3.89)$$

Additionally, for any $s' \in \mathcal{S}$ and any $a_1, a_2 \in \mathcal{A}$ with $a_1 \neq a_2$, we make note of the following decomposition:

$$\begin{aligned} & \hat{Q}_{s,a,u_0}^*(s', a_1) - \hat{Q}_{s,a,u_0}^*(s', a_2) \\ &= \hat{Q}^*(s', a_1) - \hat{Q}^*(s', a_2) + \hat{Q}_{s,a,u_0}^*(s', a_1) - \hat{Q}^*(s', a_1) - (\hat{Q}_{s,a,u_0}^*(s', a_2) - \hat{Q}^*(s', a_2)) \\ &\geq \hat{Q}^*(s', a_1) - \hat{Q}^*(s', a_2) - 2\|\hat{Q}^* - \hat{Q}_{s,a,u_0}^*\|_\infty \\ &\geq \hat{Q}^*(s', a_1) - \hat{Q}^*(s', a_2) - \frac{\omega}{2}, \end{aligned} \quad (3.90)$$

where the last line applies inequality (3.89). Moreover, our separation condition defined in (3.48) requires that: for any $s' \in \mathcal{S}$ and any $a_2 \neq \hat{\pi}^*(s')$, one has $\hat{Q}^*(s', \hat{\pi}^*(s')) - \hat{Q}^*(s', a_2) \geq \omega$, which taken together with (3.90) reveals that

$$\hat{Q}_{s,a,u_0}^*(s', \hat{\pi}^*(s')) - \hat{Q}_{s,a,u_0}^*(s', a_2) \geq \hat{Q}^*(s', \hat{\pi}^*(s')) - \hat{Q}^*(s', a_2) - \frac{\omega}{2} \geq \frac{\omega}{2}. \quad (3.91)$$

Since $\hat{\pi}_{s,a,u_0}^*(s') := \arg \max_{a'} \hat{Q}_{s,a,u_0}^*(s', a')$, it is seen from (3.91) that

$$\hat{\pi}_{s,a,u_0}^*(s') = \hat{\pi}^*(s'),$$

which holds true for all $s' \in \mathcal{S}$. This concludes the proof.

3.7 Appendix B: tie-breaking via reward perturbation

To extend the analysis in Section 3.3.4 to accommodate all MDPs of interest — without imposing the separation condition in $\mathcal{B}_\omega$ (cf. (3.48)) — this section proves that reward perturbation in Algorithm 3.3 yields a modified MDP satisfying the desirable separation condition while remaining close to the original empirical MDP.

For notational convenience, let us express the perturbed rewards in (3.32) in the vector form as

$$r_p := r + \zeta \quad \text{with } \zeta = [\zeta(s, a)]_{(s,a) \in \mathcal{S} \times \mathcal{A}}, \quad \zeta(s, a) \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, \xi). \quad (3.92)$$

Denote by π_p^* the optimal policy of the MDP $\mathcal{M}_p = (\mathcal{S}, \mathcal{A}, P, r_p, \gamma)$, and Q_p^* its optimal Q-function. Analogously, define $\hat{Q}_p^*$ and $\hat{\pi}_p^*$ for the perturbed empirical MDP $\hat{\mathcal{M}}_p = (\mathcal{S}, \mathcal{A}, \hat{P}, r_p, \gamma)$. The following lemma reveals that reward perturbation can effectively “break ties” in the Q-function and therefore ensure sufficient separation.

Lemma 3.10. With probability at least $1 - \delta$, one has

$$\widehat{Q}_p^*(s, \widehat{\pi}_p^*(s)) - \widehat{Q}_p^*(s, a) > \frac{\xi\delta(1-\gamma)}{3SA^2} \quad (3.93)$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ obeying $a \neq \widehat{\pi}_p^*(s)$.

Proof. Given that the reward perturbation ζ is randomly drawn from a continuous distribution, it can be expected that, with probability approaching one, any two actions will yield sufficiently different Q-values. We omit the proof details here and refer the reader to Li *et al.* (2024c, Lemma 6) for a complete argument. $\square$

Equipped with the above separation condition (3.93), setting $\omega := \frac{\xi\delta(1-\gamma)}{3SA^2}$ in Lemma 3.6 leads to

$$\|V_p^{\pi_p^*} - V_p^{\widehat{\pi}_p^*}\|_\infty \lesssim \sqrt{\frac{\log\left(\frac{SA}{(1-\gamma)\xi\delta}\right)}{N(1-\gamma)^3}}, \quad (3.94)$$

where V_p^π denotes the value function of policy π in the MDP $\mathcal{M}_p = (\mathcal{S}, \mathcal{A}, P, r_p, \gamma)$.

To finish up, we still need to transfer the result (3.94) into a bound on the policy learning error $\|V^{\pi_p^*} - V^{\widehat{\pi}_p^*}\|_\infty$ in the original, unperturbed MDP $\mathcal{M}$. Towards this end, observe that

$$\begin{aligned} V^{\widehat{\pi}_p^*} - V^* &= (V^{\widehat{\pi}_p^*} - V_p^{\widehat{\pi}_p^*}) + (V_p^{\widehat{\pi}_p^*} - V_p^{\pi_p^*}) + (V_p^{\pi_p^*} - V^{\pi_p^*}) + (V^{\pi_p^*} - V^*) \\ &\geq (V^{\widehat{\pi}_p^*} - V_p^{\widehat{\pi}_p^*}) + (V_p^{\widehat{\pi}_p^*} - V_p^{\pi_p^*}) + (V_p^{\pi_p^*} - V^{\pi_p^*}), \end{aligned}$$

which follows from the optimality of π_p^* w.r.t. $\mathcal{M}_p$. The term $V_p^{\widehat{\pi}_p^*} - V_p^{\pi_p^*}$ has been bounded in (3.94), whereas the other two terms can be controlled using (2.27b) as follows:

$$\|V^{\pi_p^*} - V_p^{\pi_p^*}\|_\infty \leq \frac{1}{1-\gamma} \|r - r_p\|_\infty \leq \frac{1}{1-\gamma} \xi, \quad (3.95a)$$

$$\|V_p^{\widehat{\pi}_p^*} - V_p^{\pi_p^*}\|_\infty \leq \frac{1}{1-\gamma} \|r - r_p\|_\infty \leq \frac{1}{1-\gamma} \xi. \quad (3.95b)$$

Taking these collectively gives, with probability at least $1 - 3\delta$,

$$0 \geq V^{\widehat{\pi}_p^*} - V^* \geq -\left(\frac{2}{1-\gamma} \xi + O\left(\sqrt{\frac{\log\left(\frac{SA}{\xi(1-\gamma)\delta}\right)}{N(1-\gamma)^3}}\right)\right) 1_{\mathcal{S}}.$$

By taking $\xi = \frac{(1-\gamma)\varepsilon}{3S^5A^5}$ and $N \geq \frac{c_0 \log\left(\frac{SA}{(1-\gamma)\delta\varepsilon}\right)}{(1-\gamma)^3\varepsilon^2}$ for some constant $c_0 > 0$ large enough, we arrive at $\|V^{\widehat{\pi}_p^*} - V^*\|_\infty \leq \varepsilon$ as claimed in Theorem 3.3.

3.8 Notes

Sample complexity under a generative model. The generative model (or simulator) adopted in this chapter was first introduced by Kearns and Singh (1999). Since then, it has become a standard framework in the theoretical studies of RL and has been extensively investigated in subsequent work aimed at understanding and comparing the data efficiency of RL algorithms (e.g., Kearns *et al.* (2002), Kakade (2003), Lattimore and Hutter (2012), Azar *et al.* (2013), Sidford *et al.* (2018a), Sidford *et al.* (2018b), Wang (2020), Wainwright (2019c), Pananjady and Wainwright

(2020), Agarwal *et al.* (2019b), Yang and Wang (2019), Mou *et al.* (2020), Wang *et al.* (2021a), Khamaru *et al.* (2021), Sidford *et al.* (2020), Li *et al.* (2022a), Vaswani *et al.* (2022), Shi *et al.* (2023), and Zhang *et al.* (2025c)).

Several prior works have investigated the effectiveness of model-based RL for discounted infinite-horizon MDPs under a generative model. The seminal work Azar *et al.* (2013) was the first to establish the asymptotic minimax optimality of model-based RL, albeit under the requirement that the total sample size be at least on the order of $\frac{S^2 A}{(1-\gamma)^2}$. Subsequently, Agarwal *et al.* (2019b) significantly broadened the regime in which model-based methods are provably optimal. However, the results therein restrict the accuracy level to the range $\varepsilon \in (0, \frac{1}{\sqrt{1-\gamma}}]$, and therefore do not cover the full ε -regime of interest; in other words, the optimality guarantees in Agarwal *et al.* (2019b) hold only when the total sample size surpasses the order of $\frac{SA}{(1-\gamma)^2}$. Similar technical barriers had already appeared in the simpler setting of model-based policy evaluation (Pananjady and Wainwright, 2020; Agarwal *et al.*, 2019b), and had also arisen in the analysis of model-free algorithms (Sidford *et al.*, 2018b; Wainwright, 2019c) and in the studies of finite-horizon nonstationary MDPs (Yin *et al.*, 2021b). In stark contrast, however, existing lower bounds (e.g., Azar *et al.* (2013)) do not rule out the possibility of achieving nontrivial statistical accuracy below such sample size thresholds, giving rise to a gap between upper and lower bounds in the most sample-starved regime.

The works of Li *et al.* (2020) and Li *et al.* (2024c) derived the first theoretical guarantees establishing full-range minimax optimality of model-based algorithms, encompassing both discounted infinite-horizon MDPs and finite-horizon nonstationary MDPs. These results close the gap between upper and lower bounds in the sample complexity analyses and fully pin down the fundamental sample complexity limit under a generative model across the entire ε -range. More recently, the reward-perturbation step introduced in these works has been further relaxed by Zurek and Chen (2025).

Leave-one-out analyses. Finally, we remark that the construction of state-absorbing MDPs or state-action-absorbing MDPs falls under the category of “leave-one-out” type analysis, which is particularly effective in decoupling complicated statistical dependency in various high-dimensional statistical problems (Chen *et al.*, 2021c), such as high-dimensional M-estimation (El Karoui *et al.*, 2013; El Karoui, 2015; Javanmard and Montanari, 2018), phase retrieval (Chen *et al.*, 2019a; Ma *et al.*, 2020), synchronization (Zhong and Boumal, 2018; Abbe *et al.*, 2020; Ling, 2022), ranking from pairwise comparisons (Chen *et al.*, 2019b; Chen *et al.*, 2022b; Fan *et al.*, 2024; Fan *et al.*, 2025a; Fan *et al.*, 2026), matrix completion and inference (Ma *et al.*, 2020; Chen *et al.*, 2019c; Chen *et al.*, 2020b), tensor completion (Cai *et al.*, 2022), network and clustering analysis (Abbe *et al.*, 2020; Bhattacharya *et al.*, 2026; Zhang and Zhou, 2024; Agterberg and Zhang, 2025; Ke and Wang, 2025). The first application of this powerful analysis framework to MDPs should be attributed to Agarwal *et al.* (2019b), which has since been applied and refined by Pananjady and Wainwright (2020), Li *et al.* (2020), Li *et al.* (2024c), Li *et al.* (2024b), Yan *et al.* (2024), and Shi *et al.* (2023) to establish the optimality of model-based RL in various settings.

4

RL with a generative model: the model-free approach

Thus far, we have established the statistical optimality of model-based RL algorithms under a generative model. Another widely studied, yet fundamentally different, paradigm is the model-free approach (see Figure 2.4), which avoids explicit model estimation and instead seeks to learn the desired value function and/or policy directly from data. In this chapter, we discuss some of the canonical model-free algorithms, once again assuming access to a generative model.

4.1 Prelude: stochastic approximation

The algorithmic and theoretical foundations of many model-free algorithms — such as temporal difference learning and Q-learning — are deeply rooted in the stochastic approximation framework, originally developed by Robbins and Monro (1951). In this section, we briefly review the key ideas underlying stochastic approximation.

Consider the problem of finding a fixed point of a given mapping $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, that is,

$$\text{find } x \in \mathbb{R}^n \quad \text{such that } F(x) = x. \quad (4.1)$$

In many practical settings, however, the mapping $F(\cdot)$ is not directly accessible: one cannot evaluate $F(z)$ exactly at an arbitrary point $z \in \mathbb{R}^n$. Instead, for any given query point $z \in \mathbb{R}^n$, one is permitted to draw a random sample $f(z; \zeta)$ — with ζ capturing the underlying randomness — that provides an unbiased, noisy estimate of $F(z)$.

To address the challenge arising from the lack of direct access to $F(\cdot)$, Robbins and Monro (1951) proposed the following iterative scheme for solving (4.1) based on successive draws of random samples:

$$x_t = x_{t-1} + \eta_t (f(x_{t-1}; \zeta_t) - x_{t-1}), \quad t = 1, 2, \dots \quad (4.2)$$

where $x_0 \in \mathbb{R}^n$ indicates an initial guess, $0 < \eta_t < 1$ denotes the learning rate at iteration t , and $f(x_{t-1}; \zeta_t)$ is a fresh sample satisfying

$$\mathbb{E}[f(x_{t-1}; \zeta_t) \mid x_{t-1}] = F(x_{t-1}).$$

Under suitable conditions on the mapping $F(\cdot)$ (e.g., monotonicity and quasi-contractivity w.r.t. appropriate cones), the data generating process (e.g., bounded noise variance), and the learning rate sequence $\{\eta_t\}_{t \geq 0}$ (e.g., linearly or polynomial decaying stepsizes), this remarkably simple procedure (4.2) is guaranteed to converge to a solution of (4.1) (Lai, 2003; Borkar and Borkar, 2008). See, for example, Wainwright (2019b, Section 2.2) for a non-asymptotic treatment.

The stochastic approximation paradigm has had far-reaching impacts across statistics, optimization, machine learning, and many other fields. A particularly influential descendant is *stochastic gradient descent*, now a cornerstone of modern machine learning (Bottou, 2010; Bottou *et al.*, 2018).

4.2 Temporal difference learning

Equipped with the stochastic approximation framework, we are now ready to introduce several canonical model-free RL algorithms. As in Chapter 3, we begin with the simpler task of policy evaluation, previously introduced in Section 3.2. Given a policy π and a target precision level $\varepsilon > 0$, the objective is to compute an estimate of the corresponding value function V^π , denoted by $\hat{V}^\pi \in \mathbb{R}^S$, such that

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \varepsilon.$$

In sharp contrast to the model-based policy evaluation procedure (3.4), which solves the Bellman equation using an explicit estimate of the probability transition kernel $\hat{P}_\pi$, we now consider an alternative approach that entirely avoids estimating $\hat{P}_\pi$.

4.2.1 Algorithm

Arguably, the most well-known model-free policy evaluation algorithm is temporal-difference (TD) learning, originally formalized by Sutton (1988) and extensively analyzed in the literature (e.g., Tesauro *et al.* (1995), Tsitsiklis and Van Roy (1997), Bhandari *et al.* (2021), Srikant and Ying (2019), and Li *et al.* (2024d)). The core idea of TD learning is to apply stochastic approximation to solve the Bellman equation (or equivalently, to find a fixed point of the Bellman operator).

More concretely, recall that the value function of a policy π is characterized by the Bellman consistency equation (2.12). Consequently, computing V^π amounts to finding a fixed point of the mapping

$$F(V) = \mathcal{T}_\pi(V) := r_\pi + \gamma P_\pi V.$$

This reformulation situates policy evaluation within the stochastic approximation framework. In particular, it naturally suggests an iterative scheme: maintain an estimate $V_t : \mathcal{S} \rightarrow \mathbb{R}$ of the value function in each iteration t , and update it according to

$$V_t = V_{t-1} + \eta_t (\mathcal{T}_{\pi,t}(V_{t-1}) - V_{t-1}), \quad t = 1, 2, \dots \quad (4.3a)$$

provided that, at each iteration, one can construct an unbiased estimate $\mathcal{T}_{\pi,t}(V_{t-1})$ of the Bellman operator applied to the current iterate, satisfying

$$\mathbb{E}[\mathcal{T}_{\pi,t}(V_{t-1}) \mid V_{t-1}] = \mathcal{T}_\pi(V_{t-1}). \quad (4.3b)$$

When a generative model is available, this scheme (4.3) admits a particularly simple implementation: at each iteration t ,

- for each $s \in \mathcal{S}$, draw an independent next-state sample $s_t(s) \sim P_\pi(\cdot \mid s)$ (cf. (2.2)) by querying the generative model;

Algorithm 4.1: TD learning with a generative model

```

1 input: reward function  $r_\pi$  induced by policy  $\pi$ , discount factor  $\gamma$ , learning rates  $\{\eta_t\}$ ,
   number of iterations  $T$ , initialization  $V_0$ .
2 for  $t = 1, 2, \dots, T$  do
3   draw a sample  $s_t(s) \stackrel{\text{ind.}}{\sim} P_\pi(\cdot | s)$  (cf. (2.2)) for every  $s \in \mathcal{S}$ .
   /* stochastic approximation. */
4   for every  $s \in \mathcal{S}$ , update  $V_t(s)$  as
       
$$V_t(s) = (1 - \eta_t)V_{t-1}(s) + \eta_t\{r_\pi(s) + \gamma V_{t-1}(s_t(s))\}.$$


```

- update the value estimate V_t for each state $s \in \mathcal{S}$ according to

$$V_t(s) = V_{t-1}(s) + \eta_t\{r_\pi(s) + \gamma V_{t-1}(s_t) - V_{t-1}(s)\} \quad (4.4a)$$

$$= (1 - \eta_t)V_{t-1}(s) + \eta_t(r_\pi(s) + \gamma V_{t-1}(s_t)), \quad (4.4b)$$

with s_t taken to be $s_t(s)$.

A straightforward calculation shows that this update is unbiased, i.e.,

$$\mathbb{E}[V_t | V_{t-1}] = (1 - \eta_t)V_{t-1} + \eta_t(r_\pi + \gamma P_\pi V_{t-1}),$$

so that (4.4b) indeed implements a stochastic approximation procedure for solving the Bellman equation. Notably, the term

$$r_\pi(s) + \gamma V_{t-1}(s_t) - V_{t-1}(s)$$

in (4.4a) is often referred to as the “temporal difference”, as it reflects the discrepancy between the current value estimate and a one-step look-ahead estimate obtained by bootstrapping one step forward in the MDP. The complete procedure is summarized in Algorithm 4.1. This variant of TD learning is commonly termed *synchronous* TD learning, since each iteration updates all coordinates of the value function estimate simultaneously.

In contrast to the model-based approach in Algorithm 3.1, TD learning maintains only S -dimensional value iterates and does not require storing an explicit S^2 -dimensional estimate of the transition kernel induced by π . This substantially reduces memory requirements, highlighting a central appeal of model-free algorithms.

4.2.2 Theoretical guarantees

We now analyze the efficacy of TD learning through the lens of sample complexity. To this end, we need to specify the learning rate schedule $\{\eta_t\}_{t \geq 1}$, as it plays a crucial role in determining the algorithm’s performance. In this section, we choose the learning rates to satisfy

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2(1-\gamma)}{\log^2 T}t}, \quad 1 \leq t \leq T, \quad (4.5)$$

where $c_1 \geq c_2 > 0$ are some universal constants. Clearly, Condition (4.5) encompasses two commonly used learning-rate choices as special cases:

- (i) *constant learning rates*, where $\eta_t \equiv \frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}}$ for all $t \leq T$;
- (ii) *rescaled linear learning rates*, where $\eta_t = \frac{1}{1 + \frac{c_2(1-\gamma)}{\log^2 T} t}$ for all $t \leq T$.

The following theorem characterizes the ℓ_∞ -based sample complexity of TD learning under a generative model, a result first established by Li *et al.* (2024a).

Theorem 4.1 (Li *et al.*, 2024a). Consider any $\delta \in (0, 1)$, $\varepsilon \in (0, 1]$, and $\gamma \in [1/2, 1)$. Suppose that the learning rates satisfy (4.5) for some small enough universal constants $c_1 \geq c_2 > 0$, and that the total number of iterations T obeys

$$T \geq \frac{c_3 (\log^3 T) (\log \frac{ST}{\delta})}{(1 - \gamma)^3 \varepsilon^2} \quad (4.6)$$

for some sufficiently large universal constant $c_3 > 0$. If $0 \leq V_0(s) \leq \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, then with probability at least $1 - \delta$, TD learning (see Algorithm 4.1) achieves

$$\forall s \in \mathcal{S} : |V_T(s) - V^\pi(s)| \leq \varepsilon.$$

Remark 4.1 (Mean estimation guarantee). In view of the trivial upper bound $|V_T(s) - V^\pi(s)| \leq \frac{1}{1-\gamma}$, the high-probability result in Theorem 4.1 immediately implies an error bound in expectation. Specifically, choosing $\delta = \varepsilon(1 - \gamma)$ in Theorem 4.1 and decomposing according to the high-probability event, we obtain

$$\mathbb{E} \left[\max_{s \in \mathcal{S}} |V_T(s) - V^\pi(s)| \right] \leq \varepsilon(1 - \delta) + \frac{1}{1 - \gamma} \delta \leq 2\varepsilon,$$

as long as the total number of iterations exceeds

$$T \geq \frac{c_3 (\log^3 T) (\log \frac{ST}{\varepsilon(1-\gamma)})}{(1 - \gamma)^3 \varepsilon^2}.$$

Given that each iteration of TD learning draws S fresh samples (one per state), the total number of samples used by Algorithm 4.1 is ST . Consequently, Theorem 4.1 immediately implies that, to yield an accuracy level $\varepsilon \in (0, 1]$, TD learning needs no more than

$$\tilde{O} \left(\frac{S}{(1 - \gamma)^3 \varepsilon^2} \right)$$

samples, and moreover, the *last iterate* of this iterative procedure already attains the desired accuracy. This non-asymptotic result is valid provided the learning rates are properly chosen — either as a suitable constant or according to a rescaled linear schedule (see (4.5)). In comparison with the minimax lower bound presented in Theorem 3.7 (by setting $A = 1$ therein), Theorem 4.1 demonstrates that plain TD learning — without any additional algorithmic trick like iterate averaging or variance reduction — attains minimax-optimal sample complexity (up to logarithmic factors) when one draws samples from the generative model.

There is, however, an important caveat. The minimax optimality of TD learning holds only when the target accuracy resides within the range $\varepsilon \in (0, 1]$, and it is no longer guaranteed once ε becomes significantly larger than 1 (recall that the largest meaningful value of ε is $\frac{1}{1-\gamma}$). This observation carries a broader implication: while model-free algorithms may attain optimality within certain regimes, there is currently no evidence that they can achieve full-range optimality over $\varepsilon \in (0, \frac{1}{1-\gamma}]$, particularly in the most sample-constrained settings.

Remark 4.2. It is also noteworthy that while the last iterate of plain TD learning is shown to be minimax optimal (which concerns worst-case optimality), it may not necessarily enjoy local optimality. As discussed by Khamaru *et al.* (2021), additional algorithmic tricks like variance reduction may sometimes be needed to ensure local optimality.

4.2.3 Analysis: a first attempt

As a starting point, we present an initial — albeit suboptimal — proof strategy that captures some of the core ideas for analyzing the performance of TD learning; this approach will be refined shortly in Section 4.2.4.

Let us adopt the vector and matrix notation introduced in Section 2.1.5, viewing V_t, V^π and r_π as vectors in $\mathbb{R}^S$, among others. By defining an empirical transition matrix $P_t \in \{0, 1\}^{S \times S}$ such that

$$\forall (s, s') \in \mathcal{S}, \quad P_t(s, s') := \begin{cases} 1, & \text{if } s' = s_t(s), \\ 0, & \text{otherwise,} \end{cases} \quad (4.7)$$

where we recall that $s_t(s) \stackrel{\text{ind.}}{\sim} P_\pi(\cdot | s)$, we can express the TD learning update rule (4.4b) compactly in vector-matrix form as follows:

$$V_t = (1 - \eta_t)V_{t-1} + \eta_t(r_\pi + \gamma P_t V_{t-1}), \quad t \geq 1. \quad (4.8)$$

Preliminary facts. Before proceeding, we gather a few elementary facts that will be useful throughout the analysis. The following lemma asserts that the TD iterates always fall within the admissible range of value functions. The proof, based on induction, is deferred to Section 4.6.1.

Lemma 4.2. If $0 \leq V_0(s) \leq \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, then the TD learning iterates (4.8) obey

$$0 \leq V_t \leq \frac{1}{1-\gamma} \mathbf{1}_S \quad \text{and} \quad \|V_t - V^\pi\|_\infty \leq \frac{1}{1-\gamma}$$

for all $0 \leq t \leq T$, provided that $0 \leq \eta_t \leq 1$ for all $0 \leq t \leq T$.

Next, given the iterative nature of TD learning, it is convenient to introduce the following quantities that, as we shall see momentarily, capture the cumulative effect of the learning rates $\{\eta_t\}$:

$$\eta_k^{(t)} := \begin{cases} \prod_{i=1}^t (1 - \eta_i(1 - \gamma)), & \text{if } k = 0, \\ \eta_k \prod_{i=k+1}^t (1 - \eta_i(1 - \gamma)), & \text{if } 0 < k < t, \\ \eta_t, & \text{if } k = t. \end{cases} \quad (4.9)$$

Some basic properties of $\eta_k^{(t)}$ are stated below. In particular, $\eta_k^{(t)}$ can be very small when k is small. The proof is elementary and deferred to Section 4.6.2.

Lemma 4.3. Recall that the learning rates are chosen to satisfy (4.5), and suppose $c_1 c_2 \leq 1/8$. Then for any $t \geq T/(c_2 \log T)$, one has

$$\eta_k^{(t)} \leq \begin{cases} \frac{1}{T^2}, & \text{if } 0 \leq k \leq t/2, \\ \frac{2 \log^3 T}{(1-\gamma)T}, & \text{if } t/2 < k \leq t. \end{cases}$$

Step 1: decomposing the error $V_t - V^\pi$. For notational convenience, define the error sequence

$$\Delta_t := V_t - V^\pi. \quad (4.10)$$

In view of the TD learning update rule (4.8), we have the following decomposition:

$$\begin{aligned} \Delta_t &= (1 - \eta_t)(V_{t-1} - V^\pi) + \eta_t(r_\pi + \gamma P_t V_{t-1} - V^\pi) \\ &= (1 - \eta_t)\Delta_{t-1} + \eta_t\gamma(P_t V_{t-1} - P_\pi V^\pi) \\ &= (1 - \eta_t)\Delta_{t-1} + \eta_t\gamma(P_\pi \Delta_{t-1} + (P_t - P_\pi)V_{t-1}) \end{aligned} \quad (4.11)$$

$$= (I - \eta_t(I - \gamma P_\pi))\Delta_{t-1} + \eta_t\gamma(P_t - P_\pi)V_{t-1}, \quad (4.12)$$

where P_π is defined in (2.19), and the second line arises from the Bellman equation $V^\pi = r_\pi + \gamma P_\pi V^\pi$ (cf. (2.24b)). Iterating the recursion (4.12) yields

$$\Delta_t = \prod_{i=1}^t (I - \eta_i(I - \gamma P_\pi))\Delta_0 + \sum_{k=1}^t Z_k, \quad (4.13)$$

where the process $\{Z_k : 1 \leq k \leq t\}$ is defined as

$$Z_k := \gamma\eta_k \left(\prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right) (P_k - P_\pi)V_{k-1}. \quad (4.14)$$

In words, the error at iteration t decomposes into a suitably contracted version of the initial error and a martingale sum, since each Z_k has zero conditional mean given the history up to iteration $k - 1$ and hence $\{Z_k\}$ forms a martingale difference sequence (this is an immediate consequence of the definition (4.7) of P_k). In what follows, we shall look at the two terms on the right-hand side of (4.13) separately.

Step 2: bounding the first term in (4.13). We first show that the first term on the right-hand side of (4.13) is negligible. To this end, it is observed that

$$\begin{aligned} \left\| \prod_{i=1}^t (I - \eta_i(I - \gamma P_\pi))\Delta_0 \right\|_\infty &\leq \left\| \prod_{i=1}^t (I - \eta_i(I - \gamma P_\pi)) \right\|_1 \|\Delta_0\|_\infty \\ &\stackrel{(i)}{=} \left\{ \prod_{i=1}^t (1 - \eta_i(1 - \gamma)) \right\} \|\Delta_0\|_\infty \\ &= \eta_0^{(t)} \|\Delta_0\|_\infty \stackrel{(ii)}{\leq} \frac{1}{(1 - \gamma)T^2}. \end{aligned} \quad (4.15)$$

Here, (i) invokes the basic fact that, for any $0 \leq k < t$,

$$\left\| \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right\|_1 = \prod_{i=k+1}^t (1 - \eta_i(1 - \gamma)), \quad (4.16)$$

given that P_π is a probability transition matrix; and (ii) results from Lemma 4.3 and the assumption that $\|\Delta_0\|_\infty \leq \frac{1}{1-\gamma}$, provided that $t \geq T/(c_2 \log T)$. As will be shown shortly, this upper bound (4.15) is negligible relative to the other term.

Step 3: bounding the martingale sum in (4.13). We now turn to the second term on the right-hand side of (4.13). Recall that $\{Z_k\}$ forms a martingale difference sequence. In order to control the martingale sum $\sum_k Z_k$, a natural approach is to apply standard martingale concentration inequalities, such as the Azuma–Hoeffding inequality (cf. Lemma 2.8). To this end, we first note that

$$\begin{aligned}
\|Z_k\|_\infty &= \gamma \left\| \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi))(P_k - P_\pi)V_{k-1} \right\|_\infty \\
&\leq \gamma \left\| \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right\|_1 \|(P_k - P_\pi)V_{k-1}\|_\infty \\
&\stackrel{(i)}{=} \gamma \left\{ \eta_k \prod_{i=k+1}^t (1 - \eta_i(1 - \gamma)) \right\} \|(P_k - P_\pi)V_{k-1}\|_\infty \\
&\leq \gamma \eta_k^{(t)} (\|P_k\|_1 + \|P_\pi\|_1) \|V_{k-1}\|_\infty \\
&\stackrel{(ii)}{\leq} \begin{cases} \frac{4\gamma \log^3 T}{(1-\gamma)^{2T}}, & \text{if } k > t/2 \\ \frac{2\gamma}{(1-\gamma)^{T^2}}, & \text{if } 0 \leq k \leq t/2 \end{cases} =: B_{Z_k},
\end{aligned} \tag{4.17}$$

provided that $t \geq T/(c_2 \log T)$. Here, (i) follows from (4.16), whereas (ii) arises from the facts $\|P_k\|_1 = \|P_\pi\|_1 = 1$ (recall the definition (4.7) of P_k), as well as Lemmas 4.2 and 4.3.

Take $t = T$. Armed with the bound (4.17), invoking the Azuma–Hoeffding inequality in Lemma 2.8 and applying the union bound over all coordinates of $\sum_k Z_k$, we arrive at

$$\begin{aligned}
\left\| \sum_{k=1}^T Z_k \right\|_\infty &\leq \left(2 \sum_{k=1}^T B_{Z_k}^2 \log \left(\frac{2S}{\delta} \right) \right)^{1/2} \\
&\leq \gamma \left(\frac{16 \log^6 T \log \left(\frac{2S}{\delta} \right)}{(1-\gamma)^4 T} + \frac{4 \log \left(\frac{2S}{\delta} \right)}{(1-\gamma)^2 T^3} \right)^{1/2} \\
&\leq \tilde{O} \left(\frac{1}{(1-\gamma)^2 \sqrt{T}} \right)
\end{aligned} \tag{4.18}$$

with probability greater than $1 - \delta$.

Step 4: putting all pieces together. Substituting the above results (4.15) and (4.18) into (4.13), we arrive at

$$\|\Delta_T\|_\infty \leq \tilde{O} \left(\frac{1}{(1-\gamma)^2 \sqrt{T}} \right)$$

with high probability. In other words, this result tells us that, in order to yield ε accuracy (i.e., $\|\Delta_T\|_\infty \leq \varepsilon$), it suffices to take T on the order of

$$\tilde{O} \left(\frac{1}{(1-\gamma)^4 \varepsilon^2} \right) \quad \text{iterations.}$$

Discussion. While the above proof is simple and fully non-asymptotic, it yields an iteration complexity that exceeds the bound in Theorem 4.1 by a multiplicative factor of $\frac{1}{1-\gamma}$, and therefore falls short of establishing the minimax optimality of synchronous TD learning. A closer inspection reveals that this suboptimality stems primarily from the use of the Azuma–Hoeffding inequality

(cf. Lemma 2.8), which fails to exploit the underlying variance structure — an ingredient that is crucial for shaving the extra $\frac{1}{1-\gamma}$ factor. To sharpen the analysis, we would need to invoke the Freedman inequality (cf. Lemma 2.9) and leverage the intimate connection between the cumulative conditional covariance and the intrinsic variance structure of the discounted infinite-horizon MDP, as developed in the next subsection.

4.2.4 Sharp analysis: proof of Theorem 4.1

We now present a refined analysis of TD learning that, as noted above, leverages the Freedman inequality for martingales in conjunction with the intrinsic variance structure of the discounted infinite-horizon MDP.

Step 1: decomposing the error as in Section 4.2.3. As already shown in Section 4.2.3 (see (4.13) and (4.15)), we can decompose and bound the estimation error $\Delta_t = V_t - V^\pi$ at iteration t as

$$\|\Delta_t\|_\infty \leq \frac{1}{(1-\gamma)T^2} + \left\| \sum_{k=1}^t Z_k \right\|_\infty, \quad (4.19)$$

where $\{Z_k : 1 \leq k \leq t\}$ is defined in (4.14). It thus boils down to bounding the entrywise magnitude of the martingale sum $\sum_k Z_k$.

Step 2: bounding the martingale sum. We intend to apply the Freedman inequality in Lemma 2.9 to control each coordinate of the sum $\sum_{k=1}^t Z_k$. This requires examining the following two quantities appearing in the Freedman inequality:

$$\max_{1 \leq k \leq t} \|Z_k\|_\infty \quad \text{and} \quad W_t := \sum_{k=1}^t \text{Var}_{k-1}(Z_k), \quad (4.20)$$

where $\text{Var}_{k-1}(Z_k)$ represents an S -dimensional vector whose j -th entry is the variance of the j -th entry of Z_k *conditioned* on what happens up to iteration $k-1$. Throughout this step, we assume $t \geq T/(c_2 \log T)$.

1. **Step 2.1: bounding the maximum magnitude.** Recalling inequality (4.17), one immediately sees that

$$\max_{k: 1 \leq k \leq t} \|Z_k\|_\infty \leq \frac{4\gamma \log^3 T}{(1-\gamma)^2 T} =: B. \quad (4.21)$$

2. **Step 2.2: bounding the conditional entrywise variance of Z_k .** Next, we turn to W_t by first examining each conditional variance term $\text{Var}_{k-1}(Z_k)$. As can be seen from the definition (4.14) of Z_k , each coordinate of Z_k is a bilinear form of $P_k - P_\pi$ (recall the definition of P_k in (4.7)). To analyze such a bilinear form, we first make note of a useful elementary fact: for any fixed vectors $u, v \in \mathbb{R}^S$ obeying $u \geq 0$, one has

$$\begin{aligned} \text{Var}(u^\top (P_k - P_\pi) v) &= \sum_{i=1}^S u_i^2 \text{Var}((P_k - P_\pi)_{i,\cdot} v) \\ &\leq \max_i |u_i| \cdot u^\top \text{Var}_{P_\pi}(v) \end{aligned}$$

$$\leq \|u\|_1 u^\top \text{Var}_{P_\pi}(v), \quad (4.22)$$

where we remind the reader of the notation $\text{Var}_P(V)$ in (2.23). Applying this relation to each entry of Z_k defined in (4.14) gives

$$\begin{aligned} \text{Var}_{k-1}(Z_k) &\leq \gamma^2 \left\| \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right\|_1 \left\{ \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right\} \text{Var}_{P_\pi}(V_{k-1}) \\ &= \gamma^2 \eta_k^{(t)} \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \text{Var}_{P_\pi}(V_{k-1}), \end{aligned} \quad (4.23)$$

where the first inequality is a consequence of (4.22), and the second relation results from (4.16) and the definition of $\eta_k^{(t)}$ (cf. (4.9)). Although the upper bound in (4.23) appears somewhat complicated, it admits a more tractable upper bound once summed as in the next step.

Step 2.3: bounding the cumulative conditional variance W_t . To control the cumulative conditional variance W_t , we bound two partial sums of $\text{Var}_{k-1}(Z_k)$ separately, where the first is negligible and the second constitutes the dominant contribution.

- Consider first those k 's obeying $k \leq t/2$. It follows from (4.23) that

$$\begin{aligned} \sum_{k=1}^{t/2} \text{Var}_{k-1}(Z_k) &\leq \gamma^2 \sum_{k=1}^{t/2} \eta_k^{(t)} \left\| \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right\|_1 \|V_{k-1}\|_\infty^2 1_S \\ &= \gamma^2 \sum_{k=1}^{t/2} (\eta_k^{(t)})^2 \|V_{k-1}\|_\infty^2 1_S \leq \frac{\gamma^2}{2(1-\gamma)^2 T^3} 1_S, \end{aligned} \quad (4.24)$$

where the penultimate step uses (4.16) and (4.9), and the last relation follows from Lemmas 4.2-4.3 and $t \leq T$. As will be made clear shortly, this partial sum (4.24) becomes negligible compared to the remaining part.

- Next, let us turn to those k 's obeying $t/2 < k \leq t$, which account for the dominant contribution to W_t . Invoke Lemma 4.3 and (4.23) to first demonstrate that

$$\text{Var}_{k-1}(Z_k) \leq \frac{2\gamma^2 \log^3 T}{(1-\gamma)T} \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \text{Var}_{P_\pi}(V_{k-1}),$$

which in turn allows one to obtain

$$\sum_{k=t/2}^t \text{Var}_{k-1}(Z_k) \leq \frac{2\gamma^2 \log^3 T}{(1-\gamma)T} \left\{ \sum_{k=t/2+1}^t \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) \right\} \max_{k: t/2 \leq k < t} \text{Var}_{P_\pi}(V_k), \quad (4.25)$$

where the max operator is applied *entrywise*. This inequality (4.25) becomes more useful once the summation inside the curly brackets can be simplified. As it turns out, this summation within the curly brackets — which forms a nonnegative matrix given that P_π is a probability transition matrix — admits a telescoping representation:

$$\sum_{k=t/2+1}^t \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi))$$

$$\begin{aligned}
&= (I - \gamma P_\pi)^{-1} \sum_{k=t/2+1}^t \left[\prod_{i=k+1}^t (I - \eta_i(I - \gamma P_\pi)) - \prod_{i=k}^t (I - \eta_i(I - \gamma P_\pi)) \right] \\
&= (I - \gamma P_\pi)^{-1} - (I - \gamma P_\pi)^{-1} \prod_{i=t/2+1}^t (I - \eta_i(I - \gamma P_\pi)) \\
&\leq (I - \gamma P_\pi)^{-1},
\end{aligned}$$

where the inequality in the last line holds entrywise. Here, the last relation is valid since $(I - \gamma P_\pi)^{-1}$ and $I - \eta_i(I - \gamma P_\pi)$ are both nonnegative matrices. Substituting it into (4.25) and recognizing that the variance terms are all nonnegative lead to

$$\sum_{k=t/2}^t \text{Var}_{k-1}(Z_k) \leq \frac{2\gamma^2 \log^3 T}{(1-\gamma)T} (I - \gamma P_\pi)^{-1} \max_{k: t/2 \leq k < t} \text{Var}_{P_\pi}(V_k). \quad (4.26)$$

Step 2.4: Exploiting the variance structure. The variance upper bound in (4.26) remains intricate, as it involves the entrywise maximum of the variances of the iterates along the TD trajectory. In this step, we further simplify this expression by taking advantage of certain variance structure.

Recall the following key inequality (cf. (3.14)) that captures the intrinsic variance structure of the discounted infinite-horizon MDP:

$$0 \leq (I - \gamma P_\pi)^{-1} \text{Var}_{P_\pi}(V^\pi) \leq \frac{2}{\gamma^2(1-\gamma)^2} 1_S, \quad (4.27)$$

which has played a crucial role in establishing optimal performance for model-based algorithms in Chapter 3. As it turns out, this inequality will also be instrumental here, provided that we can link $\max_k \text{Var}_{P_\pi}(V_k)$ with $\text{Var}_{P_\pi}(V^\pi)$. To do so, we invoke Lemma 2.2 to obtain

$$\text{Var}_{P_\pi}(V_k) \leq \text{Var}_{P_\pi}(V^\pi) + \frac{4}{1-\gamma} \|\Delta_k\|_\infty 1_S,$$

which combined with (4.26) and the fact $(I - \gamma P_\pi)^{-1} 1_S = \frac{1}{1-\gamma} 1_S$ gives

$$\begin{aligned}
\sum_{k=t/2}^t \text{Var}_{k-1}(Z_k) &\leq \frac{2\gamma^2 \log^3 T}{(1-\gamma)T} (I - \gamma P_\pi)^{-1} \text{Var}_{P_\pi}(V^\pi) + \frac{8\gamma^2 \log^3 T}{(1-\gamma)^3 T} \max_{k: t/2 \leq k < t} \|\Delta_k\|_\infty 1_S \\
&\leq \frac{2\gamma^2 \log^3 T}{(1-\gamma)T} \frac{2}{\gamma^2(1-\gamma)^2} 1_S + \frac{8\gamma^2 \log^3 T}{(1-\gamma)^3 T} \max_{k: t/2 \leq k < t} \|\Delta_k\|_\infty 1_S,
\end{aligned} \quad (4.28)$$

where the last inequality results from (4.27). Substituting this bound and (4.24) into the definition (4.20) of W_t then leads to

$$\|W_t\|_\infty \leq \frac{8 \log^3 T}{(1-\gamma)^3 T} \left(1 + \max_{k: t/2 \leq k < t} \|\Delta_k\|_\infty \right). \quad (4.29)$$

Notably, this deterministic bound depends on the estimation error Δ_k from iterations prior to time t , since W_t is defined as the cumulative conditional variance. As we will see momentarily, connecting $\|W_t\|_\infty$ with $\{\|\Delta_k\|_\infty\}$ induces a recursion about this error sequence that ultimately leads to a tight bound on $\{\|\Delta_k\|_\infty\}$.

Before moving on, we note that the above analysis — particularly (4.24) and (4.26) — also implies a coarse deterministic bound by using the elementary fact $\text{Var}_{P_\pi}(V_k) \leq \frac{1}{(1-\gamma)^2} 1_S$:

$$\|W_t\|_\infty \leq \frac{1}{2(1-\gamma)^2 T^3} + \frac{2 \log^3 T}{(1-\gamma)^4 T} \leq \frac{3 \log^3 T}{(1-\gamma)^4 T} =: \sigma_w^2, \quad (4.30)$$

where we invoke Lemma 4.2 and the fact $\|(I - \gamma P)^{-1}\|_1 = \frac{1}{1-\gamma}$. A coarse, deterministic bound of this kind is also needed to apply Freedman's inequality (cf. Lemma 2.9).

Step 2.5: applying Freedman's inequality and establishing a recursion. Equipped with the preceding bounds, we can apply Freedman's inequality (cf. Lemma 2.9) to each entry of $\sum_{k=1}^t Z_k$ and invoke the union bound over all S entries to show that, with probability at least $1 - \delta/T$,

$$\left\| \sum_{k=1}^t Z_k \right\|_\infty \leq \sqrt{8 \max \left\{ \|W_t\|_\infty, \frac{\sigma_w^2}{2K} \right\} \log \frac{2STK}{\delta}} + \frac{4}{3} B \log \frac{2STK}{\delta} \quad (4.31)$$

holds for any fixed integer $K \geq 1$, provided that $t \geq T/(c_2 \log T)$. In view of the condition (4.6) and the assumptions that $\gamma \in [1/2, 1)$ and $\varepsilon \in (0, 1]$, we have

$$T \geq \frac{(\log^3 T)(\log \frac{ST}{\delta})}{(1-\gamma)^3} \geq 8 \quad (4.32)$$

as long as $c_3 \geq 1$. Therefore, by taking $K = 2\lceil \log_2(\frac{1}{1-\gamma}) \rceil$, one can easily verify that

$$\frac{\sigma_w^2}{2K} \leq \frac{3 \log^3 T}{(1-\gamma)^2 T} \quad \text{and} \quad \log \frac{2STK}{\delta} \leq 2 \log \frac{ST}{\delta}. \quad (4.33)$$

Substituting the bounds (4.21), (4.29) and (4.33) into (4.31) yields

$$\left\| \sum_{i=1}^t Z_k \right\|_\infty \leq 4 \sqrt{\frac{(\log^3 T)(\log \frac{ST}{\delta})}{(1-\gamma)^3 T} \left(1 + \max_{k: t/2 \leq k < t} \|\Delta_k\|_\infty \right)} \quad (4.34)$$

with probability at least $1 - \delta/T$, as long as $t \geq T/(c_2 \log T)$. Finally, by plugging (4.34) into (4.19) and taking the union bound over $t \in [T/(c_2 \log T), T]$, we can see that with probability greater than $1 - \delta$,

$$\|\Delta_t\|_\infty \leq 5 \sqrt{\frac{(\log^3 T)(\log \frac{ST}{\delta})}{(1-\gamma)^3 T} \left(1 + \max_{i: t/2 \leq i < t} \|\Delta_i\|_\infty \right)} \quad (4.35)$$

holds simultaneously for all $T/(c_2 \log T) \leq t \leq T$. Importantly, (4.35) establishes a recursive relation governing $\{\Delta_t\}$ across iterations.

Step 3: solving the recursion (4.35). It remains to solve this recursive relation (4.35). Before continuing, let us provide a heuristic assessment of its tightness. Take $t = T$ and assume that $\|\Delta_t\|_\infty \leq 1$ for all $t \geq T/2$, then (4.35) implies that

$$\|\Delta_T\|_\infty \leq \tilde{O} \left(\sqrt{\frac{1}{(1-\gamma)^3 T}} \right),$$

which matches the advertised rate.

We now turn to a formal analysis, establishing the estimation error bound by solving the recursion (4.35). To do so, set

$$u_k := \max \left\{ \|\Delta_t\|_\infty \mid 2^k \frac{T}{c_2 \log T} \leq t \leq T \right\}, \quad (4.36)$$

where the factor 2^k is introduced to reflect the comparison between $\|\Delta_t\|_\infty$ and $\|\Delta_{t/2}\|_\infty$ in (4.35). It is self-evident that $u_0 \leq \frac{1}{1-\gamma}$ (see Lemma 4.2) and, for $k \geq 1$, it follows immediately from (4.35) that

$$u_k \leq q\sqrt{1+u_{k-1}} \quad \text{where} \quad q := 5\sqrt{\frac{(\log^3 T)(\log \frac{ST}{\delta})}{(1-\gamma)^3 T}}. \quad (4.37)$$

In view of condition (4.6), we know that $q < 1/\sqrt{2}$ as long as c_3 is large enough. The following lemma then characterizes the evolution of this sequence $\{u_k\}$, with the proof deferred in Section 4.6.3.

Lemma 4.4. Assume that $q \in (0, 1/\sqrt{2})$ and $u_0 \geq 0$. The following upper bound holds for any $k \geq 1$:

$$u_k \leq \sqrt{2}q + (2q^2)^{1-1/2^k} u_0^{1/2^k}.$$

Therefore, taking $k \geq \max\{1, \log_2(c_2 \log \frac{1}{1-\gamma})\}$, we arrive at

$$u_0^{1/2^k} \leq \left(\frac{1}{1-\gamma}\right)^{1/2^k} \leq e^{1/c_2} \quad \text{and} \quad (2q^2)^{1-1/2^k} \leq \sqrt{2}q.$$

As a result, the bound in Lemma 4.4 simplifies to

$$u_k \leq O(q) \lesssim \sqrt{\frac{(\log^3 T)(\log \frac{ST}{\delta})}{(1-\gamma)^3 T}}, \quad \text{for } k \geq \max\left\{1, \log_2\left(c_2 \log \frac{1}{1-\gamma}\right)\right\}$$

To finish up, take $k = \lfloor \log_2(c_2 \log T) \rfloor$ — which, under our condition on T , satisfies $k \geq \max\{1, \log_2(c_2 \log \frac{1}{1-\gamma})\}$ — and use the definition (4.36) of u_k to yield

$$\|\Delta_T\|_\infty \leq u_{\lfloor \log_2(c_2 \log T) \rfloor} \lesssim \sqrt{\frac{(\log^3 T)(\log \frac{ST}{\delta})}{(1-\gamma)^3 T}}$$

with probability greater than $1 - \delta$, thus establishing Theorem 4.1.

4.3 Q-learning

Having established the appealing performance of synchronous TD learning for policy evaluation, we now turn our attention to the use of model-free algorithms for learning optimal Q-functions. In particular, we focus on the Q-learning algorithm introduced by Watkins (1989) and Watkins and Dayan (1992), one of the most celebrated model-free reinforcement learning methods.

4.3.1 Algorithm

Akin to TD learning, the Q-learning algorithm is designed as a stochastic approximation method for computing a fixed point of the Bellman operator. A key distinction is that the operator of interest is now the Bellman optimality operator (2.13), rather than the Bellman consistency operator (2.10). Unlike the latter, the Bellman optimality operator is nonlinear and nonsmooth (owing to the presence of the max operation), which introduces additional algorithmic and analytical challenges.

We now describe the algorithmic procedure of Q-learning. Following the stochastic approximation framework (4.2), we maintain a running estimate of the optimal Q-function, $Q_t : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, for all $t \geq 0$, and at each iteration t , we update the Q-function estimate according to

$$Q_t = (1 - \eta_t)Q_{t-1} + \eta_t \mathcal{T}_t(Q_{t-1}), \quad (4.38a)$$

Algorithm 4.2: Q-learning with a generative model

```

1 input: discount factor  $\gamma$ , reward function  $r$ , learning rates  $\{\eta_t\}$ , number of iterations  $T$ ,
   initialization  $Q_0$ .
2 for  $t = 1, 2, \dots, T$  do
3   draw a sample  $s_t(s, a) \stackrel{\text{ind.}}{\sim} P(\cdot | s, a)$  for each  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .
   /* stochastic approximation. */
4   update  $Q_t$  according to
      
$$Q_t = (1 - \eta_t)Q_{t-1} + \eta_t \mathcal{T}_t(Q_{t-1}), \quad (4.40)$$

      where for every  $(s, a)$  pair, the  $(s, a)$ -th entry of  $\mathcal{T}_t(Q_{t-1})$  is
      
$$\mathcal{T}_t(Q_{t-1})(s, a) = r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t(s, a), a'). \quad (4.41)$$


```

where we initialize it such that $0 \leq Q_0(s, a) \leq \frac{1}{1-\gamma}$ for every state-action pair (s, a) . Here, $\eta_t \in (0, 1]$ denotes the learning rate at iteration t , and $\mathcal{T}_t(Q_{t-1})$ is constructed as an unbiased estimator of $\mathcal{T}(Q_{t-1})$, where $\mathcal{T}(\cdot)$ denotes the Bellman optimality operator (2.13). With the generative model in place, $\mathcal{T}_t(Q_{t-1})$ can be constructed as follows:

- for every pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, draw an independent sample

$$s_t(s, a) \sim P(\cdot | s, a)$$

by querying the generative model;

- compute $\mathcal{T}_t(Q_{t-1})$ for every (s, a) -th entry as

$$\mathcal{T}_t(Q_{t-1})(s, a) = r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t(s, a), a'). \quad (4.38b)$$

This version of Q-learning is commonly referred to as *synchronous* Q-learning, since all entries of the Q-function are updated simultaneously at each iteration. In addition, the corresponding estimate for the optimal value function, $V_t : \mathcal{S} \rightarrow \mathbb{R}$, can be computed as

$$\forall s \in \mathcal{S} : \quad V_t(s) := \max_{a \in \mathcal{A}} Q_t(s, a) \quad (4.39)$$

at iteration t . The complete description of the Q-learning procedure is summarized in Algorithm 4.2.

4.3.2 Theoretical guarantees

Having established the minimax optimality of TD learning (at least for $\varepsilon \leq 1$), which can be viewed as a special case of Q-learning with number of actions $A = 1$, it is natural to conjecture that Q-learning should likewise attain near-optimal performance. However, this intuition breaks down when $A \geq 2$, as the multi-action setting fundamentally alters the situation. We make this precise via the following ℓ_∞ -based sample complexity bound for Q-learning.

To state the result, we need to specify the learning rate schedule. In particular, we choose learning rates satisfying

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^3 T}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^3 T}}, \quad 0 \leq t \leq T \quad (4.42)$$

for some universal constants $c_1 \geq c_2 > 0$. Similar to the discussions in Section 4.2.2, such choices encompass both constant and rescaled linear learning rates (when appropriately chosen). The theorem below, originally derived by Li *et al.* (2024a), provides the state-of-the-art sample complexity analysis for Q-learning.

Theorem 4.5 (Li *et al.*, 2024a). Consider any $\delta \in (0, 1)$, $\varepsilon \in (0, 1]$, $\gamma \in [1/2, 1)$, and $A \geq 2$. Suppose that the learning rates satisfy (4.42) for some small enough universal constants $c_1 \geq c_2 > 0$, and that the total number of iterations T exceeds

$$T \geq \frac{c_3 (\log^4 T) (\log \frac{SAT}{\delta})}{(1 - \gamma)^4 \varepsilon^2} \quad (4.43)$$

for some large enough universal constant $c_3 > 0$. If the initialization obeys $0 \leq Q_0(s, a) \leq \frac{1}{1-\gamma}$ for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, then with probability at least $1 - \delta$, Q-learning (cf. Algorithm 4.2) achieves

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad |Q_T(s, a) - Q^*(s, a)| \leq \varepsilon.$$

Remark 4.3 (Mean estimation guarantee). The high-probability guarantee in Theorem 4.5 can be readily converted into a bound on the mean estimation error. In light of the crude upper bound $|Q_T(s, a) - Q^*(s, a)| \leq \frac{1}{1-\gamma}$, one can take $\delta = \varepsilon(1 - \gamma)$ to obtain

$$\mathbb{E} \left[\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |Q_T(s, a) - Q^*(s, a)| \right] \leq \varepsilon(1 - \delta) + \frac{1}{1 - \gamma} \delta \leq 2\varepsilon,$$

provided that the total number of iterations obeys

$$T \geq \frac{c_3 (\log^4 T) (\log \frac{SAT}{\varepsilon(1-\gamma)})}{(1 - \gamma)^4 \varepsilon^2}.$$

In essence, Theorem 4.5 develops a non-asymptotic upper bound on the iteration complexity of Q-learning in the presence of the generative model. Recognizing that SA independent samples are drawn in each iteration — yielding a total of SAT samples overall — we can readily translate this result into the following sample complexity bound

$$\tilde{O} \left(\frac{SA}{(1 - \gamma)^4 \varepsilon^2} \right), \quad (4.44)$$

which suffices for Q-learning to achieve ε -accuracy (for $0 < \varepsilon \leq 1$) in the entrywise sense.

We provide further discussion about the learning rate choices. When $T \asymp \frac{(\log^4 T) (\log \frac{SAT}{\delta})}{(1-\gamma)^4 \varepsilon^2}$, Condition (4.42) permits adopting the constant learning rates

$$\eta_t \equiv \frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^3 T}} = \tilde{O}((1 - \gamma)^3 \varepsilon^2), \quad 0 \leq t \leq T,$$

depending almost solely on the discount factor γ and the target accuracy ε . Interestingly, both constant and rescaled linear learning rates satisfying (4.42) lead to the same sample complexity bound (in an order-wise sense), offering flexibility and robustness for practical use.

Nevertheless, the sample complexity bound in (4.44), regardless of the target accuracy level ε , fails to match the minimax lower bound stated in Theorem 3.7. More specifically, the sample complexity bound of Q-learning incurs an additional $\tilde{O}(\frac{1}{1-\gamma})$ multiplicative factor compared to what is achievable by its model-based counterpart (cf. Theorem 3.3 (Li *et al.*, 2024c)). In sharp contrast to the $A = 1$ regime — where TD learning attains near-optimal performance — the presence of multiple actions ($A \geq 2$) appears to degrade the sample optimality of Q-learning.

4.3.3 Tightness and suboptimality

The gap between the upper bound in Theorem 4.5 and the minimax lower bound in Theorem 3.7 naturally raises the following questions: can the sample complexity upper bound for Q-learning be further improved, or is there an inherent bottleneck that prevents plain Q-learning from achieving optimal data efficiency? As it turns out, the performance guarantees in Theorem 4.5 are sharp, in the sense that Q-learning is provably suboptimal as soon as there are at least two actions. This is confirmed by the lower bound below, originally established by Li *et al.* (2024a).

Theorem 4.6 (Li *et al.*, 2024a). Assume that $3/4 \leq \gamma < 1$ and $T \geq \frac{c_3}{(1-\gamma)^2}$ for some sufficiently large constant $c_3 > 0$. Suppose that the initialization is $Q_0 \equiv 0$, and that the learning rates are taken to be either (i) $\eta_t = \frac{1}{1+c_\eta(1-\gamma)t}$ for any constant $c_\eta > 0$ for all $t \geq 0$, or (ii) $\eta_t \equiv \eta \in (0, 1)$ for all $t \geq 0$. There exists a γ -discounted MDP with $S = 4$ and $A = 2$ such that Q-learning (cf. Algorithm 4.2) obeys

$$\max_{s \in \mathcal{S}} \mathbb{E} \left[|V_T(s) - V^*(s)|^2 \right] \geq \frac{c_{\text{lb}}}{(1-\gamma)^4 T \log^2 T} \quad (4.45)$$

for some universal constant $c_{\text{lb}} > 0$.

It is noteworthy that establishing an *algorithm-specific* lower bound requires techniques fundamentally different from those used in information-theoretic analyses (e.g., Fano's lemma is no longer useful here). Although the hard MDP instance constructed in the proof of this theorem is partially inspired by the one in Section 3.4 for deriving minimax lower bounds, the technical approach is drastically different. We defer the proof of this theorem to Section 4.8, which requires carefully tracking the dynamics of Q-learning throughout its execution.

We now discuss the implications of the algorithm-dependent lower bound established in Theorem 4.6. This result shows that plain Q-learning cannot achieve ε -accuracy, in the sense that $\max_s \mathbb{E} [|V_T(s) - V^*(s)|^2] \leq \varepsilon^2$, unless the number of iterations already exceeds

$$\tilde{\Omega} \left(\frac{1}{(1-\gamma)^4 \varepsilon^2} \right),$$

when either constant or rescaled linear learning rates are used. As a consequence, the upper bound for Q-learning derived in Theorem 4.5 is tight in its dependence on the effective horizon $\frac{1}{1-\gamma}$ when $A \geq 2$. Crucially, this lower bound reveals an inherent suboptimality of plain Q-learning: its horizon dependence is provably worse than the minimax rate by a multiplicative factor of $\frac{1}{1-\gamma}$ (up to logarithmic factors) once there are at least two actions.

We note that the potential suboptimality of Q-learning has been recognized in the literature since its early development. A commonly cited explanation attributes such limitations to the over-estimation of Q-values arising from the use of the max operator applied to empirical estimates (see, e.g., Thrun and Schwartz (1993), Hasselt (2010), and Van Hasselt *et al.* (2016)). Informally speaking, even in the idealistic scenario where Q_{t-1} forms a nearly unbiased estimate of Q^* , taking the maximum over actions can introduce an inherent positive bias, that is,

$$\mathbb{E} \left[\max_{a' \in \mathcal{A}} Q_{t-1}(s, a') \right] \geq \max_{a' \in \mathcal{A}} \mathbb{E} [Q_{t-1}(s, a')] \approx \max_{a' \in \mathcal{A}} \mathbb{E} [Q^*(s, a')],$$

where $\max_{a' \in \mathcal{A}} Q_{t-1}(s, a')$ is a key term in the empirical Bellman operator $\mathcal{T}_t(Q_{t-1})$. In this sense, Theorems 4.5 and 4.6 taken together can be viewed as quantifying the impact of over-estimation on the sample complexity and convergence rate of plain Q-learning. A more quantitative comparison between over-estimation and under-estimation errors is provided towards the end of Section 4.3.5.

4.3.4 Analysis: a first attempt

Before presenting the proof of Theorem 4.5, we first illustrate how to establish a looser upper bound $\tilde{O}(\frac{1}{(1-\gamma)^5 \varepsilon^2})$ on the iteration complexity, which exceeds the advertised bound in Theorem 4.5 by a multiplicative factor of $\frac{1}{1-\gamma}$. Although suboptimal, this result features a significantly simpler proof while revealing key recursive structures of the Q-learning iterates that underpin the sharper analysis, as well as some central ideas for handling such recursions.

Akin to the analysis of TD learning, we adopt the vector and matrix notation introduced in Section 2.1.5, allowing us to express the Q-learning update rule (4.38) as

$$Q_t = (1 - \eta_t)Q_{t-1} + \eta_t(r + \gamma P_t V_{t-1}). \quad (4.46)$$

Here, the empirical transition matrix $P_t \in \{0, 1\}^{S \times A \times S}$ is defined as

$$\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}: \quad P_t((s, a), s') := \begin{cases} 1, & \text{if } s' = s_t(s, a), \\ 0, & \text{otherwise,} \end{cases} \quad (4.47)$$

and the vector $V_t \in \mathbb{R}^S$ is defined as $V_t(s) = \max_a Q_t(s, a)$ for all $s \in \mathcal{S}$. For notational convenience, we define π_t throughout this section to be the greedy policy induced by Q_t , namely, for each $s \in \mathcal{S}$,

$$\pi_t(s) := \min \left\{ a \in \mathcal{A} \mid Q_t(s, a) = \max_{a'} Q_t(s, a') \right\}, \quad (4.48)$$

which selects the smallest-indexed action that attains the maximum value of Q_t in state s . The use of the minimum ensures a well-defined choice when multiple actions achieve the maximum value. It is straightforward to verify that

$$Q_t(s, \pi_t(s)) = V_t(s) \quad \text{and} \quad P V_t = P^{\pi_t} Q_t \geq P^{\pi} Q_t \quad (4.49)$$

for any policy π , where P^{π} is defined in (2.19).

In what follows, we first gather a few preliminary facts, before delving into the proof.

Preliminary facts. We begin with a lemma showing that, when initialized properly, each coordinate of both the Q-function and the value function estimates remains within $[0, \frac{1}{1-\gamma}]$ throughout the execution of the Q-learning algorithm. Its proof is deferred to Section 4.7.1.

Lemma 4.7. Assume that $0 \leq Q_0(s, a) \leq \frac{1}{1-\gamma}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Then for every $t \geq 0$, we have

$$0 \leq Q_t \leq \frac{1}{1-\gamma} 1_{SA} \quad \text{and} \quad 0 \leq V_t \leq \frac{1}{1-\gamma} 1_S, \quad (4.50a)$$

provided that $0 \leq \eta_t \leq 1$ for all $t \geq 0$. As an immediate consequence,

$$\|Q_t - Q^*\|_{\infty} \leq \frac{1}{1-\gamma} \quad \text{for all } t \geq 0. \quad (4.50b)$$

Additionally, with a slight abuse of notation, we define

$$\eta_i^{(t)} := \begin{cases} \prod_{j=1}^t (1 - \eta_j), & \text{if } i = 0, \\ \eta_i \prod_{j=i+1}^t (1 - \eta_j), & \text{if } 0 < i < t, \\ \eta_t, & \text{if } i = t, \end{cases} \quad (4.51)$$

which, as we shall see shortly, captures the cumulative effect of the learning rates over time. Note that this definition of $\eta_i^{(t)}$ differs from the one in (4.9) for TD learning in that it does not include the factor $1 - \gamma$. Set

$$\beta := \frac{c_4(1 - \gamma)}{\log T} \quad (4.52)$$

for some universal constant $c_4 > 0$. The following lemma summarizes several useful properties of $\eta_i^{(t)}$, with its proof deferred to Section 4.7.2.

Lemma 4.8. Let β be defined in (4.52). For any $t \geq T/(c_2 \log T)$, the quantity $\eta_i^{(t)}$ defined in (4.51) satisfies

$$\eta_i^{(t)} \leq \begin{cases} \frac{1}{2T^2}, & \text{if } 0 \leq i \leq (1 - \beta)t, \\ \frac{2 \log^4 T}{(1 - \gamma)T}, & \text{if } i > (1 - \beta)t, \end{cases}$$

provided that $c_1 c_2 \leq c_4/8$ and $c_3 \geq 1/c_1$. In addition, it holds that

$$\sum_{i=\tau}^t \eta_i^{(t)} = \begin{cases} 1 - \prod_{j=\tau}^t (1 - \eta_j), & \text{if } 1 \leq \tau \leq t, \\ 1, & \text{if } \tau = 0. \end{cases}$$

In short, the upper bound in Lemma 4.8 treats the two intervals $[0, (1 - \beta)t]$ and $((1 - \beta)t, t]$ separately, particularly since $\eta_i^{(t)}$ is exceedingly small — and hence largely negligible — over the first interval. The choice of β will be made clear momentarily in the proof steps below.

Step 1: decomposing the error $V_t - V^*$. Define the error sequence

$$\Delta_t := Q_t - Q^*. \quad (4.53)$$

By virtue of the update rule (4.46), we can decompose

$$\begin{aligned} \Delta_t &= (1 - \eta_t)Q_{t-1} + \eta_t(r + \gamma P_t V_{t-1}) - Q^* \\ &= (1 - \eta_t)(Q_{t-1} - Q^*) + \eta_t(r + \gamma P_t V_{t-1} - Q^*) \\ &= (1 - \eta_t)\Delta_{t-1} + \eta_t\gamma(P_t V_{t-1} - P V^*) \\ &= (1 - \eta_t)\Delta_{t-1} + \eta_t\gamma\{P(V_{t-1} - V^*) + \underbrace{(P_t - P)V_{t-1}}_{=: \varepsilon_t}\}, \end{aligned} \quad (4.54)$$

where the third line uses the Bellman optimality equation $Q^* = r + \gamma P V^*$ (cf. (2.25)).

Given the intimate connection between the term $V_{t-1} - V^*$ in (4.54) with Δ_t (cf. (4.53)), we seek to relate $P(V_{t-1} - V^*)$ to some linear mapping of Δ_t , which is oftentimes more convenient to work with. To do so, use the definition (4.48) of π_t to express

$$P(V_{t-1} - V^*) = P^{\pi_{t-1}} Q_{t-1} - P^{\pi^*} Q^*, \quad (4.55)$$

where we recall the definition of P^π in (2.19). In view of (4.49) and the optimality of policy π^* , we have $P^{\pi_{t-1}} Q_{t-1} \geq P^{\pi^*} Q_{t-1}$ and $P^{\pi^*} Q^* \geq P^{\pi_{t-1}} Q^*$, which combined with (4.55) give

$$P(V_{t-1} - V^*) \leq P^{\pi_{t-1}} Q_{t-1} - P^{\pi_{t-1}} Q^* = P^{\pi_{t-1}} \Delta_{t-1}; \quad (4.56a)$$

$$P(V_{t-1} - V^*) \geq P^{\pi^*} Q_{t-1} - P^{\pi^*} Q^* = P^{\pi^*} \Delta_{t-1}. \quad (4.56b)$$

In words, (4.56) connects $P(V_{t-1} - V^*)$ with linear mappings of Δ_{t-1} via a suitable calibration of policies. Substituting (4.56) into (4.54) yields

$$\Delta_t \leq (1 - \eta_t)\Delta_{t-1} + \eta_t\gamma(P^{\pi_{t-1}} \Delta_{t-1} + \varepsilon_t), \quad (4.57a)$$

$$\Delta_t \geq (1 - \eta_t)\Delta_{t-1} + \eta_t\gamma(P^{\pi^*}\Delta_{t-1} + \varepsilon_t). \quad (4.57b)$$

The reader may immediately notice some similarity between (4.57) and the iterative form in (4.11) for TD learning. A key difference, however, is that in the upper bound (4.57a), the policy π_{t-1} (and hence $P^{\pi_{t-1}}$) is time-varying, which makes the analysis of Q-learning significantly more challenging; in contrast, in the TD learning case, the policy is frozen to be π (see (4.11)), which simplifies the analysis substantially.

Given that $0 < \eta_t < 1$, applying (4.57) recursively yields

$$\Delta_t \leq \eta_0^{(t)}\Delta_0 + \sum_{i=1}^t \eta_i^{(t)}\gamma(P^{\pi_{i-1}}\Delta_{i-1} + \varepsilon_i), \quad (4.58a)$$

$$\Delta_t \geq \eta_0^{(t)}\Delta_0 + \sum_{i=1}^t \eta_i^{(t)}\gamma(P^{\pi^*}\Delta_{i-1} + \varepsilon_i), \quad (4.58b)$$

where $\eta_i^{(t)}$ is defined in (4.51).

Step 2: partitioning the time horizon. In the summation appearing in the upper bound (4.58a), we would like to partition the time horizon into two intervals: $[0, (1 - \beta)t)$ and $[(1 - \beta)t, t]$ (with β defined in (4.52)), and analyze their contributions separately. Let us decompose (4.58a) as

$$\Delta_t \leq \underbrace{\eta_0^{(t)}\Delta_0 + \sum_{i=1}^{(1-\beta)t} \eta_i^{(t)}\gamma(P^{\pi_{i-1}}\Delta_{i-1} + \varepsilon_i)}_{=: \zeta_t} + \underbrace{\sum_{i=(1-\beta)t+1}^t \eta_i^{(t)}\gamma\varepsilon_i}_{=: \xi_t} + \sum_{i=(1-\beta)t+1}^t \eta_i^{(t)}\gamma P^{\pi_{i-1}}\Delta_{i-1}. \quad (4.59)$$

To explain the rationale behind this partition, observe that for large enough t , the terms involving Δ_i from the earlier interval $[0, (1 - \beta)t)$ become progressively less influential and can often be treated as negligible. This allows us to focus on analyzing the contributions from the more recent, dominant terms, corresponding to indices i closer to the current iteration.

In the ensuing steps, we shall bound ζ_t and ξ_t in (4.59) separately.

Step 3: bounding the term ζ_t . We now develop an upper bound concerning ζ_t (cf. (4.59)). For any t satisfying $T/(c_2 \log T) \leq t \leq T$, it holds that

$$\begin{aligned} \|\zeta_t\|_\infty &\leq \eta_0^{(t)}\|\Delta_0\|_\infty + t \max_{i \leq (1-\beta)t} \eta_i^{(t)} \max_{1 \leq i \leq (1-\beta)t} (\|P^{\pi_{i-1}}\Delta_{i-1}\|_\infty + \|\varepsilon_i\|_\infty) \\ &\stackrel{(i)}{\leq} \eta_0^{(t)}\|\Delta_0\|_\infty + \frac{3t}{1-\gamma} \max_{i \leq (1-\beta)t} \eta_i^{(t)} \\ &\stackrel{(ii)}{\leq} \frac{3t+1}{2T^2(1-\gamma)} \leq \frac{2}{(1-\gamma)T}. \end{aligned} \quad (4.60)$$

Here, (i) follows since (recall the definition of ε_i in (4.54))

$$\|P^{\pi_{i-1}}\Delta_{i-1}\|_\infty \leq \|P^{\pi_{i-1}}\|_1 \|\Delta_{i-1}\|_\infty = \|\Delta_{i-1}\|_\infty \leq \frac{1}{1-\gamma}, \quad (4.61a)$$

$$\|\varepsilon_i\|_\infty \leq (\|P_i\|_1 + \|P\|_1)\|V_{i-1}\|_\infty \leq \frac{2}{1-\gamma}, \quad (4.61b)$$

where we use $\|P^{\pi_{i-1}}\|_1 = \|P_i\|_1 = \|P\|_1 = 1$ as all these matrices are probability transition matrices; and (ii) results from Lemma 4.8. As we shall see momentarily, the magnitude in (4.60) is negligible.

Step 4: bounding the term ξ_t . Turning to the term ξ_t (cf. (4.59)), we make the observation that

$$\mathbb{E}[\varepsilon_i \mid \varepsilon_1, \dots, \varepsilon_{i-1}] = 0 \quad \text{and} \quad \max_i \|\varepsilon_i\|_\infty \leq \frac{2}{1-\gamma} =: B_\varepsilon,$$

where we have used (4.61b). The fact that ξ_t is a weighted sum of $\{\varepsilon_i\}$ motivates the use of martingale concentration inequalities to control ξ_t . To this end, applying the Azuma–Hoeffding inequality (see Theorem 2.8) to each coordinate of ξ_t , followed by a union bound over all coordinates and all t obeying $T/(c_2 \log T) \leq t \leq T$, we can show that with probability at least $1 - \delta$,

$$\begin{aligned} \|\xi_t\|_\infty &\leq \left(2 \sum_{i=(1-\beta)t+1}^t (\eta_i^{(t)} \gamma)^{1/2} B_\varepsilon^2 \log \frac{2SAT}{\delta} \right)^{1/2} \\ &\leq \left(8 \left[\max_{i: (1-\beta)t < i \leq t} \eta_i^{(t)} \right] \left[\sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \right] \frac{\log \frac{2SAT}{\delta}}{(1-\gamma)^2} \right)^2 \\ &\leq \sqrt{\frac{16 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} \end{aligned} \quad (4.62)$$

holds simultaneously for all t obeying $T/(c_2 \log T) \leq t \leq T$, where the last inequality follows from Lemma 4.8. As will be shown shortly, this term contributes significantly to the final bound.

Step 5: establishing a key recursion. With (4.60) and (4.62) in place, we proceed to derive key recursion for Δ_t . Define

$$\psi_t := \max_{s,a} \Delta_t(s, a)$$

to be the maximum entry of Δ_t . Substituting the preceding bounds (4.60) and (4.62) into (4.59), and taking the maximum over all entries on both sides, we obtain that with probability exceeding $1 - \delta$,

$$\psi_t \leq \frac{2}{(1-\gamma)T} + \sqrt{\frac{16 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} + \sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \gamma \psi_{i-1} \quad (4.63)$$

holds simultaneously for all t obeying $T/(c_2 \log T) \leq t \leq T$, where we have made use of the basic property

$$P^{\pi_{i-1}} \Delta_{i-1} \leq \psi_{i-1} P^{\pi_{i-1}} \mathbf{1}_{SA} = \psi_{i-1} \mathbf{1}_{SA}.$$

To analyze the recursion (4.63), we rely on the following intuition. At a high level, the expectation of ψ_i may decrease over time as the algorithm's performance improves; at the same time, it may not change too rapidly over a short time window, allowing us to replace the terms within such a window by their maximum value with little loss in tightness. Building upon these intuitions, we use (4.63) to derive

$$\begin{aligned} \psi_t &\leq \frac{2}{(1-\gamma)T} + \sqrt{\frac{16 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} + \gamma \left(\sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \right) \max_{i \in [(1-\beta)t, t]} \psi_i, \\ &\leq \sqrt{\frac{64 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} + \gamma \max_{k: (1-\beta)t \leq k \leq T} \psi_k, \end{aligned}$$

where the last inequality invokes Lemma 4.8 (i.e., $\sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \leq 1$). Since this inequality holds for all t obeying $T/(c_2 \log T) \leq t \leq T$, taking the maximum over all iterations from t to T implies that, with probability exceeding $1 - \delta$,

$$\max_{k \in [t, T]} \psi_k \leq \sqrt{\frac{64 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} + \gamma \max_{k \in [(1-\beta)t, T]} \psi_k \quad (4.64)$$

holds simultaneously for all t satisfying $T/(c_2 \log T) \leq t \leq T$.

Step 6: solving the recursive relation. It then remains to solve the recursive relation (4.64). By defining the scalar sequence

$$u_i := \max_{k \in [(1-\beta)^i T, T]} \psi_k, \quad i \geq 0,$$

one can use (4.64) to show that, with probability greater than $1 - \delta$,

$$u_i \leq \sqrt{\frac{64 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} + \gamma u_{i+1}, \quad (4.65)$$

holds simultaneously for all i satisfying $(1-\beta)^i \geq 1/(c_2 \log T)$ (since in (4.64) we require $t \geq T/(c_2 \log T)$). By setting

$$i^* = \max \{i \mid (1-\beta)^i \geq 1/(c_2 \log T)\},$$

we can apply (4.65) recursively to reach

$$u_0 \leq i^* \left(\sqrt{\frac{64 \log^4 T \log \frac{2SAT}{\delta}}{(1-\gamma)^3 T}} \right) + \gamma^{i^*} u_{i^*}. \quad (4.66)$$

To solve this recursive relation, we make note of the basic facts that (recall the choice of β in (4.52))

$$u_{i^*} \leq \max_{t \geq 0} \|\Delta_t\|_\infty \leq \frac{1}{1-\gamma} \quad \text{and} \quad i^* \lesssim \frac{\log T \log \log T}{1-\gamma},$$

which result in

$$\gamma^{i^*} u_{i^*} \lesssim \gamma^{\frac{\log T \log \log T}{1-\gamma}} \frac{1}{1-\gamma} \lesssim \frac{1}{T^2} \frac{1}{1-\gamma} \lesssim \frac{1}{T}.$$

These allow one to conclude that, with probability at least $1 - \delta$,

$$\begin{aligned} \max_{s,a} \Delta_T(s,a) = u_0 &\lesssim \frac{\log T \log \log T}{1-\gamma} \left(\frac{1}{(1-\gamma)T} + \sqrt{\frac{\log^4 T \log \frac{SAT}{\delta}}{(1-\gamma)^3 T}} \right) + \frac{1}{T} \\ &\lesssim \sqrt{\frac{\log^7 T \log \frac{SAT}{\delta}}{(1-\gamma)^5 T}}. \end{aligned}$$

Remark 4.4. At a high level, the proof reveals that the final estimation error Δ_T arises primarily from contributions over $\tilde{O}(\frac{1}{1-\gamma})$ intervals of exponentially decreasing lengths, namely, $[(1-\beta)T, T)$, $[(1-\beta)^2 T, (1-\beta)T)$, $[(1-\beta)^3 T, (1-\beta)^2 T)$, and so on. Each such interval contributes $\tilde{O}(\frac{1}{\sqrt{(1-\gamma)^3 T}})$ to the final estimation error. The careful construction of these intervals is a key ingredient in establishing the above error bound. As we shall see, this idea will be further refined and play an important role in the sharper analysis.

Step 7: putting all this together. Repeating the above arguments tailored to the lower bound (4.58b) immediately reveals that, with probability at least $1 - \delta$,

$$\max_{s,a} \Delta_T(s,a) \geq -O\left(\sqrt{\frac{\log^7 T \log \frac{SAT}{\delta}}{(1-\gamma)^5 T}}\right).$$

Taking the above upper and lower bounds together, we conclude that with probability exceeding $1 - 2\delta$,

$$\|Q_T - Q^*\|_\infty = \|\Delta_T\|_\infty \lesssim \sqrt{\frac{\log^7 T \log \frac{SAT}{\delta}}{(1-\gamma)^5 T}}. \quad (4.67)$$

Discussion. As alluded to previously, the bound (4.67), while admitting a concise proof, is loose by a factor of $\frac{1}{1-\gamma}$. The gap stems from the use of the Azuma–Hoeffding inequality in controlling ξ_t , which relies solely on coarse bounds on $\{\varepsilon_i\}$ and fails to capture the variance structure inherent in the discounted infinite-horizon MDP. To obtain sharper guarantees, it is therefore essential to exploit this variance structure through more refined martingale concentration techniques, such as the Freedman inequality (see Theorem 2.9). However, even after incorporating this refinement, significant challenges remain: the resulting recursive relation becomes considerably more intricate and requires new analytical ideas to handle. These developments are the focus of the next subsection.

4.3.5 Sharp analysis: proof of Theorem 4.5

We now present the full proof of Theorem 4.5, building in part on the ideas developed in the suboptimal analysis in Section 4.3.4. Before proceeding, we record below a direct consequence of the assumption (4.43) in conjunction with the range $\varepsilon \leq 1$:

$$T \geq \frac{c_3 (\log^4 T) (\log \frac{SAT}{\delta})}{(1-\gamma)^4}, \quad (4.68)$$

a condition that will be invoked repeatedly in the subsequent analysis. In what follows, we focus on establishing an upper bound on Δ_t ; the corresponding lower bound can be obtained via nearly identical arguments.

Step 1: decomposing the error as in Section 4.3.4. Armed with (4.59) and (4.60) established in Section 4.3.4, the Q-estimation error $\Delta_t = Q_t - Q^*$ at any iteration satisfying $T/(c_2 \log T) \leq t \leq T$ obeys

$$\Delta_t \leq \frac{2}{(1-\gamma)T} 1_{SA} + \xi_t + \sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \gamma P^{\pi_{i-1}} \Delta_{i-1}, \quad (4.69)$$

where π_i , β and ξ_t are defined in (4.48), (4.52) and (4.59), respectively. It then comes down to controlling ξ_t and solving the resulting recursive relation.

Step 2: bounding the term ξ_t . We first develop a tighter bound on the term ξ_t (cf. (4.59)). Recall that ξ_t can be expressed as

$$\xi_t = \sum_{i=(1-\beta)t+1}^t Z_i, \quad \text{with } Z_i := \eta_i^{(t)} \gamma \varepsilon_i = \eta_i^{(t)} \gamma (P_i - P) V_{i-1},$$

where $\{Z_i\}$ forms a martingale difference sequence. This motivates the use of Freedman's inequality (cf. Theorem 2.9) to control ξ_t . To do so, we need to look at

$$\max_{i: (1-\beta)t < i \leq t} \|Z_i\|_\infty, \quad \text{and} \quad W_t := \sum_{i=(1-\beta)t+1}^t \text{Var}_{i-1}(Z_i),$$

where $\text{Var}_k(Z)$ is an (SA) -dimensional vector whose entries are the conditional variances of the corresponding coordinates of Z , conditional on what happens up to iteration k . We now develop upper bounds for both objects for any t obeying $T/(c_2 \log T) \leq t \leq T$.

- Regarding the maximum magnitudes, we observe that

$$\begin{aligned} \max_{i: (1-\beta)t < i \leq t} \|Z_i\|_\infty &\leq \max_{i: (1-\beta)t < i \leq t} \|\eta_i^{(t)} \gamma (P_i - P) V_{i-1}\|_\infty \\ &\leq \max_{i: (1-\beta)t < i \leq t} \eta_i^{(t)} \gamma (\|P_i\|_1 + \|P\|_1) \|V_{i-1}\|_\infty \\ &\leq \frac{4\gamma \log^4 T}{(1-\gamma)^2 T} =: B_z, \end{aligned} \tag{4.70}$$

where the last inequality follows from Lemmas 4.7 and 4.8, in conjunction with the basic fact $\|P_i\|_1 = \|P\|_1 = 1$.

- Next, we turn to the variance term W_t . Adopting the notation $\text{Var}_P(\cdot)$ from (2.22), we can deduce that

$$\begin{aligned} W_t &= \gamma^2 \sum_{i=(1-\beta)t+1}^t (\eta_i^{(t)})^2 \text{Var}_P(V_{i-1}) \\ &\leq \gamma^2 \left(\max_{i \in [(1-\beta)t, t]} \eta_i^{(t)} \right) \left(\sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \right) \max_{i \in [(1-\beta)t, t]} \text{Var}_P(V_i) \\ &\leq \frac{2\gamma^2 \log^4 T}{(1-\gamma)T} \max_{i \in [(1-\beta)t, t]} \text{Var}_P(V_i), \end{aligned} \tag{4.71}$$

where the last inequality is a consequence of Lemma 4.8.

- Moreover, we find it helpful to record a deterministic upper bound on W_t , which is also useful when invoking Theorem 2.9. By Lemma 4.7, we have $\text{Var}_P(V_i) \leq (1-\gamma)^{-2}$, which combined with (4.71) gives

$$\|W_t\|_\infty \leq \frac{2\gamma^2 \log^4 T}{(1-\gamma)T} \cdot \frac{1}{(1-\gamma)^2} = \frac{2\gamma^2 \log^4 T}{(1-\gamma)^3 T} =: \sigma_w^2.$$

With these bounds in mind, applying Freedman's inequality in Theorem 2.9 by setting $K = \lceil 2 \log_2 \frac{1}{1-\gamma} \rceil$, and taking a union bound over all SA coordinates of ξ_t and all iterations of interest, we see that with probability at least $1 - \delta$,

$$\begin{aligned} |\xi_t| &\leq \sqrt{8 \left(\|W_t\|_\infty + \frac{\sigma_w^2}{2K} 1_{SA} \right) \log \left(\frac{8SAT \log \frac{1}{1-\gamma}}{\delta} \right)} + \frac{4}{3} B_z \log \left(\frac{8SAT \log \frac{1}{1-\gamma}}{\delta} \right) 1_{SA} \\ &\leq \sqrt{\frac{48\gamma^2 (\log^4 T) (\log \frac{SAT}{\delta})}{(1-\gamma)T} \left(\max_{i \in [(1-\beta)t, t]} \text{Var}_P(V_i) + 1_{SA} \right)} \end{aligned}$$

holds simultaneously for all t obeying $T/(c_2 \log T) \leq t \leq T$. Here, the second inequality follows from the preceding bounds (4.70), (4.71) and

$$\frac{\sigma_w^2}{2^K} \leq \frac{2\gamma^2 \log^4 T}{(1-\gamma)T},$$

together with the condition (4.68) on T , provided that $c_3 > 0$ is sufficiently large.

Step 3: establishing a recursive relation for Δ_t . We now leverage the bound on ξ_t to derive a recursive relation for Δ_t . Recall the sample size condition (4.68) and assume that $c_3 > 0$ is sufficiently large and $\gamma \geq 1/2$. Substitution of the bound on ξ_t derived in Step 2 into (4.69), along with a little algebra, shows that with probability at least $1 - \delta$,

$$\Delta_t \leq \sqrt{\varphi_{\beta,t}} + \sum_{i=(1-\beta)t+1}^t \eta_i^{(t)} \gamma P^{\pi_{i-1}} \Delta_{i-1} \quad (4.72a)$$

holds for all t satisfying $T/(c_2 \log T) \leq t \leq T$, where we define

$$\varphi_{\beta,t} := \frac{64\gamma^2 (\log^4 T) (\log \frac{SAT}{\delta})}{(1-\gamma)T} \left(\max_{i \in [(1-\beta)t, t]} \text{Var}_P(V_i) + 1_{SA} \right). \quad (4.72b)$$

Additionally, given that $\beta := c_4(1-\gamma)/\log T \leq 1/8$ for large enough T , we know that for any k obeying $2t/3 \leq k \leq t$, one has $t/2 \leq 7k/8 \leq (1-\beta)k$. As a result, (4.72) implies that with probability at least $1 - \delta$,

$$\Delta_k \leq \sqrt{\varphi_{\beta,k}} + \sum_{i=(1-\beta)k+1}^k \eta_i^{(k)} \gamma P^{\pi_{i-1}} \Delta_{i-1} \leq \sqrt{\varphi_t} + \sum_{i=(1-\beta)k}^{k-1} \eta_{i+1}^{(k)} \gamma P^{\pi_i} \Delta_i \quad (4.73a)$$

holds simultaneously for all t and k obeying $T/(c_2 \log T) \leq k \leq t \leq T$ and $k \geq 2t/3$, where φ_t is defined as

$$\varphi_t := \frac{64\gamma^2 (\log^4 T) (\log \frac{SAT}{\delta})}{(1-\gamma)T} \left(\max_{i \in [t/2, t]} \text{Var}_P(V_i) + 1_{SA} \right). \quad (4.73b)$$

Note that the difference between φ_t and $\varphi_{\beta,t}$ lies in the interval over which the maximum is taken: $[t/2, t]$ vs. $[(1-\beta)t, t]$.

Before continuing, we remark that in comparison to (4.63) in the looser analysis, we now retain the variance term $\text{Var}_P(V_i)$ in φ_t (rather than bounding it crudely) as well as the transition matrix P^{π_i} , which are crucial for exploiting the variance structure in subsequent steps.

Step 4: iterating the recursion for $\tilde{O}(\frac{1}{1-\gamma})$ times. Recall that in the suboptimal analysis in Section 4.3.4 (see Remark 4.4), we have observed that the estimation error Δ_t is primarily driven by contributions from $\tilde{O}(\frac{1}{1-\gamma})$ intervals, namely $[(1-\beta)t, t)$, $[(1-\beta)^2 t, (1-\beta)t)$, $[(1-\beta)^3 t, (1-\beta)^2 t)$, and so on. This observation motivates us to invoke the recursive relation (4.73) for $\tilde{O}(\frac{1}{1-\gamma})$ times, so as to make explicit the contributions from these intervals.

To do so, we first find it convenient to introduce the normalized weights defined by

$$\alpha_i^{(t)} := \frac{\eta_{i+1}^{(t)}}{\sum_{j=(1-\beta)t}^{t-1} \eta_{j+1}^{(t)}},$$

which clearly satisfy (see Lemma 4.8)

$$\alpha_i^{(t)} \geq \eta_{i+1}^{(t)} \quad \text{and} \quad \sum_{i=(1-\beta)t}^{t-1} \alpha_i^{(t)} = 1. \quad (4.74)$$

As we will see shortly, this normalization helps streamline the presentation of the proof. Let us use this notation to rewrite (4.73) as

$$\Delta_k \leq \sum_{i_1=(1-\beta)k}^{k-1} \left(\alpha_{i_1}^{(k)} \sqrt{\varphi_t} + \eta_{i_1+1}^{(k)} \gamma P^{\pi_{i_1}} \Delta_{i_1} \right) \quad (4.75)$$

for all t and k obeying $T/(c_2 \log T) \leq k \leq t \leq T$ and $k \geq 2t/3$. Since $(1-\beta)^2 t \geq 2t/3$, applying this relation (4.75) once more yields

$$\begin{aligned} \Delta_t &\leq \sum_{i_1=(1-\beta)t}^{t-1} \left(\alpha_{i_1}^{(t)} \sqrt{\varphi_t} + \eta_{i_1+1}^{(t)} \gamma P^{\pi_{i_1}} \Delta_{i_1} \right) \\ &\leq \sum_{i_1=(1-\beta)t}^{t-1} \left[\alpha_{i_1}^{(t)} \sqrt{\varphi_t} + \eta_{i_1+1}^{(t)} \gamma P^{\pi_{i_1}} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \left(\alpha_{i_2}^{(i_1)} \sqrt{\varphi_t} + \eta_{i_2+1}^{(i_1)} \gamma P^{\pi_{i_2}} \Delta_{i_2} \right) \right] \\ &\leq \sum_{i_1=(1-\beta)t}^{t-1} \alpha_{i_1}^{(t)} \sqrt{\varphi_t} + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} (\gamma P^{\pi_{i_1}}) \sqrt{\varphi_t} \\ &\quad + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \left\{ \prod_{k=1}^2 (\gamma P^{\pi_{i_k}}) \right\} \Delta_{i_2} \\ &= \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} (I + \gamma P^{\pi_{i_1}}) \sqrt{\varphi_t} + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \left\{ \prod_{k=1}^2 (\gamma P^{\pi_{i_k}}) \right\} \Delta_{i_2}, \end{aligned} \quad (4.76)$$

where we adopt the following matrix product notation (note the order of multiplication):

$$\prod_{k=1}^j A_k = \begin{cases} A_1 \cdots A_j, & \text{if } j \geq 1, \\ I, & \text{if } j = 0. \end{cases} \quad (4.77)$$

Here, the second inequality results from (4.75), the third inequality follows since $\eta_{i_1+1}^{(t)} \leq \alpha_{i_1}^{(t)}$ (cf. (4.74)), whereas the last identity uses the normalization condition $\sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_2}^{(i_1)} = 1$ (cf. (4.74)).

Thus far, we have illustrated the effect of two iterations of the recursive relation (4.75). We now proceed to unroll this recursion $\tilde{O}(\frac{1}{1-\gamma})$ times, in a manner analogous to the derivation of (4.76). To this end, we begin with several preparatory steps.

- Introduce the following notation:

$$H := \frac{\log T}{1-\gamma}, \quad \alpha_{\{i_k\}_{k=1}^H} := \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} \cdots \alpha_{i_H}^{(i_{H-1})} \geq 0 \quad (4.78)$$

for any indices $t > i_1 > i_2 > \cdots > i_H$. By (4.74), we have

$$\alpha_{\{i_k\}_{k=1}^H} \geq \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \cdots \eta_{i_H+1}^{(i_{H-1})}.$$

- Define the index set

$$\mathcal{I}_t := \left\{ (i_1, \dots, i_H) \mid (1-\beta)t \leq i_1 \leq t-1, \right. \\ \left. \forall 1 \leq j < H : (1-\beta)i_j \leq i_{j+1} \leq i_j - 1 \right\}, \quad (4.79)$$

which, according to (4.74), satisfies

$$\sum_{(i_1, \dots, i_H) \in \mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} = 1. \quad (4.80)$$

- Given that $\beta = c_4(1-\gamma)/\log T$, the above choice of H obeys

$$(1-\beta)^H = \left(1 - \frac{c_4(1-\gamma)}{\log T}\right)^{\frac{\log T}{1-\gamma}} \geq \frac{2}{3},$$

provided that c_4 is sufficiently small. Hence, for any $(i_1, \dots, i_H) \in \mathcal{I}_t$, we have

$$i_1 > i_2 > \dots > i_H \geq (1-\beta)^H t \geq \frac{2t}{3}.$$

With these preparations in place, applying (4.75) recursively for H times yields — analogous to the derivation of (4.76) — that with probability exceeding $1 - \delta$,

$$\begin{aligned} \Delta_t &\leq \sum_{(i_1, \dots, i_H) \in \mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} \left\{ \left(\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^h P^{\pi_{i_k}} \right) \sqrt{\varphi_t} + \gamma^H \left(\prod_{k=1}^H P^{\pi_{i_k}} \right) |\Delta_{i_H}| \right\} \\ &\leq \max_{(i_1, \dots, i_H) \in \mathcal{I}_t} \left\{ \underbrace{\left(\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^h P^{\pi_{i_k}} \right) \sqrt{\varphi_t}}_{=: \beta_1} + \underbrace{\gamma^H \left(\prod_{k=1}^H P^{\pi_{i_k}} \right) |\Delta_{i_H}|}_{=: \beta_2} \right\} \end{aligned} \quad (4.81)$$

holds simultaneously for all $t \geq 3T/(2c_2 \log T)$, where we adopt the matrix product notation in (4.77), and the last line follows from (4.80).

We are now left with two terms to control: β_1 and β_2 , which we handle next.

Step 5: bounding the term β_2 . Let us begin by bounding the term β_2 in (4.81), which is much easier to cope with than β_1 . Observing that $\prod_{k=1}^H P^{\pi_{i_k}}$ is a probability transition matrix, we can deduce that

$$\begin{aligned} \|\beta_2\|_\infty &\leq \gamma^H \left\| \prod_{k=1}^H P^{\pi_{i_k}} \right\|_1 \|\Delta_{i_H}\|_\infty = \gamma^H \|\Delta_{i_H}\|_\infty \\ &\stackrel{(i)}{\leq} \frac{1}{1-\gamma} \gamma^H \stackrel{(ii)}{\leq} \frac{1}{(1-\gamma)T}. \end{aligned} \quad (4.82)$$

Here, (i) follows from the crude bound (4.50b); to justify (ii), recall from (4.78) that $H = \frac{\log T}{1-\gamma}$, so that

$$\gamma^H = (1 - (1-\gamma))^{\frac{\log T}{1-\gamma}} \leq e^{-\log T} = \frac{1}{T},$$

where the inequality uses the elementary bound $\gamma^{1/(1-\gamma)} \leq e^{-1}$ valid for all $0 < \gamma < 1$. In words, under our choice of H , the term β_2 becomes negligible due to the factor γ^H , and consequently β_1 dominates.

Step 6: bounding the term β_1 . We now turn attention to β_1 . Recalling the definitions of β_1 and φ_t in (4.81) and (4.73b), respectively, we see that β_1 involves terms of the form

$$P^{\text{tmp}} \sqrt{\max_i \text{Var}_P(V_i)}$$

for some probability transition matrix P^{tmp} . Based on the analysis in previous sections (e.g., Section 3.2.4), it is often convenient to move the transition matrix P^{tmp} inside the square root operator, as this facilitates exploiting the variance structure to improve horizon dependence. To do so, we proceed to bound the entrywise square of β_1 , denoted by $|\beta_1|^2$, via Jensen's inequality. More precisely, we can demonstrate that

$$\begin{aligned} |\beta_1|^2 &= \left| \left(\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^h P^{\pi_{i_k}} \right) \sqrt{\varphi_t} \right|^2 \stackrel{(i)}{\leq} \left| \sum_{h=0}^{H-1} \gamma^{h/2} \cdot \gamma^{h/2} \sqrt{\left(\prod_{k=1}^h P^{\pi_{i_k}} \right) \varphi_t} \right|^2 \\ &\stackrel{(ii)}{\leq} \sum_{h=0}^{H-1} \gamma^h \cdot \sum_{h=0}^{H-1} \gamma^h \left(\prod_{k=1}^h P^{\pi_{i_k}} \right) \varphi_t \\ &\stackrel{(iii)}{\leq} \frac{1}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \left(\prod_{k=1}^h P^{\pi_{i_k}} \right) \frac{64\gamma^2(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)T} \left(\max_{i \in [t/2, t]} \text{Var}_P(V_i) + 1_{SA} \right) \\ &\stackrel{(iv)}{\leq} \frac{64\gamma^2(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^2 T} \sum_{h=0}^{H-1} \gamma^h \left(\prod_{k=1}^h P^{\pi_{i_k}} \right) \max_{i \in [t/2, t]} \text{Var}_P(V_i) + \frac{64\gamma^2(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^3 T} 1_{SA}. \end{aligned} \quad (4.83)$$

Here, (i) follows from Jensen's inequality and the fact that $\prod_{k=1}^h P^{\pi_{i_k}}$ is a probability transition matrix; (ii) applies the Cauchy-Schwarz inequality; (iii) uses the definition of φ_t in (4.73b) and $\sum_{h=0}^{H-1} \gamma^h \leq \frac{1}{1-\gamma}$; and (iv) follows since $\prod_{k=1}^h P^{\pi_{i_k}} 1_{SA} = 1_{SA}$ and $\sum_{h=0}^{H-1} \gamma^h \leq \frac{1}{1-\gamma}$.

Step 7: exploiting the variance structure. The remaining task is to derive a concise yet tight bound on the summation in (4.83). As it turns out, accomplishing this requires a careful exploitation of the underlying variance structure of the MDP. Towards this, we present the following lemma, which plays a central role in our sharp analysis of Q-learning and may be of independent interest as well. The proof is deferred to Section 4.7.3.

Lemma 4.9. Suppose $T/(c_2 \log T) \leq t \leq T$. Then for any $(i_1, \dots, i_H) \in \mathcal{I}_t$ (cf. (4.79)), one has

$$\left\| \sum_{h=0}^{H-1} \gamma^h \left(\prod_{k=1}^h P^{\pi_{i_k}} \right) \max_{i \in [t/2, t]} \text{Var}_P(V_i) \right\|_{\infty} \leq \frac{4(1 + 2 \max_{i \in [t/2, t]} \|\Delta_i\|_{\infty})}{\gamma^2(1-\gamma)^2}.$$

Taking this lemma and (4.83) collectively, we immediately obtain

$$\|\beta_1\|_{\infty}^2 \leq \frac{320(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^4 T} \left(1 + 2 \max_{i \in [t/2, t]} \|\Delta_i\|_{\infty} \right). \quad (4.84)$$

Step 8: establishing a simpler recursion for Δ_t . Substituting the above bounds (4.82) and (4.84) on β_1 and β_2 into (4.81) reveals that, with probability at least $1 - \delta$,

$$\Delta_t \leq \frac{1}{(1-\gamma)T} 1_{SA} + \sqrt{\frac{320(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^4 T} \left(1 + 2 \max_{i \in [t/2, t]} \|\Delta_i\|_{\infty} \right) 1_{SA}}$$

$$\leq 30 \sqrt{\frac{(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^4 T}} \left(1 + \max_{i \in [t/2, t]} \|\Delta_i\|_\infty\right) 1_{SA}, \quad (4.85)$$

holds simultaneously for all t obeying $T/(c_2 \log T) \leq t \leq T$, where the second inequality follows from the sample size condition (4.68).

Repeating similar arguments (which we omit here for the sake of brevity) also shows that with probability at least $1 - \delta$,

$$\Delta_t \geq -30 \sqrt{\frac{(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^4 T}} \left(1 + \max_{i \in [t/2, t]} \|\Delta_i\|_\infty\right) 1_{SA} \quad (4.86)$$

holds simultaneously for all t obeying $T/(c_2 \log T) \leq t \leq T$. Combine (4.85) and (4.86) and apply the union bound to demonstrate that, with probability exceeding $1 - 2\delta$,

$$\|\Delta_t\|_\infty \leq 30 \sqrt{\frac{(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^4 T}} \left(1 + \max_{i \in [t/2, t]} \|\Delta_i\|_\infty\right). \quad (4.87)$$

holds simultaneously for all t obeying $T/(c_2 \log T) \leq t \leq T$. This forms the core recursive relation governing the dynamics of Δ_t .

Step 9: solving the recursion (4.87). The careful reader may immediately notice that the recursive relation (4.87) closely resembles (4.35) derived for the TD learning case, except for a difference in the prefactor. As a result, the same arguments can be applied to solve the recursion (4.87). In fact, invoking the recursion argument from the TD learning setting (see Step 3 in Section 4.2.4, particularly Lemma 4.4) readily yields

$$\|\Delta_T\|_\infty \lesssim \sqrt{\frac{(\log^4 T)(\log \frac{SAT}{\delta})}{(1-\gamma)^4 T}}$$

under the sample size condition (4.43), thus completing the proof.

Discussion: over-estimation vs. under-estimation. Before concluding this subsection, we single out a difference between analyzing the over-estimation and under-estimation errors in Q-learning. Recall that the starting point of our analysis is the decomposition (4.57), where the lower inequality (4.57b) for Δ_t closely resembles the recursion (4.11) arising in the TD learning setting. In fact, the lower error can be controlled by repeating the analysis for the TD learning case, yielding

$$\Delta_T \geq -\tilde{O}\left(\sqrt{\frac{1}{(1-\gamma)^3 T}}\right) 1_{SA}.$$

This under-estimation bound is smaller than the corresponding over-estimation bound (4.85) by a factor of $\frac{1}{\sqrt{1-\gamma}}$. In other words, over-estimation poses a more significant issue in Q-learning than under-estimation — an asymmetry that is reflected in the proof.

4.4 Extension: asynchronous Q-learning

As noted previously, the Q-learning algorithm presented in Algorithm 4.2 is commonly referred to as *synchronous* Q-learning, reflecting the fact that the Q-values for all state-action pairs are

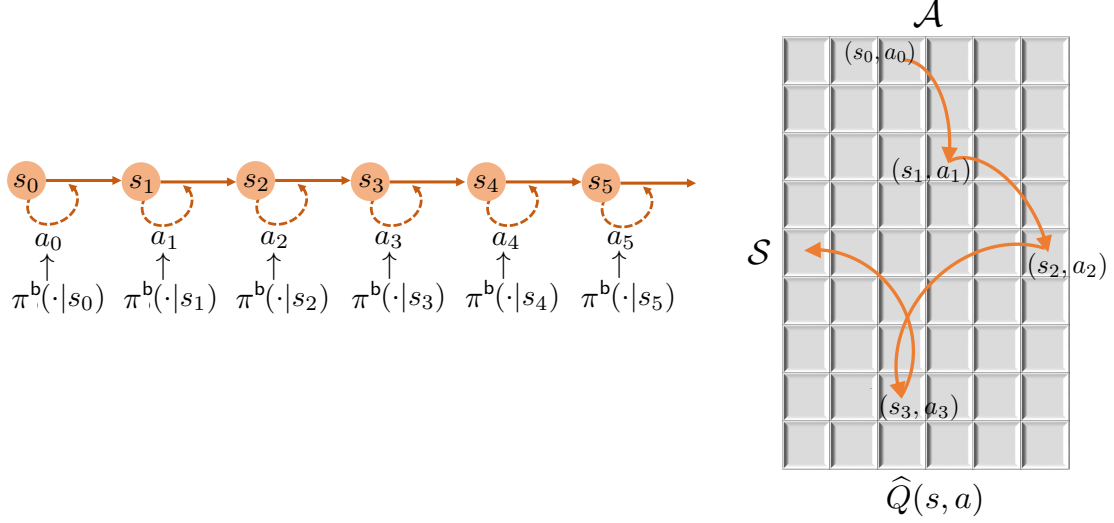


Figure 4.1: In asynchronous Q-learning, one observes a single Markovian trajectory $\{(s_t, a_t)\}_{t \geq 0}$ induced by a fixed behavior policy π^b , and updates the Q-estimate sequentially according to it.

updated simultaneously at each iteration. Another important practical variant is *asynchronous* Q-learning (Tsitsiklis, 1994; Jaakkola *et al.*, 1993), in which only a single state–action pair is updated per iteration (typically following a Markovian trajectory, to be made precise shortly). Although the sampling mechanism in asynchronous Q-learning often deviates substantially from that of a generative model, its core algorithmic ideas and theoretical guarantees remain closely aligned with those of the synchronous counterpart. For this reason, we introduce this variant in this chapter as a natural extension of synchronous Q-learning.

4.4.1 Learning from Markovian samples

In contrast to the generative model setting, asynchronous Q-learning is most commonly applied in settings where samples are collected along a Markovian trajectory. To be precise, suppose that we have access to a sample trajectory $\{(s_t, a_t, r_t)\}_{t=0}^\infty$, generated by the MDP of interest $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$ under a stationary *behavior policy* π^b — so called because this policy dictates the actual “behavior” in generating this sample trajectory. Starting from an initial state s_0 , this trajectory evolves sequentially as

$$a_t \sim \pi^b(\cdot | s_t), \quad r_t = r(s_t, a_t), \quad s_{t+1} \sim P(\cdot | s_t, a_t), \quad t \geq 0; \quad (4.88)$$

see Figure 4.1 for an illustration. When π^b is stationary, the process $\{(s_t, a_t)\}_{t=0}^\infty$ can be viewed as a sample path of a time-homogeneous Markov chain. Notably, the behavior policy π^b may differ substantially from the target optimal policy π^* . Throughout this section, we assume that asynchronous Q-learning operates on this single Markovian trajectory, with sample complexity naturally measured by the length of the observed trajectory required to achieve a prescribed level of accuracy.

4.4.2 Algorithm: asynchronous Q-learning

We now introduce the asynchronous Q-learning algorithm, in which only the state–action pair visited along the Markovian trajectory is updated at each iteration. Specifically, given the Markovian trajectory $\{s_t, a_t, r_t\}_{t=0}^\infty$ generated by the behavior policy π^b , the asynchronous Q-learning algorithm maintains a Q-function estimate $Q_t : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ at each time $t \geq 0$ and updates only one entry at each iterate as follows:

$$\begin{aligned} Q_t(s_{t-1}, a_{t-1}) &= (1 - \eta_t)Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t \mathcal{T}_t(Q_{t-1})(s_{t-1}, a_{t-1}), \\ Q_t(s, a) &= Q_{t-1}(s, a), \quad \forall (s, a) \neq (s_{t-1}, a_{t-1}), \end{aligned} \quad (4.89)$$

whereas $\eta_t \in (0, 1)$ represents the learning rate, and $\mathcal{T}_t$ denotes the empirical Bellman optimality operator w.r.t. the t -th sample transition (s_{t-1}, a_{t-1}, s_t) of the Markovian trajectory, defined as

$$\mathcal{T}_t(Q)(s_{t-1}, a_{t-1}) := r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'). \quad (4.90)$$

It is self-evident that the term $\mathcal{T}_t(Q)(s_{t-1}, a_{t-1})$ at iteration t can be viewed as a stochastic, unbiased estimate of the deterministic quantity $\mathcal{T}(Q_{t-1})(s_{t-1}, a_{t-1})$, where $\mathcal{T}(\cdot)$ is the Bellman optimality operator defined in (2.13). Consequently, the update of $Q_t(s_{t-1}, a_{t-1})$ follows the paradigm of stochastic approximation, and asynchronous Q-learning can also be interpreted as a stochastic approximation scheme for computing the fixed point of the Bellman optimality operator. The corresponding value function estimate $V_t : \mathcal{S} \rightarrow \mathbb{R}$ at time t is again given by

$$\forall s \in \mathcal{S} : \quad V_t(s) := \max_{a \in \mathcal{A}} Q_t(s, a).$$

It is worth emphasizing that, at each time t , only a single entry — the one corresponding to the sampled state–action pair (s_{t-1}, a_{t-1}) — is updated, while all other entries remain unchanged; see Figure 4.1 for an illustration. At a high level, this procedure resembles coordinate descent in optimization, with coordinates selected according to a Markov chain (Sun *et al.*, 2020). In addition, although the initial estimate Q_0 can be chosen arbitrarily, we often set $Q_0(s, a) = 0$ for all (s, a) unless otherwise noted. The complete asynchronous Q-learning algorithm is summarized in Algorithm 4.3.

4.4.3 Assumptions and key quantities

In contrast to the generative model setting — where one can obtain the same number of samples across all state–action pairs — it is generally unclear whether a Markovian sample trajectory will visit every state–action pair sufficiently often. Since insufficient visitation may result in a lack of critical information, we impose a standard assumption and introduce several key quantities, commonly used in the analysis of asynchronous Q-learning, to ensure adequate coverage of the state–action space by the sample trajectory.

Recall that when the behavior policy π^b is stationary, the sequence $\{(s_t, a_t)\}_{t=0}^\infty$ forms a time-homogeneous Markov chain; in what follows, we denote by μ_{π^b} its stationary distribution. The following assumption concerns the uniform ergodicity of this chain.

Assumption 4.1. The Markov chain $\{(s_t, a_t)\}_{t=0}^\infty$ induced by the behavior policy π^b (see (4.88)) is assumed to be uniformly ergodic.

Algorithm 4.3: Asynchronous Q-learning

```

1 input: discount factor  $\gamma$ , reward function  $r$ , learning rates  $\{\eta_t\}$ , number of iterations  $T$ ,
   initialization  $Q_0$ .
2 for  $t = 1, 2, \dots, T$  do
3   draw action  $a_{t-1} \sim \pi_b(s_{t-1})$ , observe reward  $r(s_{t-1}, a_{t-1})$ , and draw next state
      $s_t \sim P(\cdot | s_{t-1}, a_{t-1})$ .
   /* stochastic approximation. */
4   update  $Q_t$  according to
       
$$Q_t(s_{t-1}, a_{t-1}) = (1 - \eta_t)Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t \mathcal{T}_t(Q_{t-1})(s_{t-1}, a_{t-1}),$$

       
$$Q_t(s, a) = Q_{t-1}(s, a), \quad \forall (s, a) \neq (s_{t-1}, a_{t-1}),$$

   where the  $(s_{t-1}, a_{t-1})$ -th entry of  $\mathcal{T}_t(Q)$  is given by
       
$$\mathcal{T}_t(Q)(s_{t-1}, a_{t-1}) = r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a').$$


```

At a high level, uniform ergodicity ensures that, regardless of the initial state, the Markov chain eventually “forgets” its starting point and converges to the stationary distribution μ_{π^b} at a geometric rate. Consequently, aside from an initial burn-in period, the state–action visitation frequencies can be well approximated by the stationary distribution μ_{π^b} . See Paulin (2015, Section 1.2) for a formal definition of uniform ergodicity and its further implications.

With uniform ergodicity in place, two key quantities characterize, respectively, the information bottleneck in the stationary distribution and the rate of convergence toward it, both of which are critical in determining the performance of asynchronous Q-learning.

- The first key quantity is the *minimum state–action occupancy probability* of the sample trajectory, defined formally as

$$\mu_{\min} := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mu_{\pi^b}(s, a). \quad (4.91)$$

This metric captures the information bottleneck induced by the least-visited state-action pair along the trajectory. In particular, a smaller $\mu_{\min}$ indicates poorer coverage of the state-action space and, consequently, a higher chance that certain state-action pairs are insufficiently visited, which can adversely affect learning performance.

- The second key quantity is the *mixing time* associated with the sample trajectory, defined and denoted by

$$t_{\text{mix}} := \min \left\{ t \mid \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \text{TV}(P^t(\cdot | s, a), \mu_{\pi^b}) \leq \frac{1}{4} \right\}, \quad (4.92)$$

where we recall that $\text{TV}(\cdot, \cdot)$ indicates the total-variation distance between two probability measures. Here, $P^t(\cdot | s, a)$ represents the distribution of (s_t, a_t) when the sample trajectory is initialized at $(s_0, a_0) = (s, a)$. In words, the mixing time characterizes how quickly the Markov chain can “forget” its initial state, quantifying (the order of) the time required for the distribution of the trajectory to approach its stationary distribution.

4.4.4 Theoretical guarantees

With the above assumption and key quantities in place, we are now ready to present the state-of-the-art theoretical guarantees for asynchronous Q-learning, originally established by Li *et al.*, 2024a. For simplicity, the theorem below adopts a time-invariant learning rate schedule given by

$$\eta_t \equiv \eta = \frac{c_1 \log^3 T}{(1 - \gamma)T\mu_{\min}}, \quad 0 \leq t \leq T \quad (4.93a)$$

for some universal constants $0 < c_1 \leq 1$.

Theorem 4.10 (Li *et al.*, 2024a). Consider any $\delta \in (0, 1)$, $\varepsilon \in (0, 1]$, and $\gamma \in [1/2, 1)$. Suppose that Assumption 4.1 holds, and that the total number of iterations T exceeds

$$T \geq \frac{c_2 \log^2(\frac{SAT}{\delta})}{\mu_{\min}} \max \left\{ \frac{\log^3 T}{(1 - \gamma)^4 \varepsilon^2}, \frac{t_{\text{mix}}}{1 - \gamma} \right\} \quad (4.93b)$$

for some sufficiently large universal constant $c_2 > 0$. If the initialization obeys $0 \leq Q_0(s, a) \leq \frac{1}{1 - \gamma}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, then asynchronous Q-learning (cf. Algorithm 4.3) with learning rates (4.93a) achieves

$$\max_{(s, a) \in \mathcal{S} \times \mathcal{A}} |Q_T(s, a) - Q^*(s, a)| \leq \varepsilon$$

with probability at least $1 - \delta$.

Remark 4.5. As before, this high-probability result can be readily translated into a bound on the expected ℓ_∞ -based estimation error with essentially the same sample complexity. We omit the details here and refer the reader to Li *et al.*, 2024a for details.

Given that each iteration of asynchronous Q-learning uses a single sample transition, Theorem 4.10 demonstrates that, with high probability, the total sample size needed for asynchronous Q-learning to yield entrywise ε accuracy is

$$\tilde{O} \left(\frac{1}{\mu_{\min}(1 - \gamma)^4 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1 - \gamma)} \right), \quad (4.94)$$

provided that the learning rates are suitably chosen (see (4.93a)). We next provide several remarks on the two terms appearing in this bound.

- *Dependency on the minimum state-action occupancy.* The first term in (4.94) mirrors the sample complexity of synchronous Q-learning. More specifically, recall that synchronous Q-learning achieves a sample complexity of $\tilde{O}(\frac{SA}{(1 - \gamma)^4 \varepsilon^2})$ under a generative model (cf. (4.44)). In comparison, the key distinction here is that the total number of state-action pairs (i.e., SA) is replaced by $1/\mu_{\min}$ for asynchronous Q-learning, in order to account for the non-uniform visitation frequencies induced by the Markovian sampling process. Moreover, when the target accuracy ε is sufficiently small, this term becomes the dominant contribution to the sample complexity of asynchronous Q-learning. The sharpness of this term has been further confirmed by a matching algorithm-dependent lower bound established in Li *et al.* (2024a, Theorem 5).
- *Dependency on the mixing time.* The second term in (4.94) is nearly independent of the target accuracy (up to logarithmic factors) and can be interpreted as a burn-in time. Roughly speaking, it reflects the time required for asynchronous Q-learning to effectively emulate

its synchronous counterpart despite the temporal dependence inherent in Markovian data. Notably, the mixing time of the Markov chain seems to only influence the length of this burn-in phase and becomes largely immaterial once the trajectory is sufficiently close to its stationary distribution.

For the sake of brevity, we omit the full proof of Theorem 4.10, as a substantial portion of the argument closely parallels that of Theorem 4.5. Instead, we highlight a key observation that bridges the analysis of asynchronous and synchronous Q-learning. Specifically, over any time window of length $T_0 = \tilde{\Theta}(\frac{t_{\text{mix}}}{\mu_{\min}})$, classical concentration inequalities for Markov chains (e.g., Paulin (2015, Theorem 3.14)) guarantee that each state-action pair (s, a) is visited at least on the order of

$$T_0 \mu_{\pi^b}(s, a) \geq T_0 \mu_{\min} = \tilde{\Theta}(t_{\text{mix}})$$

times with high probability. Consequently, it can be anticipated that within each such window of length $\tilde{\Theta}(\frac{t_{\text{mix}}}{\mu_{\min}})$, the progress made by asynchronous Q-learning is at least comparable to — if not exceeding — that of $\tilde{\Theta}(t_{\text{mix}})$ iterations of synchronous Q-learning. This observation suggests a natural coupling between the two algorithms. When the total number of iterations T is large enough, the trajectory of asynchronous Q-learning can be partitioned into windows of length $\tilde{\Theta}(\frac{t_{\text{mix}}}{\mu_{\min}})$, each of which can be compared to $\tilde{\Theta}(t_{\text{mix}})$ iterations of synchronous Q-learning. Once this correspondence is established, the performance guarantees for asynchronous Q-learning follow by leveraging the analysis developed for the synchronous setting. In particular, since synchronous Q-learning requires $\tilde{O}(\frac{1}{(1-\gamma)^4 \varepsilon^2})$ iterations to reach ε -accuracy, this argument suggests that asynchronous Q-learning requires $\tilde{O}(\frac{1}{\mu_{\min}(1-\gamma)^4 \varepsilon^2})$ iterations to attain the same level of accuracy, provided that T is large enough. We refer the interested reader to Li *et al.* (2024a) for the complete proof.

Before concluding this section, we briefly mention several minor extensions of the above results. First, careful readers may observe that the learning rate schedule in (4.93a) requires prior knowledge of the stopping time T and, more importantly, the minimum state-action occupancy probability $\mu_{\min}$. Fortunately, this requirement can be eliminated by adopting a more data-driven, adaptive learning rate schedule without sacrificing performance; see Li *et al.* (2022b, Section III.D) for details. Moreover, it is worth emphasizing that “mixing” is not strictly necessary for the success of asynchronous Q-learning. The analysis can be readily extended to settings where the Markov chain underlying the sample trajectory does not mix but instead admits a finite cover time — namely, that within any time window of length exceeding t_{cover} , the event that all state-action pairs are visited at least once occurs with probability at least $1/2$. In this case, the analysis proceeds by partitioning the dynamics of asynchronous Q-learning into time windows of length $\tilde{\Theta}(t_{\text{cover}})$ (rather than those determined by the mixing time) and analyzing each window separately through comparisons with the synchronous counterpart. We refer the reader to Li *et al.* (2022b, Theorem 2) and the accompanying discussion therein for further details.

4.5 Extension: finite-horizon nonstationary MDPs

We now briefly discuss synchronous Q-learning for finite-horizon nonstationary MDPs. Both the algorithm and its analysis closely parallel those in the discounted infinite-horizon setting. We therefore omit the details here and refer interested readers to Li *et al.* (2021a) for a complete treatment.

In a nutshell, the algorithm maintains a sequence of Q-function estimates $Q_{t,h} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ for each iteration t and time step h . At each iteration t , the algorithm updates these estimates backward from $h = H$ down to 1 via

$$Q_{t,h} = (1 - \eta_t)Q_{t-1,h} + \eta_t \mathcal{T}_{t,h}(Q_{t,h+1}), \quad (4.95)$$

where $0 < \eta_t < 1$ denotes the learning rate for the t -th iteration, and $\mathcal{T}_{t,h}$ is the empirical Bellman operator constructed from samples drawn from the generative model at step h of this iteration. More concretely, $\mathcal{T}_{t,h}$ is defined as follows:

- for each tuple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, draw an independent next-state sample

$$s_{t,h}(s, a) \stackrel{\text{ind.}}{\sim} P_h(\cdot | s, a);$$

- compute $\mathcal{T}_{t,h}$ entrywise according to

$$\mathcal{T}_{t,h}(Q_{t,h+1})(s, a) := r_h(s, a) + \max_{a' \in \mathcal{A}} Q_{t,h+1}(s_{t,h}(s, a), a').$$

In addition, the value function estimate $V_{t,h} : \mathcal{S} \rightarrow \mathbb{R}$ associated with the t -th iterate and step h is defined as

$$\forall s \in \mathcal{S} : \quad V_{t,h}(s) := \max_{a \in \mathcal{A}} Q_{t,h}(s, a). \quad (4.96)$$

To ensure the desired convergence guarantees, we also assume that the learning rates satisfy

$$\frac{1}{1 + \frac{c_1 T}{H \log^2 T}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2 t}{H \log^2 T}}, \quad 0 \leq t \leq T \quad (4.97)$$

for sufficiently small universal constants $c_1 \geq c_2 > 0$.

The next theorem establishes the ℓ_∞ -based sample complexity of this synchronous Q-learning algorithm for finite-horizon MDPs.

Theorem 4.11 (Li *et al.* (2021a)). Consider any $\delta \in (0, 1)$ and $\varepsilon \in (0, 1]$. Suppose that the learning rates obey (4.97) and the total number of iterations obeys

$$T \geq \frac{c_3 H^4 (\log^3 T) (\log \frac{SAT}{\delta})}{\varepsilon^2} \quad (4.98)$$

for some sufficiently large universal constant $c_3 > 0$. If the initialization obeys $0 \leq Q_{0,h}(s, a) \leq H + 1 - h$ for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, then the aforementioned synchronous Q-learning algorithm achieves

$$\forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H] : \quad |Q_{T,h}(s, a) - Q_h^*(s, a)| \leq \varepsilon$$

with probability greater than $1 - \delta$.

Recalling that each iteration collects SAH independent samples, Theorem 4.11 indicates that synchronous Q-learning achieves a sample complexity of

$$\tilde{O}\left(\frac{SAH^5}{\varepsilon^2}\right).$$

for finite-horizon nonstationary MDPs. Notably, this sample complexity bound is a factor of H larger than the minimax-optimal complexity (3.87) attained by the model-based approach in Section 3.5, revealing a suboptimality gap for Q-learning analogous to that observed in the discounted infinite-horizon setting.

4.6 Appendix A: proof of auxiliary lemmas in Section 4.2

4.6.1 Proof of Lemma 4.2

We would like to apply induction to establish that, for any $0 \leq t \leq T$,

$$0 \leq V_t \leq \frac{1}{1-\gamma} 1_S. \quad (4.99)$$

Regarding the base case with $t = 0$, this property is automatically guaranteed by our assumption. Next, suppose that (4.99) holds true for the $(t-1)$ -th iteration. Recalling that $0 \leq \eta_t \leq 1$, $0 \leq r_\pi \leq 1_S$, $P_t \geq 0$ and $\|P_t\|_1 = 1$, one can invoke the update rule (4.8) to show that

$$V_t = (1 - \eta_t)V_{t-1} + \eta_t(r_\pi + \gamma P_t V_{t-1}) \geq 0$$

and

$$\begin{aligned} V_t &= (1 - \eta_t)V_{t-1} + \eta_t(r_\pi + \gamma P_t V_{t-1}) \\ &\leq (1 - \eta_t) \|V_{t-1}\|_\infty 1_S + \eta_t(\|r_\pi\|_\infty + \gamma \|P_t\|_1 \|V_{t-1}\|_\infty) 1_S \\ &\leq (1 - \eta_t) \frac{1}{1-\gamma} 1_S + \eta_t \left(1 + \frac{\gamma}{1-\gamma}\right) 1_S = \frac{1}{1-\gamma} 1_S. \end{aligned}$$

This establishes (4.99) for the t -th iteration, which completes the induction argument for proving (4.99). Moreover, taking together (4.99) and the fact that $0 \leq V^\pi \leq \frac{1}{1-\gamma} 1_S$ immediately reveals that

$$\|V_t - V^\pi\|_\infty \leq \frac{1}{1-\gamma}.$$

4.6.2 Proof of Lemma 4.3

Suppose that $t \geq T/(c_2 \log T)$. We divide into two cases.

- For any $0 \leq k \leq t/2$, it follows from the definition (4.9) of $\eta_k^{(t)}$ as well as the condition $0 < \eta_k < 1$ that

$$\begin{aligned} \eta_k^{(t)} &\leq \prod_{i=t/2+1}^t (1 - \eta_i(1 - \gamma)) \stackrel{(i)}{\leq} \left(1 - \frac{\log^2 T}{2c_1 T}\right)^{t/2} \stackrel{(ii)}{\leq} \left(1 - \frac{\log^2 T}{2c_1 T}\right)^{\frac{T}{2c_2 \log T}} \\ &= \left\{ \left(1 - \frac{\log^2 T}{2c_1 T}\right)^{\frac{2c_1 T}{\log^2 T}} \right\}^{\frac{\log T}{4c_1 c_2}} \stackrel{(iii)}{\leq} \frac{1}{T^2}. \end{aligned} \quad (4.100a)$$

Here, (i) follows from a direct consequence of (4.5), i.e.,

$$(1 - \gamma)\eta_i \geq \frac{1 - \gamma}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \geq \frac{1 - \gamma}{\frac{2c_1(1-\gamma)T}{\log^2 T}} = \frac{\log^2 T}{2c_1 T}$$

as long as $T \geq \frac{\log^2 T}{c_1(1-\gamma)}$ (as ensured by (4.6)); (ii) is valid provided that $t \geq T/(c_2 \log T)$; and (iii) holds as long as $c_1 c_2 \leq 1/8$ and $T \geq c_1^{-1} \log^2 T$ (again ensured by (4.6)).

- When it comes to the case with $k > t/2$, we can apply the definition (4.9) and the condition (4.5) to reach

$$\eta_k^{(t)} \leq \eta_k \leq \frac{1}{c_2(1 - \gamma)k / \log^2 T} < \frac{2 \log^3 T}{(1 - \gamma)T}, \quad (4.100b)$$

where the last inequality holds since $k > t/2 \geq T/(2c_2 \log T)$.

4.6.3 Proof of Lemma 4.4

Define

$$a := \sqrt{2}q \quad \text{and} \quad b_k := (2q^2)^{1-1/2^k} u_0^{1/2^k}.$$

It is straightforward to check that

$$b_k = q\sqrt{2b_{k-1}}, \quad k \geq 2. \quad (4.101)$$

In what follows, we intend to prove $u_k \leq a + b_k$ ($k \geq 1$) by induction.

- For the base case with $k = 1$, we have

$$u_1 \leq q\sqrt{1 + u_0} \leq q(1 + \sqrt{u_0}) < \sqrt{2}q + q\sqrt{2u_0} = a + b_1,$$

where we use the relation that $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$.

- Now assume that $u_{k-1} \leq a + b_{k-1}$. Then it holds that

$$u_k \leq q\sqrt{1 + u_{k-1}} \leq q\sqrt{1 + a + b_{k-1}}.$$

Since $q \leq 1/\sqrt{2}$, we have $a = \sqrt{2}q \leq 1$, and hence

$$u_k \leq q\sqrt{2 + b_{k-1}} \leq \sqrt{2}q + q\sqrt{b_{k-1}} \leq a + b_k,$$

where in the last step we used (4.101).

This completes the induction step and hence establishes Lemma 4.4.

4.7 Appendix B: proof of auxiliary lemmas in Section 4.3

4.7.1 Proof of Lemma 4.7

The proof is similar to that of Lemma 4.2. Specifically, we would like to show by induction that, for every $t \geq 0$,

$$0 \leq Q_t \leq \frac{1}{1-\gamma} 1_{SA} \quad \text{and} \quad 0 \leq V_t \leq \frac{1}{1-\gamma} 1_S. \quad (4.102)$$

We focus on proving the bounds for Q_t ; once these are established, the bounds for V_t follow immediately from the relation $V_t(s) = \max_a Q_t(s, a)$.

The base case $t = 0$ follows directly from the assumption of the lemma. Now suppose that (4.102) holds for the $(t-1)$ -th iterate. Since $0 \leq \eta_t \leq 1$, $0 \leq r \leq 1_{SA}$, $P_t \geq 0$, and each row of P_t sums to one, the update rule (4.46) yields

$$Q_t = (1 - \eta_t)Q_{t-1} + \eta_t(r + \gamma P_t V_{t-1}) \geq 0$$

and

$$\begin{aligned} Q_t &= (1 - \eta_t)Q_{t-1} + \eta_t(r + \gamma P_t V_{t-1}) \\ &\leq (1 - \eta_t)\|Q_{t-1}\|_\infty 1_{SA} + \eta_t\left(\|r\|_\infty + \gamma\|P_t\|_1\|V_{t-1}\|_\infty\right) 1_{SA} \\ &\leq (1 - \eta_t)\frac{1}{1-\gamma} 1_{SA} + \eta_t\left(1 + \frac{\gamma}{1-\gamma}\right) 1_{SA} = \frac{1}{1-\gamma} 1_{SA}. \end{aligned}$$

This verifies (4.102) for the t -th iterate and establishes (4.50a).

The other claim (4.50b) follows immediately by combining (4.50a) with the fact that $0 \leq Q^* \leq \frac{1}{1-\gamma} 1_{SA}$.

4.7.2 Proof of Lemma 4.8

From assumption (4.42) and the lower bound (4.43) on T , we have

$$\frac{\log^3 T}{2c_1(1-\gamma)T} \leq \eta_t \leq \frac{\log^3 T}{c_2(1-\gamma)t}, \quad (4.103)$$

provided that $c_3 \geq 1/c_1$. We look at the following two cases separately.

- For any $0 \leq i \leq (1-\beta)t$, applying the lower bound in (4.103) yields

$$\begin{aligned} \eta_i^{(t)} &\leq \left(1 - \frac{\log^3 T}{2c_1(1-\gamma)T}\right)^{\beta t} \leq \left(1 - \frac{\log^3 T}{2c_1(1-\gamma)T}\right)^{\frac{c_4(1-\gamma)T}{c_2 \log^2 T}} \\ &= \left(\left(1 - \frac{\log^3 T}{2c_1(1-\gamma)T}\right)^{\frac{2c_1(1-\gamma)T}{\log^3 T}}\right)^{\frac{c_4 \log T}{2c_1 c_2}} < \frac{1}{2T^2}, \end{aligned} \quad (4.104)$$

provided that $c_1 c_2 \leq c_4/8$, where the second inequality arises from the choice (4.52) of β and the condition $t \geq T/(c_2 \log T)$.

- For the remaining indices $i > (1-\beta)t \geq t/2$ (recall the definition of β in (4.52)), we invoke the upper bound in (4.103) to obtain

$$\eta_i^{(t)} \leq \eta_i \leq \frac{\log^3 T}{c_2(1-\gamma)i} < \frac{2 \log^3 T}{c_2(1-\gamma)t} \leq \frac{2 \log^4 T}{(1-\gamma)T}, \quad (4.105)$$

with the proviso that $t \geq T/(c_2 \log T)$.

Moreover, a telescoping argument reveals that

$$\begin{aligned} \sum_{i=0}^t \eta_i^{(t)} &= \prod_{j=1}^t (1 - \eta_j) + \eta_1 \prod_{j=2}^t (1 - \eta_j) + \eta_2 \prod_{j=3}^t (1 - \eta_j) + \cdots + \eta_{t-1} (1 - \eta_t) + \eta_t \\ &= \prod_{j=2}^t (1 - \eta_j) + \eta_2 \prod_{j=3}^t (1 - \eta_j) + \cdots + \eta_{t-1} (1 - \eta_t) + \eta_t \\ &= \cdots = (1 - \eta_t) + \eta_t = 1. \end{aligned}$$

In fact, this argument readily shows that, for any $1 \leq \tau \leq t$,

$$\sum_{i=\tau}^t \eta_i^{(t)} = 1 - \prod_{j=\tau}^t (1 - \eta_j).$$

4.7.3 Proof of Lemma 4.9

This lemma, inspired by Azar *et al.* (2013, Lemma 8) yet substantially more intricate, exploits the variance structure and plays a key role in removing one factor of $(1-\gamma)^{-1}$ in the sample complexity analysis.

We first attempt to replace $\max_{i \in [t/2, t]} \text{Var}_P(V_i)$ with a simpler variance term. Towards this, we make note of the following inequality:

$$\max_{i \in [t/2, t]} \text{Var}_P(V_i) - \text{Var}_P(V^*) \leq \frac{4}{1-\gamma} \max_{i \in [t/2, t]} \|\Delta_i\|_\infty 1_{SA}, \quad (4.106)$$

which is an immediate consequence of Lemmas 2.2 and 4.7. This allows one to upper-bound

$$\begin{aligned} 0 &\leq \sum_{h=0}^{H-1} \gamma^h \left(\prod_{k=1}^h P^{\pi_{i_k}} \right) \max_{i \in [t/2, t)} \text{Var}_P(V_i) \\ &\leq \sum_{h=0}^{H-1} \left(\prod_{k=1}^h \gamma P^{\pi_{i_k}} \right) \text{Var}_P(V^*) + \frac{4 \max_{i \in [t/2, t)} \|\Delta_i\|_\infty}{(1-\gamma)^2} 1_{SA}, \end{aligned} \quad (4.107)$$

where the last inequality holds since $P^\pi 1_{SA} = 1_{SA}$ and $\sum_{h=0}^{H-1} \gamma^h \leq \frac{1}{1-\gamma}$.

Consequently, it suffices to control the first term on the right-hand side of (4.107), which requires leveraging the interplay between the transition matrices and the variance term. To provide some intuition, recall from (2.22) that the variance term satisfies

$$\begin{aligned} \text{Var}_P(V^*) &= P(V^* \circ V^*) - (PV^*) \circ (PV^*) \\ &= P^{\pi^*}(Q^* \circ Q^*) - \frac{1}{\gamma^2}(Q^* - r) \circ (Q^* - r) \\ &= P^{\pi^*}(Q^* \circ Q^*) - \frac{1}{\gamma^2}Q^* \circ Q^* - \frac{1}{\gamma^2}r \circ r + \frac{2}{\gamma^2}Q^* \circ r \\ &\leq P^{\pi^*}(Q^* \circ Q^*) - \frac{1}{\gamma}Q^* \circ Q^* + \frac{2}{\gamma^2}Q^* \circ r \\ &\leq \frac{1}{\gamma}(\gamma P^{\pi^*}(Q^* \circ Q^*) - Q^* \circ Q^*) + \frac{2}{\gamma^2(1-\gamma)} 1_{SA}, \end{aligned} \quad (4.108)$$

where the second identity applies the Bellman equation $Q^* = r + \gamma PV^*$ (see (2.25)), and the last line uses $\|Q^*\|_\infty \leq \frac{1}{1-\gamma}$ and $\|r\|_\infty \leq 1$. Informally, when policy $\pi_{i_{h+1}}$ is near-optimal, one may approximate $\text{Var}_P(V^*)$ by replacing π^* with $\pi_{i_{h+1}}$ to yield

$$\text{Var}_P(V^*) \lesssim \frac{1}{\gamma} \left(\gamma P^{\pi_{i_{h+1}}}(Q^* \circ Q^*) - Q^* \circ Q^* \right).$$

The key advantage of this approximation is that it converts the sum of interest into a telescoping structure:

$$\sum_h \left(\prod_{k=1}^h \gamma P^{\pi_{i_k}} \right) \text{Var}_P(V^*) \lesssim \sum_h \frac{1}{\gamma} \left\{ \left(\prod_{k=1}^{h+1} \gamma P^{\pi_{i_k}} \right) (Q^* \circ Q^*) - \left(\prod_{k=1}^h \gamma P^{\pi_{i_k}} \right) (Q^* \circ Q^*) \right\}.$$

In this form, the intermediate terms cancel, considerably simplifying the aggregate variance bound.

We now make this argument rigorous. Continue the bound in (4.108) to derive

$$\text{Var}_P(V^*) \leq \frac{1}{\gamma}(\gamma \Theta_1 + \Theta_2) + \frac{2}{\gamma^2(1-\gamma)} 1_{SA}, \quad (4.109)$$

where we define

$$\begin{aligned} \Theta_1 &:= P^{\pi^*}(Q^* \circ Q^*) - P^{\pi_{i_{h+1}}}(Q^* \circ Q^*), \\ \Theta_2 &:= \gamma P^{\pi_{i_{h+1}}}(Q^* \circ Q^*) - (Q^*) \circ (Q^*). \end{aligned}$$

Regarding the first term Θ_1 , we observe that

$$\begin{aligned} \|\Theta_1\|_\infty &= \|P\Pi^{\pi_{i_{h+1}}}(Q^* \circ Q^*) - P\Pi^{\pi^*}(Q^* \circ Q^*)\|_\infty \\ &\stackrel{(i)}{\leq} \|\Pi^{\pi_{i_{h+1}}}(Q^* \circ Q^*) - \Pi^{\pi^*}(Q^* \circ Q^*)\|_\infty \end{aligned}$$

$$\begin{aligned}
&= \|(\Pi^{\pi_{i_{h+1}}} Q^* - \Pi^{\pi^*} Q^*) \circ (\Pi^{\pi_{i_{h+1}}} Q^* + \Pi^{\pi^*} Q^*)\|_\infty \\
&\stackrel{(ii)}{\leq} \frac{2}{1-\gamma} \|\Pi^{\pi_{i_{h+1}}} Q^* - \Pi^{\pi^*} Q^*\|_\infty = \frac{2}{1-\gamma} \|\Pi^{\pi_{i_{h+1}}} Q^* - V^*\|_\infty \\
&\leq \frac{2}{1-\gamma} (\|\Pi^{\pi_{i_{h+1}}} Q^* - \Pi^{\pi_{i_{h+1}}} Q_{i_{h+1}}\|_\infty + \|\Pi^{\pi_{i_{h+1}}} Q_{i_{h+1}} - V^*\|_\infty) \\
&\stackrel{(iii)}{\leq} \frac{2}{1-\gamma} (\|Q^* - Q_{i_{h+1}}\|_\infty + \|V_{i_{h+1}} - V^*\|_\infty) \\
&\stackrel{(iv)}{\leq} \frac{4}{1-\gamma} \max_{i \in [t/2, t)} \|\Delta_i\|_\infty,
\end{aligned}$$

where (i) is valid since $\|Pz\|_\infty \leq \|P\|_1 \|z\|_\infty = \|z\|_\infty$, (ii) follows from the bound $\|Q^*\|_\infty \leq 1/(1-\gamma)$, (iii) results from the identity $V_{i_{h+1}} = \Pi^{\pi_{i_{h+1}}} Q_{i_{h+1}}$, and (iv) follows since, for every $s \in \mathcal{S}$,

$$\begin{aligned}
|V_t(s) - V^*(s)| &= \left| \max_a Q_t(s, a) - \max_a Q^*(s, a) \right| \\
&\leq \max_a |Q_t(s, a) - Q^*(s, a)|.
\end{aligned}$$

Substitution into (4.109) then yields

$$\text{Var}_P(V^*) \leq \frac{1}{\gamma} (\gamma P^{\pi_{i_{h+1}}} (Q^* \circ Q^*) - Q^* \circ Q^*) + \left(\frac{2}{\gamma^2(1-\gamma)} + \frac{4}{1-\gamma} \max_{i \in [t/2, t)} \|\Delta_i\|_\infty \right) 1_{SA}. \quad (4.110)$$

Equipped with (4.110), we can bound the summation on the right-hand side of (4.107) as

$$\sum_{h=0}^{H-1} \left(\prod_{k=1}^h \gamma P^{\pi_{i_k}} \right) \text{Var}_P(V^*) \leq \Gamma_1 + \Gamma_2, \quad (4.111)$$

where we define

$$\begin{aligned}
\Gamma_1 &:= \frac{1}{\gamma} \sum_{h=0}^{H-1} \left(\prod_{k=1}^h \gamma P^{\pi_{i_k}} \right) (\gamma P^{\pi_{i_{h+1}}} (Q^* \circ Q^*) - Q^* \circ Q^*), \\
\Gamma_2 &:= \left(\frac{2}{\gamma^2(1-\gamma)} + \frac{4}{1-\gamma} \max_{i \in [t/2, t)} \|\Delta_i\|_\infty \right) \sum_{h=0}^{H-1} \gamma^h \left(\prod_{k=1}^h P^{\pi_{i_k}} \right) 1_{SA}.
\end{aligned}$$

We bound these two terms separately.

- By utilizing the telescoping structure of Γ_1 , we can deduce that

$$\begin{aligned}
\Gamma_1 &= \frac{1}{\gamma} \left(\sum_{h=0}^{H-1} \prod_{k=1}^{h+1} \gamma P^{\pi_{i_k}} - \sum_{h=0}^{H-1} \prod_{k=1}^h \gamma P^{\pi_{i_k}} \right) (Q^* \circ Q^*) \\
&= \frac{1}{\gamma} \left(\prod_{k=1}^H \gamma P^{\pi_{i_k}} - I \right) (Q^* \circ Q^*) \\
&\leq \frac{1}{\gamma} \left\| \prod_{k=1}^H \gamma P^{\pi_{i_k}} - I \right\|_1 \|Q^* \circ Q^*\|_\infty 1_{SA} \\
&\leq \frac{2}{\gamma} \|Q^*\|_\infty^2 1_{SA} \leq \frac{2}{\gamma(1-\gamma)^2} 1_{SA},
\end{aligned}$$

where the penultimate relation holds since each row of $\prod_{k=1}^h P^{\pi_{i_k}}$ has unit $\|\cdot\|_1$ norm for any h .

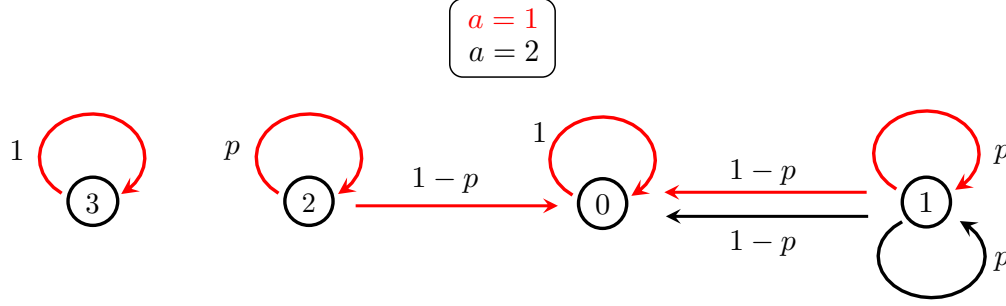


Figure 4.2: An illustration of the hard MDP in the proof of the lower bound for Q-learning.

- The fact that $\prod_{k=1}^h P^{\pi_{i_k}}$ is a probability transition matrix gives

$$\begin{aligned} \Gamma_2 &= \left(\frac{2}{\gamma^2(1-\gamma)} + \frac{4}{1-\gamma} \max_{i \in [t/2, t]} \|\Delta_i\|_\infty \right) \sum_{h=0}^{H-1} \gamma^h 1_{SA} \\ &\leq \left(\frac{2}{\gamma^2(1-\gamma)^2} + \frac{4}{(1-\gamma)^2} \max_{i \in [t/2, t]} \|\Delta_i\|_\infty \right) 1_{SA}. \end{aligned}$$

Combining the preceding bounds on Γ_1 and Γ_2 with (4.111) yields

$$\sum_{h=0}^{H-1} \prod_{k=1}^h \gamma P^{\pi_{i_k}} \text{Var}_P(V^*) \leq \frac{1}{(1-\gamma)^2} \left(\frac{4}{\gamma^2} + 4 \max_{i \in [t/2, t]} \|\Delta_i\|_\infty \right) 1_{SA}.$$

This, taken together with (4.107), completes the proof.

4.8 Appendix C: proof of the lower bound in Theorem 4.6

We now present the proof of the algorithm-dependent lower bound in Theorem 4.6 for Q-learning. For brevity, we focus on the case of time-invariant learning rates, i.e.,

$$\eta_t \equiv \eta \in (0, 1), \quad t \geq 0$$

throughout the execution of the algorithm. The proof for the case of rescaled linear learning rates is similar and can be found in Li *et al.* (2024a). Without loss of generality, we assume

$$\log T \leq \frac{1}{1-\gamma} \tag{4.112}$$

throughout the proof; otherwise the lower bound in Theorem 4.6 becomes looser than the minimax lower bound $\frac{1}{(1-\gamma)^{3T}}$ established in Azar *et al.* (2013) (see Theorem 3.7).

4.8.1 Step 1: construction of a hard MDP

We begin by constructing a hard MDP instance $\mathcal{M}_{\text{hard}}$ with 4 states and 2 actions as follows; see Figure 4.2 for an illustration. The subsequent analysis will focus on tracking the Q-learning dynamics for this specific construction.

- *State space.* The state space of this MDP contains 4 states, namely,

$$\mathcal{S} = \{0, 1, 2, 3\}.$$

- *Action space.* To streamline the presentation, we allow the action space to vary across states. Let $\mathcal{A}_s$ represent the action space associated with each state $s \in \mathcal{S}$, which we specify as follows:

$$\mathcal{A}_0 = \{1\}, \quad \mathcal{A}_1 = \{1, 2\}, \quad \mathcal{A}_2 = \{1\}, \quad \mathcal{A}_3 = \{1\}.$$

In other words, only state 1 admits more than one action, while all other states are associated with a single action.

- *Probability transition kernel and reward function.* The probability transition kernel and reward function of $\mathcal{M}_{\text{hard}}$ are chosen as

$$\begin{array}{llll} \text{from state 0 :} & P(0|0, 1) = 1, & & r(0, 1) = 0, \\ \text{from state 1 :} & P(1|1, 1) = p, & P(0|1, 1) = 1 - p, & r(1, 1) = 1, \\ & P(1|1, 2) = p, & P(0|1, 2) = 1 - p, & r(1, 2) = 1, \\ \text{from state 2 :} & P(2|2, 1) = p, & P(0|2, 1) = 1 - p, & r(2, 1) = 1, \\ \text{from state 3 :} & P(3|3, 1) = 1, & & r(3, 1) = 1, \end{array}$$

for some parameter $p \in (0, 1)$. States 0 and 3 are absorbing, while from states 1 and 2 the process either remains in the same state or transitions to state 0. Moreover, state 0 yields zero immediate reward; once it is entered, no further rewards can be obtained.

In fact, the MDP construction bears some similarity with that in Section 3.4.1 for the minimax lower bounds, with the key distinction being that the action space is now augmented to include at least two actions in order to capture the over-estimation issue of Q-learning. Motivated by the construction in Section 3.4.1 (see (3.67)), we choose the parameter p as

$$p = \frac{4\gamma - 1}{3\gamma}. \quad (4.113)$$

A careful reader may also notice the inclusion of state 3, which is isolated and absorbing. As we shall see later, this state is introduced to simplify a small part of the analysis.

To facilitate the subsequent convergence analysis of Q-learning on this MDP, we first characterize its optimal value function and Q-function. The proof is deferred to the end of this section.

Lemma 4.12. The optimal value function and Q-function of the MDP $\mathcal{M}_{\text{hard}}$ constructed above satisfy

$$\begin{aligned} V^*(0) &= Q^*(0, 1) = 0; \\ V^*(1) &= Q^*(1, 1) = Q^*(1, 2) = V^*(2) = Q^*(2, 1) = \frac{1}{1 - \gamma p} = \frac{3}{4(1 - \gamma)}; \\ V^*(3) &= Q^*(3, 1) = \frac{1}{1 - \gamma}. \end{aligned}$$

Proof of Lemma 4.12. Since state 0 is absorbing with zero reward — so no further rewards are obtained once it is entered — we have

$$V^*(0) = Q^*(0, 1) = 0.$$

We now turn to states 1 and 2. Since actions 1 and 2 in state 1 are equivalent in both rewards and transition probabilities, it is clear that

$$Q^*(1, 1) = Q^*(1, 2) = V^*(1). \quad (4.115)$$

Also, it follows from the Bellman optimality equation (2.14) that

$$Q^*(1, 1) = r(1, 1) + \gamma P(0 | 1, 1) V^*(0) + \gamma P(1 | 1, 1) V^*(1).$$

Combining this with (4.115) and rearranging terms, we reach

$$V^*(1) = \frac{r(1, 1) + \gamma P(0 | 1, 1) V^*(0)}{1 - \gamma P(1 | 1, 1)} = \frac{1}{1 - \gamma p} = \frac{3}{4(1 - \gamma)}.$$

where the second identity follows since $V^*(0) = 0$, and the third identity relies on the choice (4.113) of p . The derivations for $V^*(2)$ and $Q^*(2, 1)$ are omitted as they follow from identical arguments.

When it comes to state 3, we invoke Bellman's equation to obtain

$$V^*(3) = Q^*(3, 1) = r(3, 1) + \gamma P(3 | 3, 1) V^*(3) = 1 + \gamma V^*(3).$$

A little algebra then leads to $V^*(3) = \frac{1}{1-\gamma}$. □

4.8.2 Step 2: preliminary calculation of Q-learning iterates

We now embark on the analysis of Q-learning for the MDP instance constructed in Step 1. As before, we let P_t encode the empirical transitions such that for any triple (s, a, s') ,

$$P_t(s' | s, a) := \begin{cases} 1, & \text{if } s_t(s, a) = s', \\ 0, & \text{otherwise,} \end{cases} \quad (4.116)$$

where $s_t(s, a)$ denotes the sample collected during the t -th iteration w.r.t. the (s, a) pair. Given that state 2 has a singleton action space, we simplify the notation by writing

$$P_t(s' | 2) := P_t(s' | 2, 1).$$

In what follows, we first introduce some notation and basic properties concerning the learning rates, and then slightly expand the Q-learning updates for each state.

Preliminary facts about learning rates. Let us introduce the following notation about learning rates which will be convenient for analyzing the Q-learning iterations:

$$\eta_k^{(t)} := \begin{cases} \eta(1 - \eta(1 - \gamma p))^{t-k}, & \text{if } 1 \leq k \leq t, \\ (1 - \eta(1 - \gamma p))^t, & \text{if } k = 0. \end{cases} \quad (4.117)$$

As can be straightforwardly verified analogous to Lemma 4.8 (which we omit for brevity),

$$\eta_0^{(t)} + (1 - \gamma p) \sum_{k=1}^t \eta_k^{(t)} = 1 \quad (4.118a)$$

$$(1 - \eta(1 - \gamma p))^{t-\tau} + (1 - \gamma p) \sum_{k=\tau+1}^t \eta_k^{(t)} = 1 \quad (4.118b)$$

for any given integer $0 \leq \tau < t$.

State 0. To begin, since $Q_0 = V_0 = 0$ and state 0 is absorbing, the Q-learning update rule implies that, for all $t \geq 1$,

$$\begin{aligned} V_t(0) &= Q_t(0, 1) = (1 - \eta)Q_{t-1}(0, 1) + \eta(r(0, 1) + \gamma V_{t-1}(0)) \\ &= (1 - \eta(1 - \gamma))Q_{t-1}(0, 1) = \dots \\ &= (1 - \eta(1 - \gamma))^t Q_0(0, 1) = 0. \end{aligned} \quad (4.119)$$

State 1. For state 1, the Q-learning iterates satisfy

$$Q_t(1, 1) = (1 - \eta)Q_{t-1}(1, 1) + \eta\{1 + \gamma P_t(1 | 1, 1)V_{t-1}(1)\}, \quad (4.120a)$$

$$Q_t(1, 2) = (1 - \eta)Q_{t-1}(1, 2) + \eta\{1 + \gamma P_t(1 | 1, 2)V_{t-1}(1)\}. \quad (4.120b)$$

State 2. The Q-learning update combined with (4.119) gives

$$\begin{aligned} V_t(2) &= Q_t(2, 1) = (1 - \eta)Q_{t-1}(2, 1) + \eta\{r(2, 1) + \gamma P_t(2 | 2)V_{t-1}(2) + \gamma P_t(0 | 2)V_{t-1}(0)\} \\ &= (1 - \eta)V_{t-1}(2) + \eta\{1 + \gamma P_t(2 | 2)V_{t-1}(2)\} \\ &= (1 - \eta(1 - \gamma p))V_{t-1}(2) + \eta\{1 + \gamma(P_t(2 | 2) - p)V_{t-1}(2)\}. \end{aligned} \quad (4.121)$$

Therefore, for any t and $\tau < t$, applying (4.121) recursively leads to

$$V_t(2) = (1 - \eta(1 - \gamma p))^{t-\tau} V_\tau(2) + \sum_{k=\tau+1}^t \eta_k^{(t)} + \sum_{k=\tau+1}^t \eta_k^{(t)} \gamma (P_k(2 | 2) - p) V_{k-1}(2), \quad (4.122)$$

where $\eta_k^{(t)}$ has been defined in (4.117). Recalling from (4.118b) that

$$\sum_{k=\tau+1}^t \eta_k^{(t)} = \frac{1 - (1 - \eta(1 - \gamma p))^{t-\tau}}{1 - \gamma p},$$

we can rearrange terms in (4.122) to reach

$$V_t(2) = \frac{1}{1 - \gamma p} - (1 - \eta(1 - \gamma p))^{t-\tau} \left[\frac{1}{1 - \gamma p} - V_\tau(2) \right] + \sum_{k=\tau+1}^t \eta_k^{(t)} \gamma (P_k(2 | 2) - p) V_{k-1}(2). \quad (4.123)$$

In particular, in the special case where $\tau = 0$ (so that $V_\tau(2) = V_0(2) = 0$), this identity simplifies to

$$V_t(2) = \frac{1 - \eta_0^{(t)}}{1 - \gamma p} + \sum_{k=1}^t \eta_k^{(t)} \gamma (P_k(2 | 2) - p) V_{k-1}(2), \quad (4.124)$$

where we recall the definition of $\eta_0^{(t)}$ in (4.117).

Moreover, we know that $\mathbb{E}[P_k(2 | 2)] = p$ by construction, which together with the identity (4.124) and Lemma 4.12 leads to

$$\mathbb{E}[V_t(2)] = \frac{1 - \eta_0^{(t)}}{1 - \gamma p} = \frac{1 - (1 - \frac{4\eta}{3}(1 - \gamma))^t}{\frac{4}{3}(1 - \gamma)}; \quad (4.125a)$$

$$V^*(2) - \mathbb{E}[V_t(2)] = \frac{\eta_0^{(t)}}{1 - \gamma p} = \frac{(1 - \frac{4\eta}{3}(1 - \gamma))^t}{\frac{4}{3}(1 - \gamma)}. \quad (4.125b)$$

Similarly, applying the above argument to (4.123) and rearranging terms, we obtain, for any integer $0 \leq \tau < T$,

$$\begin{aligned} V^*(2) - \mathbb{E}[V_T(2)] &= (1 - \eta(1 - \gamma p))^{T-\tau} \left(\frac{1}{1 - \gamma p} - \mathbb{E}[V_\tau(2)] \right) \\ &= (1 - \eta(1 - \gamma p))^{T-\tau} (V^*(2) - \mathbb{E}[V_\tau(2)]). \end{aligned} \quad (4.126)$$

State 3. When it comes to state 3, it holds that

$$\begin{aligned} V_t(3) &= Q_t(3, 1) = (1 - \eta)Q_{t-1}(3, 1) + \eta\{r(3, 1) + \gamma V_{t-1}(3)\} \\ &= (1 - \eta(1 - \gamma))V_{t-1}(3) + \eta. \end{aligned} \quad (4.127)$$

Recognizing that $V^*(3) = \frac{1}{1-\gamma}$ (see Lemma 4.12) obeys

$$V^*(3) = (1 - \eta(1 - \gamma))V^*(3) + \eta,$$

we can combine these two identities to arrive at

$$\begin{aligned} V_t(3) - V^*(3) &= (1 - \eta(1 - \gamma))(V_{t-1}(3) - V^*(3)) \\ &= (1 - \eta(1 - \gamma))^T (V_0(3) - V^*(3)) \\ &= -\frac{1}{1 - \gamma} (1 - \eta(1 - \gamma))^T. \end{aligned} \quad (4.128)$$

Roadmap. In what follows, we divide the analysis into three parts based on the learning rate η : small ($0 < \eta \leq \frac{4}{(1-\gamma)T}$), large ($\eta \geq \frac{1}{(1-\gamma)^2 T}$), and moderate ($\frac{4}{(1-\gamma)T} < \eta \leq \frac{1}{(1-\gamma)^2 T}$). Each regime is tackled in its own subsection below. In particular, the regimes with small and large learning rates will be analyzed by lower bounding the mean and variance of $V^*(2) - V_T(2)$, respectively, for which it suffices to analyze this state that has a single available action. The moderate learning rate regime presents the main technical challenge. For this regime, we analyze the dynamics of state 1, which, by admitting two actions, captures the over-estimation issue; the dynamics of state 3 will also be invoked later to facilitate the analysis.

4.8.3 Step 3: the regime with small learning rates

Suppose now the learning rate satisfies

$$0 < \eta \leq \frac{4}{(1 - \gamma)T},$$

which, as we shall see, is fairly small. In this regime, it suffices to look at the dynamics associated with state 2 (which has a single action). Our strategy is to show that, under such a small learning rate, the bias $V^*(2) - \mathbb{E}[V_t(2)]$ does not decay sufficiently quickly and therefore remains large after T iterations.

Towards this end, we first recall from (4.125) that

$$V^*(2) - \mathbb{E}[V_T(2)] = \frac{\eta_0^{(T)}}{1 - \gamma p},$$

which reduces the problem to deriving a lower bound on $\eta_0^{(T)}$. Observe that

$$\eta(1 - \gamma p) \leq 1 - \gamma p = \frac{4(1 - \gamma)}{3} \leq \frac{1}{3}, \quad (4.129a)$$

$$\eta(1 - \gamma p) = \frac{4\eta(1 - \gamma)}{3} < \frac{16}{3T}, \quad (4.129b)$$

provided that $\gamma \geq 3/4$ and $\eta \leq \frac{4}{(1-\gamma)T}$. As a consequence,

$$\log \eta_0^{(T)} = T \log(1 - \eta(1 - \gamma p)) \geq -1.5T\eta(1 - \gamma p) \geq -8,$$

where the first inequality holds due to the elementary fact $\log(1 - x) \geq -1.5x$ for all $0 \leq x \leq 0.5$. Taking the above results collectively yields

$$V^*(2) - \mathbb{E}[V_T(2)] = \frac{\eta_0^{(T)}}{1 - \gamma p} \geq \frac{e^{-8}}{1 - \gamma p} = \frac{3}{4e^8(1 - \gamma)},$$

thus indicating that

$$\begin{aligned} \mathbb{E}[(V^*(2) - V_T(2))^2] &\geq (V^*(2) - \mathbb{E}[V_T(2)])^2 \geq \frac{9}{16e^{16}(1 - \gamma)^2} \\ &\geq \frac{9c_3}{16e^{16}(1 - \gamma)^4 T}, \end{aligned} \quad (4.130)$$

provided that $T \geq c_3/(1 - \gamma)^2$. This establishes the advertised lower bound (4.45) for this regime.

4.8.4 Step 4: the regime with large learning rates

Next, consider the regime with

$$\eta \geq \frac{1}{(1 - \gamma)^2 T},$$

corresponding to large learning rates. In this regime, it is again sufficient to look at state 2. In contrast to the small learning rate regime where we analyze the bias, we now aim to show that the variance of $V_T(2)$ is large. This is motivated by the intuition that larger learning rates can lead to greater variability in estimating optimal values.

Before proceeding, we first note that if $\eta > \frac{4}{(1-\gamma)T}$ (which covers the range $\eta \geq \frac{1}{(1-\gamma)^2 T}$ for $\gamma \geq 3/4$), then one has

$$\begin{aligned} \eta_0^{(T)} &= \left(1 - \frac{4\eta}{3}(1 - \gamma)\right)^T \leq \left(1 - \frac{4\eta}{3}(1 - \gamma)\right)^{\frac{3}{4\eta(1-\gamma)} \frac{4(1-\gamma)\eta T}{3}} \\ &\leq e^{-\frac{4(1-\gamma)\eta T}{3}} < e^{-16/3} < \frac{1}{100}, \end{aligned} \quad (4.131)$$

where the last line makes use of the elementary inequality $(1 - x)^{1/x} \leq e^{-1}$ for all $0 < x < 1$. This combined with (4.125b) yields an upper bound on the bias:

$$|V^*(2) - \mathbb{E}[V_T(2)]| = \frac{\eta_0^{(T)}}{\frac{4}{3}(1 - \gamma)} < \frac{1}{100(1 - \gamma)}. \quad (4.132)$$

We now turn to the variance of $V_T(2)$. Recalling the decomposition (4.123) of $V_T(2)$, we intend to exploit martingale properties to lower-bound its variance. Specifically, consider first a martingale sequence $\{Z_k\}_{0 \leq k \leq T}$ adapted to a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_T$, satisfying $\mathbb{E}[Z_{k+1} | \mathcal{F}_k] = 0$ and $\mathbb{E}[Z_k | \mathcal{F}_k] = Z_k$ for all $0 \leq k \leq T$. In addition, consider any $0 \leq \tau < T$, and let W_0 be a random

variable such that $\mathbb{E}[W_0 | \mathcal{F}_\tau] = W_0$. Then, by the law of total variance and standard martingale properties, we have

$$\begin{aligned}
\text{Var}\left(W_0 + \sum_{k=\tau+1}^T Z_k\right) &= \mathbb{E}\left[\text{Var}\left(W_0 + \sum_{k=\tau+1}^T Z_k \middle| \mathcal{F}_{T-1}\right)\right] + \text{Var}\left(\mathbb{E}\left[W_0 + \sum_{k=\tau+1}^T Z_k \middle| \mathcal{F}_{T-1}\right]\right) \\
&\stackrel{(i)}{=} \mathbb{E}[\text{Var}(Z_T | \mathcal{F}_{T-1})] + \text{Var}\left(W_0 + \sum_{k=\tau+1}^{T-1} Z_k\right) = \dots \\
&= \sum_{k=\tau+1}^T \mathbb{E}[\text{Var}(Z_k | \mathcal{F}_{k-1})] + \text{Var}(W_0) \\
&\geq \sum_{k=\tau+1}^T \mathbb{E}[\text{Var}(Z_k | \mathcal{F}_{k-1})],
\end{aligned} \tag{4.133}$$

where (i) arises from the above assumptions on W_0 and $\{Z_k\}$. In words, this decomposes the variance of a martingale sum by summing the expected conditional variances over time. Consequently, for any $0 \leq \tau < T-1$, applying inequality (4.133) to the decomposition (4.123) (with τ replaced by $\tau+1$) yields

$$\begin{aligned}
\text{Var}(V_T(2)) &\geq \mathbb{E}\left[\sum_{k=\tau+2}^T \text{Var}\left(\eta_k^{(T)} \gamma (P_k(2|2) - p) V_{k-1}(2) \middle| V_{k-1}(2)\right)\right] \\
&= \sum_{k=\tau+2}^T (\eta_k^{(T)} \gamma)^2 p(1-p) \mathbb{E}[(V_{k-1}(2))^2] \\
&\geq \frac{(1-\gamma)(4\gamma-1)}{9} \sum_{k=\tau+2}^T (\eta_k^{(T)})^2 (\mathbb{E}[V_{k-1}(2)])^2,
\end{aligned} \tag{4.134}$$

where the identity arises from the fact that $P_k(2|2)$ is a Bernoulli random variable with mean p , and the last line uses the choice (4.113) of p and the elementary fact $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$. As we will see shortly, the rationale for replacing $\mathbb{E}[(V_{k-1}(2))^2]$ with $(\mathbb{E}[V_{k-1}(2)])^2$ is that the mean $\mathbb{E}[V_{k-1}(2)]$ becomes quite large beyond a certain time threshold and is easier to compute.

To further control (4.134), we develop a lower bound on $\mathbb{E}[V_{k-1}(2)]$. Given (4.132) and the facts that $V_0(2) = 0$ and $V^*(2) = \frac{3}{4(1-\gamma)}$ (see Lemma 4.12), one is allowed to choose $0 < \tau \leq T$ such that

$$\mathbb{E}[V_\tau(2)] \geq \frac{1}{2(1-\gamma)} > \mathbb{E}[V_{\tau-1}(2)]. \tag{4.135}$$

It is also seen from (4.125b) that $\mathbb{E}[V_k(2)]$ is monotonically increasing in k , and hence

$$\mathbb{E}[V_k(2)] \geq \mathbb{E}[V_\tau(2)] \geq \frac{1}{2(1-\gamma)} \quad \text{for all } k \geq \tau;$$

namely, $\mathbb{E}[V_k(2)]$ becomes large once k exceeds the threshold τ . Combine this inequality with (4.134) to yield

$$\begin{aligned}
\text{Var}(V_T(2)) &\geq \frac{(1-\gamma)(4\gamma-1)}{9} \cdot \frac{1}{4(1-\gamma)^2} \sum_{k=\tau+2}^T (\eta_k^{(T)})^2 \\
&= \frac{4\gamma-1}{36(1-\gamma)} \frac{\eta^2(1-(1-\eta(1-\gamma p))^{2(T-\tau-1)})}{1-(1-\eta(1-\gamma p))^2},
\end{aligned} \tag{4.136}$$

where the last line arises from (4.117) and calculates the sum of the geometric sequence.

Recalling that $\eta(1 - \gamma p) \leq 1/3$ (cf. (4.129a)), one can deduce that

$$\begin{aligned} (1 - \eta(1 - \gamma p))^{2(T-\tau-1)} &\leq \frac{81}{16} (1 - \eta(1 - \gamma p))^{2(T-\tau+1)} \\ &= \frac{81}{16} \frac{(V^*(2) - \mathbb{E}[V_T(2)])^2}{(V^*(2) - \mathbb{E}[V_{\tau-1}(2)])^2} \\ &\leq \frac{81(1 - \gamma)^2}{4} (V^*(2) - \mathbb{E}[V_T(2)])^2 \leq \frac{1}{2}, \end{aligned}$$

where the second line follows from (4.126), and the two inequalities in the last line arise from (4.135) and (4.132), respectively. Moreover, it is also readily seen that

$$1 - (1 - \eta(1 - \gamma p))^2 \leq 2\eta(1 - \gamma p) = \frac{8\eta(1 - \gamma)}{3}.$$

Substitute the above two inequalities into (4.136) to yield

$$\begin{aligned} \mathbb{E}[(V^*(2) - V_T(2))^2] &\geq \text{Var}(V_T(2)) \geq \frac{\eta(4\gamma - 1)}{192(1 - \gamma)^2} \geq \frac{4\gamma - 1}{192(1 - \gamma)^4 T} \\ &\geq \frac{1}{86(1 - \gamma)^4 T}, \end{aligned} \tag{4.137}$$

with the proviso that $\eta \geq \frac{1}{(1 - \gamma)^2 T}$ and $\gamma \geq 3/4$. This proves the claimed lower bound (4.45) for this regime.

4.8.5 Step 5: the regime with moderate learning rates

We now turn to the regime of moderate learning rates, namely,

$$\frac{4}{(1 - \gamma)T} < \eta < \frac{1}{(1 - \gamma)^2 T}.$$

This is the regime in which Q-learning is most effective, and at the same time, the most challenging to analyze when developing lower bounds. Throughout this case, we make note of the property

$$\eta_0^{(T)} \leq \frac{1}{100}, \tag{4.138}$$

which has been justified previously in (4.131). We shall focus attention on the dynamics w.r.t. state 1 and its associated value function $V_t(1)$. Note that the presence of 2 available actions at this state is deliberately designed to capture the over-estimation issue inherent in Q-learning.

Step 5.1: decoupling the impact of initialization. It is slightly more convenient to work with a Q-learning sequence initialized at the true optimal value. We therefore introduce an auxiliary sequence governed by the following update rule:

$$\hat{Q}_t(a) = (1 - \eta)\hat{Q}_{t-1}(a) + \eta\{1 + \gamma P_t(1 | 1, a)\hat{V}_{t-1}\} \tag{4.139}$$

with

$$\hat{V}_{t-1} := \max_a \hat{Q}_{t-1}(a) \quad \text{and} \quad \hat{Q}_0(a) := Q^*(1, a) = \frac{1}{1 - \gamma p},$$

where $Q^*(1, a)$ has been computed in Lemma 4.12. Evidently, $\{\hat{Q}_t(a)\}$ corresponds to a Q-learning sequence initialized at the ground truth. To decouple the effect of initialization from the original dynamics of $Q_t(1, a)$, we show that the discrepancy resulting from the different initializations of $\hat{Q}_t(a)$ and $Q_t(1, a)$ vanishes geometrically, i.e.,

$$Q_t(1, a) \geq \hat{Q}_t(a) - \frac{1}{1-\gamma}(1-\eta(1-\gamma))^t, \quad a \in \{1, 2\}. \quad (4.140)$$

The proof is postponed to Section 4.8.6.

In light of this bound (4.140), we focus on bounding the auxiliary sequence $\hat{Q}_t$ in the sequel, and defer translating the result back to $Q_t(1, a)$ until the final sub-step of the proof. Before proceeding, we record a crude lower bound on $\hat{V}_t$:

$$\mathbb{P}\left\{\hat{V}_t \geq \frac{1}{4(1-\gamma)}\right\} \geq \frac{1}{2} \quad \text{for all } t \leq T. \quad (4.141)$$

In other words, it shows that $\hat{V}_t$ has the correct order with nontrivial probability. The proof is postponed to Section 4.8.6.

Step 5.2: deriving recursive relations about estimation errors. We now study the over-estimation error of the auxiliary sequence $\hat{Q}_t$ (cf. (4.139)). Define the error sequence

$$\Delta_t(a) := \hat{Q}_t(a) - Q^*(1, a), \quad a = 1, 2; \quad (4.142a)$$

$$\Delta_{t,\max} := \max_a \Delta_t(a). \quad (4.142b)$$

Our objective is to show that $\Delta_{T,\max}$ is large in expectation (so that $\hat{Q}_t(a)$ overestimates $Q^*(1, a)$), thus putting the “over-estimation” issue that fundamentally slows convergence on a rigorous footing.

To do so, a key step lies in establishing recursive relations about $\Delta_{t,\max}$. According to the update rule (4.139) and the Bellman optimality equation $Q^*(1, a) = 1 + \gamma p V^*(1)$, we have

$$\begin{aligned} \Delta_t(a) &= (1-\eta)\hat{Q}_{t-1}(a) + \eta(1+\gamma P_t(1|1, a)\hat{V}_{t-1}) - Q^*(1, a) \\ &= (1-\eta)\Delta_{t-1}(a) + \eta(1+\gamma P_t(1|1, a)\hat{V}_{t-1} - Q^*(1, a)) \\ &= (1-\eta)\Delta_{t-1}(a) + \eta\gamma(P_t(1|1, a)\hat{V}_{t-1} - pV^*(1)) \\ &= (1-\eta)\Delta_{t-1}(a) + \eta\gamma(p(\hat{V}_{t-1} - V^*(1)) + (P_t(1|1, a) - p)\hat{V}_{t-1}) \\ &= (1-\eta)\Delta_{t-1}(a) + \eta\gamma(p\Delta_{t-1,\max} + (P_t(1|1, a) - p)\hat{V}_{t-1}), \end{aligned}$$

where the last line holds since $\hat{V}_{t-1} - V^*(1) = \max_a (\hat{Q}_{t-1}(a) - V^*(1)) = \max_a \Delta_{t-1}(a)$. Applying this relation recursively yields

$$\Delta_t(a) = \sum_{k=1}^t \eta(1-\eta)^{t-k} \gamma (p\Delta_{k-1,\max} + (P_k(1|1, a) - p)\hat{V}_{k-1}),$$

where we use the initialization $\Delta_0(a) = 0$ (by construction). Letting

$$\xi_t(a) := \sum_{k=1}^t \eta(1-\eta)^{t-k} \gamma (P_k(1|1, a) - p)\hat{V}_{k-1}, \quad (4.143a)$$

$$\xi_{t,\max} := \max_a \xi_t(a), \quad (4.143b)$$

one can take the maximum over a in the above display about $\Delta_t(a)$ to derive the following identity regarding the maximum error:

$$\Delta_{t,\max} = \sum_{k=1}^t \eta(1-\eta)^{t-k} \gamma p \Delta_{k-1,\max} + \xi_{t,\max}. \quad (4.144)$$

Step 5.3: lower bounding $\mathbb{E}[\xi_{t,\max}]$. As can be seen from the recursion (4.144), a crucial step is to control $\xi_{t,\max}$, which involves the maximum of multiple random quantities. Towards this end, we develop a lower bound on $\mathbb{E}[\xi_{t,\max}]$ as follows; the proof is deferred to Section 4.8.7.

Lemma 4.13. Define a threshold

$$\hat{\tau} := \max \left\{ \tau \mid (1 - \eta(1 - \gamma p))^{T-\tau+1} \leq \frac{6}{7} \right\}. \quad (4.145)$$

Then there exists some universal constant $c > 0$ such that

$$\mathbb{E}[\xi_{t,\max}] \geq \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \quad \text{for all } t \geq \hat{\tau}. \quad (4.146)$$

Remark 4.6. Property (4.138) guarantees that $\hat{\tau}$ (cf. (4.145)) is well-defined.

Remark 4.7. Note that $\mathbb{E}[\xi_{t,\max}]$ is strictly positive due to the presence of the max operator together with the availability of more than one action. In contrast, when there is only a single action a , one readily has $\mathbb{E}[\xi_{t,\max}] = \mathbb{E}[\xi_t(a)] = 0$. This distinction gives rise to a fundamental difference in sample complexity between the cases $A = 1$ and $A \geq 2$, leading respectively to the horizon dependency $\frac{1}{(1-\gamma)^3}$ and $\frac{1}{(1-\gamma)^4}$.

Substituting (4.146) into (4.144) reveals that, for all $t \geq \hat{\tau}$,

$$\mathbb{E}[\Delta_{t,\max}] \geq \sum_{k=1}^t \eta(1-\eta)^{t-k} \gamma p \mathbb{E}[\Delta_{k-1,\max}] + \frac{c}{\sqrt{(1-\gamma)^2 T \log T}}, \quad (4.147)$$

leaving us with a deterministic recursive relation to solve.

Step 5.4: solving the recursion (4.147). To solve the above recursion, it is convenient to introduce the following auxiliary scalar sequence:

$$x_t = (1-\eta)x_{t-1} + \eta \left(\gamma p x_{t-1} + \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \right) \quad (4.148)$$

initialized at $x_{\hat{\tau}} = 0$, where $\hat{\tau}$ has been defined in (4.145). A direct comparison reveals that

$$\mathbb{E}[\Delta_{T,\max}] \geq x_T, \quad (4.149)$$

which follows readily from the iterative expansion of (4.148):

$$x_t = \sum_{k=\hat{\tau}}^t \eta(1-\eta)^{t-k} \gamma p x_{k-1} + \sum_{k=\hat{\tau}}^T \eta(1-\eta)^{t-k} \frac{c}{\sqrt{(1-\gamma)^2 T \log T}}.$$

A straightforward calculation shows that the sequence (4.148) obeys

$$x_T = \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \sum_{k=\hat{\tau}}^T \eta(1-\eta(1-\gamma p))^{T-k}$$

$$\begin{aligned}
&= \frac{c[1 - (1 - \eta(1 - \gamma p))^{T-\hat{\tau}+1}]}{(1 - \gamma p)\sqrt{(1 - \gamma)^{2T} \log T}} \\
&= \frac{3c[1 - (1 - \eta(1 - \gamma p))^{T-\hat{\tau}+1}]}{4\sqrt{(1 - \gamma)^{4T} \log T}} \geq \frac{3c}{28\sqrt{(1 - \gamma)^{4T} \log T}},
\end{aligned}$$

where the second equality follows by calculating the sum of the geometric sequence, and the last inequality holds as long as $(1 - \eta(1 - \gamma p))^{T-\hat{\tau}+1} \leq 6/7$ (see (4.145)). Combining this with (4.149) yields

$$\mathbb{E}[\Delta_{T,\max}] \geq x_T \geq \frac{3c}{28\sqrt{(1 - \gamma)^{4T} \log T}}. \quad (4.150)$$

Step 5.5: putting all this together. Thus far, we have established the claimed lower bound for the auxiliary sequence $\hat{Q}_t$. To complete the argument, it remains to connect it to the original sequence Q_t . To do so, combine the lower bound (4.150) with (4.140) to yield

$$\begin{aligned}
\mathbb{E}[V_T(1) - V^*(1)] &\geq \mathbb{E}\left[\Delta_{T,\max} - \frac{1}{1 - \gamma}(1 - \eta(1 - \gamma))^T\right] \\
&\geq \frac{3c}{28\sqrt{(1 - \gamma)^{4T} \log T}} - \frac{(1 - \eta(1 - \gamma))^T}{1 - \gamma}.
\end{aligned}$$

As it turns out, the last term in the above display can be easily handled with the aid of $V_T(3)$. Specifically, taking this result and (4.128) collectively, we arrive at

$$\begin{aligned}
\mathbb{E}[V_T(1) - V^*(1)] + \mathbb{E}[V^*(3) - V_T(3)] &= \mathbb{E}[V_T(1) - V^*(1)] + \frac{(1 - \eta(1 - \gamma))^T}{1 - \gamma} \\
&\geq \frac{3c}{28\sqrt{(1 - \gamma)^{4T} \log T}}.
\end{aligned}$$

Apply the elementary inequality $\max\{a, b\} \geq \frac{1}{2}(a + b)$ to obtain

$$\begin{aligned}
\max\left\{\mathbb{E}[|V_T(1) - V^*(1)|], \mathbb{E}[|V_T(3) - V^*(3)|]\right\} &\geq \frac{1}{2}\left\{\mathbb{E}[V_T(1) - V^*(1)] + \mathbb{E}[V^*(3) - V_T(3)]\right\} \\
&\geq \frac{3c}{56\sqrt{(1 - \gamma)^{4T} \log T}},
\end{aligned}$$

yielding

$$\begin{aligned}
\max_s \mathbb{E}[|V_T(s) - V^*(s)|^2] &\geq \max_s \left(\mathbb{E}[|V_T(s) - V^*(s)|]\right)^2 \\
&\geq \left(\frac{3c}{56\sqrt{(1 - \gamma)^{4T} \log T}}\right)^2 = \frac{9c^2}{56^2(1 - \gamma)^{4T} \log^2 T}.
\end{aligned}$$

We have thus established the desired lower bound for the moderate learning-rate regime, which, together with the bounds in (4.130) and (4.137) for the other two regimes, completes the proof of Theorem 4.6.

4.8.6 Proof of auxiliary properties

Proof of property (4.140). We prove this claim by induction. The base case $t = 0$ follows trivially. Assume now the hypothesis holds for $t - 1$. By combining the inductive hypothesis (4.140) for the

$(t-1)$ -th iteration with the definitions of V_{t-1} and $\hat{V}_{t-1}$, we obtain

$$\begin{aligned} V_{t-1}(1) &= \max_a Q_{t-1}(1, a) \geq \max_a \hat{Q}_{t-1}(a) - \frac{(1 - \eta(1 - \gamma))^{t-1}}{1 - \gamma} \\ &= \hat{V}_{t-1} - \frac{(1 - \eta(1 - \gamma))^{t-1}}{1 - \gamma}. \end{aligned}$$

Given the update rules of $Q_t(1, a)$ and $\hat{Q}_t(a)$, their difference evolves as

$$\begin{aligned} Q_t(1, a) - \hat{Q}_t(a) &= (1 - \eta)(Q_{t-1}(1, a) - \hat{Q}_{t-1}(a)) + \eta\gamma P_t(1 | 1, a)(V_{t-1}(1) - \hat{V}_{t-1}) \\ &\geq -(1 - \eta)\frac{(1 - \eta(1 - \gamma))^{t-1}}{1 - \gamma} - \eta\gamma P_t(1 | 1, a)\frac{(1 - \eta(1 - \gamma))^{t-1}}{1 - \gamma} \\ &\geq -(1 - \eta)\frac{(1 - \eta(1 - \gamma))^{t-1}}{1 - \gamma} - \eta\gamma\frac{(1 - \eta(1 - \gamma))^{t-1}}{1 - \gamma} \\ &= -\frac{(1 - \eta(1 - \gamma))^t}{1 - \gamma}, \end{aligned}$$

where the first inequality invokes the induction hypothesis for the $(t-1)$ -th iteration, and the second follows since $P_t(1 | 1, a) \leq 1$. This completes the inductive step and, in turn, establishes the result for all $t \geq 0$.

Proof of property (4.141). We find it convenient to introduce a second auxiliary sequence defined by the recursion:

$$\bar{Q}_t = (1 - \eta)\bar{Q}_{t-1} + \eta\{1 + \gamma P_t(1 | 1, 1)\bar{Q}_{t-1}\}, \quad (4.151a)$$

$$\bar{Q}_0 = V^*(1) = \frac{1}{1 - \gamma p}. \quad (4.151b)$$

In words, this sequence corresponds to the Q-learning iterates in a simplified setting with only a single available action, effectively removing the max operator. Given that $\hat{V}_t = \max_a \hat{Q}_t(a) \geq \hat{Q}_t(1)$, it is straightforward to verify that, for all $t \geq 0$,

$$\hat{Q}_t(1) \geq (1 - \eta)\hat{Q}_{t-1}(1) + \eta\{1 + \gamma P_t(1 | 1, 1)\hat{Q}_{t-1}(1)\} \geq \bar{Q}_t. \quad (4.152)$$

We use the sequence $\bar{Q}_t$ to lower-bound $\hat{V}_t$, as its simpler structure allows us to more easily characterize the evolution of the lower envelope of the Q-iterates. First, observe that $\bar{Q}_t$ admits the decomposition:

$$\begin{aligned} \bar{Q}_t &= (1 - \eta(1 - \gamma p))\bar{Q}_{t-1} + \eta\{1 + \gamma(P_t(1 | 1, 1) - p)\bar{Q}_{t-1}\} \\ &= (1 - \eta(1 - \gamma p))^t \bar{Q}_0 + \sum_{k=1}^t \eta(1 - \eta(1 - \gamma p))^{t-k} \{1 + \gamma(P_k(1 | 1, 1) - p)\bar{Q}_{k-1}\} \\ &= \eta_0^{(t)} \frac{1}{1 - \gamma p} + \sum_{k=1}^t \eta_k^{(t)} + \sum_{k=1}^t \eta_k^{(t)} \gamma (P_k(1 | 1, 1) - p) \bar{Q}_{k-1} \\ &= \frac{1}{1 - \gamma p} + \sum_{k=1}^t \underbrace{\eta_k^{(t)} \gamma (P_k(1 | 1, 1) - p) \bar{Q}_{k-1}}_{=: z_k}, \end{aligned} \quad (4.153)$$

where the last line results from (4.118a). It then comes down to controlling $\sum_k z_k$.

By construction, the sequence $\{z_k\}$ forms a martingale satisfying

$$\mathbb{E}[z_k | P_{k-1}(1 | 1, 1), \dots, P_1(1 | 1, 1)] = 0 \quad \text{and} \quad |z_k| \leq \max_{1 \leq k \leq t} \eta_k^{(t)} \cdot \frac{\gamma p}{1 - \gamma},$$

where the last inequality follows from the basic property $0 \leq \bar{Q}_{k-1} \leq \frac{1}{1-\gamma}$ (akin to Lemma 4.7) and the fact that $|P_k(1 | 1, 1) - p| \leq \max\{p, 1 - p\} = p$ since $p = \frac{4\gamma-1}{3\gamma}$ and $\gamma \geq 3/4$. In order to invoke Freedman's inequality to control (4.153), we need to bound the conditional variance. Given these properties and the fact that $P_k(1 | 1, 1)$ is a Bernoulli random variable with mean p , we obtain

$$\begin{aligned} \sum_{k=1}^t \text{Var}\left(z_k | P_{k-1}(1 | 1, 1), \dots, P_1(1 | 1, 1)\right) &= \sum_{k=1}^t (\eta_k^{(t)})^2 p(1-p)(\gamma \bar{Q}_{k-1})^2 \\ &\leq \max_{1 \leq k \leq t} \eta_k^{(t)} \cdot \sum_{k=1}^t \eta_k^{(t)} \cdot \frac{1}{3(1-\gamma)} \leq \frac{\max_{1 \leq k \leq t} \eta_k^{(t)}}{4(1-\gamma)^2}. \end{aligned}$$

Here, the penultimate inequality relies on the fact $0 \leq \bar{Q}_{k-1} \leq \frac{1}{1-\gamma}$ (akin to Lemma 4.7) and the choice of p , whereas the last inequality results from the following condition (derived through (4.118a))

$$\sum_{k=1}^t \eta_k^{(t)} = (1 - \eta_0^{(t)}) \frac{1}{1 - \gamma p} \leq \frac{1}{1 - \gamma p} = \frac{3}{4(1 - \gamma)}.$$

Freedman's inequality (see Lemma 2.9 with $K_0 = 1$) then implies that

$$\left| \sum_{k=1}^t z_k \right| \leq \sqrt{\frac{4 \max_{1 \leq k \leq t} \eta_k^{(t)}}{(1-\gamma)^2} \log \frac{2}{\delta}} + \frac{4 \max_{1 \leq k \leq t} \eta_k^{(t)}}{3(1-\gamma)} \log \frac{2}{\delta} \quad (4.154)$$

with probability at least $1 - \delta$.

To continue, observe that

$$\max_{1 \leq k \leq t} \eta_k^{(t)} \leq \eta \leq \frac{1}{(1-\gamma)^2 T} \leq \frac{1}{50}, \quad \text{for all } t \geq 1, \quad (4.155)$$

provided that $T \geq \frac{50}{(1-\gamma)^2}$. Setting $\delta = 1/2$ in (4.154) then yields

$$\sum_{k=1}^t z_k \geq -\frac{1}{2(1-\gamma)}$$

with probability at least $1/2$. Combine this with the decomposition (4.153) and the property (4.152) to conclude

$$\hat{V}_t \geq \hat{Q}_t(1) \geq \bar{Q}_t \geq \frac{1}{1-\gamma p} - \frac{1}{2(1-\gamma)} = \frac{1}{4(1-\gamma)}$$

with probability at least $1/2$, where the last identity relies on the choice (4.113) of p . This completes the proof of (4.141).

4.8.7 Proof of Lemma 4.13

In light of the definition (4.143), it is convenient to express, for every t ,

$$\mathbb{E}[\xi_{t,\max}] = \mathbb{E}\left[\frac{\xi_t(1) + \xi_t(2) + |\xi_t(1) - \xi_t(2)|}{2}\right] = \mathbb{E}\left[\frac{|\xi_t(1) - \xi_t(2)|}{2}\right]$$

$$= \frac{1}{2} \mathbb{E} \left[\left| \sum_{k=1}^t \eta(1-\eta)^{t-k} \gamma (P_k(1|1,1) - P_k(1|1,2)) \widehat{V}_{k-1} \right| \right],$$

where we have used the fact that $\mathbb{E}[\xi_t(a)] = 0$. Importantly, the summation within the absolute value operator is a martingale sum. To be more precise, define

$$\zeta_t := \sum_{k=1}^t \underbrace{\eta(1-\eta)^{t-k} \gamma (P_k(1|1,1) - P_k(1|1,2)) \widehat{V}_{k-1}}_{=: z_k},$$

where $\{z_k\}$ forms a martingale sequence since

$$\mathbb{E}[z_k | \{P_j(1|1,1), P_j(1|1,2)\}_{1 \leq j < k}] = 0.$$

In order to control $\mathbb{E}[|\zeta_t|]$, we first seek to analyze the tail of $|\zeta_t|$. To do so, we claim that, for any $0 < \delta < 1$,

$$\mathbb{P} \left\{ |\zeta_t| \geq \sqrt{\frac{8 \log \frac{2}{\delta}}{3(1-\gamma)} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2} + \frac{4\eta \log \frac{2}{\delta}}{3(1-\gamma)} \right\} \leq \delta. \quad (4.156)$$

To validate this bound, we first notice that

$$|z_k| \leq \max_{k: 1 \leq k \leq t} \eta(1-\eta)^{t-k} \cdot \frac{1}{1-\gamma} \leq \frac{\eta}{1-\gamma}, \quad (4.157)$$

where we have used $|\widehat{V}_{k-1}| \leq \frac{1}{1-\gamma}$ according to Lemma 4.7. In addition, we can further bound the aggregate condition variance as

$$\begin{aligned} & \sum_{k=1}^t \text{Var}(z_k | P_{k-1}(1|1,1), P_{k-1}(1|1,2), \dots, P_1(1|1,1), P_1(1|1,2)) \\ &= \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2 \cdot 2p(1-p) \cdot (\gamma \widehat{V}_{k-1})^2 \\ &\leq \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2 \cdot \frac{2}{3(1-\gamma)}, \end{aligned}$$

where the last inequality comes from the fact that $|\widehat{V}_{k-1}| \leq \frac{1}{1-\gamma}$, the assumption that $\gamma \geq 3/4$ and the choice $p = \frac{4\gamma-1}{3\gamma}$. These bounds, taken together with Freedman's inequality (see Lemma 2.9 with $K_0 = 1$), justify the tail bound (4.156).

Next, we make note of a crude deterministic bound:

$$|\zeta_t| \leq \sum_{k=1}^t |z_k| \leq \left[\sum_{k=1}^t \eta(1-\eta)^{t-k} \right] \cdot \frac{1}{1-\gamma} \leq \frac{1}{1-\gamma}, \quad (4.158)$$

which is a consequence of the facts that $|\widehat{V}_{k-1}| \leq \frac{1}{1-\gamma}$ (see Lemma 4.7) and $\sum_{k=1}^t \eta(1-\eta)^{t-k} \leq \eta \cdot \frac{1}{1-(1-\eta)} = 1$. This immediately implies that

$$\delta_0 := \frac{(1-\gamma)^2}{2} \mathbb{E}[|\zeta_t|^2] \leq \frac{1}{2}.$$

Thus, setting $\delta = \delta_0 \in (0, 1)$ in (4.156) reveals that, with probability at least $1 - \delta_0$,

$$|\zeta_t| \leq B := \sqrt{\frac{8 \log \frac{2}{\delta_0}}{3(1-\gamma)} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2} + \frac{4\eta \log \frac{2}{\delta_0}}{3(1-\gamma)}. \quad (4.159)$$

When $T \geq \frac{1}{(1-\gamma)^2}$, one can guarantee that

$$\begin{aligned} \mathbb{E}[\xi_{t,\max}] &= \frac{1}{2} \mathbb{E}[|\zeta_t|] \geq \frac{1}{2} \mathbb{E}[|\zeta_t| \mathbf{1}(|\zeta_t| \leq B)] \geq \frac{1}{2B} \mathbb{E}[|\zeta_t|^2 \mathbf{1}(|\zeta_t| \leq B)] \\ &= \frac{1}{2B} \left\{ \mathbb{E}[|\zeta_t|^2] - \mathbb{E}[|\zeta_t|^2 \mathbf{1}(|\zeta_t| > B)] \right\} \\ &\stackrel{(i)}{\geq} \frac{1}{2B} \left\{ \mathbb{E}[|\zeta_t|^2] - \frac{1}{(1-\gamma)^2} \mathbb{P}\{|\zeta_t| > B\} \right\} \\ &\geq \frac{1}{2B} \left\{ \mathbb{E}[|\zeta_t|^2] - \frac{\delta}{(1-\gamma)^2} \right\} \stackrel{(ii)}{\geq} \frac{1}{4B} \mathbb{E}[|\zeta_t|^2]. \end{aligned} \quad (4.160)$$

Here, (i) follows from (4.158), while (ii) holds due to the choice of δ_0 .

It thus suffices to lower bound $\mathbb{E}[|\zeta_t|^2]$ for $t \geq \hat{\tau}$, where we recall the definition

$$\hat{\tau} := \max \left\{ \tau \mid (1 - \eta(1 - \gamma p))^{T-\tau+1} \leq \frac{6}{7} \right\}. \quad (4.161)$$

According to (4.138), it satisfies $1 \leq \hat{\tau} \leq T$. First, observe that for any $t \geq \hat{\tau}$, the martingale properties about $\{z_k\}$ give

$$\begin{aligned} \mathbb{E}[|\zeta_t|^2] &\geq \sum_{k=1}^t \mathbb{E} \left[\text{Var} \left(\eta(1-\eta)^{t-k} \gamma (P_k(1|1,1) - P_k(1|1,2)) \hat{V}_{k-1} \mid \hat{V}_{k-1} \right) \right] \\ &\geq \sum_{k=1}^t \mathbb{E} \left[\text{Var} \left(\eta(1-\eta)^{t-k} \gamma (P_k(1|1,1) - P_k(1|1,2)) \hat{V}_{k-1} \mid \hat{V}_{k-1} \right) \right. \\ &\quad \cdot \mathbf{1} \left\{ \hat{V}_{k-1} \geq \frac{1}{4(1-\gamma)} \right\} \Big] \\ &\geq \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2 \gamma^2 \cdot \frac{2p(1-p)}{16(1-\gamma)^2} \mathbb{P} \left\{ \hat{V}_{k-1} \geq \frac{1}{4(1-\gamma)} \right\}, \end{aligned}$$

where the first line relies on (4.133), and the last step makes use of the constraint $\hat{V}_{k-1} \geq \frac{1}{4(1-\gamma)}$.

To further control the right-hand side, we take $\tau' := \max \{t - \frac{1}{\eta}, 1\}$ and show that

$$\begin{aligned} \mathbb{E}[|\zeta_t|^2] &\stackrel{(i)}{\geq} \frac{1}{2} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2 \gamma^2 \cdot \frac{2p(1-p)}{16(1-\gamma)^2} \\ &\geq \frac{1}{48(1-\gamma)} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2 \\ &\geq \frac{1}{48(1-\gamma)} \sum_{k=\tau'}^t \eta^2 [(1-\eta)^{t-k}]^2 \stackrel{(ii)}{\geq} \frac{\eta}{1200(1-\gamma)}, \end{aligned} \quad (4.162)$$

where (i) arises from (4.141), and (ii) is valid as long as the following property holds for $t \geq \hat{\tau}$ (which will be proven towards the end of this subsection)

$$\sum_{k=\tau'}^t \eta^2 [(1-\eta)^{t-k}]^2 \geq \frac{1}{24} \eta. \quad (4.163)$$

We are now well equipped to control $\mathbb{E}[\xi_{t,\max}]$ using the property (4.160). Recalling the expression of B in (4.159), we see that bounding $\mathbb{E}[|\zeta_t|^2]/B$ reduces to controlling the following two quantities:

$$\frac{\mathbb{E}[|\zeta_t|^2]}{\frac{\eta}{1-\gamma} \log \frac{2}{\delta_0}} \quad \text{and} \quad \frac{\mathbb{E}[|\zeta_t|^2]}{\sqrt{\frac{\log \frac{2}{\delta_0}}{1-\gamma} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2}}. \quad (4.164)$$

- For the first term in (4.164), recalling that $\delta_0 = \frac{(1-\gamma)^2}{2} \mathbb{E}[|\zeta_t|^2]$, we can demonstrate that

$$\begin{aligned} \log \frac{1}{\delta_0} &= -\log \frac{(1-\gamma)^2}{2} \mathbb{E}[|\zeta_t|^2] \leq -\log \frac{(1-\gamma)\eta}{2400} \\ &\leq \log(9600T), \end{aligned} \quad (4.165)$$

where the first inequality is due to (4.162), and the second inequality arises since $\eta \geq \frac{4}{(1-\gamma)T}$ (given the range of the learning rates in this case). Combine this with (4.162) to yield

$$\frac{\mathbb{E}[|\zeta_t|^2]}{\frac{\eta}{1-\gamma} \log \frac{2}{\delta_0}} \gtrsim \frac{1}{\log T}.$$

- Turning to the second term in (4.164), we observe that

$$\begin{aligned} \frac{\mathbb{E}[|\zeta_t|^2]}{\sqrt{\frac{\log \frac{2}{\delta_0}}{1-\gamma} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2}} &\stackrel{(i)}{\gtrsim} \sqrt{\frac{1}{(1-\gamma) \log T} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2} \\ &\stackrel{(ii)}{\gtrsim} \sqrt{\frac{\eta}{(1-\gamma) \log T}} \stackrel{(iii)}{\gtrsim} \frac{1}{\sqrt{(1-\gamma)^2 T \log T}}. \end{aligned}$$

Here, (i) follows from (4.162) and (4.165) since

$$\mathbb{E}[|\zeta_t|^2] \gtrsim \frac{1}{1-\gamma} \sum_{k=1}^t \eta^2 [(1-\eta)^{t-k}]^2 \quad \text{and} \quad \log \frac{2}{\delta_0} \lesssim \log T;$$

(ii) arises from (4.163); and (iii) relies on the fact $\eta \gtrsim \frac{1}{(1-\gamma)T}$ (given the range of the learning rates considered in this step).

Substituting the above relations into (4.160) and using the expression of B in (4.159), we reach

$$\mathbb{E}[\xi_{t,\max}] \geq \frac{1}{4B} \mathbb{E}[|\zeta_t|^2] \geq \frac{c}{\sqrt{(1-\gamma)^2 T \log T}}$$

for some numerical constant $c > 0$. Thus, this proves property (4.146), as long as the claim (4.163) can be established.

Proof of the claim (4.163). Recall that $\tau' := \max\{t - \frac{1}{\eta}, 1\}$. For any k obeying $\tau' \leq k \leq t$, one has

$$\eta(1-\eta)^{t-k} \geq \eta(1-\eta)^{t-\tau'} \geq \eta(1-\eta)^{1/\eta} \geq \frac{1}{3}\eta,$$

where the last inequality is valid as long as $\eta \leq 0.1$. Moreover, it is seen that for any $t \geq \hat{\tau}$,

$$\sum_{k=\tau'}^t \eta(1-\eta)^{t-k} = 1 - (1-\eta)^{t-\tau'+1} \geq 1 - \max\left\{(1-\eta)^{1/\eta+1}, (1-\eta)^t\right\}$$

$$\geq 1 - \max \left\{ e^{-1}, (1 - \eta)^{\hat{\tau}} \right\} \geq \frac{1}{8}.$$

Here, the first identity can be derived via the sum of a geometric sequence; to see why the last inequality holds, note that according to the definition (4.161) and property (4.138),

$$\begin{aligned} (1 - \eta(1 - \gamma p))^{T - \hat{\tau}} &> \frac{6}{7} \\ \implies (1 - \eta(1 - \gamma p))^{\hat{\tau}} &< \frac{(1 - \eta(1 - \gamma p))^T}{6/7} = \frac{\eta_0^{(T)}}{6/7} \leq \frac{1/100}{6/7} \leq \frac{7}{8}, \end{aligned}$$

which in turn implies that

$$(1 - \eta)^{\hat{\tau}} \leq (1 - \eta(1 - \gamma p))^{\hat{\tau}} \leq \frac{7}{8}. \quad (4.166)$$

Putting the preceding two bounds together yields, for any $t \geq \hat{\tau}$,

$$\begin{aligned} \sum_{k=\tau'}^t \eta^2 \left[(1 - \eta)^{t-k} \right]^2 &\geq \min_{k: \tau' \leq k \leq t} \left\{ \eta(1 - \eta)^{t-k} \right\} \sum_{k=\tau'}^t \eta(1 - \eta)^{t-k} \\ &\geq \frac{1}{3} \eta \cdot \frac{1}{8} \geq \frac{1}{24} \eta \end{aligned}$$

as claimed. $\square$

4.9 Notes

The asymptotic convergence of TD learning and Q-learning — namely, their convergence behavior as the number of iterations tends to infinity, with all other parameters held fixed — has been extensively studied since its early development, e.g., Tsitsiklis and Van Roy (1997), Tsitsiklis (1994), Jaakkola *et al.* (1993), Szepesvári (1997), Borkar and Meyn (2000), Devraj and Meyn (2021), and Li *et al.* (2023e). Recently, substantial research efforts have shifted toward understanding the non-asymptotic performance of model-free methods, which elucidates how key problem parameters — in addition to the number of iterations — affect algorithmic performance. For TD learning, finite-time, finite-sample analyses have been developed in a growing body of recent work, including but not limited to Bhandari *et al.* (2018), Lakshminarayanan and Szepesvári (2018), Srikant and Ying (2019), Xu *et al.* (2019), Doan *et al.* (2019), Cai *et al.* (2019), Chen *et al.* (2020c), Mou *et al.* (2020), Kaledin *et al.* (2020), Chen *et al.* (2021d), Khamaru *et al.* (2021), Li *et al.* (2024a), Chen *et al.* (2022d), Patil *et al.* (2023), Li *et al.* (2024d), Mou *et al.* (2024), Samsonov *et al.* (2024b), Srikant (2025), Chen *et al.* (2025c), and Durmus *et al.* (2025). In parallel, finite-time, finite-sample performance guarantees for Q-learning and its variants have been established in, e.g., Kearns and Singh (1999), Even-Dar and Mansour (2003), Murphy (2005), Beck and Srikant (2012), Azar *et al.* (2011), Wainwright (2019b), Wainwright (2019c), Chen *et al.* (2020c), Qu and Wierman (2020), Xu and Gu (2020), Li *et al.* (2022b), Chen *et al.* (2024b), Chen *et al.* (2022d), Li *et al.* (2024a), Jin *et al.* (2024b), and Jiao *et al.* (2026).

Since TD learning and Q-learning can be viewed as instances of stochastic approximation, a notable portion of the existing theoretical analyses builds upon classical techniques from stochastic approximation theory, along with various modern extensions of these techniques. Prominent examples include the ODE method (e.g., Borkar and Meyn (2000), Borkar *et al.* (2025), and Meyn (2022)), Stein’s method (e.g., Srikant and Ying (2019) and Srikant (2025)), Lyapunov-drift approach (e.g.,

Chen *et al.* (2024b), Cayci *et al.* (2023), Chen *et al.* (2025d), and Chen and Maguluri (2026)), and extensions to two-time-scale settings (e.g., Konda and Tsitsiklis (2004)), among others. See also the tutorial Meyn (2023) for the stability of Q-learning with linear function approximation. The continued development of these techniques has provided versatile tools for analyzing not only TD learning and Q-learning, but also a range of model-free variants, including TD learning with gradient correction (TDC), gradient TD learning, Zap Q-learning, SARSA, momentum Q-learning, double Q-learning, and so on (Borkar and Meyn, 2000; Sutton *et al.*, 2009; Maei *et al.*, 2009; Gupta *et al.*, 2019; Liu *et al.*, 2015; Devraj and Meyn, 2017; Xu *et al.*, 2019; Xu and Liang, 2021; Zou *et al.*, 2019; Weng *et al.*, 2020b; Weng *et al.*, 2020a; Xiong *et al.*, 2020). It is worth noting, however, that these stochastic approximation techniques alone may be insufficient to establish the sharp performance guarantees for TD learning and Q-learning in Theorems 4.1 and 4.5. Achieving these tight guarantees requires additional proof ideas that carefully exploit the intrinsic variance structure of the MDP — a key ingredient underlying the state-of-the-art analysis in Li *et al.* (2024a).

In addition, while the sample inefficiency and overestimation bias of Q-learning have long been observed in practice — motivating algorithmic variants such as double Q-learning (Hasselt, 2010) — formal lower bounds on the iteration complexity of Q-learning were not established until Li *et al.* (2024a). Notably, Wainwright (2019b) conjectured, based on numerical evidence, that the tight dependence of the sample complexity of Q-learning on the effective horizon scales as $\tilde{O}(\frac{1}{(1-\gamma)^4})$. However, prior to Li *et al.* (2024a), neither matching upper nor lower bounds confirming this scaling were available.

With the aim of improving the sample complexity of Q-learning while preserving its model-free nature, more sophisticated algorithmic techniques have been proposed and analyzed in the literature. For example, a variance-reduced variant of Q-learning — drawing inspiration from variance-reduction methods in stochastic optimization, particularly stochastic variance-reduced gradient (SVRG) (Johnson and Zhang, 2013) — has been shown to achieve the minimax-optimal sample complexity for all $\varepsilon \in (0, 1]$ (Wainwright, 2019c). At a high level, this approach partitions the iterative process into epochs. At the beginning of each epoch, a batch of samples is used to construct increasingly more accurate reference estimates of the value function. Subsequent stochastic approximation updates incorporate these references by combining them with fresh stochastic samples to reduce variability, thereby accelerating convergence. We refer interested readers to Wainwright (2019c), Li *et al.* (2022b), Mou *et al.* (2022), and Xia *et al.* (2024) for further details on both synchronous and asynchronous Q-learning with variance reduction. Notably, however, no existing model-free algorithm is known to attain minimax optimality when $\varepsilon \gg 1$. This suggests the inherent statistical suboptimality of model-free approaches in the most sample-hungry regime. Finally, we note that variance-reduction techniques have also been incorporated into TD learning for policy evaluation; see, for example, Khamaru *et al.* (2021), Xu *et al.* (2020b), and Li *et al.* (2023c).

5

Online RL

We now turn to online reinforcement learning, where an agent interacts with an unknown environment sequentially to collect data and learn a near-optimal policy. In contrast to the generative model setting — where arbitrary state–action pairs can be queried on demand — an agent in online RL must collect data by executing policies and observing trajectories in real time. This restriction couples data collection with decision-making, necessitating carefully designed exploration strategies that acquire informative samples without sacrificing long-term performance. Consequently, a central challenge of online RL lies in optimally balancing exploration and exploitation under a limited sampling budget.

5.1 Problem formulation

In this chapter, we focus primarily on online episodic RL formulated in finite-horizon nonstationary MDPs, a setting that has been studied most extensively and for which the statistical limits have been characterized even in the most sample-limited regime. Its counterpart in discounted infinite-horizon MDPs remains not fully solved to date (see, e.g., Lattimore and Hutter (2012), Strehl *et al.* (2006), Dong *et al.* (2019), Zhang *et al.* (2021f), Szita and Szepesvári (2010), Zhou *et al.* (2021b), and Ji and Li (2023)) and calls for further investigation.

Sampling protocol. To be precise, we model the unknown environment as a finite-horizon nonstationary MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$. For ease of exposition, we assume that the reward function $r = \{r_h\}_{h=1}^H$ is fully known to the agent, an assumption that can be relaxed substantially without fundamentally altering the underlying statistical limits and incurs only minor additional technical complications. By contrast, the transition kernel $P = \{P_h\}_{h=1}^H$ is assumed to be unknown *a priori*, constituting the primary source of uncertainty in the online RL setting studied in this chapter. The agent is permitted to interact with the MDP over K episodes, each of length H .

In each episode $k = 1, \dots, K$, the sampling protocol of an online RL algorithm proceeds as follows:

- (i) *Initialization.* The initial state s_1^k of each episode k is drawn independently according to

$$s_1^k \stackrel{\text{i.i.d.}}{\sim} \rho, \quad k = 1, \dots, K, \quad (5.1a)$$

where $\rho \in \Delta(\mathcal{S})$ denotes an (unknown) initial state distribution.

- (ii) *Policy selection.* Based on the information collected in episodes 1 through $k - 1$, the agent selects a policy $\pi^k = \{\pi_h^k\}_{h=1}^H$.
- (iii) *Sampling.* The agent executes policy π^k in the MDP $\mathcal{M}$, generating a sample trajectory $\{(s_h^k, a_h^k, r_h^k)\}_{1 \leq h \leq H}$, where s_h^k , a_h^k and $r_h^k = r_h(s_h^k, a_h^k)$ denote the state, action and immediate reward at step h of this episode, respectively. Here, starting from s_1^k , the trajectory is rolled out sequentially for $h = 1, \dots, H$ according to

$$a_h^k \sim \pi_h^k(\cdot | s_h^k) \quad \text{and} \quad s_{h+1}^k \sim P_h(\cdot | s_h^k, a_h^k). \quad (5.1b)$$

For notational convenience, we set $r_{H+1} = 0$ throughout this chapter. Consequently, learning the terminal transition kernel P_H is unnecessary, as all actions yield zero rewards at step $H + 1$. Since each trajectory has length H , the total sample size can be understood as

$$T = KH.$$

Exploration-exploitation tradeoff. A fundamental challenge in online RL arises from the exploration–exploitation tradeoff: the agent must collect informative samples to learn the unknown transition kernel, while simultaneously seeking policies that maximize cumulative rewards. Compared to the generative model setting, at least two aspects significantly complicate matters.

- *Coverage of the state-action space.* Effective learning typically requires adequate coverage of the state–action space, which is difficult to guarantee under online sampling protocols. In particular, the agent does not know *a priori* which regions of the state–action space are reachable through sample rollouts, nor how to reach them. This contrasts sharply to the generative model setting, where one can query arbitrary state–action pairs at any step of the MDP.
- *Complicated statistical dependency across time.* Samples in online RL are collected sequentially and depend on the agent’s past decisions. This creates complex statistical dependencies across both samples and time. In contrast, the samples studied in Chapters 3 and 4 under the generative model are statistically independent, a feature that greatly simplifies both algorithm design and theoretical analysis.

Performance metric: regret. To evaluate the performance of an online RL algorithm, arguably one of the most widely used metrics is the (cumulative) regret over all K episodes:

$$\text{Regret}(K) := \sum_{k=1}^K \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right). \quad (5.2)$$

This metric quantifies the aggregate performance gap — expressed in terms of the resulting value functions — between the agent and an oracle that always acts optimally. Importantly, regret

provides a natural way to formalize the exploration–exploitation tradeoff: it captures both the cost of exploration, since exploratory actions may yield lower rewards, and the cost of insufficient exploitation, as inaccurate model estimates arising from limited data can adversely affect decision-making. Achieving low regret implies that the algorithm can efficiently identify near-optimal policies while keeping the cumulative performance loss under control.

The main focus of this chapter is the design of online RL algorithms with provably small cumulative regret $\text{Regret}(K)$.

Connection between regret and PAC statistical guarantees. Careful readers may observe that the policies $\{\pi^k\}_{1 \leq k \leq K}$ deployed across episodes are generally different, which naturally raises the question of whether one can instead learn a single policy that attains near-optimal values. Indeed, this question was the primary focus of Chapters 3–4 and falls within the PAC (probably approximately correct) learning framework, which seeks a single policy estimate $\hat{\pi}$ satisfying

$$V_1^*(\rho) - V_1^{\hat{\pi}}(\rho) \leq \varepsilon.$$

Here, we recall that ρ denotes the state distribution generating the initial states $\{s_1^k\}_{1 \leq k \leq K}$ (see (5.1a)). Fortunately, regret guarantees can often be readily translated into corresponding PAC learning guarantees (see, e.g., Jin *et al.* (2018)).

To illustrate this connection, consider an online RL algorithm that, for any MDP with S states, A actions and horizon H , satisfies the expected regret bound

$$\mathbb{E}[\text{Regret}(K)] \leq C_{S,A,H} K^{1-\alpha} \quad (5.3a)$$

for some $\alpha \in [0, 1]$ and some quantity $C_{S,A,H}$ that may depend on (S, A, H) . Here, the expectation is taken over all randomness arising from the initial states, sample transitions, and the learning algorithm’s internal randomization. Now construct the policy estimate $\hat{\pi}$ by selecting one of the deployed policies uniformly at random:

$$\hat{\pi} \sim \text{Unif}(\pi^1, \pi^2, \dots, \pi^K). \quad (5.3b)$$

Then with probability at least $1 - \delta$,

$$V_1^*(\rho) - V_1^{\hat{\pi}}(\rho) \leq C_{S,A,H+1} K^{-\alpha} \delta^{-1}. \quad (5.3c)$$

Proof of property (5.3c). Let us first consider the special case where the initial state distribution ρ is supported on a single state s , namely, $\rho(s) = 1$ and $\rho(s') = 0$ for all $s' \neq s$. Markov’s inequality asserts that, with probability at least $1 - \delta$,

$$\begin{aligned} V_1^*(s) - V_1^{\hat{\pi}}(s) &\leq \delta^{-1} \mathbb{E}[V_1^*(s) - V_1^{\hat{\pi}}(s)] = \frac{1}{K\delta} \sum_{k=1}^K \mathbb{E}[V_1^*(s) - V_1^{\pi^k}(s)] \\ &= \frac{1}{K\delta} \mathbb{E}[\text{Regret}(K)] \leq C_{S,A,H} K^{-\alpha} \delta^{-1}, \end{aligned}$$

where the first equality follows from the definition (5.3b) of $\hat{\pi}$.

We now extend the argument to an arbitrary initial state distribution ρ . To this end, we augment the MDP by introducing an additional step 0, consisting of a single state s_0 , a single action a_0 , zero immediate reward (i.e., $r_0(s_0, a_0) = 0$), and transition kernel $P_0(\cdot | s_0, a_0) = \rho$. Since there is only one available action, this additional step involves no decision-making; accordingly, with a

slight abuse of notation, we continue to denote by π the obvious extension of a policy π for the original MDP and write V_0^π for its value function starting from step 0. The online RL algorithm can therefore be applied to the augmented MDP without modification, yielding

$$\mathbb{E}[\text{Regret}(K)] = \sum_{k=1}^K \mathbb{E}[V_0^\star(s_0) - V_0^{\pi^k}(s_0)] \leq C_{S,A,H+1} K^{1-\alpha} \quad (5.4)$$

Given that the augmented MDP has a singleton initial state, the preceding argument applies directly, implying that

$$V_0^\star(s_0) - V_0^{\hat{\pi}}(s_0) \leq C_{S,A,H+1} K^{-\alpha} \delta^{-1} \quad (5.5)$$

with probability at least $1 - \delta$. To finish up, observe that for every policy π , one has $V_0^\pi(s_0) = V_1^\pi(\rho)$, as the additional initial step incurs zero reward and admits only a single action. This together with (5.5) completes the proof. $\square$

This simple argument demonstrates that sublinear regret naturally yields sample-efficient PAC statistical guarantees. For example, if an online RL algorithm enjoys the regret bound $\mathbb{E}[\text{Regret}(K)] \leq \tilde{O}(\sqrt{SAH^3K})$ for any MDP with S states, A actions and horizon length H , then the randomly selected policy $\hat{\pi}$ achieves (by choosing $\delta = \frac{1}{\log(SAH)}$ in (5.3c))

$$V_1^\star(s) - V_1^{\hat{\pi}}(s) \leq \tilde{O}\left(\sqrt{\frac{SAH^3}{K}}\right)$$

with probability $1 - o(1)$; consequently, to ensure $V_1^\star(s) - V_1^{\hat{\pi}}(s) \leq \varepsilon$ with probability approaching 1, it suffices to run this online RL algorithm for

$$K = \tilde{\Theta}\left(\frac{SAH^3}{\varepsilon^2}\right) \text{ episodes.}$$

5.2 A detour: the UCB algorithm for multi-armed bandits

Before embarking on the design and analysis of online RL algorithms, we take a brief detour to study multi-armed bandits, a special case of online RL where state transitions are absent or $S = 1$. In this setting, there is no transition kernel to estimate; instead, we consider the scenario where the observed rewards are noisy, so that model uncertainty arises solely from the stochasticity of the rewards. As it turns out, several key ideas developed in the multi-armed bandit literature pave the way for the design of efficient online RL algorithms.

5.2.1 Setting: stochastic bandits

Consider a standard stochastic bandit model, where an agent interacts sequentially with the unknown environment over T rounds. The key components of this bandit model are described as follows.

- *Actions.* The agent has access to a set of A arms (i.e., actions), denoted by $\mathcal{A} = \{1, \dots, A\}$. One can picture a gambler standing in front of a row of A slot machines, each with its own arm; at each round, the gambler needs to decide which machine to play.

- *Mean rewards.* Each arm $a \in \mathcal{A}$ is associated with a time-invariant mean reward $r(a) \in [0, 1]$. In the gambling analogy, each slot machine has its own payout rate. The optimal mean reward is denoted by

$$r^* := \max_{a \in \mathcal{A}} r(a), \quad (5.6a)$$

with the optimal arm defined correspondingly as

$$a^* := \arg \max_{a \in \mathcal{A}} r(a). \quad (5.6b)$$

Crucially, these mean rewards are *unknown* to the agent and can only be learned through sequential interaction with the environment over time.

- *Realized rewards.* At each round t , nature generates a noisy reward $R_{a,t}$ independently for each arm $a \in \mathcal{A}$, given by

$$R_{a,t} = r(a) + z_{a,t}, \quad \text{with } z_{a,t} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \quad (5.7)$$

While we assume i.i.d. standard Gaussian noise for simplicity of exposition, this assumption can be significantly relaxed without affecting the key algorithmic principles and insights (see, e.g., Lattimore and Szepesvári (2020)).

- *Observations (bandit feedback).* The agent operates under bandit feedback, meaning that it does not observe the full collection $\{R_{a,t}\}_{a \in \mathcal{A}}$ of noisy rewards at each round. Instead, at round t , the agent selects a single arm $a_t \in \mathcal{A}$ — potentially based on previous actions and observations — and observes only the corresponding realized reward $R_{a_t,t}$.

Naturally, the agent would aim to maximize its cumulative payoff over T rounds. To evaluate the performance of a bandit algorithm (which determines how actions are selected over time), we introduce the (cumulative) regret as

$$\text{Regret}(T) := Tr^* - \sum_{t=1}^T r(a_t), \quad (5.8a)$$

which measures the total expected loss in reward relative to always playing the optimal arm. Here, we employ the mean reward $r(a_t)$ rather than the realized reward $R_{a_t,t}$ to marginalize out observation noise, which ensures non-negativity of regret. Notably, $\text{Regret}(T)$ is itself a random variable, as the sequence of actions $\{a_t\}$ is a function of the history of noisy rewards.

Moreover, the regret admits a simple yet useful decomposition. Let $N_T(a)$ denote the number of times arm a is pulled over T rounds. Then the regret can be expressed in terms of $\{N_T(a)\}$ and the suboptimality gaps:

$$\text{Regret}(T) = \sum_{t=1}^T \Delta(a_t) = \sum_{a \in \mathcal{A}} \Delta(a) N_T(a) \quad \text{with } \Delta(a) := r^* - r(a). \quad (5.8b)$$

Given that the mean rewards are unknown *a priori*, the agent must balance exploring undersampled arms with exploiting those currently deemed favorable. Despite its apparent simplicity, this classical bandit formulation already encapsulates a central challenge of sequential decision-making under uncertainty: the interplay between learning, decision-making, and data collection in real time.

5.2.2 Optimism in the face of uncertainty

We now introduce a core design principle underlying a broad family of bandit algorithms. On the one hand, if an arm a has been pulled only a small number of times, the corresponding estimate of its mean reward r_a can be highly volatile and therefore unreliable; this motivates playing the arm more often in order to improve the accuracy of the reward estimate. On the other hand, if the estimated reward of arm a is lower than that of other arms, this arm is likely suboptimal, suggesting that it should be pulled less frequently. In summary,

- (i) high uncertainty (i.e., low confidence) in an arm's reward estimate can incentivize further exploration of that arm;
- (ii) a large reward estimate encourages more frequent pulls of that arm.

One strategy to unify these two competing principles is to introduce an optimistic estimate of the form:

$$\hat{r}(a) = \text{empirical reward estimate for arm } a + \text{its uncertainty},$$

and then select the arm with the highest value of $\hat{r}(a)$, provided that this quantity is constructed appropriately. Since $\hat{r}(a)$ acts as an optimistic estimate of the reward $r(a)$ (in the sense that it intentionally biases the estimate upward to account for uncertainty), this idea is commonly referred to as the *optimism principle* in the face of uncertainty, originally introduced by Lai and Robbins (1985).

While conceptually natural, this paradigm raises a fundamental question: how should the uncertainty bonus be calibrated relative to the empirical reward estimate? Evidently, the choice of this trade-off is crucial for achieving optimal regret and is, in general, far from obvious.

5.2.3 The upper confidence bound (UCB) algorithm

Somewhat surprisingly, a simple yet elegant approach — known as the upper confidence bound (UCB) algorithm, originally developed by Lai and Robbins (1985) and Lai (1987) — is sufficient to achieve the optimal exploration-exploitation trade-off.

More precisely, the UCB algorithm operates by selecting, at each round t , the arm with the largest upper confidence bound on its mean reward, where these confidence bounds are constructed using the samples observed up to and including time t . Specifically, for each round t , the agent pulls the arm

$$a_t = \arg \max_{a \in \mathcal{A}} r_{t-1}^{\text{ucb}}(a), \quad (5.9)$$

where $r_t^{\text{ucb}}(a)$ represents the UCB estimate for arm a , defined as

$$r_t^{\text{ucb}}(a) = \bar{r}_t(a) + \sqrt{\frac{2 \log \frac{T^2}{\delta}}{N_t(a)}}. \quad (5.10a)$$

Here, $0 < \delta < 1$ is some confidence parameter (whose role will become clear shortly), $N_t(a)$ denotes the number of times arm a has been pulled up to round t , and $\bar{r}_t(a)$ represents the empirical mean reward of arm a computed from all observations collected so far:

$$\bar{r}_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^t R_{a,i} \mathbb{1}(\text{arm } a \text{ is pulled at round } i). \quad (5.10b)$$

Algorithm 5.1: The upper confidence bound (UCB) algorithm for online stochastic bandits

```

1 input: number of rounds  $T$ , confidence parameter  $\delta$ .
2 initialize  $N(a) \leftarrow 0$ ,  $r^{\text{sum}}(a) \leftarrow 0$ , and  $\bar{r}(a) \leftarrow 0$  for all  $a \in \mathcal{A}$ .
   // pull arms based on UCBs.
3 for  $t = 1, \dots, T$  do
   /* add bonus. */
4   compute  $a_t \in \arg \max_{a \in \mathcal{A}} r^{\text{ucb}}(a)$ , where
      
$$r^{\text{ucb}}(a) \leftarrow \bar{r}(a) + \sqrt{\frac{2 \log \frac{T^2}{\delta}}{N(a)}}.$$

5   pull arm  $a_t$  at round  $t$ , and receive reward  $R_{a_t, t}$ .
6   set  $N(a_t) \leftarrow N(a_t) + 1$ ,  $r^{\text{sum}}(a_t) \leftarrow r^{\text{sum}}(a_t) + R_{a_t, t}$ , and  $\bar{r}(a_t) \leftarrow r^{\text{sum}}(a_t)/N(a_t)$ .

```

The complete procedure of this algorithm is summarized in Algorithm 5.1.

The specific form of the uncertainty term $\sqrt{\frac{2 \log(T^2/\delta)}{N_t(a)}}$ in (5.10a) — often dubbed the exploration bonus — is motivated by the use of concentration inequalities to quantify statistical uncertainty. To make this precise, recall from standard Gaussian tail bounds that a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$ satisfies (see, e.g., Vershynin (2018, Chapter 2))

$$\mathbb{P}(Z \leq \mu - \tau) \leq \exp\left(-\frac{\tau^2}{2\sigma^2}\right)$$

for any $\tau \geq 0$, and as a result,

$$\mathbb{P}\left(Z + \sigma\sqrt{2 \log \frac{1}{\delta}} \geq \mu\right) > 1 - \delta$$

for any $0 < \delta < 1$. Now observe that, under the Gaussian reward model (5.7), one has $\bar{r}_t(a) - r(a) = \mathcal{N}(0, \frac{1}{N_t(a)})$. Applying the above Gaussian tail bound thus yields

$$\mathbb{P}(r_t^{\text{ucb}}(a) \geq r(a)) = \mathbb{P}\left(\bar{r}_t(a) + \sqrt{\frac{2 \log \frac{T^2}{\delta}}{N_a(t)}} \geq r(a)\right) > 1 - \frac{\delta}{T^2}, \quad (5.11a)$$

and similarly,

$$\mathbb{P}\left(r(a) + \sqrt{\frac{2 \log \frac{T^2}{\delta}}{N_t(a)}} \geq \bar{r}_t(a)\right) > 1 - \frac{\delta}{T^2}. \quad (5.11b)$$

Consequently, the exploration bonus serves as a high-probability upper bound on the reward estimation error, while $r_t^{\text{ucb}}(a)$ acts as a statistically valid upper confidence bound on the unknown mean reward $r(a)$ — hence the name of the algorithm. This interpretation makes clear that the UCB algorithm makes decision according to an optimistic, yet statistically plausible, view of their mean rewards.

5.2.4 Theoretical guarantees

To rigorously demonstrate the efficacy of the UCB algorithm, we present its regret guarantees in the following theorem; the proof is provided in Section 5.2.5.

Theorem 5.1. Consider any $0 < \delta < 1$, and suppose that $T \geq A \geq 2$. Recall that $\Delta(a) = r^* - r(a) \in [0, 1]$ for all $a \in \mathcal{A}$. Then the expected regret of the UCB algorithm (see Algorithm 5.1) satisfies

$$\mathbb{E}[\text{Regret}(T)] \leq 8\sqrt{AT \log\left(\frac{T^2}{\delta}\right)}. \quad (5.12a)$$

In addition, it also holds that

$$\mathbb{E}[\text{Regret}(T)] \leq 10 \log\left(\frac{T^2}{\delta}\right) \cdot \sum_{a: \Delta(a) > 0} \frac{1}{\Delta(a)}. \quad (5.12b)$$

Remark 5.1. These results can also be stated as high-probability regret bounds (Lattimore and Szepesvári, 2020). However, since bandits are not the primary focus of this monograph and are introduced mainly to motivate certain algorithmic ideas, we restrict our attention to the expected regret bounds for brevity.

This theorem provides two complementary results: a worst-case regret bound (5.12a) that holds uniformly over all problem instances, and a more refined instance-dependent bound (5.12b) that explicitly captures the role of the suboptimality gaps $\{\Delta(a)\}_{a \in \mathcal{A}}$. A few remarks about these regret guarantees are in order.

- Assuming that the number of arms is $A = O(1)$, then the worst-case regret bound (5.12a) simplifies to

$$\mathbb{E}[\text{Regret}(T)] \leq \tilde{O}(\sqrt{T}).$$

In the language of online learning (Shalev-Shwartz, 2012), this implies that UCB is a *no-regret* learning algorithm, meaning that the average regret per round, $\frac{1}{T}\mathbb{E}[\text{Regret}(T)]$, vanishes asymptotically as $T \rightarrow \infty$. In other words, although the agent may initially incur losses while exploring uncertain arms, its long-term performance eventually approaches that of the optimal arm. Encouragingly, the regret bound (5.12a) is minimax-optimal up to logarithmic factors, revealing that the optimism principle achieves essentially optimal worst-case performance. In fact, the logarithmic factor in (5.12a) can be shaved through more refined algorithmic and analysis techniques (see Audibert and Bubeck (2009)). See, e.g., Lattimore and Szepesvári (2020, Section 15.2) for a matching lower bound.

- The instance-dependent regret bound (5.12b), which is often tighter than the worst-case bound (5.12a), reveals the intrinsic difficulty of a bandit instance through the suboptimality gaps. In particular, this regret bound scales inversely with the *harmonic mean* of the gaps:

$$\mathbb{E}[\text{Regret}(T)] \leq \tilde{O}\left(\frac{A}{\text{harmonic-mean}(\{\Delta(a)\}_{a \neq a^*})}\right),$$

where we define, for any sequence of positive numbers $\{x_i\}_{1 \leq i \leq n}$,

$$\text{harmonic-mean}(\{x_i\}_{1 \leq i \leq n}) := \frac{n}{x_1^{-1} + \dots + x_n^{-1}}.$$

For instance, when $\Delta(a) \asymp 1$ holds for all suboptimal arms and when $A = O(1)$, the expected regret bound reduces to

$$\mathbb{E}[\text{Regret}(T)] \leq \tilde{O}(1),$$

illustrating that, for favorable problem instances, the regret can be considerably smaller than the worst-case $\tilde{O}(\sqrt{T})$ scaling.

Intuitive explanation. Before presenting the proof of Theorem 5.1, we take a moment to provide some intuition for the regret bounds. Recall the regret decomposition in (5.8b):

$$\mathbb{E}[\text{Regret}(T)] = \sum_{a \in \mathcal{A}} \Delta(a) \mathbb{E}[N_T(a)], \quad (5.13)$$

where, by a slight abuse of notation, $N_T(a)$ denotes the total number of pulls of arm a over the T rounds. This suggests controlling $\mathbb{E}[N_T(a)]$ for each arm a . As it turns out, a key step of the proof is to show that, for any suboptimal arm a ,

$$\mathbb{E}[N_T(a)] \leq \tilde{O}\left((\Delta(a))^{-2}\right).$$

This scaling is intuitive. If $N_T(a)$ becomes significantly larger than $(\Delta(a))^{-2}$, then the empirical estimate $\bar{r}(a)$ concentrates sharply around the true mean $r(a)$, making it distinguishable from that of the optimal arm a^* . More precisely, recall that the reward estimation error $\bar{r}_a - r(a) \sim \mathcal{N}(0, 1/N_T(a))$ and is therefore on the order of $1/\sqrt{N_T(a)}$; once this uncertainty is much smaller than the suboptimality gap $\Delta(a) = r^* - r(a)$, the learner can reliably identify arm a as suboptimal and will no longer select it. Consequently, the summand $\Delta(a)\mathbb{E}[N_T(a)]$ in (5.13) obeys

$$\Delta(a)\mathbb{E}[N_T(a)] \leq \Delta(a)\tilde{O}\left((\Delta(a))^{-2}\right) = \tilde{O}(1/\Delta(a)).$$

Summing over all arms suggests the claimed instance-dependent regret bound (5.12b).

5.2.5 Analysis: proof of Theorem 5.1

To begin with, we observe that the first A rounds constitute an initialization phase, during which each arm is pulled once and the corresponding noisy reward is observed. Indeed, if an arm a has never been selected, then $N(a) = 0$ and hence $r^{\text{ucb}}(a) = \infty$. Consequently, the algorithm always selects an unsampled arm whenever one exists, so that each arm must be pulled exactly once during the first A rounds, yielding an initial empirical mean estimate for every arm.

In what follows, we prove Theorem 5.1 with the aid of the regret decomposition (5.13), focusing on $t \geq A$. For notational clarity, we use $N_t(a)$, $r_t^{\text{ucb}}(a)$, and $\bar{r}_t(a)$ to denote the values of the running variables $N(a)$, $r^{\text{ucb}}(a)$, and $\bar{r}(a)$, respectively, at the end of round t .

Step 1: connecting $N_t(a)$ with the arm-pulling probability. Consider any given $a \in \mathcal{A}$. We first identify a condition under which the UCB estimate $r_t^{\text{ucb}}(a)$ of a suboptimal arm a is smaller than that of the optimal arm a^* (cf. (5.6b)), and hence arm a will not be pulled at the beginning of round $t + 1$. More specifically, if

$$N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} \quad (5.14)$$

holds for $t \geq A$, then with probability at least $1 - 2\delta/T^2$,

$$r_t^{\text{ucb}}(a) = \bar{r}_t(a) + \sqrt{\frac{2 \log \frac{T^2}{\delta}}{N_t(a)}} \stackrel{(i)}{\leq} r(a) + 2\sqrt{\frac{2 \log \frac{T^2}{\delta}}{N_t(a)}} \stackrel{(ii)}{<} r(a^*) \stackrel{(iii)}{\leq} r_t^{\text{ucb}}(a^*),$$

where (i) results from (5.11b), (ii) is a consequence of Condition (5.14), and (iii) follows from (5.11a). As a result, for any $t \geq A$ and any given $a \in \mathcal{A}$,

$$\begin{aligned} \mathbb{P}\left(a_{t+1} = a, N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right) &= \mathbb{P}\left(a_{t+1} = a \mid N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right) \mathbb{P}\left(N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right) \\ &\leq \mathbb{P}\left(a_{t+1} = a \mid N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right) \leq \frac{2\delta}{T^2}. \end{aligned} \quad (5.15)$$

Step 2: deriving an instance-dependent bound on $\mathbb{E}[N_T(a)]$. Equipped with the properties derived in Step 1, we can demonstrate that, for any arm a with $\Delta(a) > 0$,

$$\begin{aligned} \mathbb{E}[N_T(a)] &= 1 + \mathbb{E}\left[\sum_{t=A}^{T-1} \mathbb{1}(a_{t+1} = a)\right] \\ &= 1 + \mathbb{E}\left[\sum_{t=A}^{T-1} \mathbb{1}\left(a_{t+1} = a, N_t(a) \leq \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right)\right] + \mathbb{E}\left[\sum_{t=A}^{T-1} \mathbb{1}\left(a_{t+1} = a, N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right)\right] \\ &\leq 1 + \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} + \sum_{t=A}^{T-1} \mathbb{P}\left(a_{t+1} = a, N_t(a) > \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2}\right) \\ &\leq 1 + \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} + \sum_{t=A}^{T-1} \frac{2\delta}{T^2} \\ &\leq \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} + 2. \end{aligned} \quad (5.16)$$

Here, the first inequality follows because arm a is pulled exactly once during the first A rounds, and thereafter the event

$$\left\{a_{t+1} = a, N_t(a) \leq \frac{8 \log(T^2/\delta)}{(\Delta(a))^2}\right\}$$

can occur at most $\frac{8 \log(T^2/\delta)}{(\Delta(a))^2}$ times, as each occurrence increments $N_t(a)$ by one. The penultimate line in (5.16) results from (5.15), whereas the last inequality in (5.16) holds since $\delta \leq 1$ and $T \geq A \geq 2$.

Step 3: establishing the instance-dependent regret bound. Substituting (5.16) into the regret decomposition (5.13) establishes the advertised instance-dependent regret bound (i.e., (5.12b)):

$$\mathbb{E}[\text{Regret}(T)] \leq \sum_{a: \Delta(a) > 0} \Delta(a) \left(\frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} + 2 \right) \leq \sum_{a: \Delta(a) > 0} \frac{10 \log \frac{T^2}{\delta}}{\Delta(a)},$$

provided that $\Delta(a) \leq 1$ holds for all $a \in \mathcal{A}$.

Step 4: establishing the worst-case regret bound. With the instance-dependent bound (5.16) on $\mathbb{E}[N_T(a)]$ in place, we can proceed to establish the claimed worst-case regret bound (5.12a). Considering any threshold $\Delta > 0$ (to be specified shortly), we make the following observations.

- For any arm a with $\Delta(a) \geq \Delta$ (i.e., large suboptimality gap), it follows from (5.16) that

$$\mathbb{E}[N_T(a)] \leq \frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} + 2.$$

- For the set of arms with $\Delta(a) < \Delta$ (i.e., small suboptimality gap), it holds trivially that

$$\sum_a \mathbb{E}[N_T(a)] \leq T.$$

Combine these bounds with the regret decomposition (5.13) to reach

$$\begin{aligned} \mathbb{E}[\text{Regret}(T)] &\leq \sum_{a: \Delta(a) \geq \Delta} \Delta(a) \left(\frac{8 \log \frac{T^2}{\delta}}{(\Delta(a))^2} + 2 \right) + \Delta \sum_{a: \Delta(a) < \Delta} \mathbb{E}[N_T(a)] \\ &\leq \frac{8A \log \frac{T^2}{\delta}}{\Delta} + 2A + T\Delta, \end{aligned}$$

where the last line also uses the assumption that $\Delta(a) \leq 1$ for each $a \in \mathcal{A}$. Choosing $\Delta = \sqrt{\frac{4A \log(T^2/\delta)}{T}}$, we arrive at

$$\mathbb{E}[\text{Regret}(T)] \leq 6\sqrt{AT \log \frac{T^2}{\delta}} + 2A \leq 8\sqrt{AT \log \frac{T^2}{\delta}},$$

provided that $T \geq A$. This completes the proof.

5.3 A model-based online RL algorithm: UCB-VI

Recall that in the generative model setting, model-based algorithms have been shown to achieve minimax-optimal sample complexity, even in the most data-hungry regime. Such results provide strong incentives for studying the model-based paradigm in the online RL setting as well. Given the proven efficiency of the optimism principle in multi-armed bandits (see Section 5.2), it is natural to investigate how this principle can be effectively integrated into model-based online RL algorithms.

5.3.1 Algorithmic ideas

To begin with, the model-based policy learning algorithm introduced previously consists of a model estimation stage, where an empirical estimate of the probability transition kernel is constructed from collected data, followed by a planning stage that computes a policy by treating the empirical model as if it were the true environment. This naturally suggests a paradigm that integrates model-based policy learning with online decision-making and sampling, which operates as follows in each episode.

- *Model estimation.* Use all previously collected data to construct an empirical estimate $\hat{P}$ of the transition kernel.
- *Policy update.* Compute an optimal policy $\hat{\pi}$ w.r.t. the updated empirical model $\hat{P}$, for instance via dynamic programming.

- *Sampling.* Execute the updated policy $\hat{\pi}$ to obtain a new sample trajectory at the beginning of the next episode.

While conceptually intuitive, this approach lacks one crucial component: exploration. By focusing exclusively on exploitation — optimizing solely w.r.t. the current empirical model — the agent may fail to sufficiently explore regions of the state-action space that are poorly understood yet potentially highly rewarding.

To address this issue, a natural strategy is to incorporate the optimism principle in the face of uncertainty into the above framework, namely, to guide exploration by acting according to an optimistic view of the world. Compared with stochastic bandits (Section 5.2), where uncertainty arises primarily from noisy rewards, the online RL formulation considered herein faces uncertainty stemming from unknown transition kernels. Consequently, there are multiple objects around which one may construct confidence regions, including the transition kernel itself and the corresponding optimal value and Q-functions. The former (i.e., confidence regions for the transition kernel) led to some of the earliest UCB-style algorithms for online RL, most notably UCRL and UCRL2 (Auer and Ortner, 2006; Jaksch *et al.*, 2010), whereas the latter (i.e., optimism over the optimal value and Q-function) led to the development of UCB-VI (Azar *et al.*, 2017), the first online RL algorithm achieving asymptotically near-optimal regret. Since the primary focus of this work is sample efficiency, we concentrate on the UCB-VI algorithm and its variants.

In a nutshell, the UCB-VI algorithm — short for *upper confidence bound value iteration* and originally proposed by Azar *et al.* (2017) — implements the optimism principle by constructing upper confidence bounds on the optimal Q-function in each episode, and subsequently performing policy updates and data collection in accordance with these optimistic Q-value estimates rather than the unadjusted empirical estimates. To be more precise, in each episode k , the algorithm proceeds as follows, assuming the Q-value estimates are initialized optimistically as $Q_h^{\text{ucb}}(s, a) = H$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

- *Sampling.* Compute the current policy $\pi^{\text{ucb}} = \{\pi_h^{\text{ucb}}\}_{1 \leq h \leq H}$ w.r.t. the optimistic Q-estimates $\{Q_h^{\text{ucb}}\}_{1 \leq h \leq H}$ via a greedy rule, i.e.,

$$\pi_h^{\text{ucb}}(s) \leftarrow \arg \max_a Q_h^{\text{ucb}}(s, a), \quad \forall (s, h) \in \mathcal{S} \times [H].$$

Armed with the updated policy estimate π^{ucb} , the agent executes π^{ucb} in the MDP at the beginning of the episode, collecting a sample trajectory

$$(s_1^k, a_1^k, \dots, s_H^k, a_H^k, s_{H+1}^k).$$

- *Model estimation.* Compute an empirical estimate $\hat{P} = \{\hat{P}_h \in \mathbb{R}^{\mathcal{S} \times \mathcal{A} \times \mathcal{S}}\}_{1 \leq h \leq H}$ of the transition probability kernel using all previously collected samples. More specifically, the empirical model $\hat{P}$ is constructed such that, for all $(s, a, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$,

$$\hat{P}_h(s' | s, a) = \frac{N_h(s, a, s')}{N_h(s, a)}, \quad (5.17)$$

where $N_h(s, a)$ denotes the number of visits to the state-action-step tuple (s, a, h) up to the current episode, and $N_h(s, a, s')$ denotes the number of such visits that result in a transition to state s' at step $h + 1$. Throughout this section, we often use the notation $\hat{P}_{s,a,h} := [\hat{P}_h(s' | s, a)]_{s' \in \mathcal{S}}$ to denote the empirical transition vector associated with the state-action pair (s, a) at step h . If the tuple (s, a, h) has not been visited, one may simply set $\hat{P}_{s,a,h}(s') = 1/S$ for all $s' \in \mathcal{S}$, so that $\hat{P}_{s,a,h}$ remains a valid probability distribution.

Algorithm 5.2: The upper confidence bound value iteration (UCB-VI) algorithm for episodic online RL

```

1 input: number of episodes  $K$ , reward function  $r$ , confidence parameter  $\delta$ , constants  $c_1, c_2 > 0$ .
2 initialize  $N_h(s, a, s') \leftarrow 0$ ,  $N_h(s, a) \leftarrow 0$ ,  $Q_h^{\text{ucb}}(s, a) \leftarrow H$ ,  $V_h^{\text{ucb}}(s) \leftarrow H$ ,  $V_{H+1}^{\text{ucb}}(s) \leftarrow 0$ ,
    $\hat{P}_{s,a,h}(s') \leftarrow 1/S$  for all  $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ .
3 for  $k = 1, \dots, K$  do
   /* policy update. */
4   compute policy  $\pi^{\text{ucb}}$  such that  $\pi_h^{\text{ucb}}(s) \leftarrow \arg \max_a Q_h^{\text{ucb}}(s, a)$  for all  $(s, h) \in \mathcal{S} \times [H]$ .
5   observe the initial state  $s_1 \stackrel{\text{ind.}}{\sim} \rho$  of a new episode.
6   for  $h = 1, \dots, H$  do
   /* sampling. */
7   take  $a_h = \pi_h^{\text{ucb}}(s_h)$ , observe  $s_{h+1}$ ; set  $(s, a, s') \leftarrow (s_h, a_h, s_{h+1})$ .
   /* empirical model update. */
8    $N_h(s, a) \leftarrow N_h(s, a) + 1$ ;  $N_h(s, a, s') \leftarrow N_h(s, a, s') + 1$ .
9    $\hat{P}_{s,a,h}(\tilde{s}) \leftarrow \frac{N_h(s,a,\tilde{s})}{N_h(s,a)}$  for all  $\tilde{s} \in \mathcal{S}$ .
   /* optimistic updates of Q-functions & values. */
10  for  $h = H, \dots, 1$  do
11    for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
12      compute the bonus  $b_h(s, a)$  according to (5.20); update
      
$$Q_h^{\text{ucb}}(s, a) \leftarrow \min \{ r_h(s, a) + \langle \hat{P}_{s,a,h}, V_{h+1}^{\text{ucb}} \rangle + b_h(s, a), H \}, \quad (5.18a)$$

      
$$V_h^{\text{ucb}}(s) \leftarrow \max_a Q_h^{\text{ucb}}(s, a). \quad (5.18b)$$


```

- *Optimistic Q-function estimation.* Maintain upper confidence estimates of the optimal Q-function and value function. More concretely, working backward for $h = H, \dots, 1$, compute for all state-action pairs

$$\begin{aligned}
Q_h^{\text{ucb}}(s, a) &\leftarrow \min \{ r_h(s, a) + \langle \hat{P}_h(\cdot | s, a), V_{h+1}^{\text{ucb}} \rangle + b_h(s, a), H \} \\
V_h^{\text{ucb}}(s) &\leftarrow \max_a Q_h^{\text{ucb}}(s, a)
\end{aligned}$$

from the initial value $V_{H+1}^{\text{ucb}} = 0$, where Q_h^{ucb} (resp. V_h^{ucb}) indicates the running estimate for the Q-function (resp. value function), and $b_h(s, a) \geq 0$ is some suitably chosen exploration bonus that accounts for statistical uncertainty. Here, the iterate $Q_h^{\text{ucb}}(s, a)$ is explicitly capped at H to reflect the known feasible range $[0, H]$.

The full procedure is described in Algorithm 5.2.

At a high level, the UCB-VI algorithm makes decisions adaptively by favoring actions with either large estimated Q-values or high uncertainty in the Q-value estimates, thereby balancing exploitation with exploration in a principled manner. From a computational standpoint, each episode requires running a value-iteration-type procedure, whose complexity is roughly comparable to that of solving a fully specified finite-horizon MDP. Consequently, the overall computational complexity of UCB-VI is polynomial in the problem parameters.

5.3.2 Choices of exploration bonuses

A crucial design element of UCB-VI is the exploration bonus $b_h(s, a)$, which needs to be carefully constructed to ensure a valid, yet tight, upper confidence bound on the optimal Q-function (i.e., $Q_h^{\text{ucb}}(s, a) \geq Q_h^*(s, a)$). As is evident from the algorithm description, the primary source of uncertainty in Q-function estimation in each iteration of dynamic programming arises from the term $\langle \hat{P}_h(\cdot | s, a), V_{h+1}^{\text{ucb}} \rangle$ appearing in (5.18a), whereas the ideal quantity would be the population-level counterpart $\langle P_h(\cdot | s, a), V_{h+1}^{\text{ucb}} \rangle$. Two common choices for designing $b_h(s, a)$ are obtained by applying the Hoeffding inequality or the Bernstein inequality to control the deviation

$$\langle \hat{P}_h(\cdot | s, a) - P_h(\cdot | s, a), V_{h+1}^{\text{ucb}} \rangle, \quad (5.19)$$

which we introduce below. Throughout, we let $\delta > 0$ denote a confidence parameter (so that $1 - \delta$ is the target success probability), and recall that $N_h(s, a)$ is the visitation count of the state-action-step tuple (s, a, h) .

- *Hoeffding-style bonus.* This type of bonuses is comparatively simple, as it depends mainly on the range of the value function. More specifically, it typically takes the following form:

$$b_h(s, a) = \tilde{\Theta} \left(\sqrt{\frac{H^2}{N_h(s, a)}} \right).$$

This choice of upper confidence width is obtained by bounding the deviation term in (5.19) via the Hoeffding inequality (see Lemma 2.4) together with the fact that, by construction, the optimistic value estimate $V_{h+1}^{\text{ucb}}(s)$ falls within $[0, H]$.

- *Bernstein-style bonus.* In contrast, the Bernstein-type bonus incorporates variance information to form tighter confidence bounds, often leading to improved statistical guarantees compared to their Hoeffding-based counterparts. By applying the Bernstein inequality (see Lemma 2.5) and using the fact that $\|V_{h+1}^{\text{ucb}}\|_\infty \leq H$, we see that the deviation term (5.19) can be bounded, with high probability, by

$$\tilde{O} \left(\sqrt{\frac{\text{Var}_{P_{s,a,h}}(V_{h+1}^{\text{ucb}})}{N_h(s, a)}} + \frac{H}{N_h(s, a)} \right),$$

where we recall the definition of the variance term in (2.41a). Nevertheless, since the true transition kernel $P_{s,a,h}$ is unknown, this variance term $\text{Var}_{P_{s,a,h}}(\cdot)$ cannot be evaluated directly in practice. A natural remedy is to replace it with the empirical variance $\text{Var}_{\hat{P}_{s,a,h}}(\cdot)$ computed under the empirical model. This yields the following Bernstein-style exploration bonus:

$$b_h(s, a) = \tilde{\Theta} \left(\sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}}(V_{h+1}^{\text{ucb}})}{N_h(s, a)}} + \frac{H}{N_h(s, a)} \right)$$

arising from the empirical Bernstein inequality stated in Lemma 2.6.

We shall shortly see — to be formalized in Lemma 5.5 — that the above exploration bonuses suffice to guarantee, with high probability, that $Q_h^{\text{ucb}}(s, a) \geq Q_h^*(s, a)$ holds for every state-action-step tuple (s, a, h) , thereby offering valid upper confidence bounds.

Motivated by insights from the generative model setting, which highlight the importance of exploiting variance structure for data-efficient design, this monograph focuses on adopting and

analyzing Bernstein-style bonuses within UCB-VI and its variants. To facilitate precise discussion and analysis, we shall work with the following concrete choice throughout this chapter:

$$b_h(s, a) = c_1 \sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}}(V_{h+1}^{\text{ucb}}) \log \frac{1}{\delta'}}}{N_h(s, a)} + c_2 \frac{H \log \frac{1}{\delta'}}{N_h(s, a)}, \quad (5.20a)$$

where $c_1, c_2 > 0$ are some universal constants, and

$$\delta' = \frac{\delta}{100S^2A^2H^2K^2} \quad (5.20b)$$

is introduced for notational convenience in the analysis.

Remark 5.2. We note that the form of (5.20) differs from the Bernstein-style bonus proposed in the original work of Azar *et al.* (2017); instead, we adopt a simplified choice adopted by Zhang *et al.* (2021e) and Zhang *et al.* (2025e) in their design of MVP, an algorithm we will soon introduce in Section 5.4.

5.3.3 Theoretical guarantees

As alluded to previously, the development of the UCB-VI algorithm (Azar *et al.*, 2017) marked a milestone in online RL theory, leading to the first algorithm provably achieving *asymptotically* minimax-optimal regret. Its statistical guarantees are formally stated in the following theorem.

Theorem 5.2 (Azar *et al.* (2017)). Consider any $K \geq 1$ and any $0 < \delta < 1$. Then with probability at least $1 - \delta$, the UCB-VI algorithm in Algorithm 5.2 with Bernstein-style bonuses (5.20) achieves

$$\text{Regret}(K) \leq \tilde{O}(\sqrt{SAH^3K} + S^2AH^3), \quad (5.21)$$

provided that the constants $c_1, c_2 > 0$ are suitably chosen. Here, in the definition of $\text{Regret}(K)$, π^k and s_1^k denote, respectively, the sampling policy deployed and the initial state observed in episode k of Algorithm 5.2.

Remark 5.3. It is noteworthy that the regret bound presented in Azar *et al.* (2017) differs from that in Theorem 5.2 in its dependence on the horizon H , as the former focused on finite-horizon *stationary* MDPs. Intriguingly, this also unveils that the UCB-VI algorithm is asymptotically optimal for finite-horizon stationary MDPs as well.

Notably, once the number of episodes is sufficiently large, i.e.,

$$K \gtrsim S^3AH^3, \quad (5.22)$$

the lower-order term S^2AH^3 becomes negligible, and the regret bound in Theorem 5.2 simplifies to

$$\text{Regret}(K) \leq \tilde{O}(\sqrt{SAH^3K}). \quad (5.23)$$

As we will rigorize in Section 5.6, this matches the minimax lower bound up to logarithmic factors. Consequently, UCB-VI is asymptotically optimal in terms of regret.

Additionally, when $K \gtrsim S^3AH^3$, the above asymptotic regret bound (5.23) can be readily converted into a PAC-style sample complexity guarantee via the regret-to-PAC conversion in (5.3)

from Section 5.1. Specifically, if one outputs a policy $\hat{\pi}$ drawn uniformly at random from those sampling policies deployed over the K episodes, then

$$V_1^*(\rho) - V_1^{\hat{\pi}}(\rho) \leq \varepsilon$$

holds with probability approaching one, provided that the total number of collected samples satisfies

$$T = KH \geq \tilde{\Omega}\left(\frac{SAH^4}{\varepsilon^2}\right)$$

and ε is sufficiently small. Somewhat surprisingly, this PAC-style sample complexity coincides, up to logarithmic factors, with the minimax-optimal sample complexity limit under the generative model setting (see Section 3.5). In some sense, this result suggests that, at least in the asymptotic regime and from a minimax perspective, online exploration via optimism incurs little additional statistical cost compared to sampling from an idealized generative model.

Nevertheless, the statistical optimality of UCB-VI is guaranteed only after a substantial burn-in period of order $K \gtrsim S^3AH^3$, which can be prohibitively large in practical scenarios where the available number of episodes is limited. This limitation of the existing theoretical guarantees for UCB-VI has motivated a line of subsequent work aimed at reducing or eliminating the burn-in cost while retaining optimal regret guarantees.

5.3.4 Intuition

Before embarking on the proof, we provide some intuition for why the asymptotic regret scales as $\tilde{O}(\sqrt{SAH^3K})$.

Notation. To facilitate the discussion and subsequent analysis, we introduce augmented notation to explicitly index the iterates of Algorithm 5.2 by the episode number k as follows.

- $(s_1^k, a_1^k, \dots, s_H^k, a_H^k, s_{H+1}^k)$: the sample trajectory observed during episode k , with s_h^k (resp. a_h^k) the state (resp. action) at step h .
- π^k : the policy π^{ucb} computed and deployed in episode k ; note that π^k is computed at the beginning of episode k and depends only on data collected prior to episode k .
- $\hat{P}_{s,a,h}^k \in \mathbb{R}^S$: the latest empirical transition probability vector $\hat{P}_{s,a,h}$ computed *before* episode k .
- $N_h^k(s, a)$: the visitation count $N_h(s, a)$ for the (s, a, h) -tuple *before* episode k .
- $b_h^k(s, a) \geq 0$: the latest bonus term $b_h(s, a)$ *before* episode k .
- $Q_h^k \in \mathbb{R}^{SA}$: the Q-function estimate Q_h^{ucb} *before* episode k .
- $V_h^k \in \mathbb{R}^S$: the value function estimate V_h^{ucb} *before* episode k .

Heuristic arguments. Below, we provide a heuristic discussion of the regret analysis. To simplify the exposition, we consider the asymptotic regime with $K \rightarrow \infty$, where we may informally regard $V_h^k \approx V_h^* \approx V_h^{\pi^k}$ and $\hat{P}^k \approx P$.

Note that in each episode k , V_1^k is computed via backward induction from step H back to step 1 (see (5.18)). At each step h , the Q-estimate is inflated by a Bernstein-style bonus term:

$$b_h^k(s, a) \asymp \sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}^k}(V_{h+1}^k)}{N_h^k(s, a)}} \approx \sqrt{\frac{\text{Var}_{P_{s,a,h}}(V_{h+1}^{\pi^k})}{N_h^k(s, a)}} \quad (\text{ignoring log factors}),$$

where we omit the $O(1/N_h^k(s, a))$ lower-order term for simplicity, as the variance-dependent term typically dominates in the asymptotic regime with $K \rightarrow \infty$.

While this bonus is designed to compensate for model uncertainty, it inherently introduces an over-estimation of the value function on the order of $b_h^k(s, a)$ at each step. Accumulated across all episodes and steps, this leads to an aggregated over-estimation error of the order:

$$\mathcal{E}_{\text{over}} := \sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k) \asymp \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{\text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^{\pi^k})}{N_h^k(s_h^k, a_h^k)}}.$$

To bound this cumulative over-estimation error, a standard application of the Cauchy–Schwarz inequality yields

$$\mathcal{E}_{\text{over}} \lesssim \sqrt{\sum_{k,h} \frac{1}{N_h^k(s_h^k, a_h^k)}} \cdot \sqrt{\sum_{k,h} \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^{\pi^k})}. \quad (5.24)$$

Consequently, bounding $\mathcal{E}_{\text{over}}$ hinges on analyzing two sums, which can be accomplished using the following basic results.

Lemma 5.3. For any $K \geq 2$, Algorithm 5.2 satisfies

$$\sum_{k=1}^K \sum_{h=1}^H \frac{1}{\max\{N_h^k(s_h^k, a_h^k), 1\}} \leq 2SAH \log_2 K. \quad (5.25)$$

Lemma 5.4. Fix a deterministic policy π . Generate a sample trajectory $(s_1, a_1, \dots, s_H, a_H)$ by executing policy π in the MDP. Then one has

$$\mathbb{E} \left[\sum_{h=1}^H \text{Var}_{P_{s_h, a_h, h}}(V_{h+1}^{\pi}) \right] \leq H^2,$$

where the expectation is over the randomness of the sample trajectory.

Remark 5.4. The variance bound in Lemma 5.4 parallels the variance-based upper bounds — such as (3.14) in Section 3.2.4 — that were previously established for, and play a pivotal role in the analysis of, γ -discounted infinite-horizon MDPs.

Assuming that the cumulative variance term concentrates around its mean, the preceding two lemmas imply that with high probability,

$$\mathcal{E}_{\text{over}} \leq \tilde{O}(\sqrt{SAH} \cdot \sqrt{KH^2}) = \tilde{O}(\sqrt{SAH^3K}). \quad (5.26)$$

Since $\text{Regret}(K)$ (cf. (5.2)) is no larger than the order as the cumulative over-estimation error (as we shall shortly rigorize), this heuristic discussion suggests that, asymptotically,

$$\text{Regret}(K) \lesssim \tilde{O}(\sqrt{SAH^3K}).$$

Importantly, the cumulative variance bound in Lemma 5.4 plays a central role in achieving asymptotically optimal regret. Indeed, if one were to bound each summand crudely by H^2 without exploiting the dependence structure across the horizon, the resulting cumulative variance could scale as H^3 , thereby leading to suboptimal regret guarantees. This observation also highlights the importance of incorporating variance information into the design of exploration bonuses.

Proof of Lemma 5.3. For any given (s, a, h) , it is easily seen that

$$\sum_{k=1}^K \frac{1}{\max\{N_h^k(s_h^k, a_h^k), 1\}} \mathbb{1}\{(s, a) = (s_h^k, a_h^k)\} \leq 1 + \sum_{k=1}^{K-1} \frac{1}{k} \leq 2 \log_2 K.$$

This holds since: (i) when (s, a) is visited for the first time, we have $N_h^k(s, a) = 0$ by definition, so the corresponding summand equals 1; (ii) thereafter, each subsequent visit to (s, a) increments its associated visitation count by 1. Summing over all (s, a, h) tuples completes the proof. $\square$

Proof of Lemma 5.4. Consider first any fixed pair $(s, h) \in \mathcal{S} \times [H]$. It is easy to verify that

$$\begin{aligned} \text{Var}_{P_{s, \pi_h(s), h}}(V_{h+1}^\pi) &= \sum_{s' \in \mathcal{S}} P_{s, \pi_h(s), h}(s') \left[V_{h+1}^\pi(s') - \langle P_{s, \pi_h(s), h}, V_{h+1}^\pi \rangle \right]^2 \\ &= \sum_{s' \in \mathcal{S}} P_{s, \pi_h(s), h}(s') \left[V_{h+1}^\pi(s') - V_h^\pi(s) + r_h(s, \pi_h(s)) \right]^2 \\ &= \sum_{s' \sim P_{s, \pi_h(s), h}} \mathbb{E} \left[\left(V_{h+1}^\pi(s') - V_h^\pi(s) + r_h(s, \pi_h(s)) \right)^2 \right], \end{aligned} \quad (5.27)$$

where the second identity results from the Bellman equation. The cumulative variance over the horizon then satisfies

$$\begin{aligned} \mathbb{E} \left[\sum_{h=1}^H \text{Var}_{P_{s_h, a_h, h}}(V_{h+1}^\pi) \right] &\stackrel{(i)}{=} \sum_{h=1}^H \mathbb{E} \left[\left(V_{h+1}^\pi(s_{h+1}) - V_h^\pi(s_h) + r_h(s_h, a_h) \right)^2 \right] \\ &\stackrel{(ii)}{=} \mathbb{E} \left[\left(\sum_{h=1}^H \left[V_{h+1}^\pi(s_{h+1}) - V_h^\pi(s_h) + r_h(s_h, a_h) \right] \right)^2 \right] \\ &\stackrel{(iii)}{=} \mathbb{E} \left[\left(\sum_{h=1}^H r_h(s_h, a_h) - V_1^\pi(s_1) \right)^2 \right] \stackrel{(iv)}{\leq} H^2. \end{aligned}$$

Here, (i) follows by using (5.27) and recalling that $a_h = \pi_h(s_h)$ in the sample trajectory; (ii) comes from the Markov property of the trajectory, which ensures that the cross terms vanish; (iii) arises from telescoping together with the condition $V_{H+1}^\pi = 0$; and (iv) holds because both $\sum_{h=1}^H r_h(s_h, a_h)$ and $V_1^\pi(s_1)$ reside within the interval $[0, H]$. $\square$

5.3.5 Analysis: proof of Theorem 5.2

In this subsection, we provide the proof of Theorem 5.2. While the proof below differs somewhat from the original arguments of Azar *et al.* (2017) — though remaining similar in spirit — it provides a treatment that is more coherent with the proof of the MVP algorithm to be presented in the next section. For notational simplicity, we introduce a logarithmic term ι_b as follows:

$$\iota_b := 40000(\log_2 K)^3 \log(3SAH) \log \frac{1}{\delta'}, \quad (5.28)$$

where δ' is defined in (5.20b). We will also adopt the notation introduced in Section 5.3.4.

Before embarking on the main proof, we first observe that when $K \leq \iota_b SAH^2$, the regret bound in Theorem 5.2 follows trivially. Indeed, since the value function of any state lies within $[0, H]$, one has

$$\begin{aligned} \text{Regret}(K) &= \sum_{k=1}^K \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \leq HK \leq SAH^3 \iota_b \\ &\leq \tilde{O}(\sqrt{SAH^3 K} + S^2 AH^3), \end{aligned}$$

which already satisfies the claimed regret bound in this regime. Consequently, throughout the remainder of the proof, it suffices to assume

$$K \geq SAH^2 \iota_b. \quad (5.29)$$

Step 1: justifying optimism. We first establish that the running Q-function and value function estimates in Algorithm 5.2 serve as upper bounds for the true optimal Q-function and value function, respectively. This property provides a formal justification for the optimism principle underlying UCB-VI. Note that the same optimism property continues to hold if the exploration bonuses are replaced by confidence widths derived from alternative concentration inequalities, such as Hoeffding's inequality.

Lemma 5.5 (Optimism). Suppose that the constants $c_1, c_2 > 0$ in (5.20) obey $c_1 \geq 20/3$ and $c_2 \geq c_1^2$. With probability exceeding $1 - 2SAHK\delta'$,

$$Q_h^k(s, a) \geq Q_h^*(s, a) \quad \text{and} \quad V_h^k(s) \geq V_h^*(s) \quad (5.30)$$

hold simultaneously for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$.

Intuitively, this result holds because the exploration bonus added in each iteration is sufficiently large to dominate the uncertainty in the Bellman-type updates (5.18) arising from the use of empirical model estimates. The proof is deferred to Section 5.7.1.

Step 2: decomposing regret. In view of the optimism property established in Lemma 5.5, the regret admits the following upper bound

$$\text{Regret}(K) = \sum_{k=1}^K \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \leq \sum_{k=1}^K \left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \quad (5.31)$$

with probability at least $1 - 2SAHK\delta'$. The remainder of the analysis therefore focuses on upper bounding the right-hand side of (5.31).

Remark 5.5. To help clarify the rationale behind replacing V_1^* with V_1^k , note that $V_1^k(s_1^k) = \max_a Q_1^k(s_1^k, a)$ coincides with $Q_1^k(s_1^k, \pi_1^k(s_1^k))$ by the greedy policy update (cf. line 4 of Algorithm 5.2). Consequently, both $V_1^k(s_1^k)$ and $V_1^{\pi^k}(s_1^k)$ are, in some sense, aligned with the same policy iterate π^k . This policy alignment — particularly in terms of the selected actions — proves convenient when recursively expanding the difference $V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)$ using Bellman-type updates.

To control the right-hand side of (5.31), we begin by deriving an upper bound on the optimistic value function estimate.

Lemma 5.6. For every $1 \leq k \leq K$, one has

$$V_1^k(s_1^k) \leq \sum_{h=1}^H \left\{ \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle + b_h^k(s_h^k, a_h^k) + r_h(s_h^k, a_h^k) + \left\langle P_{s_h^k, a_h^k, h}^k - e_{s_{h+1}^k}, V_{h+1}^k \right\rangle \right\}.$$

Here and throughout, e_s denotes the s -th standard basis vector.

Intuitively, this recurrence relation is derived by recursively unrolling $V_1^k(s_1^k)$ via its update rule through the entire horizon. Crucially, both the empirical model estimation error $\hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k$ and the sampling noise $P_{s_h^k, a_h^k, h}^k - e_{s_{h+1}^k}$ are zero-mean fluctuations, whose cumulative contributions may potentially be controlled via concentration inequalities. We include the proof here given its central role in the analysis of UCB-type online RL algorithms.

Proof of Lemma 5.6. According to the update rule (5.18), the value function estimates for each $1 \leq h \leq H$ and $1 \leq k \leq K$ satisfy

$$\begin{aligned} V_h^k(s_h^k) &= \max_{a \in \mathcal{A}} Q_h^k(s_h^k, a) \stackrel{(i)}{=} Q_h^k(s_h^k, a_h^k) \\ &\leq r_h(s_h^k, a_h^k) + \hat{P}_{s_h^k, a_h^k, h}^k V_{h+1}^k + b_h^k(s_h^k, a_h^k) \\ &= \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle + b_h^k(s_h^k, a_h^k) + r_h(s_h^k, a_h^k) + \left\langle P_{s_h^k, a_h^k, h}^k - e_{s_{h+1}^k}, V_{h+1}^k \right\rangle + V_{h+1}^k(s_{h+1}^k), \end{aligned}$$

where (i) is valid due to the greedy rule for selecting a_h^k . Importantly, this recurrence relation links the value function estimate at step h (i.e., $V_h^k(s_h^k)$) to its counterpart at step $h+1$ (i.e., $V_{h+1}^k(s_{h+1}^k)$). By unrolling this relation recursively from $h=1$ to H , we arrive at

$$V_1^k(s_1^k) \leq \sum_{h=1}^H \left\{ \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle + b_h^k(s_h^k, a_h^k) + r_h(s_h^k, a_h^k) + \left\langle P_{s_h^k, a_h^k, h}^k - e_{s_{h+1}^k}, V_{h+1}^k \right\rangle \right\},$$

where the recursive expansion terminates at step H since $V_{H+1}^k = 0$. $\square$

Substituting the regret bound from Lemma 5.6 into the decomposition (5.31) reveals that, with probability at least $1 - 2SAHK\delta'$,

$$\begin{aligned} (5.31) &\leq \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle}_{=: T_1} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k)}_{=: T_2} \\ &\quad + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h}^k - e_{s_{h+1}^k}, V_{h+1}^k \right\rangle}_{=: T_3} + \underbrace{\sum_{k=1}^K \left(\sum_{h=1}^H r_h(s_h^k, a_h^k) - V_1^{\pi^k}(s_1^k) \right)}_{=: T_4}. \end{aligned} \quad (5.32)$$

Therefore, it remains to bound each of the four terms $\{T_i\}_{i=1}^4$. Before proceeding, we briefly comment on each term.

- The term T_2 , which involves the (Bernstein-style) bonuses, involves variance quantities over time that need to be carefully aggregated in order to attain asymptotically optimal regret.
- The term T_3 can be controlled relatively directly using martingale concentration inequalities, since s_{h+1}^k is conditionally independent of V_{h+1}^k given what has happened prior to episode k . For similar reasons, the term T_4 can also be handled via martingale concentration arguments.

- The most challenging term is T_1 , due to the intricate statistical dependence between $\hat{P}_{s_h, a_h, h}^k$ and V_{h+1}^k . Addressing this dependence requires considerably more delicate analysis. In fact, as we shall see, the difficulty of handling this term in the moderate-to-small K regime also motivates the design of the MVP algorithm to be introduced in Section 5.4.

In what follows, we bound these terms separately. As will become clear shortly, the resulting bounds for these terms are inherently intertwined, ultimately leading to a system of interdependent inequalities that are solved jointly in the final step. To facilitate the analysis, we also introduce a few quantities concerning cumulative variance or cumulative second moments across time:

$$T_5 := \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{\hat{P}_{s_h, a_h, h}^k} (V_{h+1}^k), \quad (5.33a)$$

$$T_6^{\text{ucb}} := \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h, a_h, h}^k} (V_{h+1}^k), \quad (5.33b)$$

$$T_6^{\text{opt}} := \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h, a_h, h}^k} (V_{h+1}^*), \quad (5.33c)$$

$$T_6^{\text{low}} := \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h, a_h, h}^k} (V_{h+1}^{\pi^k}), \quad (5.33d)$$

$$T_6^{\text{diff}} := \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h, a_h, h}^k, (V_{h+1}^k - V_{h+1}^{\pi^k})^2 \right\rangle. \quad (5.33e)$$

Here, T_5 aggregates empirical variance terms, while the remaining quantities correspond to cumulative true variances or second moments associated with objects related to value functions.

Step 3: bounding the term T_1 in (5.32). We begin with the term T_1 , which is particularly challenging to cope with due to the complicated statistical dependency between $\hat{P}_{s_h, a_h, h}^k$ and V_{h+1}^k . To see this, note that $\hat{P}_{s, a, h}^k$ is constructed from *all* previous samples associated with (s, a, h) , while V_{h+1}^k is itself computed based on past samples. Consequently, these two objects exhibit substantial statistical dependence.

One common strategy for circumventing this technical difficulty is to replace V_{h+1}^k with the optimal value function V_{h+1}^* , the latter of which is fixed and therefore statistically independent of the sampling process. This strategy has been adopted in a large body of existing analyses of model-based online RL algorithms (e.g., Azar *et al.* (2017), Dann *et al.* (2017), Zanette and Brunskill (2019), and Zhang *et al.* (2021e)). To be precise, let us decompose

$$T_1 = \underbrace{\sum_{k, h} \left\langle \hat{P}_{s_h, a_h, h}^k - P_{s_h, a_h, h}^k, V_{h+1}^* \right\rangle}_{=: W_1} + \underbrace{\sum_{k, h} \left\langle \hat{P}_{s_h, a_h, h}^k - P_{s_h, a_h, h}^k, V_{h+1}^k - V_{h+1}^* \right\rangle}_{=: W_2}. \quad (5.34)$$

The rationale behind the decomposition (5.34) is two-fold: the first term W_1 can be readily controlled via concentration inequalities given that V_{h+1}^* is fixed, whereas the second term W_2 acts as a higher-order residual that would vanishes asymptotically as the value estimate V_{h+1}^k converges to V_{h+1}^* for large enough k .

The lemma below develops upper bounds that connect W_1 and W_2 with cumulative variances. The proof is deferred to Section 5.7.2.

Lemma 5.7. With probability greater than $1 - 8S^2AHK\delta'$, one has

$$\begin{aligned} W_1 &\leq \frac{1}{8}\sqrt{SAHT_6^{\text{low}}\iota_b} + \frac{1}{8}\sqrt{SAHT_6^{\text{diff}}\iota_b} + \frac{1}{8}SAH^2\iota_b, \\ W_2 &\leq \frac{1}{8}\sqrt{S^2AHT_6^{\text{diff}}\iota_b} + \frac{1}{8}S^2AH^2\iota_b, \\ W_3 &\leq \frac{1}{8}\sqrt{SAH^3T_6^{\text{low}}\iota_b} + \frac{1}{8}\sqrt{S^2AH^3T_6^{\text{diff}}\iota_b} + \frac{1}{8}S^2AH^3\iota_b, \end{aligned}$$

where T_6^{low} and T_6^{diff} are defined in (5.33), and we also set

$$W_3 := \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h, a_h, h}^k - P_{s_h, a_h, h}, (V_{h+1}^k)^2 \right\rangle. \quad (5.35)$$

Remark 5.6. While the primary goal of this step is to bound W_1 and W_2 , the same proof techniques can be readily adapted to upper bound W_3 as well, which will appear in the subsequent analysis.

Remark 5.7. A careful reader may notice that, in addition to the difference in the variance terms, the bound on W_2 contains an extra factor in S compared to the bound on W_1 . This discrepancy arises because, in the analysis of W_1 , the value function V_{h+1}^* is fixed and statistically independent of the data, which substantially simplifies the argument. In contrast, the term W_2 involves the difference $V_{h+1}^k - V_{h+1}^*$, which exhibits intricate statistical dependence on the data and consequently incurs an additional factor of S .

Combining Lemma 5.7 with (5.34), we reach

$$T_1 = W_1 + W_2 \leq \frac{1}{4}\sqrt{SAHT_6^{\text{low}}\iota_b} + \frac{1}{4}\sqrt{S^2AHT_6^{\text{diff}}\iota_b} + \frac{1}{4}S^2AH^2\iota_b =: T_1^{\text{ub}}. \quad (5.36)$$

It is also worth noting that the variance terms T_6^{low} and T_6^{diff} remain random quantities and therefore require additional arguments to control, which we shall accomplish in the ensuing steps.

Step 4: bounding the terms T_2, T_3 and T_4 in (5.32). This step is dedicated to bounding the terms T_2, T_3 and T_4 arising in the regret decomposition (5.32), as formalized in the lemma below.

Lemma 5.8. With probability exceeding $1 - 4(\log_2 K)\delta'$, one has

$$T_2 \leq \frac{1}{2}\sqrt{SAHT_5\iota_b} + \frac{1}{2}SAH^2\iota_b, \quad (5.37a)$$

$$\begin{aligned} |T_3| &\leq \sqrt{8T_6^{\text{ucb}} \log \frac{1}{\delta'}} + 5H \log \frac{1}{\delta'} \\ &\leq 4\sqrt{T_6^{\text{low}} \log \frac{1}{\delta'}} + 4\sqrt{T_6^{\text{diff}} \log \frac{1}{\delta'}} + 5H \log \frac{1}{\delta'}, \end{aligned} \quad (5.37b)$$

$$|T_4| \leq \sqrt{2KH^2 \log \frac{1}{\delta'}}, \quad (5.37c)$$

where $T_5, T_6^{\text{ucb}}, T_6^{\text{low}}$ and T_6^{diff} are defined in (5.33).

At a high level, T_3 and T_4 are sums of zero-mean fluctuations and are bounded via martingale concentration inequalities, whereas the bound on T_2 follows from the Cauchy-Schwarz inequality alongside basic properties of the visitation counts. The proof is deferred to Section 5.7.3.

Step 5: combining the bounds on $\{T_i\}_{i=1}^4$. Taking the preceding bounds (5.36) and (5.37) together with (5.32), we obtain

$$\begin{aligned} \sum_{k=1}^K \left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \right) &\leq T_1 + T_2 + T_3 + T_4 \\ &\leq T_1^{\text{ub}} + T_2 + 4\sqrt{T_6^{\text{low}} \log \frac{1}{\delta'}} + 4\sqrt{T_6^{\text{diff}} \log \frac{1}{\delta'}} + 7\sqrt{KH^2 \log \frac{1}{\delta'}} \\ &\leq T_0^{\text{ub}}, \end{aligned} \quad (5.38a)$$

with probability at least $1 - 9S^2AHK\delta'$, where we define

$$\begin{aligned} T_0^{\text{ub}} &:= \frac{3}{8}\sqrt{SAHT_6^{\text{low}}\iota_b} + \frac{3}{8}\sqrt{S^2AHT_6^{\text{diff}}\iota_b} + \frac{1}{4}S^2AH^2\iota_b \\ &\quad + 7\sqrt{KH^2 \log \frac{1}{\delta'}} + T_2. \end{aligned} \quad (5.38b)$$

Before proceeding, we note that the same arguments used to bound $\sum_{k=1}^K (V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k))$ extend directly to $\sum_{k=1}^K (V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k))$ for any $h > 1$. Combining this observation with a union bound yields that, with probability at least $1 - 9S^2AH^2K\delta'$,

$$\sum_{k=1}^K \left(V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) \right) \leq T_0^{\text{ub}} \quad (5.39)$$

holds simultaneously for all $1 \leq h \leq H$.

Step 6: bounding the second moments in (5.33). Thus far, the bounds on T_1 , T_2 and T_3 (see Lemmas 5.7 and 5.8) depend explicitly on the cumulative variances T_5 , T_6^{low} and T_6^{diff} (cf. (5.33)). To obtain the desirable regret bounds, it remains to derive asymptotically tight upper bounds for these aggregate second moments. We now bound these quantities via the following lemma, whose proof is provided in Section 5.7.4.

Lemma 5.9. With probability at least $1 - 2\delta' \log(KH)$,

$$T_5 \leq W_3 + 2HT_2 + 2KH^2 + \sqrt{32H^2T_6^{\text{ucb}} \log \frac{1}{\delta'}} + 5H^2 \log \frac{1}{\delta'}, \quad (5.40a)$$

$$T_6^{\text{low}} \leq KH^2 + \sqrt{2H^6K \log \frac{1}{\delta'}}, \quad (5.40b)$$

$$T_6^{\text{diff}} \leq 2H^2T_0^{\text{ub}} + 18H^2 \log \frac{1}{\delta'}, \quad (5.40c)$$

where W_3 and T_0^{ub} are defined in (5.35) and (5.38), respectively.

In addition, under the condition (5.29), one further has

$$T_6^{\text{low}} \leq 2KH^2, \quad (5.41a)$$

$$T_5 \leq 6KH^2 + \frac{1}{2}\sqrt{S^2AH^3T_6^{\text{diff}}\iota_b} + \frac{7}{2}S^2AH^3\iota_b. \quad (5.41b)$$

Remark 5.8. Asymptotically (as $K \rightarrow \infty$), it is natural to expect these variance quantities to satisfy $T_5 \approx T_6^{\text{ucb}} \approx T_6^{\text{opt}} \approx T_6^{\text{low}} \lesssim KH^2$, where the last inequality follows from (5.41a) (thanks to the Markov structure of the cumulative variance). Importantly, both establishing and exploiting the bound $T_6^{\text{low}} \lesssim KH^2$ are crucial for obtaining the asymptotically optimal regret guarantees.

Step 7: putting all pieces together. We now combine the preceding (interdependent) bounds to derive a clean upper bound on the regret.

First, substituting (5.41b) into (5.37a) in Lemma 5.8 yields

$$\begin{aligned}
T_2 &\leq \frac{1}{2}\sqrt{SAHT_5\iota_b} + \frac{1}{2}SAH^2\iota_b \\
&\leq \frac{1}{2}\sqrt{6SAH^3K\iota_b} + \frac{1}{2}\sqrt{S^2AH^2\left(\frac{1}{2}\sqrt{AHT_6^{\text{diff}}\iota_b} + \frac{7}{2}SAH^2\iota_b\right)\iota_b} + \frac{1}{2}SAH^2\iota_b \\
&\leq \frac{1}{2}\sqrt{6SAH^3K\iota_b} + \frac{1}{4}S^2AH^2\iota_b + \frac{1}{8}\sqrt{AHT_6^{\text{diff}}\iota_b} + \frac{7}{8}SAH^2\iota_b + \frac{1}{2}SAH^2\iota_b \\
&\leq \frac{1}{2}\sqrt{6SAH^3K\iota_b} + \frac{1}{4}\sqrt{AHT_6^{\text{diff}}\iota_b} + \frac{13}{8}S^2AH^2\iota_b,
\end{aligned}$$

where the second inequality follows from the elementary fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and the penultimate inequality is due to the AM-GM inequality. Substituting this into the definition of T_0^{ub} (cf. (5.38)) and combining terms, we reach

$$\begin{aligned}
T_0^{\text{ub}} &\leq \frac{3}{8}\sqrt{SAHT_6^{\text{low}}\iota_b} + \sqrt{S^2AHT_6^{\text{diff}}\iota_b} + 2\sqrt{SAH^3K\iota_b} + 2S^2AH^2\iota_b \\
&\leq 3\sqrt{SAH^3K\iota_b} + \sqrt{S^2AHT_6^{\text{diff}}\iota_b} + 2S^2AH^2\iota_b,
\end{aligned} \tag{5.42}$$

where the last inequality results from (5.41a).

To conclude the proof, observe that the AM-GM inequality implies

$$\sqrt{S^2AHT_6^{\text{diff}}\iota_b} \leq \frac{1}{4H^2}T_6^{\text{diff}} + S^2AH^3\iota_b \leq \frac{1}{2}T_0^{\text{ub}} + 2S^2AH^3\iota_b,$$

where the last inequality follows from Lemma 5.9. Combining this bound with (5.42) and rearranging terms, we conclude that

$$T_0^{\text{ub}} \lesssim \sqrt{SAH^3K\iota_b} + S^2AH^3\iota_b$$

with probability at least $1 - 20S^2AH^2K\delta'$. Together with the choice of δ' in (5.28), this completes the proof of Theorem 5.2.

5.4 An improved model-based online RL algorithm: MVP

As noted in Section 5.3.3, the UCB-VI algorithm achieves asymptotically optimal regret only after an excessively large burn-in phase characterized by (5.22). How to overcome this limitation and obtain regret guarantees that are optimal for all sample sizes has therefore become a fundamentally important research direction, drawing considerable attention in the years following the introduction of UCB-VI. This objective was recently accomplished by Zhang *et al.* (2024c) and Zhang *et al.* (2025e), which established that a variant of the optimistic model-based algorithms — called *Monotonic Value Propagation* (MVP), originally proposed by Zhang *et al.* (2021e) — attains minimax-optimal regret (up to logarithmic factors) for any number K of sample trajectories.

5.4.1 Revisiting UCB-VI in the sample-starved regime

At a high level, the MVP algorithm can be viewed as a refinement of UCB-VI, requiring only a handful of modifications to the version of UCB-VI presented in Algorithm 5.2. These seemingly modest

changes enable substantially stronger regret guarantees in the sample-starved regime. To motivate these algorithmic innovations, we first revisit the main bottleneck in the analysis of UCB-VI that limits its theoretical performance guarantees in the most sample-constrained settings.

As discussed in Step 2 of Section 5.3.5, the regret of the UCB-VI algorithm admits a decomposition of the form (5.32), consisting of four terms $\{T_i\}_{1 \leq i \leq 4}$. As already noted there, the terms T_3 and T_4 can be readily handled using standard martingale concentration inequalities, while T_2 can likewise be controlled through the variance-based inequalities developed there. The primary obstacle is the term

$$T_1 = \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle, \quad (5.43)$$

whose analysis is complicated due to the intricate statistical dependence between $\hat{P}_{s, a, h}^k$ and V_{h+1}^k ; indeed, both the empirical model estimate $\hat{P}^k$ and the value estimate V^k depend on the entire collection of samples gathered prior to episode k , precluding a direct application of standard concentration arguments. In the sequel, we discuss two fundamentally different strategies for overcoming this obstacle.

Strategy 1: replacing V^k with V^* . Thus far, the analysis of UCB-VI copes with T_1 by first replacing V_{h+1}^k in (5.43) with the optimal value function V_{h+1}^* — thereby enabling the use of standard concentration techniques given that V_{h+1}^* is independent from samples — and then controlling the resulting approximation error induced by the discrepancy between V_{h+1}^k and V_{h+1}^* . Although this approach is asymptotically well justified — indeed, one expects V_{h+1}^k to converge to V_{h+1}^* as $K \rightarrow \infty$ — the gap between V_{h+1}^k and V_{h+1}^* can remain substantial and non-negligible in the regime where K is only moderate or small. As a result, the induced approximation error can be difficult to control sharply in this regime, ultimately giving rise to the large burn-in cost exhibited in Theorem 5.2.

Strategy 2: a covering-type argument. We now describe an alternative approach for analyzing this term T_1 , based on a covering-type argument. We provide a brief sketch of this argument below, as it serves to motivate both the algorithmic design and analysis of MVP.

To simplify the discussion, we keep the following argument informal. Recall that $N_h^k(s, a)$ denotes the visitation count of (s, a, h) prior to episode k . Owing to the Markovian nature of the data and the backward Bellman recursion (5.18) used to compute V_h^k , it is natural to expect that $\hat{P}_{s, a, h}^k$ and V_{h+1}^k are conditionally independent given the collection of visitation counts at step h :

$$\{N_h^k(s, a)\}_{(s, a, k) \in \mathcal{S} \times \mathcal{A} \times [K]}.$$

Now suppose, for the sake of intuition, that these visitation counts were fixed in advance and independent of the empirical transition kernels $\{\hat{P}_{s, a, h}^k\}$ (a heuristic assumption that is not technically accurate, but useful for an informal discussion). Under this heuristic viewpoint, standard concentration techniques could be applied directly to control $\sum_{k, h} (\hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}) V_{h+1}^k$. The remaining step would be to remove the artificial independence assumption by taking a union bound over all possible realizations of $N_h^k(s, a)$. The main obstacle, however, is combinatorial: the number of such visitation-count configurations grows exponentially with K , making a naive union bound prohibitively costly and hence significantly weakening the resulting regret guarantee.

Nevertheless, this heuristic argument points to a potentially useful idea: if the effective number of admissible visitation-count configurations could be reduced — say, to a quantity polynomial in K — then the covering-based argument outlined above could become viable. This observation motivates the design of MVP, which we introduce next.

5.4.2 Algorithmic ideas: doubling-based updates

We now introduce the MVP algorithm by singling out its key difference from the UCB-VI algorithm presented in Algorithm 5.2. A key distinguishing feature of MVP is its *epoch-based update* scheme for the empirical transition kernels, inspired by the *doubling-trick* framework adopted in Jaksch *et al.* (2010). More specifically, the learning process is partitioned into consecutive epochs according to a simple doubling rule: whenever there exists a tuple (s, a, h) whose visitation count reaches a power of two, the current epoch terminates. At that point, MVP reconstructs the empirical model, recomputes the UCB-based optimistic Q-functions and value functions based on the updated transition kernels, and then initiates a new epoch with the corresponding updated policy. Within each epoch, the sampling policy is determined by the optimistic Q-function estimates computed prior to the beginning of that epoch; consequently, the same policy is used to collect data throughout all episodes within the epoch.

Clearly, this epoch-based paradigm stands in sharp contrast to the original UCB-VI algorithm, which updates the empirical transition model and recomputes the Q-estimates and value estimates at every episode. An important property ensured by the careful design of the doubling update rule is that: the empirical transition probability vector associated with each (s, a, h) tuple is updated only when its visitation count equals a power of two, and therefore no more than $\log_2 K$ times over the entire learning process. As we shall see, this structural property plays a central role in substantially reducing the effective covering number required in the covering-based argument mentioned in Section 5.4.1. More specifically, it induces a modified version of the visitation-count sequence $\{N_h^k(s, a)\}_{(s,a,k) \in \mathcal{S} \times \mathcal{A} \times [K]}$, which tracks the empirical model currently in use as opposed to the most recent collection of samples. Owing to the epoch-based update scheme for the empirical models, the number of admissible configurations of this modified visitation-count sequence is dramatically smaller than that arising in the original UCB-VI algorithm.

The complete procedure of MVP is presented in Algorithm 5.3, which adopts the same variance-aware Bernstein-style exploration bonuses as those recommended in Algorithm 5.2. In terms of notation, we now let $N_h^{\text{all}}(s, a)$ indicate the total visitation count of (s, a, h) accumulated over all episodes up to the current one, and let $N_h(s, a)$ (resp. $N_h(s, a, \tilde{s})$) track the visitation count of (s, a, h) (resp. $(s, a, h, \tilde{s})$) within the current epoch. In addition, the flag `triggered` is used to indicate whether we have reached the end of the current epoch.

Remark 5.9. The name *Monotonic Value Propagation*, introduced by Zhang *et al.* (2021e), stems from the fact that the exploration bonuses (5.20) guarantee a certain monotonicity property in value propagation. For instance, the function $f(p, v, n)$ defined in (5.116) in Section 5.7.1 — which resembles an important operation in certain optimistic Bellman recursions — is non-decreasing with respect to v (which typically represents a vector of value estimates). Notably, this monotonicity property has already been incorporated into the version of UCB-VI presented in Algorithm 5.2, which differs from the version presented in the original UCB-VI paper (Azar *et al.*, 2017); see also Remark 5.2.

Algorithm 5.3: The monotonic value propagation (MVP) algorithm for episodic online RL

```

1 input: number of episodes  $K$ , reward function  $r$ , confidence parameter  $\delta$ , constants  $c_1, c_2 > 0$ .
2 initialize  $N_h^{\text{all}}(s, a) \leftarrow 0$ ,  $N_h(s, a, s') \leftarrow 0$ ,  $N_h(s, a) \leftarrow 0$ ,  $Q_h^{\text{ucb}}(s, a) \leftarrow H$ ,  $V_h^{\text{ucb}}(s) \leftarrow H$ ,
    $V_{H+1}^{\text{ucb}}(s) \leftarrow 0$ ,  $\hat{P}_{s,a,h}(s') \leftarrow 1/S$  for all  $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ .
3 for  $k = 1, \dots, K$  do
   /* policy update. */
4   update policy  $\pi^{\text{ucb}}$  such that  $\pi_h^{\text{ucb}}(s) \leftarrow \arg \max_a Q_h^{\text{ucb}}(s, a)$  for all  $(s, h) \in \mathcal{S} \times [H]$ .
5   observe the initial state  $s_1 \stackrel{\text{ind.}}{\sim} \rho$  of a new episode.
6   for  $h = 1, \dots, H$  do
     /* sampling. */
7     take  $a_h = \pi_h^{\text{ucb}}(s_h)$ , observe  $s_{h+1}$ ; set  $(s, a, s') \leftarrow (s_h, a_h, s_{h+1})$ .
8      $N_h^{\text{all}}(s, a) \leftarrow N_h^{\text{all}}(s, a) + 1$ ;  $N_h(s, a, s') \leftarrow N_h(s, a, s') + 1$ .
     /* updates using data of this epoch. */
9     if  $N_h^{\text{all}}(s, a) \in \{1, 2, \dots, 2^{\log_2 K}\}$  then
10       $N_h(s, a) \leftarrow \sum_{\tilde{s}} N_h(s, a, \tilde{s})$ . // # visits this epoch.
11       $\hat{P}_{s,a,h}(\tilde{s}) \leftarrow \frac{N_h(s,a,\tilde{s})}{N_h(s,a)}$  for all  $\tilde{s} \in \mathcal{S}$ . // empirical model.
12      set triggered = true;  $N_h(s, a, \tilde{s}) \leftarrow 0$  for all  $\tilde{s} \in \mathcal{S}$ .
     /* optimistic Q-estimation for this epoch. */
13     if triggered = true then
14       set triggered = false.
15       for  $h = H, \dots, 1$  do
16         for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
17           set the bonus  $b_h(s, a)$  according to (5.20); compute
           
$$Q_h^{\text{ucb}}(s, a) \leftarrow \min \{r_h(s, a) + \langle \hat{P}_{s,a,h}, V_{h+1}^{\text{ucb}} \rangle + b_h(s, a), H\}, \quad (5.44a)$$

           
$$V_h^{\text{ucb}}(s) \leftarrow \max_a Q_h^{\text{ucb}}(s, a). \quad (5.44b)$$


```

5.4.3 Theoretical guarantees

The epoch-based design of MVP, in conjunction with the appealing features of the optimistic model-based framework, turns out to be sufficient to overcome the suboptimal regret guarantees of UCB-VI arising in the sample-limited regime. The following theorem summarizes the statistical guarantees of MVP, showing that it achieves the minimax-optimal regret uniformly for all numbers of sample trajectories $K \geq 1$.

Theorem 5.10 (Zhang *et al.*, 2024c; Zhang *et al.*, 2025e). Consider any $K \geq 1$ and any $0 < \delta < 1$. With probability at least $1 - \delta$, the MVP algorithm (see Algorithm 5.3) with Bernstein-style bonuses (5.20) achieves

$$\text{Regret}(K) \leq \tilde{O}\left(\min\{\sqrt{SAH^3K}, HK\}\right), \quad (5.45)$$

provided that the constants $c_1, c_2 > 0$ are suitably chosen. Here, in the definition of $\text{Regret}(K)$, π^k and s_1^k denote, respectively, the policy deployed and the initial state observed in episode k of Algorithm 5.3.

As we shall confirm shortly in Section 5.6 (see Theorem 5.19), the above regret guarantee for MVP matches the fundamental minimax lower bound up to logarithmic factors, regardless of how large K

is. In contrast to UCB-VI, this means that MVP essentially eliminates the burn-in phase required to attain statistically optimal performance, with only logarithmic factors remaining. By converting this regret guarantee into a PAC-style sample complexity bound (using arguments in Section 5.1), we immediately obtain the following implication: for an *arbitrary* accuracy level $\varepsilon \in (0, H]$, the MVP algorithm is guaranteed to yield an ε -optimal policy with high probability using at most

$$\tilde{O}\left(\frac{SAH^3}{\varepsilon^2}\right) \text{ episodes,} \quad (5.46)$$

where the output policy can be drawn, for instance, uniformly at random from the policy sequence $\{\pi^k\}_{k=1}^K$. Equivalently, this corresponds to $\tilde{O}(\frac{SAH^4}{\varepsilon^2})$ sample transitions, since each sample trajectory has length H . It is noteworthy that this sample complexity is information-theoretically unimprovable even when one has access to an idealized generative model (cf. Section 3.5).

Taken together, the above results further underscore the statistical optimality of model-based methods in yet another fundamental RL setting: online RL. To date, MVP remains the only online RL algorithm known to provably attain minimax-optimal regret without incurring any burn-in cost.

A note of caution is warranted, however. At present, it remains unclear whether UCB-VI is fundamentally suboptimal or whether its large burn-in cost is merely an artifact of existing proof techniques, given the absence of algorithm-specific lower bounds for UCB-VI. Having said that, MVP enjoys an additional practical advantage over UCB-VI: a substantially lower switching cost. Owing to the doubling-based update rule, the sampling policy in MVP is no longer updated at every episode and may remain fixed for extended periods of time. This low-switching property can significantly reduce deployment and operational costs in applications where changing sampling policies is expensive.

Finally, it is noteworthy that finite-horizon nonstationary MDPs, arguably the most widely studied setting for online RL, remain the only setting in which minimax-optimal regret without a burn-in period is currently known to be achievable. Other important settings, such as discounted infinite-horizon MDPs and finite-horizon stationary MDPs, remain open. Resolving these settings will likely require new ideas in both theory and algorithm design to fully bridge the remaining gaps.

5.4.4 Analysis: proof of Theorem 5.10

Following the analysis of UCB-VI, it is convenient to introduce the following logarithmic factor:

$$\iota_b := 40000(\log_2 K)^3 \log(3SAH) \log \frac{1}{\delta'}, \quad \delta' = \frac{100\delta}{S^2 A^2 H^2 K^2}. \quad (5.47)$$

As in Section 5.3.5, we first note that the desired regret bound in Theorem 5.10 holds trivially whenever $K \leq SAH\iota_b$, since

$$\text{Regret}(K) = \sum_{k=1}^K \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \leq HK = \min \left\{ \sqrt{SAH^3 K \iota_b}, HK \right\}.$$

Consequently, it suffices to restrict our attention throughout this section to the scenario where

$$K \geq SAH\iota_b. \quad (5.48)$$

Preparation: an expanded sample set from a generative model. We now introduce a seemingly different, yet closely related, sampling mechanism that generates independent samples from a generative model prior to the start of online learning. As will become clear shortly, this auxiliary mechanism often helps simplify the exposition of the analysis.

Definition 5.1 (Expanded samples). Let $\mathcal{D}^{\text{expand}}$ be a set of *SAHK* independent samples generated as follows: for each $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, draw K independent samples $(s, a, h, s'^{(i)})$ obeying $s'^{(i)} \stackrel{\text{ind.}}{\sim} P_{s,a,h}$ for all $1 \leq i \leq K$.

Importantly, $\mathcal{D}^{\text{expand}}$ can be coupled with the original set of samples, denoted by $\mathcal{D}^{\text{orig}}$. Specifically, one first generates $\mathcal{D}^{\text{expand}}$ before online learning begins, and thereafter, whenever the online process queries (s, a, h) , one returns an unused sample from $\mathcal{D}^{\text{expand}}$ without replacement. This coupling allows us to analyze the online process through the view of $\mathcal{D}^{\text{expand}}$, which is sometimes more convenient. Unless stated otherwise, all proofs assume this coupling between $\mathcal{D}^{\text{orig}}$ and $\mathcal{D}^{\text{expand}}$.

Notation. Akin to the analysis of UCB-VI, it is convenient to introduce the following notation to index the iterates in Algorithm 5.3 explicitly by the episode number k . Some of the notation defined below is analogous to that previously used in the analysis of UCB-VI in Section 5.3.5.

- $\hat{P}_{s,a,h}^k \in \mathbb{R}^S$: the latest update of the empirical transition probability vector $\hat{P}_{s,a,h}$ before episode k .
- $b_h^k(s, a) \geq 0$: the latest update of the exploration bonus term $b_h(s, a)$ before episode k .
- $N_h^{k,\text{all}}(s, a)$: the total visitation count $N_h^{\text{all}}(s, a)$ of the (s, a, h) -tuple before episode k .
- $N_h^k(s, a)$: the visitation count $N_h(s, a)$ of the (s, a, h) -tuple of the latest doubling batch used to compute $\hat{P}_{s,a,h}$ before episode k . When $N_h^{k,\text{all}}(s, a) = 0$, set $N_h^k(s, a) = 1$ for ease of presentation.
- $V_h^k \in \mathbb{R}^S$: the latest value estimate V_h^{ucb} before episode k .
- $Q_h^k \in \mathbb{R}^{SA}$: the latest Q-estimate Q_h^{ucb} before episode k .

We shall also make use of the following notation for the empirical transition probability vector indexed by the batch number.

- For any $j \geq 2$ (resp. $j = 1$), let $\hat{P}_{s,a,h}^{(j)}$ denote the empirical transition probability vector for (s, a, h) computed from the j -th batch (resp. the first batch) of data, namely, the $\{2^{j-2} + i\}_{i=1}^{2^{j-2}}$ -th samples (resp. the 1st sample) associated with (s, a, h) . For completeness, we set the empirical transition probability vector for the 0-th batch to be $\hat{P}_{s,a,h}^{(0)}(s') = 1/S$ for every $s' \in \mathcal{S}$.
- With the coupling between $\mathcal{D}^{\text{expand}}$ and $\mathcal{D}^{\text{orig}}$ in mind, we will use $\hat{P}_{s,a,h}^{(j)}$ interchangeably to denote the empirical probability vector constructed from the j -th batch of data from either $\mathcal{D}^{\text{orig}}$ or $\mathcal{D}^{\text{expand}}$, as long as no confusion arises.

Recap: a basic regret decomposition. By repeating the arguments from Step 2 of Section 5.3.5 for UCB-VI, we obtain for MVP the same regret decomposition as in (5.32): with probability at least $1 - 2SAHK\delta'$,

$$\text{Regret}(K) \leq \underbrace{\sum_{k=1}^K \sum_{h=1}^H \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle}_{=: T_1} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k)}_{=: T_2}$$

$$+ \underbrace{\sum_{k=1}^K \sum_{h=1}^H \langle P_{s_h^k, a_h^k, h} - e_{s_{h+1}^k}, V_{h+1}^k \rangle}_{=: T_3} + \underbrace{\sum_{k=1}^K \left(\sum_{h=1}^H r_h(s_h^k, a_h^k) - V_1^{\pi^k}(s_1^k) \right)}_{=: T_4}. \quad (5.49)$$

The terms $T_2, \dots, T_4$ can be handled in a similar manner as in the analysis of UCB-VI in Section 5.3.5. As alluded to previously, obtaining a sharp bound on T_1 constitutes one of the main objectives of the analysis that follows.

Step 1: a compressed representation of visitation counts $\{N_h^k(s, a)\}$. As discussed in Section 5.4.1, a potential covering-based analysis strategy (see Strategy 2 therein) is to first condition on the visitation counts $\{N_h^k(s, a)\}$ when carrying out the concentration-of-measure analysis, and then apply a union bound over all possible configurations of these visitation counts. A crucial step in this strategy is therefore to control the total number of admissible configurations — namely, an appropriate covering number — associated with $\{N_h^k(s, a)\}$. To this end, it is useful to introduce the following notion of a *profile*, which facilitates the subsequent analysis.

Definition 5.2 (Profile). Consider any collection $\{N_h^{k, \text{all}}(s, a)\}$. For any $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, define

$$I_{s, a, h}^k := \begin{cases} \max \{j \in \mathbb{Z} : 2^{j-1} \leq N_h^{k, \text{all}}(s, a)\}, & \text{if } N_h^{k, \text{all}}(s, a) > 0, \\ 0, & \text{if } N_h^{k, \text{all}}(s, a) = 0. \end{cases}$$

The profile for the k -th episode ($1 \leq k \leq K$) and the total profile are then defined respectively as

$$\mathcal{I}^k := \{I_{s, a, h}^k\}_{(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]} \quad \text{and} \quad \mathcal{I} := \{\mathcal{I}^k\}_{k=1}^K.$$

As can be readily seen, given a total profile $\mathcal{I}$ w.r.t. $\{N_h^{k, \text{all}}(s, a)\}$, the empirical transition kernel admits the representation

$$\hat{P}_{s, a, h}^k = \hat{P}_{s, a, h}^{(I_{s, a, h}^k)}, \quad \forall (s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]. \quad (5.50)$$

Thus, a total profile fully determines all the time instances and locations when the empirical models are updated. Moreover, since each visitation count $N_h^k(s, a)$ is recomputed only when the corresponding empirical model is updated (see line 10 of Algorithm 5.3), the total profile serves as a compact representation of the entire collection $\{N_h^k(s, a)\}$.

To quantify the degree of compression offered by Definition 5.2 in representing the visitation counts $\{N_h^k(s, a)\}$, we establish the following upper bound on the total number of admissible profiles.

Lemma 5.11. Suppose that $K \geq SAH \log_2 K$. Then the number of all possible total profiles w.r.t. Algorithm 5.3 is no greater than

$$(4SAHK)^{SAH \log_2 K + 1}.$$

Proof. To establish this lemma, it is helpful to isolate the following set:

$$\mathcal{C} := \left\{ \mathcal{I} = \{\mathcal{I}^1, \dots, \mathcal{I}^K\} \mid \mathcal{I}^1 \leq \mathcal{I}^2 \leq \dots \leq \mathcal{I}^K, \right. \\ \left. \mathcal{I}^k \in \{0, 1, \dots, \log_2 K\}^{SAH} \text{ for all } 1 \leq k \leq K \right\}. \quad (5.51)$$

The monotonicity constraints imply that the total profile associated with any collection $\{N_h^k(s, a)\}$ generated by Algorithm 5.3 must belong to $\mathcal{C}$. Therefore, it suffices to establish the following bound:

$$|\mathcal{C}| \leq (4SAHK)^{SAH \log_2 K + 1}. \quad (5.52)$$

This is the main goal of the proof, which is detailed in Section 5.8.1. $\square$

In contrast to the uncompressed setting (such as in UCB-VI), where the visitation counts may admit exponentially many configurations in K , the doubling update mechanism adopted in Algorithm 5.3 yields a remarkable compression effect: the exponent governing the number of admissible configurations now grows only logarithmically with K (according to Lemma 5.11). This feature is one of the key ingredients that renders the covering-based analysis strategy effective.

Before proceeding, we state another useful property of the visitation counts. This result is analogous to Lemma 5.3, but applies to a modified version of the visitation counts tailored to the epoch-based updates.

Lemma 5.12. For any $K \geq 2$, Algorithm 5.3 satisfies

$$\sum_{k=1}^K \sum_{h=1}^H \frac{1}{\max\{N_h^k(s_h^k, a_h^k), 1\}} \leq 2SAH \log_2 K.$$

Proof. For each tuple (s, a, h) , it can be readily verified that

$$\begin{aligned} \sum_{k=1}^K \frac{1}{\max\{N_h^k(s_h^k, a_h^k), 1\}} \mathbb{1}\{(s, a) = (s_h^k, a_h^k)\} &\leq \sum_{k=1}^K \frac{2}{\max\{N_h^{k, \text{all}}(s_h^k, a_h^k), 1\}} \mathbb{1}\{(s, a) = (s_h^k, a_h^k)\} \\ &\leq 2 \log_2 K, \end{aligned}$$

where the first inequality is a direct consequence of the doubling rule, and the second inequality follows from the same proof as for Lemma 5.3. $\square$

Step 2: decomposing T_1 based on doubling updates. With the notion of profiles in place, we now proceed to analyze the term T_1 in (5.49), for which we first introduce some additional notation.

- $\mathcal{I}^{\text{true}} = \{\mathcal{I}^{1, \text{true}}, \dots, \mathcal{I}^{K, \text{true}}\}$, with $\mathcal{I}^{k, \text{true}} = \{I_{s, a, h}^{k, \text{true}}\}$, denotes the total profile induced by the true visitation counts $\{N_h^{k, \text{all}}(s, a)\}$ encountered during the online learning process;
- $k_{l, j, s, a, h}$: the episode index corresponding to the $(2^{l-1} + j)$ -th visit to (s, a, h) during the online learning process.

In view of the relation (5.50), we can rewrite T_1 in a form that makes the role of doubling updates more explicit:

$$\begin{aligned} T_1 &= \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle \\ &= \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^{(I_{s_h^k, a_h^k, h}^{k, \text{true}})} - P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle \\ &= \sum_{l=0}^{\log_2 K} \sum_{s, a, h} \left\langle \hat{P}_{s, a, h}^{(l)} - P_{s, a, h}, \sum_{k=1}^K \mathbb{1}\{(s_h^k, a_h^k) = (s, a), I_{s, a, h}^{k, \text{true}} = l\} V_{h+1}^k \right\rangle \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{l=1}^{\log_2 K} \sum_{s,a,h} \left\langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, \sum_{k=1}^K \mathbb{1} \left\{ (s_h^k, a_h^k) = (s, a), I_{s,a,h}^{k,\text{true}} = l \right\} V_{h+1}^k \right\rangle + SAH^2 \\
&= \sum_{l=1}^{\log_2 K} \sum_{j=1}^{2^{l-1}} \left\{ \sum_{s,a,h} \left\langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, V_{h+1}^{k_{l,j,s,a,h}} \right\rangle \right\} + SAH^2,
\end{aligned} \tag{5.53}$$

where, for notational convenience, we set $V_{h+1}^k = 0$ for all $k > K$. Here, the inequality in the penultimate line follows from the bound $0 \leq V_{h+1}^k(s) \leq H$ for all $s \in \mathcal{S}$. In other words, the above decomposition rearrange terms so as to collect together all terms corresponding to the same empirical transition estimates.

The above decomposition (5.53) suggests first controlling the term

$$\sum_{s,a,h} \left\langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, V_{h+1}^{k_{l,j,s,a,h}} \right\rangle. \tag{5.54}$$

As before, the main challenge lies in the complicated statistical dependency between $\hat{P}_{s,a,h}^{(l)}$ and $V_{h+1}^{k_{l,j,s,a,h}}$, which we address next.

Step 3: bounding (5.54) via coupling and covering. To obtain a tight bound on (5.54), we employ a covering argument. To this end, we need to first specify the objects to be covered and explain how the cover relates to the term of interest. Our high-level strategy is as follows, with details to follow next.

- i) For any fixed total profile $\mathcal{I}$ and fixed $l \leq \log_2 K$, develop a high-probability bound on a weighted sum of the form

$$\sum_{s,a,h} \left(\hat{P}_{s,a,h}^{(l)} - P_{s,a,h} \right) X_{h+1,s,a}, \tag{5.55}$$

where each vector $X_{h+1,s,a}$ is a deterministic function of the fixed profile $\mathcal{I}$ and the samples collected at steps $h' \geq h+1$. Here, we adopt the view of $\mathcal{D}^{\text{expand}}$ (cf. Definition 5.1) and think of $\hat{P}_{s,a,h}^{(l)}$ as the empirical distribution constructed from the l -th batch of corresponding samples in $\mathcal{D}^{\text{expand}}$. Under this view, $\hat{P}_{s,a,h}^{(l)}$ is statistically independent of $\mathcal{I}$ and, consequently, independent of $X_{h+1,s,a}$. Under this independence property, standard martingale concentration inequalities come in handy.

- ii) Taking a union bound over all possible configurations of the total profile $\mathcal{I}$ — together with Lemma 5.11 — yields a uniform control of the term (5.55), simultaneously over all possible profiles $\mathcal{I}$ and all associated sequences $\{X_{h+1,s,a}\}$.
- iii) Show that the above uniform bounds can be applied to the decomposition in (5.53) to derive a sharp bound.

Step 3.1: establishing uniform bounds on (5.55). We now carry out the above steps i) and ii). Let us first formally specify the types of vectors $\{X_{h,s,a}\}$ mentioned above in (5.55).

Assumption 5.1. For each total profile $\mathcal{I} \in \mathcal{C}$ (cf. (5.51)), assume that the set $\{\mathcal{X}_{h,\mathcal{I}}\}_{1 \leq h \leq H}$ satisfies the following properties that hold for each $1 \leq h \leq H$:

- $\mathcal{X}_{h,\mathcal{I}}$ is a set of no more than $K + 1$ non-negative vectors in $\mathbb{R}^S$, and contains the all-zero vector 0, where each vector $X \in \mathcal{X}_{h,\mathcal{I}}$ obeys $\|X\|_\infty \leq H$;
- $\mathcal{X}_{h+1,\mathcal{I}}$ is a collection derived from a *deterministic* mapping of $\mathcal{I}$, the reward function r , and the empirical distribution vectors (from the view of $\mathcal{D}^{\text{expand}}$)

$$\left\{ \hat{P}_{s,a,h'}^{(I_{s,a,h'})^k} \right\}_{h' \in (h,H], (s,a,k) \in \mathcal{S} \times \mathcal{A} \times [K]}.$$

For each collection $\{\mathcal{X}_{h,\mathcal{I}}\}$ satisfying the above assumptions, we can readily establish a uniform concentration bound by combining Freedman's martingale concentration inequality with a union bound; the proof is postponed to Section 5.8.2.

Lemma 5.13. Suppose that $K \geq SAH \log_2 K$. For each $\mathcal{I} \in \mathcal{C}$ (cf. (5.51)), consider a set $\{\mathcal{X}_{h,\mathcal{I}}\}_{1 \leq h \leq H}$ satisfying Assumption 5.1. Then with probability at least $1 - \delta'$, one has

$$\begin{aligned} & \sum_{s,a,h \in \mathcal{S} \times \mathcal{A} \times [H]} \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \rangle \\ & \leq \sum_{s,a,h \in \mathcal{S} \times \mathcal{A} \times [H]} \max \left\{ \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \rangle, 0 \right\} \\ & \leq \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}}(X_{h+1,s,a}) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} + \frac{4H}{2^{l-2}} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) \end{aligned} \quad (5.56)$$

holds simultaneously for all $\mathcal{I} \in \mathcal{C}$, all $2 \leq l \leq \log_2 K + 1$, and all sequences $\{X_{h,s,a}\}_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]}$ obeying $X_{h,s,a} \in \mathcal{X}_{h+1,\mathcal{I}}$, $\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Step 3.2: coupling. Next, we carry out iii), as in the following lemma. This result is proven by combining the uniform bound in Lemma 5.13 with the decomposition (5.53), which allows one to rigorously couple $\mathcal{D}^{\text{expand}}$ with the online learning process through the union bound. The proof is deferred to Section 5.8.3.

Lemma 5.14. Suppose that $K \geq SAH \log_2 K$. With probability exceeding $1 - \delta'$, we have

$$\begin{aligned} & \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle \leq \sum_{k=1}^K \sum_{h=1}^H \max \left\{ \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle, 0 \right\} \\ & \leq \sqrt{16(\log_2 K) \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} \\ & \quad + 49SAH^2 \log_2^3 K + 8H(\log_2 K) \log \frac{1}{\delta'} \end{aligned}$$

and

$$\begin{aligned} & \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, (V_{h+1}^k)^2 \right\rangle \\ & \leq 8H \sqrt{(\log_2 K) \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} \end{aligned}$$

$$+ 49SAH^3 \log_2^3 K + 8H^2(\log_2 K) \log \frac{1}{\delta'}.$$

Remark 5.10. Note that this lemma also develops upper bounds for two additional quantities: $\sum_{k,h} \max \{ \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, V_{h+1}^k \rangle, 0 \}$ and $\sum_{k,h} \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, (V_{h+1}^k)^2 \rangle$, which follow from similar arguments and will be useful in subsequent analyses.

Step 4: combining and solving all inter-dependent inequalities. We now collect the bounds established thus far for the terms $T_1, \dots, T_4$ in (5.49), together with several related quantities. Following the analysis of UCB-VI in Section 5.3.5, we introduce the auxiliary quantities below:

$$T_5 := \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{\hat{P}_{s_h^k, a_h^k, h}^k} (V_{h+1}^k), \quad (5.57a)$$

$$T_6^{\text{ucb}} := \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}^k} (V_{h+1}^k), \quad (5.57b)$$

$$W_3 := \sum_{k=1}^K \sum_{h=1}^H \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, (V_{h+1}^k)^2 \rangle. \quad (5.57c)$$

$$W_4 := \sum_{k=1}^K \sum_{h=1}^H \max \{ \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}^k, V_{h+1}^k \rangle, 0 \}, \quad (5.57d)$$

To begin with, it follows from Lemma 5.14 that, with probability at least $1 - \delta'$,

$$T_1 \leq W_4 \leq \sqrt{T_6^{\text{ucb}} SAH \iota_b} + SAH^2 \iota_b \quad (5.58a)$$

$$W_3 \leq H \sqrt{T_6^{\text{ucb}} SAH \iota_b} + SAH^3 \iota_b \quad (5.58b)$$

Next, repeating the same arguments as in Lemma 5.8, we can show that

$$T_2 \leq \frac{1}{2} \sqrt{SAHT_5 \iota_b} + \frac{1}{2} SAH^2 \iota_b \quad (5.58c)$$

$$|T_3| \leq \sqrt{T_6^{\text{ucb}} \iota_b} + H \iota_b \quad (5.58d)$$

$$|T_4| \leq \sqrt{KH^2 \iota_b} \quad (5.58e)$$

hold with probability at least $1 - (4 \log_2 K) \delta'$. In addition, repeating the same arguments as in the proof of Lemma 5.9 reveals that, with probability at least $1 - 2\delta' \log(KH)$,

$$T_5 \leq W_3 + 2HT_2 + \sqrt{H^2 T_6^{\text{ucb}} \iota_b} + KH^2 \iota_b, \quad (5.58f)$$

At this point, the only term that remains to be controlled is T_6^{ucb} . In the analysis of UCB-VI in Section 5.3.5, this term is handled by bounding T_6^{low} and T_6^{diff} separately. Here, we take a different approach that bounds T_6^{ucb} directly in terms of the quantities introduced above. This result is formalized in the following lemma; the proof is provided in Section 5.8.4.

Lemma 5.15. With probability at least $1 - 4SAHK\delta'$,

$$T_6^{\text{ucb}} \leq 4HT_2 + 4HW_4 + 12KH^2 + 38H^2 \log \frac{1}{\delta'}.$$

As an immediate consequence of this lemma, we have

$$T_6^{\text{ucb}} \leq 4HT_2 + 4HW_4 + KH^2\iota_b, \quad (5.58g)$$

thereby bounding T_6^{ucb} in terms of the quantities T_2 and W_4 , both of which have already been controlled.

To finish up, solving all inequalities in (5.58) and using the regret decomposition (5.49), we reach

$$\text{Regret}(K) \leq T_1 + T_2 + T_3 + T_4 \lesssim \sqrt{SAH^3K}\iota_b \quad (5.59)$$

as claimed. The details of solving (5.58) to establish (5.59) are postponed to Section 5.8.5.

5.5 Model-free online RL

Having established the statistical efficiency of model-based methods for online RL, we now shift our attention to model-free algorithms. In contrast to their model-based counterparts, the best currently known model-free approaches still fail to achieve minimax-optimal regret guarantees over the full range of sample sizes. For this reason, rather than providing a treatment as comprehensive as that for model-based methods, we restrict ourselves to a brief overview of the main ideas underlying optimistic Q-learning algorithms.

Broadly speaking, the key question is how to incorporate the principle of optimism into Q-learning in order to encourage efficient exploration, for instance by constructing suitable upper confidence estimates for the optimal Q-values. One of the most notable developments along this direction is the optimistic Q-learning algorithm — commonly referred to as Q-UCB — developed by Jin *et al.* (2018), whose complete procedure is summarized in Algorithm 5.4. Compared to the vanilla Q-learning algorithm introduced in Chapter 4, the update rule of Q-UCB incorporates two key modifications:

- First, the learning rate for each visited entry (s_h, a_h) is chosen adaptively according to the following linearly rescaled schedule:

$$\eta_n = \frac{H+1}{H+n}, \quad (5.60)$$

where n denotes the number of visits accumulated thus far for (s_h, a_h) at step h . This carefully designed learning-rate schedule plays an important role in controlling error propagation across the horizon and, consequently, in achieving a regret bound with polynomial dependence on H .

- Second, the Q-learning update now incorporates an exploration bonus b_n , analogous to the bonus used in the UCB-VI algorithm, in order to ensure that the learned Q-function remains an optimistic estimate (i.e., an upper confidence bound) of the optimal Q-function. More specifically, the optimistic Q-learning update takes the form

$$\begin{aligned} Q_h^{\text{ucb}}(s_h, a_h) &\leftarrow (1 - \eta_n)Q_h^{\text{ucb}}(s_h, a_h) + \eta_n\{r_h(s_h, a_h) + V_{h+1}^{\text{ucb}}(s_{h+1}) + b_n\}, \\ V_h^{\text{ucb}}(s_h) &\leftarrow \min \left\{ \max_{a' \in \mathcal{A}} Q_h^{\text{ucb}}(s_h, a'), H \right\}, \end{aligned}$$

where Q_h^{ucb} (resp. V_h^{ucb}) denotes the running estimate of the optimal Q-function (resp. value function), and b_n denotes the exploration bonus depending on the visitation count n described

above. A simple and natural choice of the bonus term can be derived from Hoeffding-type concentration inequalities:

$$b_n = c_b \sqrt{\frac{H^3 \log(SAT/\delta)}{n}}, \quad (5.61)$$

where $c_b > 0$ is some suitable universal constant, and $\delta \in (0, 1)$ is the usual confidence parameter.

The resulting regret guarantee for this model-free online RL algorithm, originally established by Jin *et al.* (2018), is summarized below.

Theorem 5.16 (Jin *et al.* (2018)). For any $K \geq 1$ and any $0 < \delta < 1$, the Q-UCB algorithm in Algorithm 5.4 with Hoeffding-style bonuses (5.61) achieves

$$\text{Regret}(K) \leq \tilde{O}(\sqrt{SAH^5 K}) \quad (5.62)$$

with probability at least $1 - \delta$, provided that the constant c_b is suitably chosen.

In comparison with the statistical guarantees achieved by model-based methods (see, e.g., Theorem 5.10), the regret guarantee in Theorem 5.16 remains statistically suboptimal in its dependence on the horizon length H . This dependence can be improved through more refined uncertainty quantification for the exploration bonus. For instance, incorporating Bernstein-style exploration bonuses reduces the regret bound to $\tilde{O}(\sqrt{SAH^4 K})$ for sufficiently large K (Jin *et al.*, 2018).

Moreover, additional variance-reduction techniques — often referred to as *reference-advantage decomposition* in the online model-free RL literature (Zhang *et al.*, 2020f; Li *et al.*, 2021b) — can further improve the regret guarantee to become asymptotically minimax optimal. Nevertheless, these refined guarantees still rely on an extremely large burn-in phase before statistically optimal behavior is attained. We refer interested readers to the aforementioned work for further details.

5.5.1 Analysis: proof of Theorem 5.16

We follow the proof strategy of Jin *et al.* (2018, Section 4), adapted to the notation used above. We start with recording a few useful properties of the learning rates. For notational convenience, we further introduce two sequences of related quantities defined for any integer $N \geq 0$ and $n \geq 1$:

$$\eta_n^N := \begin{cases} \eta_n \prod_{i=n+1}^N (1 - \eta_i), & \text{if } N > n, \\ \eta_n, & \text{if } N = n, \\ 0, & \text{if } N < n, \end{cases} \quad (5.64)$$

and

$$\eta_0^N := \begin{cases} \prod_{i=1}^N (1 - \eta_i) = 0, & \text{if } N > 0, \\ 1, & \text{if } N = 0. \end{cases} \quad (5.65)$$

As can be easily verified, we have

$$\sum_{n=1}^N \eta_n^N = \begin{cases} 1, & \text{if } N > 0, \\ 0, & \text{if } N = 0. \end{cases} \quad (5.66)$$

The following properties play an important role in the analysis.

Algorithm 5.4: Q-learning with upper confidence bounds (Q-UCB) for episodic online RL

```

1 input: number of episodes  $K$ , reward function  $r$ , confidence parameter  $\delta$ .
2 initialize  $N_h(s, a) \leftarrow 0$ ,  $Q_h^{\text{ucb}}(s, a) \leftarrow H$ , and  $V_{H+1}(s) \leftarrow 0$  for all  $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ .
3 for  $k = 1, \dots, K$  do
    /* policy update. */
4     update policy  $\pi$  such that  $\pi_h(s) \leftarrow \arg \max_a Q_h^{\text{ucb}}(s, a)$  for all  $(s, h) \in \mathcal{S} \times [H]$ .
5     observe the initial state  $s_1 \stackrel{\text{ind.}}{\sim} \rho$  from a new episode.
6     for  $h = 1, \dots, H$  do
        /* sampling. */
7         take  $a_h = \pi_h(s_h)$ , observe  $s_{h+1}$ ; set  $(s, a, s') \leftarrow (s_h, a_h, s_{h+1})$ .
        /* update of visitation counts. */
8          $N_h(s, a) \leftarrow N_h(s, a) + 1$ ;  $n \leftarrow N_h(s, a)$ .
        /* optimistic update of Q-functions & values. */
9         compute  $\eta_n$  and  $b_n$  according to (5.60) and (5.61).
10        update
            
$$Q_h^{\text{ucb}}(s, a) \leftarrow (1 - \eta_n)Q_h^{\text{ucb}}(s, a) + \eta_n \{r_h(s, a) + V_{h+1}^{\text{ucb}}(s') + b_n\}, \quad (5.63a)$$

            
$$V_h^{\text{ucb}}(s) \leftarrow \min \left\{ \max_{a' \in \mathcal{A}} Q_h^{\text{ucb}}(s, a'), H \right\}. \quad (5.63b)$$


```

Lemma 5.17. For the learning rates in (5.60) and the weights in (5.64), the following properties hold. For every integer $N \geq 1$,

$$\frac{1}{\sqrt{N}} \leq \sum_{n=1}^N \frac{\eta_n^N}{\sqrt{n}} \leq \frac{2}{\sqrt{N}}, \quad (5.67a)$$

$$\sum_{n=1}^N (\eta_n^N)^2 \leq \frac{2H}{N}. \quad (5.67b)$$

In addition, for every integer $n \geq 1$,

$$\sum_{N=n}^{\infty} \eta_n^N = 1 + \frac{1}{H}. \quad (5.67c)$$

We shall adopt the following set of notation explicitly indexed by the episode number k , detailed below.

- (s_h^k, a_h^k) : the state-action pair encountered and chosen at time step h in the k -th episode;
- $N_h^k(s, a)$: the visitation count $N_h(s, a)$ for the (s, a, h) -tuple before episode k ;
- $b_h^k(s, a) \geq 0$: the latest bonus term $b_h(s, a)$ before episode k ;
- $P_h^k \in \{0, 1\}^{1 \times |\mathcal{S}|}$: the empirical transition at time step h in the k -th episode, namely,

$$P_h^k(s) = \mathbb{1} \{s = s_{h+1}^k\}. \quad (5.68)$$

- $Q_h^k(s, a)$ and $V_h^k(s)$ denote respectively $Q_h^{\text{ucb}}(s, a)$ and $V_h^{\text{ucb}}(s)$ before episode k .

Recursion. We begin with unrolling the iterates and write out the Q-learning recursion. For any $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, let $t = N_h^k(s, a)$ and suppose that (s, a) was previously taken at step h in episodes $k_1, \dots, k_t < k$. Then unrolling the updates leads to

$$Q_h^k(s, a) = \eta_0^t H + \sum_{i=1}^t \eta_i^t \left\{ r_h(s, a) + V_{h+1}^{k_i}(s_{h+1}^{k_i}) + b_i \right\}, \quad (5.69)$$

where we make use of the initialization $Q_h^0 = H$. On the other hand, the Bellman optimality equation gives

$$Q_h^*(s, a) = r_h(s, a) + P_{s,a,h} V_{h+1}^*. \quad (5.70)$$

Consequently, it follows that

$$\begin{aligned} & (Q_h^k - Q_h^*)(s, a) \\ &= \eta_0^t (H - Q_h^*(s, a)) + \sum_{i=1}^t \eta_i^t \left\{ (V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) + [(P_h^{k_i} - P_{s,a,h}) V_{h+1}^*](s, a) + b_i \right\}, \end{aligned} \quad (5.71)$$

where we used $\eta_0^t + \sum_{i=1}^t \eta_i^t = 1$ (cf. (5.66)).

Next, a crucial property is that the Q-estimate provides an optimistic view of the true optimal Q-function, which parallels Lemma 5.5. To simplify presentation, throughout this analysis, we denote

$$\iota = \log(SAT/\delta).$$

Lemma 5.18. There exists a sufficiently large absolute constant in the bonus choice (5.61) such that, with probability at least $1 - \delta$,

$$0 \leq (Q_h^k - Q_h^*)(s, a) \leq \eta_0^t H + \sum_{i=1}^t \eta_i^t (V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) + \beta_t, \quad (5.72)$$

holds simultaneously for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, where

$$\beta_t \leq C \sqrt{\frac{H^3 \iota}{t}} \quad (5.73)$$

for some absolute constant $C > 0$. Then it follows that

$$Q_h^k(s, a) \geq Q_h^*(s, a) \quad \text{and} \quad V_h^k(s) \geq V_h^*(s). \quad (5.74)$$

We now proceed with the regret analysis.

Step 1: regret decomposition. Lemma 5.18 allows one to upper bound the regret as follows

$$\text{Regret}(K) := \sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)) \leq \sum_{k=1}^K (V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)), \quad (5.75)$$

with probability at least $1 - \delta$. Thus inequality (5.75) reduces our task to controlling

$$\sum_{k=1}^K (V_1^k - V_1^{\pi^k})(s_1^k). \quad (5.76)$$

Fix any $(k, h) \in [K] \times [H]$, and suppose that (s_h^k, a_h^k) was previously taken at step h in episodes with $t = N_h^k(s_h^k, a_h^k)$:

$$k_1, \dots, k_t < k. \quad (5.77)$$

Since the policy π^k is greedy with respect to Q_h^k , we have

$$\begin{aligned} (V_h^k - V_h^{\pi^k})(s_h^k) &\leq (Q_h^k - Q_h^{\pi^k})(s_h^k, a_h^k) \\ &= (Q_h^k - Q_h^*) (s_h^k, a_h^k) + P_{h, s_h^k, a_h^k} (V_{h+1}^* - V_{h+1}^{\pi^k}). \end{aligned} \quad (5.78)$$

Combining this with Lemma 5.18 gives

$$\begin{aligned} (V_h^k - V_h^{\pi^k})(s_h^k) &\leq \eta_0^t H + \sum_{i=1}^t \eta_i^t (V_{h+1}^{k_i} - V_{h+1}^*) (s_{h+1}^{k_i}) + \beta_t + P_{h, s_h^k, a_h^k} (V_{h+1}^* - V_{h+1}^{\pi^k}) \\ &= \eta_0^t H + \sum_{i=1}^t \eta_i^t (V_{h+1}^{k_i} - V_{h+1}^*) (s_{h+1}^{k_i}) + \beta_t - (V_{h+1}^k - V_{h+1}^*) (s_{h+1}^k) \\ &\quad + (V_{h+1}^k - V_{h+1}^{\pi^k}) (s_{h+1}^k) + [(P_{h, s_h^k, a_h^k} - P_h^k) (V_{h+1}^* - V_{h+1}^{\pi^k})] (s_h^k, a_h^k), \end{aligned} \quad (5.79)$$

where the last term is a martingale difference term.

Step 2: summing the one-step recursion. We sum (5.79) over $k \in [K]$. The first term satisfies

$$\sum_{k=1}^K \eta_0^{N_h^k(s_h^k, a_h^k)} H = \sum_{k=1}^K H \cdot \mathbf{1}\{N_h^k(s_h^k, a_h^k) = 0\} \leq SAH. \quad (5.80)$$

For the weighted terms involving $V_{h+1}^{k_i} - V_{h+1}^*$, we regroup the summation by the episode in which each such term is generated. If episode k' is the i -th visit to its corresponding (s, a, h) tuple, then its future coefficients are bounded by $\sum_{N=i}^\infty \eta_i^N \leq 1 + 1/H$ by (5.67c). Therefore,

$$\sum_{k=1}^K \sum_{i=1}^{N_h^k(s_h^k, a_h^k)} \eta_i^{N_h^k(s_h^k, a_h^k)} (V_{h+1}^{k_i} - V_{h+1}^*) (s_{h+1}^{k_i}) \leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^K (V_{h+1}^k - V_{h+1}^*) (s_{h+1}^k). \quad (5.81)$$

Plugging (5.80) and (5.81) into (5.79), and using

$$(V_{h+1}^k - V_{h+1}^*) (s_{h+1}^k) \leq (V_{h+1}^k - V_{h+1}^{\pi^k}) (s_{h+1}^k), \quad (5.82)$$

since $V_{h+1}^* \geq V_{h+1}^{\pi^k}$, yields

$$\begin{aligned} &\sum_{k=1}^K (V_h^k - V_h^{\pi^k})(s_h^k) \\ &\leq SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^K (V_{h+1}^k - V_{h+1}^*) (s_{h+1}^k) - \sum_{k=1}^K (V_{h+1}^k - V_{h+1}^*) (s_{h+1}^k) \\ &\quad + \sum_{k=1}^K (V_{h+1}^k - V_{h+1}^{\pi^k}) (s_{h+1}^k) + \sum_{k=1}^K \left\{ \beta_{N_h^k(s_h^k, a_h^k)} + [(P_{h, s_h^k, a_h^k} - P_h^k) (V_{h+1}^* - V_{h+1}^{\pi^k})] (s_h^k, a_h^k) \right\} \\ &\leq SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^K (V_{h+1}^k - V_{h+1}^{\pi^k}) (s_{h+1}^k) \end{aligned}$$

$$+ \sum_{k=1}^K \left\{ \beta_{N_h^k(s_h^k, a_h^k)} + [(P_{h, s_h^k, a_h^k} - P_h^k)(V_{h+1}^* - V_{h+1}^{\pi^k})](s_h^k, a_h^k) \right\}. \quad (5.83)$$

Recurring (5.83) over $h = 1, \dots, H$ and using

$$(V_{H+1}^k - V_{H+1}^{\pi^k})(s_{H+1}^k) = 0 \quad (5.84)$$

gives

$$\begin{aligned} & \sum_{k=1}^K (V_1^k - V_1^{\pi^k})(s_1^k) \\ & \leq O\left(H^2 SA + \sum_{h=1}^H \sum_{k=1}^K \beta_{N_h^k(s_h^k, a_h^k)} + \sum_{h=1}^H \sum_{k=1}^K [(P_{h, s_h^k, a_h^k} - P_h^k)(V_{h+1}^* - V_{h+1}^{\pi^k})](s_h^k, a_h^k)\right), \end{aligned} \quad (5.85)$$

where we used $(1 + 1/H)^H \leq e$.

Step 3: controlling the bonus and martingale terms. For any fixed $h \in [H]$, by Lemma 5.18 and the pigeonhole principle,

$$\begin{aligned} \sum_{k=1}^K \beta_{N_h^k(s_h^k, a_h^k)} & \leq C \sum_{s,a} \sum_{n=1}^{N_h^{K+1}(s,a)} \sqrt{\frac{H^3 \iota}{n}} \leq C \sqrt{H^3 \iota} \sum_{s,a} \sqrt{N_h^{K+1}(s,a)} \\ & \leq C \sqrt{H^3 SAK \iota}. \end{aligned} \quad (5.86)$$

Regarding the last term in (5.85), each term in the summation is a martingale difference term, and bounded by H in absolute value. By the Azuma–Hoeffding inequality and a union bound over the events already used in Lemma 5.18, with probability at least $1 - \delta$,

$$\left| \sum_{h=1}^H \sum_{k=1}^K [(P_{h, s_h^k, a_h^k} - P_h^k)(V_{h+1}^* - V_{h+1}^{\pi^k})](s_h^k, a_h^k) \right| \leq CH \sqrt{HK \iota}. \quad (5.87)$$

This follows the same argument as relation (5.164). Combining (5.85), (5.86), and (5.87) gives

$$\text{Regret}(K) \leq C \left(H^2 SA + \sqrt{H^5 SAK \iota} \right). \quad (5.88)$$

Equivalently, using $T = HK$, this is

$$\text{Regret}(K) \leq C \left(H^2 SA + \sqrt{H^4 SAT \iota} \right). \quad (5.89)$$

Finally, since the per-episode regret is at most H , the total regret is always at most $T = HK$. As in Jin *et al.* (2018), if $T \leq \sqrt{H^4 SAT \iota}$ then the desired bound follows from $\text{Regret}(K) \leq T$; otherwise $\sqrt{H^4 SAT \iota}$ dominates the lower-order term $H^2 SA$ up to an absolute constant. This establishes

$$\text{Regret}(K) \leq \tilde{O}(\sqrt{SAH^5 K}), \quad (5.90)$$

which proves Theorem 5.16.

5.6 Minimax lower bounds

To provide theoretical benchmarks for assessing the statistical optimality of online RL algorithms, a line of research has been devoted to establishing regret lower bounds (e.g., Jaksch *et al.* (2010),

Bartlett and Tewari (2009), Osband and Van Roy (2016), Jin *et al.* (2018), and Domingues *et al.* (2021)), drawing in part on techniques originally developed for multi-armed bandits. In what follows, we present a minimax regret lower bound, originally established in Jin *et al.* (2018) and Domingues *et al.* (2021), which applies uniformly over all values of K .

Theorem 5.19. Consider any $S \geq 2$, $A \geq 3$ and $H \geq 2$. There exists a collection $\mathcal{C}_{\text{mdp}}$ of finite-horizon nonstationary MDPs — each with S states, A actions and horizon H — together with an initial state distribution $\rho \in \Delta(S)$, such that for any $K \geq 1$,

$$\inf_{\hat{\pi}} \sup_{P \in \mathcal{C}_{\text{mdp}}} \mathbb{E} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi^k}(\rho)) \right] \geq c_0 \min \{ \sqrt{H^3 S A K}, H K \}$$

holds for some universal constant $c_0 > 0$, where the infimum is taken over any online learning algorithm $\hat{\pi}$. Here, $\hat{\pi}$ generates a sequence of policies $\hat{\pi} = \{\pi^k\}_{k=1}^K$, and each π^k may depend arbitrarily on the trajectories collected before episode k (but not on any observations from episode k or later episodes).

Since a regret of at most HK is trivially achievable — because the cumulative reward collected over K episodes cannot exceed HK — we shall often drop the linear term and simply write

$$\Omega(\sqrt{SAH^3K}) \quad \text{if } K \geq SAH. \quad (5.91)$$

Importantly, this lower bound taken together with Theorem 5.10, rigorously justifies the minimax optimality of model-based online RL (modulo some logarithmic factors).

The proof of Theorem 5.19 is given in Sections 5.6.1 and 5.6.2, which differs from the arguments in Jin *et al.* (2018) and Domingues *et al.* (2021). Our proof is once again built upon Fano’s lemma and proceeds in two stages: we first establish the lower bound for MDPs with two states (Section 5.6.1) by extending the corresponding bandit lower bound analysis, and then lift the resulting lower bound to MDPs with $S \geq 2$ states (Section 5.6.2).

5.6.1 Analysis: lower bound for a simpler case with $S = 2$

For clarity of exposition, we first tackle the case $S = 2$, before extending the argument to arbitrary $S \geq 2$ in Section 5.6.2. To be precise, we intend to establish the existence of a family $\mathcal{C}_{\text{mdp}}^{\text{sub}}$ of MDPs — each with $S = 2$ states, $A \geq 3$ actions, and horizon length $H \geq 2$ — as well as an initial state distribution ρ , such that

$$\inf_{\hat{\pi}} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi^k}(\rho)) \right] \geq c_{\text{lb}} \min \{ \sqrt{H^3 A K}, H K \} \quad (5.92)$$

holds for some universal constant $c_{\text{lb}} > 0$.

Construction of the MDPs. As before, we begin by constructing a family of finite-horizon nonstationary MDPs, denoted by $\mathcal{C}_{\text{mdp}}^{\text{sub}}$. Each MDP $\mathcal{M}$ in this family is specified by the following components.

- *State space.* The state space $\mathcal{S}$ consists of 2 states:

$$\mathcal{S} = \{0, 1\}.$$

- *Action space.* State 0 and state 1 are associated respectively with the following action space:

$$\mathcal{A}_0 = \{0\} \quad \text{and} \quad \mathcal{A}_1 = \{1, \dots, A\}.$$

- *Reward function.* The reward function $r = \{r_h\}_{1 \leq h \leq H}$ is given by

$$r_h(0, 0) = 0 \quad \text{and} \quad r_h(1, a) = 1 \quad \text{for every } a \in \mathcal{A}_1 \quad (5.93)$$

for each $1 \leq h \leq H$.

- *Optimal actions and probability transition kernel.* Since the action space $\mathcal{A}_0$ associated with state 0 is a singleton, it suffices to specify the optimal action at state 1 for each step $1 \leq h \leq H$. For an MDP $\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}$, we denote these optimal actions by

$$a_1^{*, \mathcal{M}}, \dots, a_H^{*, \mathcal{M}} \in \mathcal{A}_1 \quad (5.94)$$

with $a_h^{*, \mathcal{M}}$ denoting the optimal action at state 1 and step h . The probability transition kernel $P^{\mathcal{M}}$ of $\mathcal{M}$ is then defined by

$$P_h^{\mathcal{M}}(0 | 0, 0) = 1, \quad (5.95a)$$

$$P_h^{\mathcal{M}}(0 | 1, a) = q_{a_h^{*, \mathcal{M}}}(a), \quad P_h^{\mathcal{M}}(1 | 1, a) = 1 - q_{a_h^{*, \mathcal{M}}}(a) \quad (5.95b)$$

for every $h \in [H]$ and $a \in \mathcal{A}_1$. By construction, state 0 is absorbing and yields zero reward. Thus, the behavior of the MDP is largely determined by the dynamics at state 1. We will specify the function $q_{a_h^{*, \mathcal{M}}}(\cdot)$ next and verify that the resulting transition kernel is indeed consistent with the designated optimal actions (5.94).

- *Initial state distribution.* In what follows, we take the initial state distribution ρ to be concentrated on state 1, namely,

$$\rho(s) = \begin{cases} 1, & \text{if } s = 1; \\ 0, & \text{else.} \end{cases} \quad (5.96)$$

It remains to specify the functions $q_{a_h^{*, \mathcal{M}}}(\cdot)$ appearing in (5.95). For notational simplicity, consider an arbitrary sequence of optimal actions $a_1^*, \dots, a_H^*$. For each $h \in [H]$, we define

$$\text{when } a_h^* \neq 1 : \quad q_{a_h^*}(a) = \begin{cases} \frac{1}{H} - \frac{\varepsilon_0}{H^2} & \text{if } a = a_h^*, \\ \frac{1}{H} - \frac{\varepsilon_0}{2H^2} & \text{if } a = 1, \\ \frac{1}{H}, & \text{else,} \end{cases} \quad (5.97a)$$

$$\text{when } a_h^* = 1 : \quad q_{a_h^*}(a) = \begin{cases} \frac{1}{H} - \frac{\varepsilon_0}{2H^2} & \text{if } a = a_h^* = 1, \\ \frac{1}{H}, & \text{else,} \end{cases} \quad (5.97b)$$

where the parameter $0 < \varepsilon_0 \leq H$ will be specified later (see (5.108)). The consistency of the resulting transition kernel with the designated optimal actions is immediate. Indeed, since state 0 is absorbing and yields zero reward, maximizing the value function is equivalent to minimizing $q_{a_h^*}(a)$ over $a \in \mathcal{A}_1$. By (5.97), this minimum is attained uniquely at $a = a_h^*$ for every step h .

As it turns out, it suffices to consider a collection $\mathcal{C}_{\text{mdp}}^{\text{sub}}$ consisting of A MDPs. Specifically, let

$$\mathcal{C}_{\text{mdp}}^{\text{sub}} = \{\mathcal{M}_1, \dots, \mathcal{M}_A\}, \quad (5.98)$$

where for each $1 \leq i \leq A$, the MDP $\mathcal{M}_i$ is constructed as above with the optimal actions chosen to be

$$a_h^{*,\mathcal{M}_i} = i \quad \text{for all } 1 \leq h \leq H. \quad (5.99)$$

Basic properties of each MDP instance $\mathcal{M}$ from $\mathcal{C}_{\text{mdp}}^{\text{sub}}$. In the sequel, we make note of several basic properties that hold for an arbitrary MDP $\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}$. For notational convenience, we often suppress the dependence of quantities (e.g., Q_h^π , V_h^π , d_h^π , $a_h^{*,\mathcal{M}}$, $P_h^\mathcal{M}$) on the underlying MDP $\mathcal{M}$, as long as it is clear from the context.

- *Value functions and Q-functions.* Given that state 0 is absorbing with zero immediate rewards, one has

$$Q_h^\pi(0, 0) = 0 \quad \text{for all } 1 \leq h \leq H. \quad (5.100a)$$

Also, since $V_{H+1}^\pi(0) = V_{H+1}^\pi(1) = 0$, we see from (5.93) that

$$V_H^\pi(1) = \sum_{a \in \mathcal{A}_1} \pi_H(a | 1) r_H(1, a) = 1, \quad (5.100b)$$

$$V_H^\pi(0) = r_H(0, 0) = 0. \quad (5.100c)$$

Moreover, for any $1 \leq h \leq H$ and any $a \in \mathcal{A}_1$, the Bellman equation (cf. (2.40a)) together with (5.93) and (5.100a) gives

$$\begin{aligned} Q_h^\pi(1, a) &= r_h(1, a) + P_h(1 | 1, a) V_{h+1}^\pi(1) + P_h(0 | 1, a) V_{h+1}^\pi(0) \\ &= 1 + P_h(1 | 1, a) V_{h+1}^\pi(1). \end{aligned} \quad (5.100d)$$

According to the construction (5.95) and (5.97), we have

$$P_h(1 | 1, a) \geq 1 - 1/H \quad \text{for all } a \in \mathcal{A}_1, \quad (5.101)$$

which combined with (5.100d) immediately leads to

$$\begin{aligned} Q_h^\pi(1, a) &\geq 1 + (1 - 1/H) V_{h+1}^\pi(1) \quad \text{for all } a \in \mathcal{A}_1, \\ \implies V_h^\pi(1) &\geq 1 + (1 - 1/H) V_{h+1}^\pi(1). \end{aligned}$$

Iterating this recursion and using $V_H^\pi(1) = 1$ (cf. (5.100b)) yield

$$\begin{aligned} V_h^\pi(1) &\geq 1 + (1 - 1/H) \left\{ 1 + (1 - 1/H) V_{h+2}^\pi(1) \right\} \geq \dots \\ &\geq \sum_{i=0}^{H-h} (1 - 1/H)^i = \frac{1 - (1 - 1/H)^{H-h+1}}{1/H} \geq \frac{1 - e^{-\frac{H-h+1}{H}}}{1/H} \\ &\geq \frac{\frac{1}{2} \frac{H-h+1}{H}}{1/H} = \frac{H-h+1}{2}, \end{aligned} \quad (5.102)$$

where we have used the elementary inequalities $1 - x \leq \exp(-x)$ and $1 - \exp(-x) \geq x/2$ for $0 \leq x \leq 1$. Additionally, it also follows directly from (5.100d) that

$$V_h^*(1) = Q_h^*(1, a_h^*) = 1 + P_h(1 | 1, a_h^*) V_{h+1}^*(1). \quad (5.103)$$

- *Occupancy distributions.* We next lower bound the occupancy distribution. Under the initial state distribution (5.96), the occupancy distribution associated with any policy π (see the definition in (2.34)) satisfies

$$\begin{aligned} d_h^\pi(1; \rho) &= \sum_{a \in \mathcal{A}} d_{h-1}^\pi(1, a; \rho) P_{h-1}(1 | 1, a) \\ &\geq (1 - 1/H) \sum_{a \in \mathcal{A}} d_{h-1}^\pi(1, a; \rho) = (1 - 1/H) d_{h-1}^\pi(1; \rho), \end{aligned}$$

where the inequality results from (5.101). Recursively applying this bound yields

$$d_h^\pi(1; \rho) \geq (1 - 1/H)^{h-1} d_1^\pi(1; \rho) = (1 - 1/H)^{h-1} \geq e^{-1} \quad (5.104)$$

for each $h \in [H]$, where we use $(1 - 1/H)^H \geq e^{-1}$.

A lower bound on the suboptimality gap. Now, we analyze the suboptimality gap associated with policy π . Towards this, we first use (5.100d) and (5.103) to demonstrate that

$$\begin{aligned} V_h^*(1) - V_h^\pi(1) &= 1 + P_h(1 | 1, a_h^*) V_{h+1}^*(1) - \left\{ 1 + \mathbb{E}_{a_h \sim \pi_h(\cdot | 1)} [P_h(1 | 1, a_h) V_{h+1}^\pi(1)] \right\} \\ &= \mathbb{E}_{a_h \sim \pi_h(\cdot | 1)} [P_h(1 | 1, a_h^*) V_{h+1}^*(1) - P_h(1 | 1, a_h) V_{h+1}^\pi(1)] \\ &= \mathbb{E}_{a_h \sim \pi_h(\cdot | 1)} [P_h(1 | 1, a_h^*) (V_{h+1}^*(1) - V_{h+1}^\pi(1))] + \mathbb{E}_{a_h \sim \pi_h(\cdot | 1)} [(P_h(1 | 1, a_h^*) - P_h(1 | 1, a_h)) V_{h+1}^\pi(1)] \\ &\geq \left(1 - \frac{1}{H}\right) (V_{h+1}^*(1) - V_{h+1}^\pi(1)) + \frac{\varepsilon_0}{2H^2} V_{h+1}^\pi(1) \mathbb{E}_{a_h \sim \pi_h(\cdot | 1)} [\mathbb{1}(a_h \neq a_h^*)] \\ &\geq \left(1 - \frac{1}{H}\right) (V_{h+1}^*(1) - V_{h+1}^\pi(1)) + \frac{\varepsilon_0(H-h)}{4H^2} (1 - \pi_h(a_h^* | 1)), \end{aligned} \quad (5.105)$$

where the last relation follows from (5.102), and the penultimate inequality holds true since, according to (5.95) and (5.97),

$$\begin{aligned} P_h(1 | 1, a_h^*) &\geq 1 - 1/H, \\ P_h(1 | 1, a_h^*) - P_h(1 | 1, a_h) &\geq \frac{\varepsilon_0}{2H^2} \mathbb{1}(a_h \neq a_h^*). \end{aligned}$$

Applying the above property (5.105) recursively, we arrive at

$$\begin{aligned} V_1^*(1) - V_1^\pi(1) &\geq \left(1 - \frac{1}{H}\right)^H (V_{H+1}^*(1) - V_{H+1}^\pi(1)) + \sum_{h=1}^H \left(1 - \frac{1}{H}\right)^h \frac{\varepsilon_0(H-h)}{4H^2} (1 - \pi_h(a_h^* | 1)) \\ &= \sum_{h=1}^H \left(1 - \frac{1}{H}\right)^h \frac{\varepsilon_0(H-h)}{4H^2} (1 - \pi_h(a_h^* | 1)) \\ &\geq \sum_{h=1}^H \frac{\varepsilon_0(H-h)}{4eH^2} (1 - \pi_h(a_h^* | 1)), \end{aligned} \quad (5.106)$$

where the last inequality holds since $(1 - 1/H)^h \geq (1 - 1/H)^H \geq e^{-1}$. In other words, this lower bound (5.106) on the suboptimality gap hinges upon how often policy π selects suboptimal actions at each step.

A detour: a lower bound on a bandit problem. To further control the right-hand side of (5.106), we take a brief detour to establish a lower bound in a closely related but simpler bandit setting. The formulation is the following.

- Consider a collection $\mathcal{C}_{\text{bandit}}$ consisting of A bandit instances, where each bandit instance has A arms.
- For $i \in [A]$, let $\mathcal{B}_i$ denote the bandit instance whose unique optimal arm is $a_{\mathcal{B}_i}^* = i$. The mean reward of arm a in $\mathcal{B}_i$ is given by

$$r_{\mathcal{B}_i}(a) = 1 - q_i(a),$$

where $q_i(\cdot)$ is defined in (5.97). More concretely, pulling arm a in round t in the bandit instance $\mathcal{B}_i$, yields an independent Bernoulli reward as

$$R_{a,t} = \begin{cases} 1, & \text{with probability } 1 - q_i(a), \\ 0, & \text{else.} \end{cases}$$

In what follows, let $\mathbb{P}_{\mathcal{B}_i}$ denote the probability distribution under the bandit instance $\mathcal{B}_i$. As before, the learner interacts with the bandit instance in a sequential manner; at each round t , the learner selects an arm based only on the observations collected prior to round t .

We now apply Fano's lemma to develop a round-wise lower bound for this collection of bandit instances.

Lemma 5.20. Consider the collection $\mathcal{C}_{\text{bandit}}$ of bandit instances defined above. Suppose that $A \geq 3$. Then for any $t \leq \frac{H^3 A}{18\varepsilon_0^2}$, one has

$$\inf_{\hat{\pi}} \frac{1}{A} \sum_{i=1}^A \mathbb{P}_{\mathcal{B}_i}(a_{t+1} \neq i) \geq \frac{1}{4},$$

where the infimum is taken over all online bandit algorithms $\hat{\pi}$. Here, for a given algorithm $\hat{\pi}$, denote by $\{\pi^t\}_{t \geq 1}$ the resulting sequence of policies, where π^t depends only on observations collected prior to round t ; the arm a_t is then selected according to π^t .

In words, the lemma states that, at any round $t + 1$, the probability of selecting a suboptimal action — based on the observations collected through round t — is bounded away from zero. Importantly, this is a per-round guarantee that holds whenever t is not too large. As will become clear shortly, it is more convenient to develop this result in terms of the average probability of error rather than the minimax probability of error. The proof is deferred to Section 5.10.

Regret analysis. We now leverage the preceding error-probability lower bound for bandits to derive a regret lower bound for the collection $\mathcal{C}_{\text{mdp}}^{\text{sub}}$ of MDPs. Consider any online learning algorithm $\hat{\pi}$. In view of (5.106), one can derive

$$\mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(1) - V_1^{\pi^k}(1)) \right] \geq \sum_{k=1}^K \sum_{h=1}^H \frac{\varepsilon_0(H-h)}{4eH^2} \mathbb{P}_{\mathcal{M}}(a_h^k \neq a_h^{*,\mathcal{M}})$$

for any $\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}$. As a result,

$$\begin{aligned}
\inf_{\hat{\pi}} \max_{i: 1 \leq i \leq A} \mathbb{E}_{\mathcal{M}_i} \left[\sum_{k=1}^K (V_1^*(1) - V_1^{\pi^k}(1)) \right] &\geq \inf_{\hat{\pi}} \frac{1}{A} \sum_{i=1}^A \mathbb{E}_{\mathcal{M}_i} \left[\sum_{k=1}^K (V_1^*(1) - V_1^{\pi^k}(1)) \right] \\
&\geq \inf_{\hat{\pi}} \sum_{k=1}^K \sum_{h=1}^H \frac{\varepsilon_0(H-h)}{4eH^2} \left\{ \frac{1}{A} \sum_{i=1}^A \mathbb{P}_{\mathcal{M}_i}(a_h^k \neq i) \right\} \\
&\geq \sum_{k=1}^K \sum_{h=1}^H \frac{\varepsilon_0(H-h)}{4eH^2} \inf_{\hat{\pi}} \left\{ \frac{1}{A} \sum_{i=1}^A \mathbb{P}_{\mathcal{M}_i}(a_h^k \neq i) \right\} \\
&\geq \frac{1}{4} \sum_{k=1}^K \sum_{h=1}^H \frac{\varepsilon_0(H-h)}{4eH^2} = \frac{\varepsilon_0 K(H-1)}{32eH} \geq \frac{\varepsilon_0 K}{64e}, \quad (5.107)
\end{aligned}$$

provided that $K \leq \frac{AH^3}{18\varepsilon_0^2}$ and $\varepsilon_0 \leq H$.

To finish up, choosing

$$\varepsilon_0 = \min \left\{ \sqrt{\frac{H^3 A}{18K}}, H \right\}, \quad (5.108)$$

clearly satisfies $K \leq \frac{AH^3}{18\varepsilon_0^2}$ and $\varepsilon_0 \leq H$. Substituting this choice into (5.107), we arrive at

$$\begin{aligned}
\inf_{\hat{\pi}} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(1) - V_1^{\pi^k}(1)) \right] &\geq \inf_{\hat{\pi}} \frac{1}{|\mathcal{C}_{\text{mdp}}^{\text{sub}}|} \sum_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(1) - V_1^{\pi^k}(1)) \right] \\
&\geq \frac{\varepsilon_0 K}{64e} \asymp \min \{ \sqrt{H^3 A K}, H K \}. \quad (5.109)
\end{aligned}$$

This taken together with the choice of ρ (cf. (5.96)) establishes the advertised lower bound (5.92) for the case with $S = 2$.

5.6.2 Analysis: lower bound for general S (proof of Theorem 5.19)

We now demonstrate how the regret lower bound established in Section 5.6.1 for $S = 2$ can be lifted to obtain the desired lower bound for general S . Throughout this proof, we assume without loss of generality that S is even.

Construction of the MDPs and notation. We begin by constructing a family $\mathcal{C}_{\text{mdp}}$ of MDPs as follows.

- Each MDP $\mathcal{M}$ within this family is composed of $S/2$ disjoint sub-MDPs, each containing 2 states.
- For each $1 \leq j \leq S/2$, the j -th sub-MDP of $\mathcal{M}$ — denoted by $\mathcal{M}^{\text{sub},j}$ — has state space $\{j, j + S/2\}$, and is constructed according to the same structure as the MDPs in the collection $\mathcal{C}_{\text{mdp}}^{\text{sub}}$ introduced in Section 5.6.1.
- The initial state distribution $\rho \in \Delta(\mathcal{S})$ is taken to be

$$\rho(s) = \begin{cases} 2/S, & \text{if } 1 \leq s \leq S/2, \\ 0, & \text{else.} \end{cases} \quad (5.110)$$

For clarity of exposition, we introduce the augmented regret notation below to make explicit the dependence on the learning algorithm, the underlying MDP and the initial state distribution. Specifically, for a given online learning algorithm $\hat{\pi}$ (which generates a policy sequence $\{\pi^k\}$) and any $\mathcal{M} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}$ (see Section 5.6.1), we write

$$\text{Regret}_{\hat{\pi}}(K; \mathcal{M}) := \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(1) - V_1^{\pi^k}(1)) \right], \quad (5.111)$$

which denotes the *expected* regret incurred by $\hat{\pi}$ over K episodes in the MDP $\mathcal{M}$. We emphasize that this metric is in terms of expected regret, rather than realized regret, as this formulation is more convenient for the subsequent analysis.

Analysis when $K < S/2$. We now move on to the regret analysis for a simple case with $K < S/2$. In this case, $\min\{\sqrt{H^3 S A K}, HK\} = HK$ and it is sufficient to establish the lower bound of order HK . Under this condition on K , there exists a state subset $\mathcal{S}^{\text{low}} \subseteq \{1, \dots, S/2\}$ with cardinality at least $S/4$ such that: for every $j \in \mathcal{S}^{\text{low}}$, the state j is selected as the initial state in at most two of the K episodes. Equivalently, this means there are at least $S/4$ sub-MDPs that are visited no more than twice over the course of K episodes. Since each of these sub-MDPs receives at most two visits, the learner has insufficient information to reliably identify its optimal actions. Consequently, the argument developed for the two-state construction immediately yields a regret lower bound of order HK . We omit the details here for brevity.

Analysis when $K \geq S/2$. We now turn to the more challenging regime where $K \geq S/2$. For any underlying MDP $\mathcal{M} \in \mathcal{C}_{\text{mdp}}$, we first make note of the following decomposition across the $S/2$ sub-MDPs:

$$\begin{aligned} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi^k}(\rho)) \right] &= \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K \mathbb{E}_{s_1^k \sim \rho} [V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)] \right] \\ &= \sum_{k=1}^K \mathbb{E}_{\mathcal{M}} \mathbb{E}_{s_1^k \sim \rho} \left[\sum_{j=1}^{S/2} (V_1^*(j) - V_1^{\pi^k}(j)) \mathbb{1}\{s_1^k = j\} \right] \\ &= \mathbb{E}_{s_1^1, \dots, s_1^K \stackrel{\text{i.i.d.}}{\sim} \rho} \left[\underbrace{\sum_{j=1}^{S/2} \sum_{k=1}^K \mathbb{E}_{\mathcal{M}} [(V_1^*(j) - V_1^{\pi^k}(j)) \mathbb{1}\{s_1^k = j\}]}_{=: Q_j^{\mathcal{M}}} \right], \quad (5.112) \end{aligned}$$

where the penultimate line is valid since ρ (cf. (5.110)) is supported on $\{1, \dots, S/2\}$. Next, for each $1 \leq j \leq S/2$, define

$$K_j := \sum_{k=1}^K \mathbb{1}\{s_1^k = j\},$$

to be the number of episodes whose initial state lies in the j -th sub-MDP $\mathcal{M}^{\text{sub},j}$. Since the sub-MDPs are disjoint and independent, we see that: conditional on K_j , the quantity $Q_j^{\mathcal{M}}$ coincides with the regret incurred by $\hat{\pi}$ over K_j episodes in $\mathcal{M}^{\text{sub},j}$, i.e.,

$$Q_j^{\mathcal{M}} = \text{Regret}_{\hat{\pi}}(K_j; \mathcal{M}^{\text{sub},j});$$

see the notation (5.111). This allows one to further derive

$$\begin{aligned}
& \inf_{\hat{\pi}} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi_k}(\rho)) \right] \geq \inf_{\hat{\pi}} \frac{1}{|\mathcal{C}_{\text{mdp}}|} \sum_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi_k}(\rho)) \right] \\
& = \inf_{\hat{\pi}} \frac{1}{|\mathcal{C}_{\text{mdp}}|} \sum_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[\sum_{j=1}^{S/2} \text{Regret}_{\hat{\pi}}(K_j; \mathcal{M}^{\text{sub},j}) \right] \\
& = \inf_{\hat{\pi}} \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[\sum_{j=1}^{S/2} \frac{1}{|\mathcal{C}_{\text{mdp}}^{\text{sub}}|} \sum_{\mathcal{M}^{\text{sub},j} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}} \text{Regret}_{\hat{\pi}}(K_j; \mathcal{M}^{\text{sub},j}) \right] \\
& \geq \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[\sum_{j=1}^{S/2} \inf_{\hat{\pi}} \frac{1}{|\mathcal{C}_{\text{mdp}}^{\text{sub}}|} \sum_{\mathcal{M}^{\text{sub},j} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}} \text{Regret}_{\hat{\pi}}(K_j; \mathcal{M}^{\text{sub},j}) \right]. \tag{5.113}
\end{aligned}$$

It therefore remains to lower bound each term appearing in the summation on the right-hand side of (5.113). By (5.109), we have

$$\inf_{\hat{\pi}} \frac{1}{|\mathcal{C}_{\text{mdp}}^{\text{sub}}|} \sum_{\mathcal{M}^{\text{sub},j} \in \mathcal{C}_{\text{mdp}}^{\text{sub}}} \text{Regret}_{\hat{\pi}}(K_j; \mathcal{M}^{\text{sub},j}) \geq \frac{\varepsilon_0 K_j}{64e}, \tag{5.114}$$

provided that $K_j \leq \frac{AH^3}{18\varepsilon_0^2}$. Substituting this bound into (5.113) gives

$$\begin{aligned}
& \inf_{\hat{\pi}} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi_k}(\rho)) \right] \geq \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[\sum_{j=1}^{S/2} \frac{\varepsilon_0 K_j}{64e} \mathbb{1} \left\{ K_j \leq \frac{AH^3}{18\varepsilon_0^2} \right\} \right] \\
& \geq \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[\sum_{j=1}^{S/2} \frac{\varepsilon_0 K_j}{64e} \right] - \frac{\varepsilon_0}{64e} \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[\sum_{j=1}^{S/2} K_j \mathbb{1} \left\{ K_j > \frac{AH^3}{18\varepsilon_0^2} \right\} \right] \\
& = \frac{\varepsilon_0 K}{64e} - \frac{\varepsilon_0}{64e} \sum_{j=1}^{S/2} \mathbb{E}_{s_1^1, \dots, s_1^K \sim \rho} \left[K_j \mathbb{1} \left\{ K_j > \frac{AH^3}{18\varepsilon_0^2} \right\} \right]. \tag{5.115}
\end{aligned}$$

To complete the proof, it remains to control the tail term $\mathbb{E} [K_j \mathbb{1} \{K_j > \frac{AH^3}{18\varepsilon_0^2}\}]$. Since K_j is the sum of K independent Bernoulli random variables each with mean $2/S$, Lemma 2.7 implies that

$$\begin{aligned}
\mathbb{E} \left[K_j \mathbb{1} \left\{ K_j > \frac{AH^3}{18\varepsilon_0^2} \right\} \right] & \leq \mathbb{E} \left[K_j \mathbb{1} \left\{ K_j > \frac{100K}{S} \right\} \right] \\
& = \frac{100K}{S} \mathbb{P} \left\{ K_j > \frac{100K}{S} \right\} + \int_{100K/S}^{\infty} \mathbb{P} \{K_j > \tau\} d\tau \\
& \leq \frac{100K}{S} \exp \left(-\frac{19K}{S} \right) + 2 \int_{100K/S}^{\infty} \exp(-0.19\tau) d\tau \\
& \leq \frac{111K}{S} \exp \left(-\frac{19K}{S} \right) \leq \frac{0.01K}{S}
\end{aligned}$$

provided that $K \geq S/2$, where the first inequality is valid as long as $\varepsilon_0^2 \leq \frac{SAH^3}{1800K}$, the second line relies on the basic property

$$\mathbb{E}[X \mathbb{1} \{X > x\}] = x\mathbb{P}(X > x) + \int_x^{\infty} \mathbb{P}(X > \tau) d\tau.$$

Substitution into (5.115) gives

$$\begin{aligned} \inf_{\widehat{\pi}} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi_k}(\rho)) \right] &\geq \frac{\varepsilon_0 K}{64e} - \frac{\varepsilon_0(S/2)}{64e} \cdot \frac{0.01K}{S} \\ &= \frac{0.995\varepsilon_0 K}{64e}. \end{aligned}$$

To finish up, we take

$$\varepsilon_0 = \min \left\{ \sqrt{\frac{SAH^3}{1800K}}, H \right\}$$

in the above display to conclude that

$$\inf_{\widehat{\pi}} \max_{\mathcal{M} \in \mathcal{C}_{\text{mdp}}} \mathbb{E}_{\mathcal{M}} \left[\sum_{k=1}^K (V_1^*(\rho) - V_1^{\pi_k}(\rho)) \right] \gtrsim \min \{ \sqrt{SAH^3 K}, HK \}.$$

5.7 Appendix A: proof of auxiliary lemmas in Section 5.3.5

5.7.1 Proof of Lemma 5.5

We begin by introducing a scalar-valued function f , which facilitates the subsequent analysis:

$$f(p, v, n) := \langle p, v \rangle + \max \left\{ c_1 \sqrt{\frac{\text{Var}_p(v) \log \frac{1}{\delta'}}{n}}, \frac{c_1^2 H \log \frac{1}{\delta'}}{n} \right\}, \quad (5.116)$$

defined for any probability vector $p = [p_s]_{1 \leq s \leq S} \in \Delta^S$, any non-negative vector $v = [v_s]_{1 \leq s \leq S} \in \mathbb{R}^S$ obeying $\|v\|_\infty \leq H$, and any positive integer n . A key property of this function, which we rely on in this proof, is its monotonicity with respect to v , namely,

$$f(p, v, n) \text{ is non-decreasing in each entry of } v. \quad (5.117)$$

The proof of this property is postponed to the end of this subsection.

We are now positioned to prove Lemma 5.5. Consider any tuple (h, k, s, a) , and we divide into two cases.

Case 1: $N_h^k(s, a) \leq 2$. In this situation, the update rule (5.18) yields the following trivial bounds:

$$Q_h^k(s, a) = H \geq Q_h^*(s, a) \quad \text{and} \quad V_h^k(s) = H \geq V_h^*(s),$$

provided that the constant c_2 (cf. (5.20)) exceeds $c_2 \geq 2$.

Case 2: $N_h^k(s, a) > 2$. Suppose now, by way of induction, that $Q_{h+1}^k \geq Q_{h+1}^*$, which necessarily implies that $V_{h+1}^k \geq V_{h+1}^*$. If $Q_h^k(s, a) = H$, then $Q_h^k(s, a) \geq Q_h^*(s, a)$ holds trivially. Therefore, it suffices to restrict attention to those state-action pairs obeying $Q_h^k(s, a) < H$ in the sequel.

In view of the update rule (5.18) and the bonus (5.20), we have

$$Q_h^k(s, a) = r_h(s, a) + \langle \widehat{P}_{s,a,h}^k, V_{h+1}^k \rangle + c_1 \sqrt{\frac{\text{Var}_{\widehat{P}_{s,a,h}^k}(V_{h+1}^k) \log \frac{1}{\delta'}}{N_h^k(s, a)}} + c_2 \frac{H \log \frac{1}{\delta'}}{N_h^k(s, a)}$$

$$\begin{aligned}
&\geq r_h(s, a) + \frac{(c_2 - c_1^2)H \log \frac{1}{\delta'}}{N_h^k(s, a)} + f(\hat{P}_{s,a,h}^k, V_{h+1}^k, N_h^k(s, a)) \\
&\geq r_h(s, a) + f(\hat{P}_{s,a,h}^k, V_{h+1}^*, N_h^k(s, a))
\end{aligned} \tag{5.118}$$

as long as $c_2 \geq c_1^2$, where the last inequality results from the claim (5.117) and the property $V_{h+1}^k \geq V_{h+1}^*$. Moreover, applying the empirical Bernstein inequality in Lemma 2.6 yields

$$\begin{aligned}
&\mathbb{P}\left\{\left|\langle \hat{P}_{s,a,h}^k - P_{s,a,h}, V_{h+1}^* \rangle\right| > 2\sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}^k}(V_{h+1}^*) \log \frac{1}{\delta'}}{N_h^k(s, a)} + \frac{14H \log \frac{1}{\delta'}}{3N_h^k(s, a)}}\right\} \\
&\leq \mathbb{P}\left\{\left|\langle \hat{P}_{s,a,h}^k - P_{s,a,h}, V_{h+1}^* \rangle\right| > \sqrt{\frac{2\text{Var}_{\hat{P}_{s,a,h}^k}(V_{h+1}^*) \log \frac{1}{\delta'}}{N_h^k(s, a) - 1} + \frac{\frac{7}{3}H \log \frac{1}{\delta'}}{N_h^k(s, a) - 1}}\right\} \leq 2\delta'.
\end{aligned}$$

This in turn implies that, with probability exceeding $1 - 2\delta'$,

$$\begin{aligned}
f(\hat{P}_{s,a,h}^k, V_{h+1}^*, N_h^k(s, a)) &= \langle P_{s,a,h}, V_{h+1}^* \rangle + \langle \hat{P}_{s,a,h}^k - P_{s,a,h}, V_{h+1}^* \rangle \\
&\quad + \max\left\{c_1\sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}^k}(V_{h+1}^*) \log \frac{1}{\delta'}}{N_h^k(s, a)}}, \frac{c_1^2 H \log \frac{1}{\delta'}}{N_h^k(s, a)}\right\} \geq \langle P_{s,a,h}, V_{h+1}^* \rangle,
\end{aligned}$$

where the last line holds as long as $c_1 \geq 20/3$. Substitution into (5.118) then yields that, with probability at least $1 - 2\delta'$,

$$Q_h^k(s, a) \geq r_h(s, a) + \langle P_{s,a,h}, V_{h+1}^* \rangle = Q_h^*(s, a),$$

where the identity is due to the Bellman equation (2.40).

To finish up, taking the union bound over all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ concludes the proof.

Putting all this together. With the above two cases in place, one can invoke standard induction arguments to deduce that: with probability at least $1 - 4SAHK\delta'$, one has $Q_h^k(s, a) \geq Q_h^*(s, a)$ and $V_h^k = \max_a Q_h^k(s, a) \geq \max_a Q_h^*(s, a) = V_h^*(s)$ for every (s, a, h, k) , thus completing the proof.

Proof of property (5.117). Fix any index $1 \leq s \leq S$, and consider f as a function of the s -th entry v_s while freezing p, n , and all other entries of v . We begin with the following observations:

- f is a continuous function w.r.t. v ;
- f is differentiable in v_s , except for at most two possible points that make $c_1\sqrt{\frac{\text{Var}_p(v) \log \frac{1}{\delta'}}{n}} = c_1^2 \frac{H \log \frac{1}{\delta'}}{n}$ (due to the presence of non-differentiable points of the max operator).

By virtue of the continuity of f , establishing that f is non-decreasing in v_s reduces to verifying that $\frac{\partial f(p, v, n)}{\partial v_s} \geq 0$ at all differentiable points.

To this end, we compute the partial derivative at any differentiable point v_s as follows:

$$\frac{\partial f(p, v, n)}{\partial v_s} = p_s + c_1 \mathbb{1}\left\{c_1\sqrt{\frac{\text{Var}_p(v) \log \frac{1}{\delta'}}{n}} \geq c_1^2 \frac{H \log \frac{1}{\delta'}}{n}\right\} \cdot \frac{p_s(v_s - \langle p, v \rangle)\sqrt{\log \frac{1}{\delta'}}}{\sqrt{n\text{Var}_p(v)}}$$

$$= p_s + \mathbb{1}(\mathcal{A}_v) \cdot \frac{c_1 H \log \frac{1}{\delta'}}{\sqrt{n \text{Var}_p(v) \log \frac{1}{\delta'}}} \cdot \frac{p_s(v_s - \langle p, v \rangle)}{H}, \quad (5.119)$$

where we introduce the event $\mathcal{A}_v := \{\sqrt{n \text{Var}_p(v) \log \frac{1}{\delta'}} \geq c_1 H \log \frac{1}{\delta'}\}$. We now divide into two cases.

- If $v_s - \langle p, v \rangle \geq 0$, then it is trivially seen that $\frac{\partial f(p, v, n)}{\partial v_s} \geq 0$.
- If $v_s - \langle p, v \rangle < 0$, then one can continue the derivation in (5.119) and use the definition of $\mathcal{A}_v$ to reach

$$\frac{\partial f(p, v, n)}{\partial v_s} \geq p_s + p_s \frac{(v_s - \langle p, v \rangle)}{H} = p_s \frac{H + v_s - \langle p, v \rangle}{H} \geq 0,$$

where the last inequality holds since $v_s \geq 0$ and $\langle p, v \rangle \leq H$.

Combining these two cases immediately establishes property (5.117). $\square$

5.7.2 Proof of Lemma 5.7

Bounding the quantity W_1 . We begin by applying Bernstein's inequality (cf. Lemma 2.5) in conjunction with a union bound over all tuples $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$. More specifically, since $\|V_{h+1}^*\|_\infty \leq H$, it holds with probability at least $1 - 4SAHK\delta'$ that

$$\left| \left\langle \hat{P}_{s,a,h}^k - P_{s,a,h}, V_{h+1}^* \right\rangle \right| \leq \sqrt{\frac{2\text{Var}_{P_{s,a,h}}(V_{h+1}^*) \log \frac{1}{\delta'}}{N_h^k(s, a)}} + \frac{H \log \frac{1}{\delta'}}{N_h^k(s, a)} \quad (5.120)$$

simultaneously for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$. Using this relation to bound each summand in W_1 , we obtain

$$\begin{aligned} W_1 &\leq \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{2\text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^*) \log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)}} + \sum_{k=1}^K \sum_{h=1}^H \frac{H \log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)} \\ &\leq \sqrt{\sum_{k,h} \frac{2 \log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)}} \sqrt{\sum_{k,h} \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^*)} + H \log \frac{1}{\delta'} \sum_{k,h} \frac{1}{N_h^k(s_h^k, a_h^k)} \\ &\leq \sqrt{2SAHT_6^{\text{opt}}(\log_2 K) \log \frac{1}{\delta'}} + SAH^2(\log_2 K) \log \frac{1}{\delta'} \end{aligned} \quad (5.121)$$

with probability greater than $1 - 4SAHK\delta'$, where the second line invokes the Cauchy-Schwarz inequality, and the last inequality makes use of Lemma (5.3) and the definition (5.33c) of T_6^{opt} .

To finish up, note that for any two random variables X and Y ,

$$\text{Var}(X + Y) \leq 2\text{Var}(X) + 2\text{Var}(Y) \leq 2\text{Var}(X) + 2\mathbb{E}[Y^2].$$

This combined with the inequality $V_{h+1}^k \geq V_{h+1}^* \geq V_{h+1}^{\pi^k}$ (see Lemma 5.5) gives rise to

$$\begin{aligned} \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^*) &\leq 2\text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^{\pi^k}) + 2\left\langle P_{s_h^k, a_h^k, h}, (V_{h+1}^* - V_{h+1}^{\pi^k})^2 \right\rangle \\ &\leq 2\text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^{\pi^k}) + 2\left\langle P_{s_h^k, a_h^k, h}, (V_{h+1}^k - V_{h+1}^{\pi^k})^2 \right\rangle, \end{aligned}$$

which taken together with the definition (5.33b) results in

$$T_6^{\text{opt}} \leq 2T_6^{\text{low}} + 2T_6^{\text{diff}} \quad (5.122a)$$

$$\implies \sqrt{T_6^{\text{opt}}} \leq \sqrt{2T_6^{\text{low}}} + \sqrt{2T_6^{\text{diff}}}. \quad (5.122b)$$

This together with (5.121) and the definition of ι_b (cf. (5.28)) establishes the claimed bound

$$W_1 \leq \frac{1}{8} \sqrt{\iota_b SAHT_6^{\text{low}}} + \frac{1}{8} \sqrt{\iota_b SAHT_6^{\text{diff}}} + \frac{1}{8} \iota_b SAH^2.$$

Remark 5.11. Repeating the above arguments for (5.122) also yields

$$T_6^{\text{ucb}} \leq 2T_6^{\text{low}} + 2T_6^{\text{diff}} \quad (5.123a)$$

$$\implies \sqrt{T_6^{\text{ucb}}} \leq \sqrt{2T_6^{\text{low}}} + \sqrt{2T_6^{\text{diff}}}, \quad (5.123b)$$

which will prove useful in the subsequent analysis.

Bounding the quantity W_2 . We now turn our attention to the term W_2 in (5.34). Setting $\Delta_{h+1}^k := V_{h+1}^k - V_{h+1}^*$ for notational convenience and denoting by $v(s)$ the s -th entry of a vector v , we derive

$$\begin{aligned} \left\langle \hat{P}_{s,a,h}^k - P_{s,a,h}, V_{h+1}^k - V_{h+1}^* \right\rangle &= \sum_{s' \in \mathcal{S}} \left(\hat{P}_{s,a,h}^k(s') - P_{s,a,h}(s') \right) \Delta_{h+1}^k(s') \\ &\stackrel{(i)}{=} \sum_{s' \in \mathcal{S}} \left(\hat{P}_{s,a,h}^k(s') - P_{s,a,h}(s') \right) \Delta_{h+1}^k(s') - \left\{ \sum_{s' \in \mathcal{S}} \left(\hat{P}_{s,a,h}^k(s') - P_{s,a,h}(s') \right) \right\} \langle P_{s,a,h}, \Delta_{h+1}^k \rangle \\ &= \sum_{s' \in \mathcal{S}} \left(\hat{P}_{s,a,h}^k(s') - P_{s,a,h}(s') \right) \left(\Delta_{h+1}^k(s') - \langle P_{s,a,h}, \Delta_{h+1}^k \rangle \right) \\ &\leq \sum_{s' \in \mathcal{S}} \left| \hat{P}_{s,a,h}^k(s') - P_{s,a,h}(s') \right| \cdot \left| \Delta_{h+1}^k(s') - \langle P_{s,a,h}, \Delta_{h+1}^k \rangle \right|, \end{aligned} \quad (5.124)$$

where (i) follows since $\sum_{s'} \hat{P}_{s,a,h}^k(s') = \sum_{s'} P_{s,a,h}(s') = 1$. The Bernstein inequality (cf. Lemma 2.5) combined with the union bound tells us that, with probability greater than $1 - 4S^2 AHK\delta'$,

$$\begin{aligned} \left| \hat{P}_{s,a,h}^k(s') - P_{s,a,h}(s') \right| &= \left| \left\langle \hat{P}_{s,a,h}^k - P_{s,a,h}, e_{s'} \right\rangle \right| \\ &\leq \sqrt{\frac{2\text{Var}_{P_{s,a,h}}(e_{s'}) \log \frac{1}{\delta'}}{N_h^k(s, a)}} + \frac{2 \log \frac{1}{\delta'}}{3N_h^k(s, a)} \\ &\leq \sqrt{\frac{2P_{s,a,h}(s') \log \frac{1}{\delta'}}{N_h^k(s, a)}} + \frac{\log \frac{1}{\delta'}}{N_h^k(s, a)} \end{aligned} \quad (5.125)$$

holds simultaneously for all $(s, a, s', h, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \times [K]$, which uses the fact that $\text{Var}_{P_{s,a,h}}(e_{s'}) = P_{s,a,h}(s')(1 - P_{s,a,h}(s')) \leq P_{s,a,h}(s')$. Substitution into (5.124) then leads to

$$\begin{aligned} &\left\langle \hat{P}_{s,a,h}^k - P_{s,a,h}, V_{h+1}^k - V_{h+1}^* \right\rangle \\ &\leq \sum_{s' \in \mathcal{S}} \left(\sqrt{\frac{2P_{s,a,h}(s') \log \frac{1}{\delta'}}{N_h^k(s, a)}} + \frac{\log \frac{1}{\delta'}}{N_h^k(s, a)} \right) \left| \Delta_{h+1}^k(s') - \langle P_{s,a,h}, \Delta_{h+1}^k \rangle \right| \end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} \sqrt{\frac{2S \log \frac{1}{\delta'}}{N_h^k(s, a)}} \sqrt{\sum_{s' \in \mathcal{S}} P_{s, a, h}(s') \left| \Delta_{h+1}^k(s') - \langle P_{s, a, h}, \Delta_{h+1}^k \rangle \right|^2} + \sum_{s' \in \mathcal{S}} \frac{\log \frac{1}{\delta'}}{N_h^k(s, a)} \left| \Delta_{h+1}^k(s') - \langle P_{s, a, h}, \Delta_{h+1}^k \rangle \right| \\
&\stackrel{(iii)}{\leq} \sqrt{\frac{2S \log \frac{1}{\delta'}}{N_h^k(s, a)}} \sqrt{\text{Var}_{P_{s, a, h}}(V_{h+1}^k - V_{h+1}^*)} + \frac{SH \log \frac{1}{\delta'}}{N_h^k(s, a)},
\end{aligned}$$

where (ii) follows from the Cauchy-Schwarz inequality, and (iii) results from the definition of $\text{Var}_{P_{s, a, h}}(V)$ and the fact that $\Delta_{h+1}^k(s) \in [0, H]$ for all $s \in \mathcal{S}$ (since both $V_{h+1}^k(s)$ and $V_{h+1}^*(s)$ lie in $[0, H]$). Repeating similar arguments as in (5.121), we arrive at

$$\begin{aligned}
W_2 &\leq \sqrt{2S^2 AH (\log_2 K) \log \frac{1}{\delta'}} \sqrt{\sum_{k, h} \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k - V_{h+1}^*)} + S^2 AH^2 (\log_2 K) \log \frac{1}{\delta'} \quad (5.126) \\
&\leq \frac{1}{8} \sqrt{\iota_b S^2 AHT_6^{\text{diff}}} + \frac{1}{8} \iota_b S^2 AH^2
\end{aligned}$$

with probability at least $1 - 4S^2 AHK\delta'$, where we recall the definition of T_6^{diff} in (5.33e). Here, the last line follows since, according to the property $V_{h+1}^k \geq V_{h+1}^* \geq V_{h+1}^{\pi^k}$ (see Lemma 5.5),

$$\begin{aligned}
\text{Var}_{P_{s, a, h}}(V_{h+1}^k - V_{h+1}^*) &\leq \langle P_{s, a, h}, (V_{h+1}^k - V_{h+1}^*)^2 \rangle \leq \langle P_{s, a, h}, (V_{h+1}^k - V_{h+1}^{\pi^k})^2 \rangle, \\
&\implies \sum_{k, h} \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k - V_{h+1}^*) \leq T_6^{\text{diff}}. \quad (5.127)
\end{aligned}$$

Bounding the quantity W_3 . Akin to the decomposition (5.34), we split W_3 into two components:

$$W_3 = \underbrace{\sum_{k, h} \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, (V_{h+1}^*)^2 \rangle}_{=: W_4} + \underbrace{\sum_{k, h} \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, (V_{h+1}^k)^2 - (V_{h+1}^*)^2 \rangle}_{=: W_5}. \quad (5.128)$$

The resulting quantities W_4 and W_5 can then be controlled using arguments analogous to those employed for W_1 and W_2 , respectively. Additionally, it is helpful to record the following basic variance bound, quoted from Chen *et al.* (2021b, Lemma 30).

Lemma 5.21. Let X be a random variable, and denote by $C_{\max}$ the largest possible magnitude of X . Then we have $\text{Var}(X^2) \leq 4C_{\max}^2 \text{Var}(X)$.

Proof of Lemma 5.21. In view of the elementary inequality $\text{Var}(U + V) \leq 2\text{Var}(U) + 2\text{Var}(V)$, we can derive

$$\begin{aligned}
\text{Var}(X^2) &\leq 2\text{Var}((X - \mathbb{E}[X])X) + 2\text{Var}(\mathbb{E}[X]X) \\
&\leq 2\mathbb{E}[(X - \mathbb{E}[X])^2 X^2] + 2(\mathbb{E}[X])^2 \text{Var}(X) \\
&\leq 2C_{\max}^2 \mathbb{E}[(X - \mathbb{E}[X])^2] + 2C_{\max}^2 \text{Var}(X) = 4C_{\max}^2 \text{Var}(X)
\end{aligned}$$

as claimed. $\square$

Regarding the term W_4 in (5.128), repeating the arguments in (5.121) yields, with probability greater than $1 - 4SAHK\delta'$,

$$W_4 \leq \sqrt{2SAH (\log_2 K) \log \frac{1}{\delta'}} \sqrt{\sum_{k, h} \text{Var}_{P_{s_h^k, a_h^k, h}}((V_{h+1}^*)^2)} + SAH^3 (\log_2 K) \log \frac{1}{\delta'}$$

$$\begin{aligned}
&\leq \sqrt{8SAH^3T_6^{\text{opt}}(\log_2 K) \log \frac{1}{\delta'}} + SAH^3(\log_2 K) \log \frac{1}{\delta'} \\
&\leq \frac{1}{16} \sqrt{\iota_b SAH^3T_6^{\text{low}}} + \frac{1}{16} \sqrt{\iota_b SAH^3T_6^{\text{diff}}} + \frac{1}{16} \iota_b SAH^3.
\end{aligned}$$

Here, the last inequality follows from (5.122), while the penultimate inequality applies Lemma 5.21 and the property $\|V_{h+1}^*\|_\infty \leq H$ to obtain

$$\text{Var}_{P_{s_h^k, a_h^k, h}}((V_{h+1}^*)^2) \leq 4H^2 \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^*).$$

When it comes to the term W_5 in (5.128), repeat the arguments for (5.126) to derive

$$W_5 \leq \sqrt{2S^2AH(\log_2 K) \log \frac{1}{\delta'}} \sqrt{\sum_{k,h} \text{Var}_{P_{s,a,h}}((V_{h+1}^k)^2 - (V_{h+1}^*)^2)} + S^2AH^3(\log_2 K) \log \frac{1}{\delta'}. \quad (5.129)$$

Furthermore, recalling the property $H \geq V_{h+1}^k(s) \geq V_{h+1}^*(s) \geq V_{h+1}^{\pi^k}(s) \geq 0$ that holds for all $s \in \mathcal{S}$ (see Lemma 5.5), one can deduce that

$$\begin{aligned}
\text{Var}_{P_{s,a,h}}((V_{h+1}^k)^2 - (V_{h+1}^*)^2) &\leq \left\langle P_{s,a,h}, ((V_{h+1}^k)^2 - (V_{h+1}^*)^2)^2 \right\rangle \\
&\leq \left\langle P_{s,a,h}, ((V_{h+1}^k)^2 - (V_{h+1}^{\pi^k})^2)^2 \right\rangle \\
&= \left\langle P_{s,a,h}, ((V_{h+1}^k - V_{h+1}^{\pi^k}) \circ (V_{h+1}^k + V_{h+1}^{\pi^k}))^2 \right\rangle \\
&\leq 4H^2 \left\langle P_{s,a,h}, (V_{h+1}^k - V_{h+1}^{\pi^k})^2 \right\rangle.
\end{aligned}$$

Taking this inequality together with the definition (5.33), we arrive at

$$\begin{aligned}
W_5 &\leq \sqrt{8S^2AH^3T_6^{\text{diff}}(\log_2 K) \log \frac{1}{\delta'}} + S^2AH^3(\log_2 K) \log \frac{1}{\delta'} \\
&\leq \frac{1}{16} \sqrt{\iota_b S^2AH^3T_6^{\text{diff}}} + \frac{1}{16} \iota_b S^2AH^3.
\end{aligned}$$

Substituting the preceding bounds on W_4 and W_5 into (5.128), we conclude that

$$W_3 \leq W_4 + W_5 \leq \frac{1}{8} \sqrt{\iota_b SAH^3T_6^{\text{low}}} + \frac{1}{8} \sqrt{\iota_b S^2AH^3T_6^{\text{diff}}} + \frac{1}{8} \iota_b S^2AH^3$$

as claimed.

5.7.3 Proof of Lemma 5.8

Bounding the term T_2 . We begin by analyzing T_2 (cf. (5.32)). Recalling the definition of the bonus term (5.20), we have

$$T_2 = \sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k) = c_1 \sum_{k,h} \sqrt{\frac{\text{Var}_{\hat{P}_{s_h^k, a_h^k, h}^k}(V_{h+1}^k) \log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)}} + c_2 \sum_{k,h} \frac{H \log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)}. \quad (5.130)$$

Applying the Cauchy-Schwarz inequality to the first sum in (5.130) yields

$$T_2 \leq c_1 \sqrt{\sum_{k,h} \frac{\log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)}} \sqrt{\sum_{k,h} \text{Var}_{\hat{P}_{s_h^k, a_h^k, h}^k}(V_{h+1}^k)} + c_2 H \log \frac{1}{\delta'} \sum_{k,h} \frac{1}{N_h^k(s_h^k, a_h^k)}$$

$$\begin{aligned}
&= c_1 \sqrt{\sum_{k,h} \frac{\log \frac{1}{\delta'}}{N_h^k(s_h^k, a_h^k)}} \sqrt{T_5} + c_2 H \log \frac{1}{\delta'} \sum_{k,h} \frac{1}{N_h^k(s_h^k, a_h^k)} \\
&\leq c_1 \sqrt{SAH(\log_2 K) \left(\log \frac{1}{\delta'} \right) T_5} + c_2 SAH^2(\log_2 K) \log \frac{1}{\delta'} \\
&\leq \frac{1}{2} \sqrt{\iota_b SAHT_5} + \frac{1}{2} \iota_b SAH^2,
\end{aligned}$$

where the second relation uses the definition (5.33a) of T_5 , the penultimate line invokes the visitation count property (5.25), and the last line holds as long as $\iota_b \geq 4 \max\{c_1^2, c_2\}(\log_2 K) \log \frac{1}{\delta'}$.

Bounding the term T_3 . Next, we bound the term T_3 defined in (5.32). Recall that the value function estimate V_{h+1}^k is computed prior to the start of episode k . Consequently, conditional on the state-action pair (s_h^k, a_h^k) and the history up to the beginning of episode k , the next-state indicator vector $e_{s_{h+1}^k}$ is statistically independent of V_{h+1}^k and has conditional mean $P_{s_h^k, a_h^k, h}$. Furthermore, since the entries of V_{h+1}^k lie within $[0, H]$, one has the deterministic bound:

$$|\langle P_{s_h^k, a_h^k, h} - e_{s_{h+1}^k}, V_{h+1}^k \rangle| \leq H.$$

Also, this combined with the definition of T_6 in (5.33b) immediately reveals that $T_6 \leq KH^2$.

Therefore, using the definition of T_6 in (5.33b), one can invoke the Freedman inequality (see Lemma 2.9 with $K_0 = \log_2 K$) to show that, with probability at least $1 - 2(\log_2 K)\delta'$,

$$\begin{aligned}
|T_3| &\leq \sqrt{8 \max\{T_6, H^2\} \log \frac{1}{\delta'}} + \frac{4}{3} H \log \frac{1}{\delta'} \\
&\leq \sqrt{8T_6 \log \frac{1}{\delta'}} + \sqrt{8} H \log \frac{1}{\delta'} + \frac{4}{3} H \log \frac{1}{\delta'} \\
&\leq \sqrt{8T_6 \log \frac{1}{\delta'}} + 5H \log \frac{1}{\delta'},
\end{aligned}$$

where we have used the elementary inequality $\sqrt{\max\{a, b\}} \leq \sqrt{a} + \sqrt{b}$. Furthermore, repeating the same arguments as in (5.122) reveals that

$$T_6 \leq 2T_6^{\text{low}} + 2T_6^{\text{diff}}, \quad (5.131)$$

which together with the above bound on T_3 and the fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ readily yields

$$|T_3| \leq 4\sqrt{T_6^{\text{low}} \log \frac{1}{\delta'}} + 4\sqrt{T_6^{\text{diff}} \log \frac{1}{\delta'}} + 5H \log \frac{1}{\delta'}.$$

Bounding the term T_4 . We now turn attention to the term T_4 , i.e.,

$$T_4 = \sum_{k=1}^K \underbrace{\left(\sum_{h=1}^H r_h(s_h^k, a_h^k) - V_1^{\pi^k}(s_1^k) \right)}_{=: E_k}, \quad (5.132)$$

Note that conditional on π^k , E_k is a zero-mean random variable bounded in magnitude by H . An application of the Azuma-Hoeffding inequality (cf. Lemma 2.8) guarantees that, with probability exceeding $1 - 2\delta'$,

$$|T_4| = \left| \sum_{k=1}^K E_k \right| \leq \sqrt{2KH^2 \log \frac{1}{\delta'}}.$$

5.7.4 Proof of Lemma 5.9

To facilitate the analysis, we introduce the following auxiliary quantity:

$$T_8 := \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h} - e_{s_{h+1}^k}, (V_{h+1}^k)^2 \right\rangle. \quad (5.133)$$

Given that this quantity will appear in the upper bounds of other terms, we first derive a bound for T_8 .

Bounding the term T_8 . According to Lemma 5.21, $\text{Var}(X^2) \leq 4H^2 \text{Var}(X)$ holds for any random variable X with support on $[-H, H]$. Therefore, invoking Freedman's inequality (see Lemma 2.9 with $K_0 = \log_2(KH)$) reveals that, with probability at least $1 - 2\delta' \log(KH)$,

$$\begin{aligned} |T_8| &\leq \sqrt{8 \max_{k,h} \left\{ \sum_{s_h^k, a_h^k, h} \text{Var}_{P_{s_h^k, a_h^k, h}}^k \left((V_{h+1}^k)^2 \right), H^4 \right\} \log \frac{1}{\delta'} + \frac{4}{3} H^2 \log \frac{1}{\delta'}} \\ &\leq \sqrt{8 \max_{k,h} \left\{ \sum_{s_h^k, a_h^k, h} \text{Var}_{P_{s_h^k, a_h^k, h}}^k \left((V_{h+1}^k)^2 \right), H^4 \right\} \log \frac{1}{\delta'} + \sqrt{8H^4 \log \frac{1}{\delta'}} + \frac{4}{3} H^2 \log \frac{1}{\delta'}} \\ &\leq \sqrt{8 \sum_{k,h} \text{Var}_{P_{s_h^k, a_h^k, h}}^k \left((V_{h+1}^k)^2 \right) \log \frac{1}{\delta'} + 5H^2 \log \frac{1}{\delta'}} \end{aligned} \quad (5.134)$$

$$\begin{aligned} &\leq \sqrt{32H^2 \sum_{k,h} \text{Var}_{P_{s_h^k, a_h^k, h}}^k (V_{h+1}^k) \log \frac{1}{\delta'} + 5H^2 \log \frac{1}{\delta'}} \\ &= \sqrt{32H^2 T_6^{\text{ucb}} \log \frac{1}{\delta'} + 5H^2 \log \frac{1}{\delta'}}, \end{aligned} \quad (5.135)$$

where the second line is valid since $\sqrt{\max\{a, b\}} \leq \sqrt{a} + \sqrt{b}$.

Bounding the term T_5 (cf. (5.33a)). Regarding the term T_5 , expanding the variance terms leads to

$$\begin{aligned} T_5 &= \sum_{k=1}^K \sum_{h=1}^H \left(\left\langle \hat{P}_{s_h^k, a_h^k, h}^k, (V_{h+1}^k)^2 \right\rangle - \left(\left\langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle \right)^2 \right) \\ &= \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, (V_{h+1}^k)^2 \right\rangle + \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h} - e_{s_{h+1}^k}, (V_{h+1}^k)^2 \right\rangle \\ &\quad + \sum_{k=1}^K \sum_{h=1}^H (V_{h+1}^k(s_{h+1}^k))^2 - \sum_{k=1}^K \sum_{h=1}^H \left(\left\langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle \right)^2 \\ &\leq W_3 + T_8 + \sum_{k=1}^K \sum_{h=2}^H (V_h^k(s_h^k))^2 - \sum_{k=1}^K \sum_{h=2}^H \left(\left\langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle \right)^2, \end{aligned} \quad (5.136)$$

where W_3 and T_8 are defined in (5.35) and (5.133), respectively. Here, in the last line we reindex the first sum (by replacing $h+1$ with h) and make use of the fact that $V_{H+1}^k = 0$.

In addition, we make the observation that

$$(V_h^k(s_h^k))^2 - \left(\left\langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle \right)^2 = \left(V_h^k(s_h^k) + \left\langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle \right) \left(V_h^k(s_h^k) - \left\langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \right\rangle \right)$$

$$\begin{aligned}
&\leq 2H \max \left\{ V_h^k(s_h^k) - \langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \rangle, 0 \right\} \\
&= 2H \max \left\{ Q_h^k(s_h^k, a_h^k) - \langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \rangle, 0 \right\},
\end{aligned}$$

where the inequality holds since $V_h^k(s) \leq H$ for all $s \in \mathcal{S}$, and the last identity holds since, due to the greedy policy choice, $V_h^k(s_h^k) = \max_a Q_h^k(s_h^k, a) = Q_h^k(s_h^k, a_h^k)$. Substitution back into (5.136) yields

$$\begin{aligned}
T_5 &\leq W_3 + T_8 + 2H \sum_{k=1}^K \sum_{h=2}^H \max \left\{ Q_h^k(s_h^k, a_h^k) - \langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \rangle, 0 \right\} \\
&\leq W_3 + T_8 + 2H \sum_{k=1}^K \sum_{h=2}^H b_h^k(s_h^k, a_h^k) + 2H \sum_{k=1}^K \sum_{h=2}^H r_h(s_h^k, a_h^k) \\
&\leq W_3 + T_8 + 2HT_2 + 2KH^2,
\end{aligned}$$

where the second line is a direct consequence of the update rule (5.18), and the last line uses the definition (5.37a) of T_2 combined with the assumption that $r_h(s, a) \leq 1$. Taking this collectively with (5.135) establishes the claimed bound (5.40a).

Bounding the term T_6^{low} (cf. (5.33d)). Let $\mathcal{F}_k$ denote the σ -field generated by all events prior to episode k , and define

$$U_k := \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^{\pi^k}).$$

It then follows directly from Lemma 5.4 that

$$\mathbb{E}[U_k \mid \mathcal{F}_k] \leq H^2.$$

Moreover, since $\|V_{h+1}^{\pi}\|_{\infty} \leq H$, we trivially have $U_k \leq H^3$. Applying the Azuma–Hoeffding inequality (Lemma 2.8) therefore yields that, with probability at least $1 - 2\delta'$,

$$T_6^{\text{low}} = \sum_{k=1}^K U_k \leq \mathbb{E} \left[\sum_{k=1}^K U_k \right] + \sqrt{2H^6 K \log \frac{1}{\delta'}} \leq KH^2 + \sqrt{2H^6 K \log \frac{1}{\delta'}}.$$

Bounding the term T_6^{diff} (cf. (5.33e)). Given that $H \geq V_{h+1}^k(s) \geq V_{h+1}^{\star}(s) \geq V_{h+1}^{\pi^k}(s) \geq 0$ for all $s \in \mathcal{S}$ (see Lemma 5.5), we can derive

$$\begin{aligned}
T_6^{\text{diff}} &= \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h}, (V_{h+1}^k - V_{h+1}^{\pi^k})^2 \right\rangle \\
&\leq H \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h}, V_{h+1}^k - V_{h+1}^{\pi^k} \right\rangle \\
&= H \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h} - e_{s_{h+1}^k}, V_{h+1}^k - V_{h+1}^{\pi^k} \right\rangle}_{=: J_1} + H \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \right)}_{=: J_2}, \quad (5.137)
\end{aligned}$$

leaving us with two terms to control.

- The first term J_1 in (5.137) can be controlled using the same argument as for T_8 . Repeating the arguments in (5.134) tells us that, with probability at least $1 - 2\delta' \log(KH)$,

$$\begin{aligned}
|J_1| &\leq \sqrt{8 \sum_{k,h} \text{Var}_{P_{s_h^k, a_h^k, h}^k} (V_{h+1}^k - V_{h+1}^{\pi^k}) \log \frac{1}{\delta'} + 5H \log \frac{1}{\delta'}} \\
&\leq \sqrt{8 \sum_{k,h} \langle P_{s_h^k, a_h^k, h}^k, (V_{h+1}^k - V_{h+1}^{\pi^k})^2 \rangle \log \frac{1}{\delta'} + 5H \log \frac{1}{\delta'}} \\
&= \sqrt{8T_{6,\text{diff}} \log \frac{1}{\delta'} + 5H \log \frac{1}{\delta'}} \\
&\leq \frac{1}{2H} T_{6,\text{diff}} + 4H \log \frac{1}{\delta'} + 5H \log \frac{1}{\delta'} \\
&\leq \frac{1}{2H} T_{6,\text{diff}} + 9H \log \frac{1}{\delta'}, \tag{5.138}
\end{aligned}$$

where the penultimate inequality applies the AM-GM inequality.

- When it comes to the second term J_2 in (5.137), it follows from (5.39) that, with probability at least $1 - 9S^2AH^2K\delta'$,

$$|J_2| \leq HT_0^{\text{ub}}. \tag{5.139}$$

The above two bounds taken collectively demonstrate that

$$T_6^{\text{diff}} \leq HJ_1 + HJ_2 \leq \frac{1}{2}T_{6,\text{diff}} + 9H^2 \log \frac{1}{\delta'} + H^2T_0^{\text{ub}},$$

thus implying that, with probability at least $1 - 10S^2AH^2K\delta'$,

$$T_6^{\text{diff}} \leq 2H^2T_0^{\text{ub}} + 18H^2 \log \frac{1}{\delta'}.$$

Establishing the bounds (5.41). Under the condition (5.29), it follows from (5.40b) that

$$T_6^{\text{low}} \leq KH^2 + \sqrt{2H^6K \log \frac{1}{\delta'}} \leq 2KH^2.$$

Moreover, combining (5.40a) with Lemma 5.7 and (5.123) (an inequality in the proof of Lemma 5.7 in Section 5.7.2) leads to

$$\begin{aligned}
T_5 &\leq 2HT_2 + 2KH^2 + \frac{1}{4}\sqrt{\iota_bSAH^3T_6^{\text{low}}} + \frac{1}{4}\sqrt{\iota_bS^2AH^3T_6^{\text{diff}}} + \frac{1}{4}\iota_bS^2AH^3 \\
&\stackrel{(i)}{\leq} 2HT_2 + 2KH^2 + \frac{1}{4}\sqrt{2\iota_bSAH^5K} + \frac{1}{4}\sqrt{\iota_bS^2AH^3T_6^{\text{diff}}} + \frac{1}{4}\iota_bS^2AH^3 \\
&\stackrel{(ii)}{\leq} \sqrt{\iota_bSAH^3T_5} + 3KH^2 + \frac{3}{2}\iota_bSAH^3 + \frac{1}{4}\sqrt{\iota_bS^2AH^3T_6^{\text{diff}}} + \frac{1}{4}\iota_bS^2AH^3 \\
&\stackrel{(iii)}{\leq} \frac{1}{2}T_5 + 3KH^2 + \frac{1}{4}\sqrt{\iota_bS^2AH^3T_6^{\text{diff}}} + \frac{7}{4}\iota_bS^2AH^3,
\end{aligned}$$

and as a result,

$$\implies T_5 \leq 6KH^2 + \frac{1}{2}\sqrt{\iota_bS^2AH^3T_6^{\text{diff}}} + \frac{7}{2}\iota_bS^2AH^3.$$

Here, (i) follows from (5.41a), (ii) invokes Lemma 5.8 and the AM-GM inequality $\frac{1}{4}\sqrt{2\iota_bSAH^5K} \leq KH^2 + \frac{1}{2}\iota_bSAH^3$, while (iii) applies the AM-GM inequality $\sqrt{\iota_bSAH^3T_5} \leq \frac{1}{2}T_5 + \frac{1}{2}\iota_bSAH^3$.

5.8 Appendix B: proof of auxiliary lemmas in Section 5.4.4

5.8.1 Proof of Lemma 5.11

Recalling the set $\mathcal{C}$ defined in (5.51), it suffices to prove (5.52).

Towards this end, set

$$M = \log_2 K \quad \text{and} \quad N = SAH,$$

we find it helpful to introduce the following useful sets:

$$\begin{aligned} \mathcal{C}^{\text{distinct}}(l) &:= \left\{ \mathcal{I} = \{\mathcal{I}^1, \dots, \mathcal{I}^l\} \mid \mathcal{I}^1 \leq \dots \leq \mathcal{I}^l, \mathcal{I}^\tau \in \{0, 1, \dots, M\}^N \text{ and } \mathcal{I}^\tau \neq \mathcal{I}^{\tau+1} \ (\forall \tau) \right\}; \\ \mathcal{C}^{\text{distinct}} &:= \bigcup_{l \geq 1} \mathcal{C}^{\text{distinct}}(l). \end{aligned}$$

In words, $\mathcal{C}^{\text{distinct}}(l)$ can be viewed as the set of non-decreasing length- l paths in $\{0, 1, \dots, M\}^N$, with all points on a path being distinct; $\mathcal{C}^{\text{distinct}}$ thus consists of all such paths regardless of the length.

We first establish a connection between $|\mathcal{C}|$ and $|\mathcal{C}^{\text{distinct}}|$. Define the operator $\text{Proj} : \mathcal{C} \rightarrow \mathcal{C}^{\text{distinct}}$ that maps each $\mathcal{I} \in \mathcal{C}$ to $\mathcal{I}^{\text{distinct}} \in \mathcal{C}^{\text{distinct}}$, where $\mathcal{I}^{\text{distinct}}$ is composed of all distinct elements in $\mathcal{I}$ (in other words, this operator simply removes redundancy in $\mathcal{I}$). Let us look at the following set

$$\mathcal{B}(\mathcal{I}^{\text{distinct}}) := \{\mathcal{I} \in \mathcal{C} \mid \text{Proj}(\mathcal{I}) = \mathcal{I}^{\text{distinct}}\}$$

for each $\mathcal{I}^{\text{distinct}} \in \mathcal{C}^{\text{distinct}}$. Since $\mathcal{I}^{\text{distinct}}$ is a non-decreasing path with all its points being distinct, there are at most $MN + 1$ elements in each $\mathcal{I}^{\text{distinct}}$. Hence, the size of $\mathcal{B}(\mathcal{I}^{\text{distinct}})$ is at most the number of solutions to the following equations

$$\sum_{i=1}^{MN+1} x_i = K \quad \text{and} \quad x_i \in \mathbb{N} \text{ for all } 1 \leq i \leq MN + 1.$$

Elementary combinatorial arguments then reveal that

$$|\mathcal{B}(\mathcal{I}^{\text{distinct}})| \leq \binom{K + MN}{MN} \leq (K + MN)^{MN} \leq (2K)^{MN}$$

for each $\mathcal{I}^{\text{distinct}} \in \mathcal{C}^{\text{distinct}}$, provided that $K \geq MN = SAH \log_2 K$. We then arrive at

$$|\mathcal{C}| \leq |\mathcal{C}^{\text{distinct}}| \cdot (2K)^{MN}. \quad (5.140)$$

Everything then boils down to bounding $|\mathcal{C}^{\text{distinct}}|$. To do so, let us first look at the set $\mathcal{C}^{\text{distinct}}(MN + 1)$, as each path in $\mathcal{C}^{\text{distinct}}$ cannot have length more than $MN + 1$. For each $\mathcal{I}^{\text{distinct}} = \{\tilde{\mathcal{I}}^1, \tilde{\mathcal{I}}^2, \dots, \tilde{\mathcal{I}}^{MN+1}\} \in \mathcal{C}^{\text{distinct}}(MN + 1)$, it is easily seen that

- $\tilde{\mathcal{I}}^1 = [0, 0, \dots, 0]^\top$ and $\tilde{\mathcal{I}}^{MN+1} = [M, M, \dots, M]^\top$.
- For each $1 \leq \tau \leq MN$, $\tilde{\mathcal{I}}^\tau$ and $\tilde{\mathcal{I}}^{\tau+1}$ differ only in one element (i.e., their Hamming distance is 1).

In other words, we can view $\mathcal{I}^{\text{distinct}}$ as an MN -step path from $[0, 0, \dots, 0]^\top$ to $[M, M, \dots, M]^\top$, with each step moving in one dimension. Clearly, each step has at most N directions to choose from, meaning that there are at most N^{MN} such paths. This implies that

$$|\mathcal{C}^{\text{distinct}}(MN + 1)| \leq N^{MN}.$$

To finish up, we further observe that for each $\mathcal{I}^{\text{distinct}} \in \mathcal{C}^{\text{distinct}}$, there exists some $\tilde{\mathcal{I}}^{\text{distinct}} \in \mathcal{C}^{\text{distinct}}(MN+1)$ such that $\mathcal{I}^{\text{distinct}} \subseteq \tilde{\mathcal{I}}^{\text{distinct}}$. This observation together with basic combinatorial arguments yields

$$|\mathcal{C}^{\text{distinct}}| \leq 2^{MN+1} |\mathcal{C}^{\text{distinct}}(MN+1)| \leq (2N)^{MN+1},$$

which taken collectively with (5.140) leads to the advertised bound

$$|\mathcal{C}| \leq (2K)^{MN} |\mathcal{C}^{\text{distinct}}| \leq (4KN)^{MN+1} \leq (4KN)^{MN+1}.$$

5.8.2 Proof of Lemma 5.13

Let us begin by considering any fixed total profile $\mathcal{I} \in \mathcal{C}$, any fixed integer l obeying $2 \leq l \leq \log_2 K + 1$, and any given feasible sequence $\{X_{h,s,a}\}$. Recall that

- (i) $\hat{P}_{s,a,h}^{(l)}$ is computed based on the l -th batch of data comprising 2^{l-2} independent samples from $\mathcal{D}^{\text{expand}}$ (see Definition 5.1);
- (ii) each $X_{h+1,s,a}$ is given by a deterministic function of $\mathcal{I}$ and the empirical models for steps $h' \in [h+1, H]$.

As a consequence, Lemma 2.9 combined with Definition 5.1 tells us that, with probability greater than $1 - \delta'$,

$$\begin{aligned} & \sum_{s,a,h} \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \rangle \\ & \leq \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}}(X_{h+1,s,a}) \log \frac{3 \log_2(SAHK)}{\delta'}} + \frac{4H}{2^{l-2}} \log \frac{3 \log_2(SAHK)}{\delta'}, \end{aligned} \quad (5.141)$$

where we view the left-hand side of (5.141) as a martingale sequence from $h = H$ back to $h = 1$.

Moreover, given that each $X_{h,s,a}$ has at most $K+1$ different choices (since in Assumption 5.1 we assume $|\mathcal{X}_{h,\mathcal{I}}| \leq K+1$), there are no more than $(K+1)^{SAH} \leq (2K)^{SAH}$ possible choices of the feasible sequences $\{X_{h,s,a}\}_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]}$. In addition, it has been shown in Lemma 5.11 that there are no more than $(4SAHK)^{2SAH \log_2 K}$ possibilities of the total profile $\mathcal{I}$. Taking the union bound over all these choices and replacing δ' in (5.141) with

$$\frac{\delta'}{((4SAHK)^{2SAH \log_2 K} (2K)^{SAH \log_2 K})},$$

we can demonstrate that with probability at least $1 - \delta'$,

$$\begin{aligned} & \sum_{s,a,h} \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \rangle \\ & \leq \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}}(X_{h+1,s,a}) B_0} + \frac{4H}{2^{l-2}} B_0 \\ & \leq \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}}(X_{h+1,s,a}) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'}\right)} + \frac{4H}{2^{l-2}} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'}\right) \end{aligned} \quad (5.142)$$

holds simultaneously for all $\mathcal{I} \in \mathcal{C}$, all $2 \leq l \leq \log_2 K + 1$, and all feasible sequences $\{X_{h,s,a}\}_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]}$. Here, in the second line above we take

$$B_0 = 2SAH \log_2 K \log(4SAHK) + SAH \log(2K) + \log \frac{3 \log_2^2(SAHK)}{\delta'}.$$

Finally, recalling our assumption $0 \in \mathcal{X}_{h+1,\mathcal{I}}$, we see that for every total profile $\mathcal{I}$ and its associated feasible sequence $\{X_{h,s,a}\}$,

$$\begin{aligned} & \sum_{s,a,h} \max \left\{ \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \rangle, 0 \right\} \\ & \in \left\{ \sum_{s,a,h} \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, \tilde{X}_{h+1,s,a} \rangle \mid \tilde{X}_{h+1,s,a} \in \mathcal{X}_{h+1,\mathcal{I}}, \forall (s,a,h) \right\} \end{aligned}$$

holds true. Consequently, the uniform upper bound on the right-hand side of (5.142) continues to be a valid upper bound on $\sum_{s,a,h} \max \left\{ \langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \rangle, 0 \right\}$. This completes the proof.

5.8.3 Proof of Lemma 5.14

We begin by making the following claim, which we shall establish towards the end of this subsection.

Claim 5.1. With probability exceeding $1 - \delta'$,

$$\begin{aligned} & \sum_{s,a,h} \left\langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, V_{h+1}^{k_{l,j,s,a,h}} \right\rangle \\ & \leq \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}} (V_{h+1}^{k_{l,j,s,a,h}})} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) + \frac{4H}{2^{l-2}} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) \end{aligned} \quad (5.143)$$

holds simultaneously for all $l = 1, \dots, \log_2 K$ and all $j = 1, \dots, 2^{l-1}$, where we recall that $k_{l,j,s,a,h}$ stands for the episode index of the sample that visits (s,a,h) for the $(2^{l-1} + j)$ -th time in the online learning process.

Assuming the validity of Claim 5.1 for the moment, we can combine this claim with the decomposition (5.53) and apply the Cauchy-Schwarz inequality to reach

$$\begin{aligned} & \sum_{k=1}^K \sum_{h=1}^H \left\langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle \leq \sum_{l=1}^{\log_2 K} \sum_{j=1}^{2^{l-1}} \sum_{s,a,h} \left\langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, V_{h+1}^{k_{l,j,s,a,h}} \right\rangle + SAH^2 \\ & \leq \sum_{l=1}^{\log_2 K} \sum_{j=1}^{2^{l-1}} \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}} (V_{h+1}^{k_{l,j,s,a,h}})} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) \\ & \quad + \sum_{l=1}^{\log_2 K} \sum_{j=1}^{2^{l-1}} \frac{4H}{2^{l-2}} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) + SAH^2 \\ & \leq \sum_{l=1}^{\log_2 K} \sqrt{16 \sum_{j=1}^{2^{l-1}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}} (V_{h+1}^{k_{l,j,s,a,h}})} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) \end{aligned}$$

$$\begin{aligned}
& + \sum_{l=1}^{\log_2 K} 8H \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) + SAH^2 \\
& \leq \sqrt{16(\log_2 K) \sum_{l=1}^{\log_2 K} \sum_{j=1}^{2^{l-1}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}}(V_{h+1}^{k_{l,j,s,a,h}}) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} \\
& \quad + \left(48SAH^2 \log_2^3 K + 8H(\log_2 K) \log \frac{1}{\delta'} \right) + SAH^2 \\
& \leq \sqrt{16(\log_2 K) \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} \\
& \quad + 49SAH^2 \log_2^3 K + 8H(\log_2 K) \log \frac{1}{\delta'}.
\end{aligned}$$

Here, the last inequality is valid due to our assumption $V_{h+1}^k = 0$ ($\forall k > K$) and the identity

$$\begin{aligned}
\sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k) &= \sum_{l=1}^{\log_2 K} \sum_{s,a,h} \sum_{j=1}^{2^{l-1}} \text{Var}_{P_{s,a,h}}(V_{h+1}^{k_{l,j,s,a,h}}) \\
&\quad + \sum_{k=1}^K \sum_{h=1}^H \mathbb{1}\{N_h^{k,\text{all}}(s_h^k, a_h^k) = 0\} \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k).
\end{aligned}$$

This establishes our advertised bound on $\sum_{k,h} \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle$, provided that Claim 5.1 is valid.

Before proceeding to the proof of Claim 5.1, we note that the other two quantities $\sum_{k,h} \max \{ \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle, 0 \}$ and $\sum_{k,h} \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, (V_{h+1}^k)^2 \rangle$ can be upper bounded using exactly the same arguments, which we omit for the sake of brevity. In particular, the latter quantity further satisfies

$$\begin{aligned}
& \sum_{k=1}^K \sum_{h=1}^H \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, (V_{h+1}^k)^2 \rangle \\
& \leq \sqrt{16(\log_2 K) \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}((V_{h+1}^k)^2) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} \\
& \quad + 49SAH^3 \log_2^3 K + 8H^2(\log_2 K) \log \frac{1}{\delta'} \\
& \leq 8H \sqrt{(\log_2 K) \sum_{k=1}^K \sum_{h=1}^H \text{Var}_{P_{s_h^k, a_h^k, h}}(V_{h+1}^k) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} \\
& \quad + 49SAH^3 \log_2^3 K + 8H^2(\log_2 K) \log \frac{1}{\delta'},
\end{aligned}$$

where the last inequality follows from Lemma 5.21 and the fact that $0 \leq V_{h+1}^k(s) \leq H$ for all $s \in \mathcal{S}$.

Proof of Claim 5.1. To invoke Lemma 5.13 to prove this claim, we need to choose the set $\{\mathcal{X}_{h,\mathcal{I}}\}$ properly to include the true value function estimates $\{V_h^k\}$. To do so, we find it helpful to first introduce an auxiliary algorithm tailored to each total profile. Specifically, for each $\mathcal{I} \in \mathcal{C}$ (cf. (5.51)), consider the following updates operating upon the expanded sample set $\mathcal{D}^{\text{expand}}$.

Algorithm 5.5: Monotonic Value Propagation for a given total profile $\mathcal{I} \in \mathcal{C}$ (MVP($\mathcal{I}$))

```

1 initialization: set  $V_{H+1}^{k,\mathcal{I}}(s) \leftarrow H$  for all  $s \in \mathcal{S}$  and  $1 \leq k \leq K$ .
2 for  $k = 1, 2, \dots, K$  do
3   for  $h = H, H-1, \dots, 1$  do
4     for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
        $j \leftarrow I_{s,a,h}^k, \quad n \leftarrow 2^{j-2},$ 
        $b_h(s, a) \leftarrow c_1 \sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}^{(j)}}(V_{h+1}^{k,\mathcal{I}}) \log \frac{1}{\delta'}}}{\max\{n, 1\}} + c_2 \frac{H \log \frac{1}{\delta'}}{\max\{n, 1\}},$ 
        $Q_h^{k,\mathcal{I}}(s, a) \leftarrow \min \left\{ r_h(s, a) + \langle \hat{P}_{s,a,h}^{(j)}, V_{h+1}^{k,\mathcal{I}} \rangle + b_h(s, a), H \right\},$ 
        $V_h^{k,\mathcal{I}}(s) \leftarrow \max_a Q_h^{k,\mathcal{I}}(s, a).$ 

```

If we construct

$$\mathcal{X}_{h,\mathcal{I}} := \{V_h^{k,\mathcal{I}} \mid 1 \leq k \leq K\} \cup \{0\}, \quad \forall h \in [H] \text{ and } \mathcal{I} \in \mathcal{C}, \quad (5.144)$$

then it can be easily seen that $\{\mathcal{X}_{h,\mathcal{I}}\}$ satisfies the properties stated right before Lemma 5.13. As a consequence, applying Lemma 5.13 yields

$$\begin{aligned} & \sum_{s,a,h} \left\langle \hat{P}_{s,a,h}^{(l)} - P_{s,a,h}, X_{h+1,s,a} \right\rangle \\ & \leq \sqrt{\frac{8}{2^{l-2}} \sum_{s,a,h} \text{Var}_{P_{s,a,h}}(X_{h+1,s,a}) \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right)} + \frac{4H}{2^{l-2}} \left(6SAH \log_2^2 K + \log \frac{1}{\delta'} \right) \end{aligned} \quad (5.145)$$

simultaneously for all $l = 1, \dots, \log_2 K$, all $\mathcal{I} \in \mathcal{C}$, and all sequences $\{X_{h,s,a}\}$ obeying $X_{h,s,a} \in \mathcal{X}_{h,\mathcal{I}}$, $\forall (s, a, h)$.

To finish up, denote by $\mathcal{I}^{\text{true}}$ the true total profile resulting from the online learning process. Given the way we couple $\mathcal{D}^{\text{expand}}$ and $\mathcal{D}^{\text{original}}$ (see the remark right after Definition 5.1), we can easily see that the true value function estimate $\{V_h^k\}$ obeys

$$V_h^k = V_h^{k,\mathcal{I}^{\text{true}}} \in \mathcal{X}_{h,\mathcal{I}^{\text{true}}}, \quad 1 \leq k \leq K. \quad (5.146)$$

The claimed result then follows immediately from (5.146) and the uniform bound (5.145). $\square$

5.8.4 Proof of Lemma 5.15

Let us decompose the term T_6^{ucb} (cf. (5.57b)) as follows

$$\begin{aligned} T_6^{\text{ucb}} &= \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h}, (V_{h+1}^k)^2 \right\rangle - \sum_{k=1}^K \sum_{h=1}^H \left(\left\langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle \right)^2 \\ &\stackrel{(i)}{=} \sum_{k=1}^K \sum_{h=1}^H \left\langle P_{s_h^k, a_h^k, h} - e_{s_{h+1}^k}, (V_{h+1}^k)^2 \right\rangle + \sum_{k=1}^K \sum_{h=2}^H (V_h^k(s_h^k))^2 - \sum_{k=1}^K \sum_{h=1}^H \left(\left\langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \right\rangle \right)^2 \end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} T_8 + 2H \sum_{k=1}^K \sum_{h=1}^H \max \left\{ V_h^k(s_h^k) - \langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle, 0 \right\} \\
&\stackrel{(iii)}{=} T_8 + 2H \sum_{k=1}^K \sum_{h=1}^H \max \left\{ Q_h^k(s_h^k, a_h^k) - \langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle, 0 \right\}.
\end{aligned}$$

Here, (i) replaces $h+1$ with h in the second sum and uses $V_{H+1}^k = 0$; in (ii), we recall the definition of T_8 in (5.133) and invoke the basic property

$$\begin{aligned}
(V_h^k(s_h^k))^2 - \left(\langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle \right)^2 &= \left(V_h^k(s_h^k) - \langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle \right) \left(V_h^k(s_h^k) + \langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle \right) \\
&\leq 2H \max \left\{ V_h^k(s_h^k) - \langle P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle, 0 \right\};
\end{aligned}$$

and (iii) follows since $V_h^k(s_h^k) = \max_a Q_h^k(s_h^k, a) = Q_h^k(s_h^k, a_h^k)$.

In order to leverage the update rule (5.44), we need to introduce $\hat{P}_{s_h^k, a_h^k, h}^k$ into the max operator, which results in

$$\begin{aligned}
T_6^{\text{ucb}} &\leq T_8 + 2H \sum_{k=1}^K \sum_{h=1}^H \max \left\{ Q_h^k(s_h^k, a_h^k) - \langle \hat{P}_{s_h^k, a_h^k, h}^k, V_{h+1}^k \rangle, 0 \right\} \\
&\quad + 2H \sum_{k=1}^K \sum_{h=1}^H \max \left\{ \langle \hat{P}_{s_h^k, a_h^k, h}^k - P_{s_h^k, a_h^k, h}, V_{h+1}^k \rangle, 0 \right\} \\
&\leq T_8 + 2H \sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k) + 2H \sum_{k=1}^K \sum_{h=1}^H r_h(s_h^k, a_h^k) + 2HW_4 \\
&\leq T_8 + 2HT_2 + 2KH^2 + 2HW_4,
\end{aligned} \tag{5.147}$$

where we recall the definition of W_4 in (5.57d). This taken together with the upper bound (5.135) on T_8 immediately establishes that

$$\begin{aligned}
T_6^{\text{ucb}} &\leq 2HT_2 + 6KH^2 + \sqrt{32H^2 T_6^{\text{ucb}} \log \frac{1}{\delta'}} + 3H^2 \log \frac{1}{\delta'} + 2HW_4 \\
&\leq 2HT_2 + 6KH^2 + \frac{1}{2} T_6^{\text{ucb}} + 16H^2 \log \frac{1}{\delta'} + 3H^2 \log \frac{1}{\delta'} + 2HW_4,
\end{aligned}$$

where the last line applied the AM-GM inequality. Rearranging terms then yields, with probability exceeding $1 - 4SAHK\delta'$,

$$T_6^{\text{ucb}} \leq 4HT_2 + 4HW_4 + 12KH^2 + 38H^2 \log \frac{1}{\delta'}.$$

5.8.5 Solving the inequalities in (5.58)

To begin with, substituting (5.58a) into (5.58g) and invoking the AM-GM inequality give

$$\begin{aligned}
T_6^{\text{ucb}} &\leq 4HT_2 + 4H \sqrt{T_6^{\text{ucb}} SAH \iota_b} + 4SAH^3 \iota_b + KH^2 \iota_b \\
&\leq 4HT_2 + \frac{1}{2} T_6^{\text{ucb}} + 8SAH^3 \iota_b + 4SAH^3 \iota_b + KH^2 \iota_b,
\end{aligned}$$

which together with a little algebra yields

$$T_6^{\text{ucb}} \leq 8HT_2 + 24SAH^3 \iota_b + 2KH^2 \iota_b. \tag{5.148}$$

Next, it follows from (5.58f) and (5.58b) that

$$\begin{aligned} T_5 &\leq 2H\sqrt{T_6^{\text{ucb}}SAH\iota_b} + 2HT_2 + SAH^3\iota_b + KH^2\iota_b \\ &\leq \frac{1}{2}T_6^{\text{ucb}} + 2SAH^3\iota_b + 2HT_2 + SAH^3\iota_b + KH^2\iota_b \\ &\leq 4HT_2 + 12SAH^3\iota_b + KH^2\iota_b + 2SAH^3\iota_b + 2HT_2 + SAH^3\iota_b + KH^2\iota_b, \end{aligned}$$

where the second inequality arises from the AM-GM inequality, and the third inequality results from (5.148). Rearrange terms to arrive at

$$T_5 \leq 6HT_2 + 15SAH^3\iota_b + 2KH^2\iota_b. \quad (5.149)$$

In addition, it is also seen from (5.58c) that

$$T_2 \leq \frac{1}{12H}T_5 + 12SAH^2\iota_b + \frac{1}{2}SAH^2\iota_b, \quad (5.150)$$

which combined with (5.149) reveals that

$$T_5 \leq \frac{1}{2}T_5 + 90SAH^3\iota_b + 2KH^2\iota_b, \quad \implies \quad T_5 \leq 180SAH^3\iota_b + 4KH^2\iota_b. \quad (5.151)$$

Armed with inequality (5.151), we can immediately combine it with (5.58c) and (5.148) to yield

$$T_2 \lesssim \sqrt{SAH^3K\iota_b} + SAH^2\iota_b, \quad (5.152)$$

$$T_6^{\text{ucb}} \lesssim SAH^3\iota_b + KH^2\iota_b. \quad (5.153)$$

Substitution into (5.58a) then leads to

$$\begin{aligned} T_1 &\lesssim \sqrt{SAH^3K\iota_b} + SAH^2\iota_b, \\ T_3 &\lesssim \sqrt{SAH^3\iota_b} + \sqrt{KH^2\iota_b}. \end{aligned}$$

Putting these two bounds collectively with (5.152) and (5.58e), we reach

$$\text{Regret}(K) \leq T_1 + T_2 + T_3 + T_4 \lesssim \sqrt{SAH^3K\iota_b} + SAH^2\iota_b.$$

5.9 Appendix C: proof of auxiliary lemmas in Section 5.5

5.9.1 Proof of Lemma 5.17

We first establish (5.67b). For any $1 \leq n \leq N$,

$$\eta_n^N = \frac{H+1}{H+n} \prod_{i=n+1}^N \frac{i-1}{H+i} = \frac{H+1}{H+N} \prod_{i=n}^{N-1} \frac{i}{H+i} \leq \frac{H+1}{H+N} \leq \frac{2H}{N}. \quad (5.154)$$

Since $\sum_{n=1}^N \eta_n^N = 1$, this intermediate bound implies

$$\sum_{n=1}^N (\eta_n^N)^2 \leq \left(\max_{1 \leq n \leq N} \eta_n^N \right) \sum_{n=1}^N \eta_n^N \leq \frac{2H}{N}. \quad (5.155)$$

We next verify (5.67c). For a fixed $n \geq 1$, define

$$a_N := \prod_{i=n+1}^N \frac{i-1}{H+i}, \quad N \geq n, \quad (5.156)$$

with the convention $a_n = 1$. Then $\eta_n^N = \frac{H+1}{H+n}a_N$. Moreover,

$$a_N = \frac{H+N}{H}a_N - \frac{H+N+1}{H}a_{N+1}, \quad (5.157)$$

so the partial sums telescope:

$$\sum_{N=n}^L a_N = \frac{H+n}{H} - \frac{H+L+1}{H}a_{L+1}. \quad (5.158)$$

Since the last term vanishes as $L \rightarrow \infty$, we have

$$\sum_{N=n}^{\infty} \eta_n^N = \frac{H+1}{H+n} \cdot \frac{H+n}{H} = 1 + \frac{1}{H}. \quad (5.159)$$

It remains to establish (5.67a). When $N = 1$, the statement holds trivially since

$$\sum_{n=1}^N \frac{\eta_n^N}{\sqrt{n}} = \eta_1^1 = 1.$$

Now suppose that $N \geq 2$. Making use of the basic relation $\eta_n^N = (1-\eta_N)\eta_n^{N-1}$ for all $n = 1, \dots, N-1$, we observe the following identity:

$$\sum_{n=1}^N \frac{\eta_n^N}{\sqrt{n}} = \frac{\eta_N}{\sqrt{N}} + (1-\eta_N) \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{\sqrt{n}}. \quad (5.160)$$

We prove the property by induction. Suppose for the moment that the property holds for $N-1$, namely,

$$\frac{1}{\sqrt{N-1}} \leq \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{\sqrt{n}} \leq \frac{2}{\sqrt{N-1}}. \quad (5.161)$$

Then it is readily seen from (5.160) that

$$\begin{aligned} \sum_{n=1}^N \frac{\eta_n^N}{\sqrt{n}} &= \frac{\eta_N}{\sqrt{N}} + (1-\eta_N) \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{\sqrt{n}} \\ &\geq \frac{\eta_N}{\sqrt{N}} + \frac{1-\eta_N}{\sqrt{N-1}} \geq \frac{\eta_N}{\sqrt{N}} + \frac{1-\eta_N}{\sqrt{N}} = \frac{1}{\sqrt{N}}, \end{aligned} \quad (5.162)$$

where the first inequality comes from (5.161). Similarly,

$$\begin{aligned} \sum_{n=1}^N \frac{\eta_n^N}{\sqrt{n}} &= \frac{\eta_N}{\sqrt{N}} + (1-\eta_N) \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{\sqrt{n}} \\ &\leq \frac{\eta_N}{\sqrt{N}} + \frac{2(1-\eta_N)}{\sqrt{N-1}} \\ &= \frac{H+1}{\sqrt{N}(H+N)} + \frac{2\sqrt{N-1}}{H+N} \\ &\leq \frac{H+1}{\sqrt{N}(H+N)} + \frac{2\sqrt{N}}{H+N} \\ &= \frac{1}{\sqrt{N}} \left(\frac{H+1}{H+N} + \frac{2N}{H+N} \right) \leq \frac{2}{\sqrt{N}}, \end{aligned}$$

where the last inequality follows since $H \geq 1$. This completes the induction and hence the proof of Lemma 5.17.

5.9.2 Proof of Lemma 5.18

For each fixed $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, let k_i be the episode in which (s, a) is taken at step h for the i -th time. The sequence

$$Z_i := [(P_h^{k_i} - P_{s,a,h})V_{h+1}^*](s, a) \quad (5.163)$$

forms a martingale difference sequence with $|Z_i| \leq H$. Therefore, by the Azuma–Hoeffding inequality (cf. Lemma 2.4), and a union bound over all (s, a, h) and all visit counts $1 \leq t \leq K$, with probability at least $1 - \delta$,

$$\left| \sum_{i=1}^t \eta_i^t Z_i \right| \leq C_1 H \sqrt{\iota \sum_{i=1}^t (\eta_i^t)^2} \leq C_2 \sqrt{\frac{H^3 \iota}{t}} \quad (5.164)$$

holds simultaneously for all (s, a, h, t) , where the last inequality is due to (5.67b). On the same event, if the constant c in (5.61) is chosen sufficiently large, then

$$\left| \sum_{i=1}^t \eta_i^t Z_i \right| \leq \sum_{i=1}^t \eta_i^t b_i, \quad (5.165)$$

because (5.67a) gives

$$\sum_{i=1}^t \eta_i^t b_i = c \sqrt{H^3 \iota} \sum_{i=1}^t \frac{\eta_i^t}{\sqrt{i}} \geq c \sqrt{\frac{H^3 \iota}{t}}. \quad (5.166)$$

The same calculation also gives

$$\beta_t = 2 \sum_{i=1}^t \eta_i^t b_i \leq 4c \sqrt{\frac{H^3 \iota}{t}}. \quad (5.167)$$

We prove (5.72) by combining (5.71), (5.165) and (5.167).

We next prove the optimism property (5.74) by induction. The property holds trivially at the base case $k = 0$ due to the optimistic initialization. We next assume (5.74) holds for all episodes up to k . By the recursion (5.71) and (5.165), it holds that

$$(Q_h^k - Q_h^*)(s, a) \geq \eta_0^t (H - Q_h^*(s, a)) + \sum_{i=1}^t \eta_i^t (V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) \geq 0. \quad (5.168)$$

It then follows that

$$\max_{a \in \mathcal{A}} Q_h^k(s, a) \geq \max_{a \in \mathcal{A}} Q_h^*(s, a) = V_h^*(s), \quad (5.169)$$

and the clipping $V_h^k(s) = \min\{H, \max_a Q_h^k(s, a)\}$ preserves the inequality because $V_h^*(s) \leq H$. This completes the induction and hence the proof.

5.10 Appendix D: proof of Lemma 5.20

For notational convenience, we denote by a_t (resp. $R_t = R_{a_t, t}$) the arm selected (resp. the corresponding observed reward) at round t , and define

$$x_t := \{a_1, R_1, \dots, a_t, R_t\}$$

as the trajectory observed by the end of the t -th round. The distribution of x_t under $\mathcal{B}_i$ is denoted by $\mathbb{P}_{\mathcal{B}_i}^{(t)}$. Also, we denote the conditional reward distribution under $\mathcal{B}_i$ as

$$P_i(R|a) := \begin{cases} 1 - q_i(a), & \text{if } R = 1, \\ q_i(a), & \text{if } R = 0. \end{cases} \quad (5.170)$$

Now, consider any online learning algorithm $\hat{\pi}$ that generates a sequence of policies $\{\pi^t\}_{t \geq 0}$, where π^t depends only on the past trajectory x_{t-1} . With a slight abuse of notation, in this proof we write

$$\pi^t(\cdot | x_{t-1})$$

for the action-selection distribution at round t . Here, $\pi^t(\cdot | \cdot)$ denotes the underlying algorithmic rule, while the realized action-selection distribution at round t relies on the observed history x_{t-1} . As a result, under the bandit instance $\mathcal{B}_i$, the distribution of the trajectory x_t admits the following factorization:

$$\mathbb{P}_{\mathcal{B}_i}^{(t)}(x_t) = \prod_{\tau=1}^t \pi^\tau(a_\tau | x_{\tau-1}) P_i(R_\tau | a_\tau).$$

where $P_i(\cdot | a_t)$ is defined in (5.170). This implies that

$$\frac{\mathbb{P}_{\mathcal{B}_i}^{(t)}(x_t)}{\mathbb{P}_{\mathcal{B}_1}^{(t)}(x_t)} = \prod_{\tau=1}^t \frac{P_i(R_\tau | a_\tau)}{P_1(R_\tau | a_\tau)},$$

meaning that the likelihood ratio decomposes into a product of per-round reward likelihood ratios. Exploiting this decomposition together with the Markov structure of the trajectory yields

$$\text{KL}(\mathbb{P}_{\mathcal{B}_i}^{(t)} \| \mathbb{P}_{\mathcal{B}_1}^{(t)}) = \sum_{\tau=1}^t \mathbb{E}_{\mathcal{B}_i} [\text{KL}(P_i(\cdot | a_\tau) \| P_1(\cdot | a_\tau))], \quad (5.171)$$

where the expectation in each summand is taken w.r.t. the randomness of a_τ under the bandit instance $\mathcal{B}_i$.

Next, we bound each KL divergence term appearing in the summation on the right-hand side of (5.171). Recalling that the conditional reward distributions (5.170) are Bernoulli and letting $\text{KL}(p \| q)$ denote the KL divergence between Bernoulli distributions with means p and q , respectively, we can invoke Lemma 2.12 to show that:

- If $a = i$, then it holds that

$$\text{KL}(P_i(\cdot | a) \| P_1(\cdot | a)) = \text{KL}(q_i(a) \| q_1(a)) \leq \frac{(q_i(a) - q_1(a))^2}{q_1(a)(1 - q_1(a))} \leq \frac{(\varepsilon_0/H^2)^2}{1/(4H)} = \frac{4\varepsilon_0^2}{H^3},$$

where the last line is a consequence of the properties (see (5.97))

$$\begin{aligned} |q_i(a) - q_1(a)| &= \frac{\varepsilon_0}{H^2}, \\ q_1(a)(1 - q_1(a)) &\geq \left(\frac{1}{H} - \frac{\varepsilon_0}{2H^2}\right) \left(1 - \frac{1}{H} + \frac{\varepsilon_0}{2H^2}\right) \geq \frac{1}{4H} \end{aligned}$$

that hold as long as $\varepsilon_0 \leq H$.

- If $a \neq i$, then $q_i(a) = q_1(a)$ (see (5.97)) and, as a consequence,

$$\text{KL}(P_i(\cdot | a) \| P_1(\cdot | a)) = 0.$$

These two properties about $\text{KL}(P_i(\cdot | a) \| P_1(\cdot | a))$ taken together yield

$$\sum_{i=1}^A \text{KL}(P_i(\cdot | a) \| P_1(\cdot | a)) = \text{KL}(P_a(\cdot | a) \| P_1(\cdot | a)) \leq \frac{4\varepsilon_0^2}{H^3},$$

which combined with (5.171) leads to

$$\frac{1}{A} \sum_{i=1}^A \text{KL}(\mathbb{P}_{\mathcal{B}_i}^{(t)} \parallel \mathbb{P}_{\mathcal{B}_1}^{(t)}) = \frac{1}{A} \sum_{\tau=1}^t \mathbb{E}_{\mathcal{B}_i} \left[\sum_{i=1}^A \text{KL}(P_i(\cdot | a_\tau) \parallel P_1(\cdot | a_\tau)) \right] \leq \sum_{\tau=1}^t \frac{4\varepsilon_0^2}{AH^3} = \frac{4\varepsilon_0^2 t}{AH^3}. \quad (5.172)$$

With (5.172) in place, an application of Fano's lemma (cf. Lemma 2.10) reveals that: whenever $\frac{4\varepsilon_0^2 t}{AH^3} < \log(A-1)$, one necessarily has

$$\begin{aligned} \inf_{\hat{\pi}} \frac{1}{A} \sum_{i=1}^A \mathbb{P}_{\mathcal{B}_i}(a_{t+1} \neq i) &\geq \frac{\log A - \log 2 - \frac{1}{A} \sum_{i=1}^A \text{KL}(\mathbb{P}_{\mathcal{B}_i}^{(t)} \parallel \mathbb{P}_{\mathcal{B}_1}^{(t)})}{\log(A-1)} \\ &\geq \frac{\log A - \log 2 - \frac{4\varepsilon_0^2 t}{AH^3}}{\log(A-1)} \\ &\geq \frac{\log A - \log 2 - \frac{2}{9}}{\log(A-1)} \geq \frac{\log(3/2) - \frac{2}{9}}{\log(2)} \geq \frac{1}{4}, \end{aligned}$$

where the last line is valid since $t \leq \frac{AH^3}{18\varepsilon_0^2}$ and $A \geq 3$. This concludes the proof.

5.11 Notes

Multi-armed bandit problems have deep roots in the statistics literature. First formulated by Thompson (1933), the bandit framework, along with the notion of regret, was later popularized by Herbert Robbins (e.g., the seminal work of Robbins (1952)). The use of upper confidence bounds in bandit problems was initially suggested by Lai and Robbins (1985) and Lai (1987). Subsequent research established asymptotic guarantees for UCB-style algorithms (e.g., Lai (1987), Katehakis and Robbins (1995), Agrawal (1995), and Burnetas and Katehakis (1996)), while non-asymptotic regret analyses were later developed by, among others, Auer *et al.* (2002), Bubeck and Cesa-Bianchi (2012), Russo and Van Roy (2013), and Cappé *et al.* (2013). By adaptively tuning the exploration bonuses, one can further obtain anytime guarantees for UCB-style algorithms (i.e., without requiring prior knowledge of the time horizon) and remove the extra logarithmic factor in the regret bound (Auer *et al.*, 2002; Audibert and Bubeck, 2009; Degenne and Perchet, 2016). We refer the interested reader to Lattimore and Szepesvári (2020) for a comprehensive discussion of the historical development of bandit algorithms and related bibliographical remarks.

The optimism principle has subsequently become a central design paradigm in online RL. Building on ideas that originated in the bandit literature, a large body of work has developed optimistic algorithms for sequential learning and decision-making in unknown environments (e.g., Auer and Ortner (2006), Jaksch *et al.* (2010), Bartlett and Tewari (2009), Filippi *et al.* (2010), Dann and Brunskill (2015), Dann *et al.* (2017), Azar *et al.* (2017), Jin *et al.* (2018), Dann *et al.* (2019), Zanette and Brunskill (2019), and Bourel *et al.* (2020)). One of the earliest contributions in this direction is the work of Jaksch *et al.* (2010), which proposed the UCRL2 algorithm for weakly communicating MDPs. The key idea is to construct confidence regions for the transition kernel and use them to identify policies that are optimistic w.r.t. the set of plausible MDPs. This UCRL2 algorithm builds upon earlier optimistic methods, such as the UCRL algorithm proposed by Auer and Ortner (2006). When applied to episodic finite-horizon MDPs, however, the regret guarantees obtained via the UCRL2 analysis are not minimax optimal, exhibiting a gap of at least a factor of $\sqrt{H^2 S}$ relative to the minimax limit.

In comparison to UCRL2 and its variants, a more sample-efficient paradigm is to construct upper confidence bounds for the optimal Q-functions and value functions and use these confidence estimates to guide exploration. This paradigm has led to a sequence of sample-efficient algorithms for online RL, including both model-based methods (e.g., Azar *et al.* (2017), Zhang *et al.* (2021e), Ménard *et al.* (2021), and Zhang *et al.* (2025e)) and model-free methods (e.g., Jin *et al.* (2018), Bai *et al.* (2019), Yang *et al.* (2021), and Zhang *et al.* (2025b)). When combined with confidence bounds derived from empirical Bernstein inequalities, which exploit variance information, model-based algorithms based on this paradigm can attain asymptotically optimal regret guarantees. For model-free RL, however, variance-aware bonuses alone are insufficient to achieve asymptotically optimal regret, which motivates the development of more sophisticated variance-reduction techniques (Zhang *et al.*, 2020f; Li *et al.*, 2021b; Li *et al.*, 2023a; Zhang *et al.*, 2026a) (originally referred to as reference-advantage decomposition by Zhang *et al.* (2020f)). Determining how to optimally leverage optimism to achieve minimax-optimal regret without incurring high burn-in cost remained an open problem until the work Zhang *et al.* (2024c) and Zhang *et al.* (2025e), which established the minimax optimality of model-based online RL for any sample size. In contrast, the most sample-efficient model-free algorithm currently known, due to Li *et al.* (2021b), still incurs a burn-in sample size on the order of $SA\text{poly}(H)$.

Beyond the setting studied in this chapter, a variety of related problem formulations have been investigated, motivated by different practical considerations. One prominent example is reward-free (or reward-agnostic) exploration, where the goal is to collect data during an exploration phase without knowledge of the reward function, while still enabling near-optimal performance once the reward function is subsequently specified (e.g., Jin *et al.* (2020a), Zanette *et al.* (2020b), Zhang *et al.* (2020d), Zhang *et al.* (2020e), Wagenmaker *et al.* (2022), Li *et al.* (2024e), Wang *et al.* (2020b), and Zhang *et al.* (2021d)). Another direction is the study of “horizon-free RL,” which studies finite-horizon *stationary* MDPs under the assumption that the total reward accumulated within each episode is bounded by 1; the objective is to design algorithms whose regret bounds almost do not scale with the horizon length H , thereby mitigating the adverse effect of long horizon (Jiang and Agarwal, 2018; Zanette and Brunskill, 2019; Zhang *et al.*, 2021e; Li *et al.*, 2022c; Zhang *et al.*, 2024d; Wang *et al.*, 2025b). There has also been a line of work that focuses on reducing the switching cost, namely the number of times the learner changes its sampling policy throughout the learning process, motivated by practical applications in which policy deployment is costly (Bai *et al.*, 2019; Qiao and Wang, 2022; Qiao *et al.*, 2022; Huang *et al.*, 2022; Zhao *et al.*, 2024; Zhang *et al.*, 2026b). Additionally, low-regret algorithms for constrained MDPs have been developed in, e.g., Efroni *et al.* (2020), Brantley *et al.* (2020), Ding *et al.* (2021), Liu *et al.* (2021), Bura *et al.* (2022), Wei *et al.* (2022), and Wei *et al.* (2023). Furthermore, the efficient use of the optimism principle under low-complexity function approximation — particularly linear function approximation — has itself become a major research direction; we refer the reader to, e.g., Yang and Wang (2020), Jin *et al.* (2020b), Zanette *et al.* (2020a), Du *et al.* (2020), Yang *et al.* (2020b), Modi *et al.* (2020), Ayoub *et al.* (2020), Wang *et al.* (2020a), Zanette (2021), Wang *et al.* (2021b), Weisz *et al.* (2021), Wang *et al.* (2021c), Li *et al.* (2022b), Zhou *et al.* (2021a), Jin *et al.* (2021a), Du *et al.* (2021), Yin *et al.* (2022a), and He *et al.* (2023) and the references therein for details.

Beyond confidence-interval-based methods, several alternative approaches to incentivizing exploration have also been extensively studied in both theory and practice. One notable approach is *posterior sampling* (Osband *et al.*, 2013; Osband and Van Roy, 2017), with representative examples including Thompson sampling (Gopalan and Mannor, 2015; Agrawal and Jia, 2017; Ouyang *et al.*,

2017; Agrawal *et al.*, 2026), feel-Good Thompson sampling (Zhang, 2022), and learning rate randomization (Tiapkin *et al.*, 2023; Wang *et al.*, 2026). Another line of work promotes exploration by regularizing the maximum likelihood estimate of the model or value functions toward more optimistic versions, such as Hung *et al.* (2021), Mete *et al.* (2021), Yang *et al.* (2025a), and Liu *et al.* (2023b). Randomized exploration, when suitably designed has also been shown to achieve efficient exploration (Xiong *et al.*, 2022). Collectively, these developments showcase a diverse array of algorithmic ideas available for balancing exploration and exploitation.

6

Offline RL

Despite the promise of online RL, the limited capability of online data collection in various real-world applications presents a fundamental bottleneck to carrying the success of RL over to broader scenarios. For example, in applications such as clinical trials, autonomous systems, and online advertising, real-time data acquisition is expensive, high-stakes, and/or time-consuming. To circumvent this bottleneck, a plausible idea is to make more effective use of previously collected data, since historical datasets may contain information that transfers to new tasks. For instance, the state-transition dynamics encountered in past tasks may sometimes partially resemble those arising in new tasks. The potential of this data-driven paradigm has been explored and recognized in a diverse array of contexts, including but not limited to robotic manipulation (Ebert *et al.*, 2018), autonomous driving (Diehl *et al.*, 2021), and healthcare (Tang and Wiens, 2021); see Levine *et al.* (2020) and Prudencio *et al.* (2022) for recent surveys. The subfield of RL based solely on historical data, without further interaction with the environment, is now commonly referred to as *offline RL* or *batch RL* (Lange *et al.*, 2012; Levine *et al.*, 2020).

6.1 Problem formulation and challenges

To facilitate a precise technical exposition, we begin by introducing the settings and objectives that will serve as the primary focus of this chapter. We then discuss the key challenges that arise in these settings, together with some important problem parameters introduced in response to these challenges.

Data generation mechanism. The effectiveness and potential of offline RL hinge critically upon the quality of the historical dataset, denoted throughout by $\mathcal{D}$. For concreteness, this chapter focuses primarily on an independent sampling model tailored to discounted infinite-horizon MDPs to describe the sample-generating process underlying $\mathcal{D}$.

To be precise, suppose the learner has access to a batch dataset $\mathcal{D} = \{(s_i, a_i, s'_i)\}_{1 \leq i \leq N}$ comprising N sample transitions. These samples are generated according to an unknown distribution $d^b \in$

$\Delta(\mathcal{S} \times \mathcal{A})$ and the true transition kernel P of the underlying MDP, namely,

$$(s_i, a_i) \stackrel{\text{ind.}}{\sim} d^{\mathbf{b}} \quad \text{and} \quad s'_i \stackrel{\text{ind.}}{\sim} P(\cdot | s_i, a_i), \quad 1 \leq i \leq N. \quad (6.1)$$

This coincides with the sampling model proposed in the seminal work Rashidinejad *et al.* (2021). In practical applications, such samples may have been generated by executing one or more “behavior policies” in the MDP, which motivates the notation $\mathbf{b}$ in the distribution $d^{\mathbf{b}}$. As in previous chapters, we assume for simplicity that the exact reward functions are available to the learner; handling unknown rewards introduces only moderate additional technical challenges.

Goal. With the batch dataset $\mathcal{D}$ in hand, the goal of offline RL in this setting is to learn, in a data-efficient manner, a policy $\hat{\pi}$ that attains near-optimal value with respect to a prescribed test state distribution $\rho \in \Delta(\mathcal{S})$ (which is not known to the learner *a priori*). More precisely, for a prescribed precision level ε , we seek to identify an ε -optimal policy $\hat{\pi}$ satisfying

$$V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \varepsilon \quad (6.2)$$

with high probability. A desirable offline RL algorithm would achieve this target accuracy using as few samples from $\mathcal{D}$ as possible. As in Chapters 3 and 5, particular emphasis is placed on attaining minimal sample complexity uniformly across the full range of accuracy levels, namely for all $\varepsilon \in (0, \frac{1}{1-\gamma}]$.

Challenges. Unlike online RL, offline RL precludes active interaction with the environment to collect additional data. Consequently, the learner must rely entirely on the static dataset $\mathcal{D}$. The statistical efficiency of offline RL is therefore fundamentally dictated by how informative this historical dataset is for learning a near-optimal policy. Below, we highlight two fundamental challenges arising from this restricted data generation mechanism, both of which must be carefully addressed in order to achieve statistically efficient offline RL.

- *Distribution shift.* For the most part, the historical dataset is generated by one or more behavior policies that differ from the target optimal policy. As a result, the state-action distribution induced by the target policy may deviate substantially from the sampling distribution $d^{\mathbf{b}}$ underlying the dataset $\mathcal{D}$. A central challenge in offline RL is therefore how to effectively leverage past data despite this mismatch between the data distribution and the one induced by the optimal policy.
- *Limited data coverage.* Ideally, one would like the dataset $\mathcal{D}$ to contain sufficiently many samples for every state-action pair, so that the performance of all candidate policies can be accurately evaluated. In practice, however, such uniform coverage is often unrealistic, since the learner has no control over the past data collection process. Fortunately, uniform coverage may be unnecessary, given that the ultimate objective is only to identify a single near-optimal policy rather than accurately evaluate every policy.

Key parameters: single-policy concentrability. To quantitatively capture the distribution shift between the target occupancy measure and the underlying data distribution, we introduce a key quantity previously introduced in Rashidinejad *et al.* (2021).

Definition 6.1 (Single-policy concentrability for infinite-horizon MDPs). For any given initial distribution $\rho \in \Delta(\mathcal{S})$, the single-policy concentrability coefficient underlying the batch dataset $\mathcal{D}$ (see (6.1)) is defined as

$$C_\rho^\star := \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d^\star(s, a; \rho)}{d^b(s, a)}, \quad (6.3)$$

where $d^\star$ represents the discounted occupancy distribution induced by the optimal deterministic policy $\pi^\star$ (see (2.9b)), and the ratio 0/0 is interpreted as 0.

Remark 6.1. It is immediate to see that $C_\rho^\star \geq 1$, since

$$C_\rho^\star = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d^\star(s, a; \rho)}{d^b(s, a)} \geq \frac{\sum_{s,a} d^\star(s, a; \rho)}{\sum_{s,a} d^b(s, a)} = \frac{1}{1} = 1.$$

The above inequality follows from the fact that $d^\star(s, a; \rho) \leq C_\rho^\star d^b(s, a)$.

In words, $C_\rho^\star$ quantifies the distribution mismatch between the target occupancy distribution $d^\star$ (induced by the optimal policy starting from the initial distribution ρ) and the underlying data distribution d^b through the maximum density ratio. Since the value function in (2.3) is defined in a discounted manner, the occupancy distribution $d^\star$ is likewise chosen to be discounted. When $C_\rho^\star$ is close to 1, the batch dataset $\mathcal{D}$ may be viewed as containing near-expert data, meaning that its induced distribution closely aligns with that induced by the optimal policy and therefore reflects, to some extent, how an expert makes decisions. In this favorable scenario, the offline RL problem becomes intimately connected to “imitation learning,” where it may suffice to simply imitate the expert’s behavior.

Moreover, $C_\rho^\star$ is referred to as a “single-policy” concentrability coefficient because it concerns only the optimal policy $\pi^\star$. This assumption is substantially weaker than a uniform coverage requirement, or more precisely, the all-policy concentrability assumption adopted in earlier work, e.g., Munos (2007), Farahmand *et al.* (2010), Fan *et al.* (2019), and Chen and Jiang (2019). Note that the latter requires a uniform density-ratio bound over all policies and therefore demands the dataset to be highly exploratory, or equivalently, to provide adequate coverage of all state-action pairs.

Given that $C_\rho^\star$ is introduced to capture the challenges arising from both distribution mismatch and limited data coverage, we will treat it as a key problem parameter throughout the remainder of this chapter, and develop both sample complexity upper bounds and minimax lower bounds primarily in terms of $C_\rho^\star$.

Remark 6.2. A refined version of $C_\rho^\star$, called the *single-policy clipped concentrability* coefficient (Li *et al.*, 2024b), can be obtained via clipping:

$$C_{\rho, \text{clipped}}^\star := \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\min \{d^\star(s, a; \rho), \frac{1}{S}\}}{d^b(s, a)}. \quad (6.4)$$

Clearly, one has

$$C_{\rho, \text{clipped}}^\star \leq C_\rho^\star.$$

In addition, unlike $C_\rho^\star$, which always satisfies $C_\rho^\star \geq 1$, the clipped coefficient can be as low as $1/S$, a property that can sometimes lead to improved statistical guarantees, especially when d^b is close to a uniform distribution, such as the generative model setting. As it turns out, many of the upper and lower bounds developed in this chapter continue to hold with $C_\rho^\star$ replaced by $C_{\rho, \text{clipped}}^\star$; see Li *et al.* (2024b) for details.

6.2 A detour: the LCB algorithm for multi-armed bandits

Before turning to offline RL in full generality, it is instructive to first examine a much simpler setting that already captures two central challenges: distribution mismatch and limited data coverage. To build intuition, we begin with the classical multi-armed bandit problem to introduce the principle of pessimism in the face of uncertainty — a principle that will play a central role more broadly in offline RL.

6.2.1 Setting: offline stochastic bandits

The basic setup of the multi-armed bandit problem has been introduced in Section 5.2. We briefly recap the setting and describe the corresponding offline data-generation mechanism.

- *Bandits.* Consider an A -armed stochastic bandit setting, where each arm $a \in \mathcal{A} = \{1, \dots, A\}$ is associated with a mean reward $r_a \in [0, 1]$. The optimal mean reward and the corresponding optimal arm are denoted by r^* and a^* , respectively (see (5.6)).
- *Offline data.* The offline dataset $\mathcal{D}$ consists of N samples $\{(a_i, R_i)\}_{i=1}^N$ generated independently according to some joint distribution. More concretely, each action a_i is sampled as

$$a_i \stackrel{\text{i.i.d.}}{\sim} d^b, \quad i = 1, \dots, N \quad (6.5a)$$

for some action distribution $d^b \in \Delta(\mathcal{A})$, after which the corresponding (noisy) reward R_i is generated according to

$$R_i = r_{a_i} + z_i \quad \text{with } z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \quad (6.5b)$$

- *Single-policy concentrability.* Let d^* denote the distribution induced by the optimal policy, i.e., always pulling the optimal arm a^* ,

$$d^*(a) := \begin{cases} 1, & \text{if } a = a^*, \\ 0, & \text{else.} \end{cases} \quad (6.6a)$$

Then the single-policy concentrability coefficient defined in (6.3) reduces to

$$C^* := \max_{a \in \mathcal{A}} \frac{d^*(a)}{d^b(a)} = \frac{1}{d^b(a^*)}. \quad (6.6b)$$

Here, we drop the dependence on ρ in the notation C_ρ^* , since the bandit setting contains only a single state.

Goal. Armed with the offline dataset $\mathcal{D}$, our goal is to learn an arm $\hat{a}$ that achieves a suboptimality gap

$$r(a^*) - r(\hat{a})$$

as small as possible. Note that this notion of suboptimality also plays a central role in online multi-armed bandits through the definition of regret; see (5.8a) in Section 5.2.

6.2.2 Pessimism in the face of uncertainty

We now introduce some core ideas that enable statistically efficient offline learning in bandit problems. In contrast to online bandits, which rely heavily on the optimism principle (see Section 5.2) to guide exploration, offline settings do not permit active exploration. As a result, optimism-based approaches are inherently inapplicable, necessitating a fundamentally different set of ideas.

A first attempt: the empirical best arm. A natural strategy for identifying the optimal arm is to select the one with the highest empirical mean reward. Denoting by N_a the number of times arm a is pulled in $\mathcal{D}$, the empirical reward estimate for each arm $a \in \mathcal{A}$ is given by

$$\bar{r}_a := \begin{cases} \frac{1}{N_a} \sum_{t=1}^N R_t \mathbb{1}\{a_t = a\}, & \text{if } N_a > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (6.7a)$$

which coincides with the maximum likelihood estimator of the mean reward under the Gaussian noise assumption. The empirical best arm is then defined as

$$a^{\text{emp}} := \arg \max_{a \in \mathcal{A}} \bar{r}_a. \quad (6.7b)$$

Although intuitive, this empirical best-arm estimator can perform poorly when certain arms are only infrequently sampled in the dataset, making their mean rewards difficult to estimate reliably. In fact, one can construct examples showing that even in the favorable regime where $C^* \approx 1$, selecting the empirical best arm may still incur a non-vanishing suboptimality gap regardless of how large the sample size becomes. This issue arises from the high statistical uncertainty associated with poorly explored suboptimal arms, which can lead to over-estimation of their rewards with non-vanishing probability. One concrete example is given below.

Example 6.1. Suppose that $\mathcal{A} = \{1, 2\}$, and that the corresponding mean rewards are given by $r_1 = 1$ and $r_2 = 0$. Assume that $a_i = 1$ for all $i = 1, \dots, N-1$, and $a_N = 2$. Then for any $N \geq 2$, we have

$$\mathbb{E}[r^* - r_{a^{\text{emp}}}] \geq 0.075.$$

Remark 6.3. In this example, the optimal arm is arm 1, which is sampled $N-1$ times out of a total of N observations. Consequently, the corresponding C^* is exceedingly close to 1.

Proof of Example 6.1. Under the reward model, the empirical reward estimates satisfy $\bar{r}_1 = \mathcal{N}(1, \frac{1}{N-1})$ and $\bar{r}_2 = \mathcal{N}(0, 1)$, and hence

$$\mathbb{P}(\bar{r}_1 < 1) = 0.5 \quad \text{and} \quad \mathbb{P}(\bar{r}_2 \geq 1) \geq 0.15.$$

It then follows from the statistical independence between $\bar{r}_1$ and $\bar{r}_2$ that

$$\mathbb{E}[r_1 - r_{a^{\text{emp}}}] = \mathbb{P}(\bar{r}_2 > \bar{r}_1) \geq \mathbb{P}(\bar{r}_1 < 1) \mathbb{P}(\bar{r}_2 \geq 1) \geq 0.075$$

as claimed. $\square$

The lower confidence bound (LCB) algorithm. As illustrated in Example 6.1, the main issue with the empirical best-arm estimator is the systematic over-estimation of the rewards of poorly sampled arms, caused by large variance. A natural strategy toward mitigating this issue is therefore to properly adjust the empirical reward estimates by accounting for variance information.

This idea leads to the principle of pessimism (or conservatism) in the face of uncertainty: arms with fewer samples should be treated more conservatively, since their empirical rewards are more volatile and less reliable. As a concrete method, one can penalize the empirical reward estimate by a penalty function b_a that decreases as the sample count N_a (i.e., the number of times arm a is sampled) increases. Rather than selecting the arm with the largest empirical reward, the pessimism principle instead recommends choosing the arm according to the lower confidence bound:

$$a^{\text{lcb}} = \arg \max_{a \in \mathcal{A}} r_a^{\text{lcb}} \quad \text{with } r_a^{\text{lcb}} = \bar{r}_a - b_a.$$

A concrete choice of the penalty function is

$$b_a = \sqrt{\frac{2 \log(\frac{4A}{\delta})}{N_a}} \quad (6.8)$$

with a prescribed confidence parameter $\delta \in (0, 1)$, which serves as a confidence width derived from standard Gaussian tail bounds and resembles the exploration bonus (cf. (5.10a)) adopted in the UCB algorithm. Intuitively, the quantity r_a^{lcb} can be viewed as a lower confidence bound on the true reward r_a , motivating the name LCB algorithm. The full procedure is summarized in Algorithm 6.1.

Careful readers may notice that the LCB algorithm bears some resemblance to the UCB algorithm presented in Section 5.2. However, whereas UCB adds an exploration bonus to encourage exploration in online settings, the LCB approach subtracts a confidence penalty to guard against overestimation in the absence of exploration.

Remark 6.4. As a sanity check, when applying Algorithm 6.1 to Example 6.1, we can readily verify that $r_1^{\text{lcb}} \approx r_1 = 1$ for large enough N , and $r_2^{\text{lcb}} \leq r_2 = 0$. Consequently, the LCB algorithm selects arm 1 (the optimal arm) and rules out arm 2 with high probability. Part of the reason this approach works well in the present example is that the algorithm effectively “trusts” the preference of the offline dataset: since the optimal arm is sampled frequently in the data, the estimate of r_1 becomes highly reliable, allowing the algorithm to confidently distinguish it from the suboptimal arm.

6.2.3 Theoretical guarantees

We now turn to the statistical guarantees of the LCB algorithm. As it turns out, the pessimism principle described above proves highly effective in coping with the challenges inherent to offline settings. This is formalized in the following theorem, originally established by Rashidinejad *et al.* (2021).

Theorem 6.1. Suppose that the sample size satisfies $N \geq 8C^* \log(2/\delta)$. Then with probability at least $1 - \delta$, the action a^{lcb} returned by the LCB algorithm (cf. Algorithm 6.1) with penalty function (6.8) achieves

$$r^* - r_{a^{\text{lcb}}} \leq 4 \sqrt{\frac{C^* \log(4A/\delta)}{N}}. \quad (6.9)$$

Algorithm 6.1: The lower confidence bound (LCB) algorithm for multi-armed bandits

```

1 input: batch dataset  $\mathcal{D} = \{(a_i, R_i)\}_{i=1}^N$ , number of arms  $A$ , confidence parameter  $\delta \in (0, 1)$ .
2 for  $a \in \mathcal{A}$  do
3   set  $N_a \leftarrow \sum_{i=1}^N \mathbb{1}\{a_i = a\}$ . /* visitation count. */
4   compute empirical reward
      
$$\bar{r}_a = \begin{cases} \frac{1}{N_a} \sum_{i=1}^N R_i \mathbb{1}\{a_i = a\} & \text{if } N_a > 0 \\ 0 & \text{otherwise} \end{cases}.$$

      set  $b_a \leftarrow \sqrt{\frac{2 \log(4A/\delta)}{N_a}}$ . /* LCB penalty. */
5 output:  $a^{\text{lcb}} = \arg \max_a r_a^{\text{lcb}}$ , where  $r_a^{\text{lcb}} = \bar{r}_a - b_a$ .
```

Before embarking on the proof of this result, we make several remarks. To achieve any target accuracy $0 < \varepsilon < 1$ (in the sense that $r^* - r_{a^{\text{lcb}}} \leq \varepsilon$), it suffices for the LCB algorithm to use

$$\tilde{O}\left(\frac{C^*}{\varepsilon^2}\right) \quad (6.10)$$

samples from the offline dataset $\mathcal{D}$. This stands in sharp contrast to the empirical best-arm estimator, which in certain scenarios may be fundamentally incapable of achieving $o(1)$ statistical error. This theorem therefore provides rigorous evidence that the pessimism principle can significantly outperform the vanilla empirical approach.

Notably, the sample complexity of the LCB algorithm scales linearly with the single-policy concentrability coefficient, without requiring this coefficient to be known *a priori*. This unveils that the statistical efficiency of the pessimistic approach automatically adapts to the intrinsic quality of the offline dataset, as characterized by C^* , rather than the number of arms A .

Intuitive explanation. Prior to delving into the proof of Theorem 6.1, it is instructive to first develop some intuition for the result. Suppose, for simplicity, that every suboptimal arm a satisfies $r_a < r^* - \varepsilon$, so that achieving ε -accuracy amounts to ensuring that any suboptimal arm is unlikely to be selected.

Recall from the definition (6.6b) that $C^* = 1/d^b(a^*)$. Consequently, the number of observations collected for the optimal arm a^* scales as

$$N_{a^*} \asymp N d^b(a^*) \asymp N/C^*.$$

As a result, when the sample size N is sufficiently large, both the empirical estimate $\bar{r}_{a^*}$ and the corresponding lower confidence bound $r_{a^*}^{\text{lcb}}$ concentrate tightly around the true reward:

$$\bar{r}_{a^*} \approx r_{a^*}^{\text{lcb}} = r_{a^*} - O\left(\frac{1}{\sqrt{N_{a^*}}}\right) = r^* - O\left(\sqrt{\frac{C^*}{N}}\right),$$

where logarithmic factors are omitted for simplicity. Therefore, the lower confidence bound $r_{a^*}^{\text{lcb}}$ can be statistically separated from the mean reward r_a of any suboptimal arm — and hence also from its lower confidence bound r_a^{lcb} — provided that

$$\sqrt{\frac{C^*}{N}} \ll \varepsilon \quad \Longleftrightarrow \quad N \gg \frac{C^*}{\varepsilon^2}.$$

This heuristic captures the key intuition underlying the analysis.

6.2.4 Analysis: proof of Theorem 6.1

We begin by establishing two high-probability events. First, akin to the application of the Gaussian tail bounds (5.11) in the online setting, one can show that

$$r_a^{\text{lcb}} = \bar{r}_a - b_a \leq r_a \leq \bar{r}_a + b_a \quad \text{for all } 1 \leq a \leq A, \quad (6.11)$$

with probability at least $1 - \delta/2$. This guarantees that r_a^{lcb} is indeed a pessimistic estimate of the true mean reward r_a . Second, the concentration bound in Lemma 2.7 tells us that

$$N_{a^*} \geq \mathbb{E}[N_{a^*}] - \frac{1}{2}Nd^{\text{b}}(a^*) = \frac{1}{2}Nd^{\text{b}}(a^*) \quad (6.12)$$

with probability at least $1 - \exp(-Nd^{\text{b}}(a^*)/8)$. Consequently, if $N \geq 8C^* \log(2/\delta)$, then (6.12) holds with probability exceeding $1 - \delta/2$.

Conditioning on the intersection of the two events above, we proceed to compare the optimal arm a^* with the pessimistic choice a^{lcb} . It is readily seen from (6.11) that

$$r_{a^*} \leq \bar{r}_{a^*} + b_{a^*} = \bar{r}_{a^*} - b_{a^*} + 2b_{a^*}.$$

In view of the definition of a^{lcb} , we have $\bar{r}_{a^*} - b_{a^*} \leq \bar{r}_{a^{\text{lcb}}} - b_{a^{\text{lcb}}}$, which combined with (6.11) again yields

$$r_{a^*} \leq \bar{r}_{a^{\text{lcb}}} - b_{a^{\text{lcb}}} + 2b_{a^*} \leq r_{a^{\text{lcb}}} + 2b_{a^*}. \quad (6.13)$$

Therefore, to bound the suboptimality gap $r_{a^*} - r_{a^{\text{lcb}}}$, it suffices to control the penalty term b_{a^*} , which reduces to lower bounding N_{a^*} . By virtue of (6.12) and the definition $C^* = 1/d^{\text{b}}(a^*)$ (cf. (6.6b)), we have

$$N_{a^*} \geq \frac{1}{2}Nd^{\text{b}}(a^*) = \frac{N}{2C^*} > 0,$$

which further implies that

$$b_{a^*} = \sqrt{\frac{2 \log(4A/\delta)}{N_{a^*}}} \leq \sqrt{\frac{4C^* \log(4A/\delta)}{N}}.$$

Substituting this into (6.13), we arrive at

$$r_{a^*} - r_{a^{\text{lcb}}} \leq 2b_{a^*} \leq 4\sqrt{\frac{C^* \log(4A/\delta)}{N}}$$

with probability at least $1 - \delta$. This establishes Theorem 6.1.

6.3 Model-based offline RL

Thus far, we have demonstrated the effectiveness of the pessimism principle in the simpler setting of offline multi-armed bandits. This naturally motivates extending the pessimism principle to offline RL, with the goal of addressing the challenges of distribution mismatch and limited data coverage in much more general offline settings.

As it turns out, despite the different sources of uncertainty in offline bandits and offline RL — namely, noisy rewards in the former and unknown transition models in the latter — the pessimism

principle can be incorporated into both model-based and model-free approaches in a principled manner. Given the minimax optimality of model-based methods in both RL with a generative model and online RL, it is natural to first investigate model-based offline RL algorithms and their statistical optimality, which will be the focus of this section.

6.3.1 Algorithmic ideas

A natural approach to incorporating pessimism into offline RL is to penalize the value estimates of state-action pairs that receive poor coverage in the offline dataset $\mathcal{D}$. Intuitively, when a state-action pair is rarely observed, the corresponding value estimate is subject to high uncertainty and should therefore be treated with caution. Pessimistic value estimation addresses this challenge by systematically lowering value estimates in poorly covered regions of the state-action space, with the amount of penalization calibrated according to the degree of uncertainty.

In what follows, we introduce a model-based offline RL algorithm that integrates lower confidence bounds into the value estimation procedure, building on ideas originally developed in Jin *et al.* (2021b) and Rashidinejad *et al.* (2021). Commonly referred to as VI-LCB, this algorithm performs value iteration over the empirical MDP through a pessimistic Bellman operator. We now describe its core algorithmic components.

- *Model estimation.* As in the setting of model-based RL with a generative model discussed in Chapter 3, our first step is to build an empirical estimate $\hat{P}$ of the true transition kernel P underlying the MDP. Recall that the offline dataset contains N independent sample transitions $\{(s_i, a_i, s'_i)\}_{i=1}^N$ (see (6.1)). For each state-action pair (s, a) , denote by

$$N_{s,a} := \sum_{i=1}^N \mathbb{1}((s_i, a_i) = (s, a)) \quad (6.14a)$$

the number of times (s, a) appears in the dataset. Based on these visitation counts, we define the empirical transition kernel $\hat{P}$ by

$$\hat{P}(s' | s, a) = \begin{cases} \frac{1}{N_{s,a}} \sum_{i=1}^N \mathbb{1}\{(s_i, a_i, s'_i) = (s, a, s')\}, & \text{if } N_{s,a} > 0 \\ \frac{1}{S}, & \text{else} \end{cases} \quad (6.14b)$$

for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. The uniform assignment in the case $N_{s,a} = 0$ is introduced merely to ensure that $\hat{P}(\cdot | s, a)$ remains a valid probability distribution even for unseen state-action pairs.

- *The pessimistic Bellman operator.* At the core of the VI-LCB algorithm lies a modified Bellman operator whose fixed point produces conservative value estimates. Recall that, given a transition kernel P , the classical Bellman operator $\mathcal{T}(\cdot) : \mathbb{R}^{\mathcal{S}\mathcal{A}} \rightarrow \mathbb{R}^{\mathcal{S}\mathcal{A}}$ is defined as follows: for any input vector $Q \in \mathbb{R}^{\mathcal{S}\mathcal{A}}$ and its induced value vector $V \in \mathbb{R}^{\mathcal{S}}$ (i.e., $V(s) := \max_a Q(s, a)$ for all s),

$$\mathcal{T}(Q)(s, a) := r(s, a) + \gamma P_{s,a} V \quad \text{for all } (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (6.15)$$

To incorporate pessimism into the value estimation procedure, we replace the true transition kernel P with the empirical kernel $\hat{P}$ (cf. (6.14b)) and introduce an additional penalty term

that accounts for statistical uncertainty. Specifically, we define the pessimistic Bellman operator $\widehat{\mathcal{T}}_{\text{pe}}$ by

$$\widehat{\mathcal{T}}_{\text{pe}}(Q)(s, a) := \max \left\{ r(s, a) + \gamma \widehat{P}_{s,a} V - b(s, a; V), 0 \right\}, \quad (6.16)$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here, $b(s, a; V)$ stands for a penalty function designed to capture the uncertainty associated with estimating the Bellman update at (s, a) , whose precise form will be specified momentarily. The VI-LCB algorithm then seeks to compute the fixed point of this pessimistic Bellman operator.

- *Modified value iterations.* A natural scheme for computing the fixed point of the pessimistic Bellman operator $\widehat{\mathcal{T}}_{\text{pe}}$ (cf. (6.16)) is to employ a value-iteration-style iterative procedure. Specifically, analogous to standard value iteration (2.16), we repeatedly apply the operator $\widehat{\mathcal{T}}_{\text{pe}}$ starting from the initialization $\widehat{Q}_0 = 0$. For each iteration $\tau = 1, 2, \dots, \tau_{\max}$, we update

$$\begin{aligned} \widehat{Q}_\tau(s, a) &= \widehat{\mathcal{T}}_{\text{pe}}(\widehat{Q}_{\tau-1})(s, a) \\ &= \max \left\{ r(s, a) + \gamma \widehat{P}_{s,a} \widehat{V}_{\tau-1} - b(s, a; \widehat{V}_{\tau-1}), 0 \right\}, \end{aligned} \quad (6.17)$$

where for all $s \in \mathcal{S}$,

$$\widehat{V}_{\tau-1}(s) := \max_a \widehat{Q}_{\tau-1}(s, a).$$

After $\tau_{\max}$ iterations, the algorithm outputs $\widehat{Q} = \widehat{Q}_{\tau_{\max}}$ as the final Q-estimate. The resulting policy estimate $\widehat{\pi}$ is then obtained greedily with respect to $\widehat{Q}$, namely,

$$\widehat{\pi}(s) \in \arg \max_a \widehat{Q}(s, a) \quad \text{for all } s \in \mathcal{S}. \quad (6.18)$$

The whole algorithm is summarized in Algorithm 6.2.

Let us briefly provide some high-level intuition underlying this algorithmic design. When a state-action pair (s, a) is insufficiently covered in the offline dataset $\mathcal{D}$ (that is, when the visitation count $N_{s,a}$ is small), the corresponding estimate $\widehat{Q}_\tau(s, a)$ suffers from high variance and therefore becomes unreliable; if not carefully handled, such uncertainty can propagate through the value iterations and adversely affect the overall value estimation process. To mitigate this issue, the VI-LCB algorithm introduces a penalty function $b(s, a; \widehat{V}_{\tau-1})$ derived from suitable lower confidence bounds, which effectively downweights poorly visited state-action pairs and limits their negative influence on the learned value function. Importantly, when the coverage coefficient C_ρ^* is small, suppressing these poorly covered regions does not necessarily introduce significant bias in estimating the optimal value function. For instance, when the behavior policy π^b generating $\mathcal{D}$ aligns closely with the optimal policy π^* , the poorly covered regions typically correspond to suboptimal actions that are rarely visited under π^b , making it acceptable to assign conservative estimates to such pairs.

6.3.2 Choices of penalty functions

As one might expect, the properties of the fixed point of $\widehat{\mathcal{T}}_{\text{pe}}$ — and consequently the statistical performance of VI-LCB — depend critically on the choice of the penalty function. The role of $b(s, a; V)$ is to help construct a lower confidence bound for the true Bellman update $r(s, a) + \gamma \langle P_{s,a}, V \rangle$ based on its empirical counterpart (see (6.17))

$$r(s, a) + \gamma \langle \widehat{P}_{s,a}, V \rangle.$$

Accordingly, the penalty term $b(s, a; V)$ should be chosen to characterize the confidence width of the estimation error of the form

$$\langle \hat{P}_{s,a} - P_{s,a}, V \rangle,$$

where $V \in \mathbb{R}^S$ is some vector obeying $V(s) \in [0, \frac{1}{1-\gamma}]$. Similar to the design of the exploration bonus in UCB-VI (see Section 5.3.2), there are at least two commonly used choices of penalty functions for this purpose.

- *Hoeffding-style penalty.* The first type of penalty functions, adopted for instance by Rashidinejad *et al.* (2021), is motivated by Hoeffding-type concentration inequalities together with the boundedness property $Q^*(s, a) \in [0, \frac{1}{1-\gamma}]$. Specifically, one can choose

$$b(s, a; V) = \tilde{\Theta} \left(\sqrt{\frac{1}{(1-\gamma)^2 N_{s,a}}} \right),$$

which admits a relatively simple form and is often amenable to analysis. However, since it does not leverage variance information, the resulting statistical guarantees are generally suboptimal.

- *Bernstein-style penalty.* Similar to the online RL setting, employing a Bernstein-style penalty to characterize the confidence width often leads to improved statistical performance, owing to its ability to properly exploit the underlying variance structure of the MDP of interest. More specifically, the Bernstein-style penalty typically takes the form

$$b(s, a; V) = \tilde{\Theta} \left(\sqrt{\frac{\text{Var}_{\hat{P}_{s,a}}(V)}{N_{s,a}}} + \frac{1}{(1-\gamma)N_{s,a}} \right),$$

where the variance term appearing in the Bernstein inequality is replaced by its empirical counterpart in order to make the penalty computable. Compared with the Hoeffding-style counterpart, this choice yields a substantially sharper confidence width by adapting to the local variability of the value function. As we shall justify shortly, such variance-aware penalization enables VI-LCB to achieve the minimax-optimal sample complexity for offline RL.

In this section, we focus on the following concrete choice of Bernstein-style penalty functions:

$$b(s, a; V) := \min \left\{ \max \left\{ \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(V), \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} \right\}, \frac{1}{1-\gamma} \right\} + \frac{5}{N} \quad (6.19)$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $c_b > 0$ is some numerical constant, and $\delta \in (0, 1)$ is some confidence parameter. As usual, $\text{Var}_{\hat{P}_{s,a}}(V)$ denotes the variance of the vector $V \in \mathbb{R}^S$ under the empirical transition distribution $\hat{P}_{s,a}$; see (2.21).

6.3.3 Computational efficiency

Given the iterative nature of the VI-LCB algorithm in Algorithm 6.2, a natural question concerns the number of iterations required for convergence, as this directly determines the computational complexity of the algorithm.

As it turns out, the VI-LCB algorithm enjoys a rapid convergence guarantee. In particular, the pessimistic Bellman operator $\hat{\mathcal{T}}_{\text{pe}}$ inherits a γ -contraction property (to be established shortly), which implies that the iterates $\{\hat{Q}_\tau\}_{\tau \geq 0}$ converge geometrically fast to the fixed point $\hat{Q}_{\text{pe}}^*$. More precisely, we have the following result.

Algorithm 6.2: Value iteration with lower confidence bounds (VI-LCB) for discounted infinite-horizon MDPs

```

1 input: dataset  $\mathcal{D}$ , discount factor  $\gamma$ , reward function  $r$ , confidence parameter  $\delta$ , max
   iteration number  $\tau_{\max}$ .
2 initialize  $\hat{Q}_0 = 0$ ,  $\hat{V}_0 = 0$ .
   /* model estimation based on offline dataset. */
3 construct the empirical transition kernel  $\hat{P}$  according to (6.14).
   /* pessimistic value iteration. */
4 for  $\tau = 1, 2, \dots, \tau_{\max}$  do
5   for  $s \in \mathcal{S}, a \in \mathcal{A}$  do
6     compute penalty term  $b(s, a; \hat{V}_{\tau-1})$  according to (6.19).
7      $\hat{Q}_\tau(s, a) \leftarrow \max \{r(s, a) + \gamma \hat{P}_{s,a} \hat{V}_{\tau-1} - b(s, a; \hat{V}_{\tau-1}), 0\}$ .
8   for  $s \in \mathcal{S}$  do
9      $\hat{V}_\tau(s) \leftarrow \max_a \hat{Q}_\tau(s, a)$ .
10 output:  $\hat{\pi}$  s.t.  $\hat{\pi}(s) \in \arg \max_a \hat{Q}_{\tau_{\max}}(s, a)$  for every  $s \in \mathcal{S}$ .

```

Lemma 6.2. Suppose $\hat{Q}_0 = 0$. Then the iterates of Algorithm 6.2 obey

$$\hat{Q}_\tau \leq \hat{Q}_{\text{pe}}^* \quad \text{and} \quad \|\hat{Q}_\tau - \hat{Q}_{\text{pe}}^*\|_\infty \leq \frac{\gamma^\tau}{1-\gamma} \quad \text{for all } \tau \geq 0, \quad (6.20)$$

where $\hat{Q}_{\text{pe}}^*$ denotes the unique fixed point of $\hat{\mathcal{T}}_{\text{pe}}$. As a consequence, choosing $\tau_{\max} \geq \frac{\log \frac{N}{1-\gamma}}{\log(1/\gamma)}$ guarantees that

$$\|\hat{Q}_{\tau_{\max}} - \hat{Q}_{\text{pe}}^*\|_\infty \leq 1/N. \quad (6.21)$$

The proof of this lemma is deferred to Section 6.7.1. According to this lemma, $\hat{Q}_\tau$ becomes an ε -accurate approximation of $\hat{Q}_{\text{pe}}^*$ — in the sense that $\|\hat{Q}_\tau - \hat{Q}_{\text{pe}}^*\|_\infty \leq \varepsilon$ — after at most $\frac{1}{1-\gamma} \log \frac{1}{\varepsilon(1-\gamma)}$ iterations.

6.3.4 Theoretical guarantees

We now turn to the statistical guarantees of the VI-LCB algorithm. As we shall see, when equipped with the Bernstein-style penalty function in (6.19), the VI-LCB algorithm described in Algorithm 6.2 is guaranteed to produce an ε -optimal policy using a minimal number of offline samples. This is formalized in the theorem below, originally established by Li *et al.* (2024b).

Theorem 6.3 (Li *et al.* (2024b)). Suppose $\gamma \in [\frac{1}{2}, 1)$. Consider any $0 < \delta < 1$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$, and let $\rho \in \Delta(\mathcal{S})$ be any test distribution over the state space. Assume that the number of iterations satisfies $\tau_{\max} \geq \frac{1}{1-\gamma} \log \frac{N}{1-\gamma}$. Then with probability at least $1 - 2\delta$, the policy $\hat{\pi}$ returned by VI-LCB (see Algorithm 6.2) with penalty function (6.19) obeys

$$V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \varepsilon, \quad (6.22)$$

provided that $c_b > 0$ is some sufficiently large numerical constant and the total sample size exceeds

$$N \geq \frac{c_1 S C_\rho^* \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3 \varepsilon^2} \quad (6.23)$$

for some large enough universal constant $c_1 > 0$. Here, we recall that $C_\rho^\star \geq 1$ has been defined in Definition 6.1.

We now briefly discuss the implications of Theorem 6.3. Under the Bernstein-style penalty, the VI-LCB algorithm achieves ε -optimality with a sample complexity no greater than

$$\tilde{O}\left(\frac{SC_\rho^\star}{(1-\gamma)^3\varepsilon^2}\right). \quad (6.24)$$

Notably, this sample size can be substantially smaller than the ambient dimension S^2A of the transition kernel P , in which case accurately estimating the full model becomes information-theoretically impossible. Nevertheless, Theorem 6.3 demonstrates that pessimistic model-based methods can still achieve statistically efficient offline learning even in regimes where reliable model estimation is unattainable.

Importantly, the statistical guarantees in Theorem 6.3 scale linearly with the single-policy concentrability coefficient $C_\rho^\star$, a metric quantifying the degree of distribution mismatch between the behavior policy generating the offline dataset and the target optimal policy. We emphasize that Algorithm 6.2 requires no prior knowledge of $C_\rho^\star$, making the method fully adaptive to the quality and coverage of the dataset.

It is also instructive to compare the sample complexity in (6.24) with the minimax sample complexity $\tilde{\Theta}(\frac{SA}{(1-\gamma)^3\varepsilon^2})$ under a generative model (as established in Chapter 3). In the favorable scenarios where $C_\rho^\star \leq \tilde{O}(1)$, the model-based offline approach therefore enjoys an $\tilde{O}(A)$ sample size saving relative to the generative-model setting when controlling the suboptimality gap $V^\star(\rho) - V^{\hat{\pi}}(\rho)$ w.r.t. the initial state distribution ρ . At a high level, VI-LCB achieves this improvement by effectively “trusting” the offline dataset $\mathcal{D}$: it primarily relies on regions of the state-action space that are well covered by $\mathcal{D}$, while acting conservatively in poorly covered regions. Intuitively, when $\mathcal{D}$ is generated by a behavior policy with expert-level performance — so that $C_\rho^\star$ becomes small — the frequently visited state-action pairs are precisely those achieving high-quality decision-making, making it advantageous to place greater emphasis on these well-sampled regions.

Furthermore, as will be corroborated by a matching minimax lower bound established in Section 6.5, the VI-LCB algorithm achieves minimax-optimal sample complexity up to logarithmic factors. In particular, the sample complexity bound in (6.23) is valid over the entire accuracy regime $\varepsilon \in (0, \frac{1}{1-\gamma}]$, implying that *no additional burn-in sample size* is needed to attain sample optimality. Establishing such a statistical guarantee is especially challenging in the sample-starved regime.

Finally, we remark on a technical aspect of Algorithm 6.2. The algorithm reuses the entire offline dataset across all iterations, a feature that is crucial for attaining optimal sample complexity. Noteworthy, while such sample reuse is natural from a practical perspective, it introduces substantial technical challenges for the analysis by creating intricate statistical dependencies across iterations. By contrast, some earlier work, such as Rashidinejad *et al.* (2021), assumed the use of fresh samples at each iteration in order to simplify the analysis, albeit at the expense of higher sample complexity.

6.3.5 Analysis: proof of Theorem 6.3

In this subsection, we outline the proof of Theorem 6.3. In what follows, we first introduce several preliminary facts, and then decompose the proof into a few steps.

Preliminary facts. In view of Lemma 6.2, the final Q-estimate $\hat{Q}$ in Algorithm 6.2 satisfies

$$\|\hat{Q} - \hat{Q}_{\text{pe}}^*\|_\infty \leq \frac{1}{N}. \quad (6.25)$$

The following lemma records a few direct consequences of (6.25), whose proof is deferred to Section 6.7.2.

Lemma 6.4. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\|\hat{V} - \hat{V}_{\text{pe}}^*\|_\infty \leq \frac{1}{N}, \quad (6.26a)$$

$$|b(s, a; \hat{V}_{\text{pe}}^*) - b(s, a; \hat{V})| \leq \frac{3}{N}. \quad (6.26b)$$

The next lemma provides a lower bound on $N_{s,a}$, the total number of sample transitions collected from each state-action pair (s, a) . The proof is deferred to Section 6.7.3.

Lemma 6.5. Consider any $\delta \in (0, 1)$. With probability exceeding $1 - \delta$, the quantities $\{N_{s,a}\}$ defined in (6.14a) satisfy

$$\max \left\{ N_{s,a}, \frac{2}{3} \log \frac{N}{\delta} \right\} \geq \frac{Nd^b(s, a)}{12} \quad (6.27)$$

simultaneously for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

In view of Lemma 6.5, throughout the proof, unless otherwise specified, we focus on the high-probability event that

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad \max \left\{ N_{s,a}, \frac{2}{3} \log \frac{N}{\delta} \right\} \geq \frac{Nd^b(s, a)}{12}. \quad (6.28)$$

It is worth mentioning that, under the sampling model (6.1), the sample transitions used to construct $\hat{P}$ are statistically independent conditional on $\{N_{s,a}\}$. The proof proceeds in four steps, as detailed below.

Step 1: Bernstein-style inequalities and leave-one-out analysis. We first need to obtain sharp control of the quantity $(\hat{P}_{s,a} - P_{s,a})\hat{V}$. This is challenging because $\hat{P}$ and the value estimate $\hat{V}$ are statistically dependent, as the same samples are reused throughout all iterations of Algorithm 6.2. To overcome this difficulty, we employ a leave-one-out argument to decouple this statistical dependence, leading to the following Bernstein-style concentration bounds.

Lemma 6.6. Suppose that $\gamma \in [\frac{1}{2}, 1)$, and consider any $\delta \in (0, 1)$. With probability at least $1 - \delta$, we have

$$|(\hat{P}_{s,a} - P_{s,a})\tilde{V}| \leq 12 \sqrt{\frac{\log \frac{2N}{(1-\gamma)\delta}}{N_{s,a}} \text{Var}_{\hat{P}_{s,a}}(\tilde{V})} + \frac{74 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}, \quad (6.29a)$$

$$\text{Var}_{\hat{P}_{s,a}}(\tilde{V}) \leq 2\text{Var}_{P_{s,a}}(\tilde{V}) + \frac{41 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N_{s,a}}. \quad (6.29b)$$

simultaneously for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and all $\tilde{V}$ satisfying $\|\tilde{V} - \hat{V}_{\text{pe}}^*\|_\infty \leq \frac{1}{N}$ and $\|\tilde{V}\|_\infty \leq \frac{1}{1-\gamma}$.

Proof. At a high level, the proof constructs, for each state $s \in \mathcal{S}$, a finite collection of auxiliary MDPs $\{\widehat{\mathcal{M}}^{s,u}\}$ with the following two properties. First, each $\widehat{\mathcal{M}}^{s,u}$ is constructed without using any sample transition originating from the state s , and is therefore statistically independent of any $\widehat{P}_{s,a}$. Second, at least one MDP in this collection is sufficiently close to the original true MDP in terms of the resulting value function. These leave-one-out auxiliary MDPs allow us to control $(\widehat{P}_{s,a} - P_{s,a})\widetilde{V}$ by first exploiting the statistical independence between $\widehat{P}_{s,a}$ and $\{\widehat{\mathcal{M}}^{s,u}\}$, and then transferring the resulting concentration bound back to the original true MDP. See Section 6.7.4 for the formal proof. $\square$

In the rest of the proof, we also focus on the high-probability event that (6.29) holds, in addition to the event (6.28). As an immediate consequence, combining (6.29a) with the definition (6.19) of $b(s, a; V)$ yields

$$\left| (\widehat{P}_{s,a} - P_{s,a})\widetilde{V} \right| + \frac{5}{N} \leq b(s, a; \widetilde{V}) \quad \text{for all } (s, a) \in \mathcal{S} \times \mathcal{A} \quad (6.30)$$

for all $\widetilde{V}$ obeying $\|\widetilde{V} - \widehat{V}_{\text{pe}}^*\|_\infty \leq \frac{1}{N}$ and $\|\widetilde{V}\|_\infty \leq \frac{1}{1-\gamma}$, provided that the constant c_b is sufficiently large. It is worth mentioning that, the bound (6.29a) has been derived only for state-action pairs (s, a) with $N_{s,a} > 0$. For any (s, a) with $N_{s,a} = 0$, we directly have

$$\left| (\widehat{P}_{s,a} - P_{s,a})\widetilde{V} \right| = \max\{|\widehat{P}_{s,a}\widetilde{V}|, |P_{s,a}\widetilde{V}|\} \leq \max\{\|\widehat{P}_{s,a}\|_1, \|P_{s,a}\|_1\} \|\widetilde{V}\|_\infty \leq \frac{1}{1-\gamma},$$

hence (6.30) holds trivially.

Step 2: showing that $\widehat{Q}(s, a)$ is a lower bound on $Q^{\widehat{\pi}}(s, a)$. We next show that $\widehat{Q}(s, a)$, respectively $\widehat{V}(s)$, is a “pessimistic” estimate of $Q^{\widehat{\pi}}(s, a)$, respectively $V^{\widehat{\pi}}(s)$. This follows from the pessimism principle, which ensures that the algorithm seeks lower estimates in value iteration, together with the Bernstein-style bounds in Lemma 6.6, which ensure that the penalty term dominates the uncertainty arising from the empirical transition kernel.

To begin with, recall that $\widehat{Q}_{\text{pe}}^*(s, a)$ is the unique fixed point of the pessimistic Bellman operator and satisfies

$$\widehat{Q}_{\text{pe}}^*(s, a) = \max \left\{ r(s, a) + \gamma \widehat{P}_{s,a} \widehat{V}_{\text{pe}}^* - b(s, a; \widehat{V}_{\text{pe}}^*), 0 \right\}. \quad (6.31)$$

We divide the state-action pairs (s, a) into two cases.

- Case 1: $\widehat{Q}_{\text{pe}}^*(s, a) = 0$. In view of Lemma 6.2, we know that $\widehat{Q}(s, a) = 0$, which immediately gives $\widehat{Q}(s, a) \leq Q^{\widehat{\pi}}(s, a)$.
- Case 2: $\widehat{Q}_{\text{pe}}^*(s, a) > 0$. We first observe that

$$\begin{aligned} \widehat{Q}(s, a) &\stackrel{(i)}{\leq} \widehat{Q}_{\text{pe}}^*(s, a) + \frac{1}{N} \stackrel{(ii)}{=} r(s, a) - b(s, a; \widehat{V}_{\text{pe}}^*) + \gamma \widehat{P}_{s,a} \widehat{V}_{\text{pe}}^* + \frac{1}{N} \\ &\leq r(s, a) - b(s, a; \widehat{V}_{\text{pe}}^*) + \gamma \widehat{P}_{s,a} \widehat{V} + \frac{1}{N} + \gamma \|\widehat{P}_{s,a}\|_1 \|\widehat{V} - \widehat{V}_{\text{pe}}^*\|_\infty \\ &\stackrel{(iii)}{\leq} r(s, a) - b(s, a; \widehat{V}_{\text{pe}}^*) + \gamma \widehat{P}_{s,a} \widehat{V} + \frac{2}{N} \\ &\leq r(s, a) - b(s, a; \widehat{V}) + \gamma P_{s,a} \widehat{V} + \frac{2}{N} + \gamma |(\widehat{P}_{s,a} - P_{s,a})\widehat{V}| + |b(s, a; \widehat{V}_{\text{pe}}^*) - b(s, a; \widehat{V})| \end{aligned}$$

$$\stackrel{(iv)}{\leq} r(s, a) + \gamma P_{s,a} \hat{V}. \quad (6.32)$$

Here, (i) follows from (6.25); (ii) uses the fixed point characterization (6.31) and the assumption that $\hat{Q}_{\text{pe}}^*(s, a) > 0$; (iii) utilizes (6.26a); while (iv) follows from (6.26b) as well as (6.30). Combining (6.32) with the Bellman equation $Q^{\hat{\pi}} = r + \gamma P V^{\hat{\pi}}$ gives

$$\begin{aligned} Q^{\hat{\pi}}(s, a) - \hat{Q}(s, a) &\geq r(s, a) + \gamma P_{s,a} V^{\hat{\pi}} - (r(s, a) + \gamma P_{s,a} \hat{V}) \\ &= \gamma P_{s,a} (V^{\hat{\pi}} - \hat{V}). \end{aligned} \quad (6.33)$$

Suppose that there exists some (s', a') such that $Q^{\hat{\pi}}(s', a') - \hat{Q}(s', a') < 0$. Notice that (s', a') must belong to Case 2 as we already showed that $Q^{\hat{\pi}}(s, a) - \hat{Q}(s, a) \geq 0$ for any (s, a) in Case 1. Hence we have

$$\begin{aligned} \min_{s,a} [Q^{\hat{\pi}}(s, a) - \hat{Q}(s, a)] &= Q^{\hat{\pi}}(s', a') - \hat{Q}(s', a') \\ &\stackrel{(i)}{\geq} \gamma P_{s',a'} (V^{\hat{\pi}} - \hat{V}) \stackrel{(ii)}{\geq} \gamma \min_s [V^{\hat{\pi}}(s) - \hat{V}(s)] \\ &= \gamma \min_s [Q^{\hat{\pi}}(s, \hat{\pi}(s)) - \hat{Q}(s, \hat{\pi}(s))] \\ &\geq \gamma \min_{s,a} [Q^{\hat{\pi}}(s, a) - \hat{Q}(s, a)]. \end{aligned} \quad (6.34)$$

Here (i) follows from (6.33), while (ii) holds since $P_{s,a} \in \Delta(\mathcal{S})$. Since $\gamma \in (0, 1)$, (6.34) implies $\min_{s,a} [Q^{\hat{\pi}}(s, a) - \hat{Q}(s, a)] \geq 0$, which contradicts the previous assumption. Hence, in Case 2 as well, we have $Q^{\hat{\pi}}(s, a) \geq \hat{Q}(s, a)$.

Combining the two cases, we conclude that

$$Q^{\hat{\pi}}(s, a) \geq \hat{Q}(s, a) \quad \text{for all } (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (6.35)$$

As an immediate consequence, for all $s \in \mathcal{S}$ we have

$$V^*(s) \geq V^{\hat{\pi}}(s) = Q^{\hat{\pi}}(s, \hat{\pi}(s)) \geq \hat{Q}(s, \hat{\pi}(s)) = \hat{V}(s) \quad (6.36)$$

Step 3: bounding $V^*(s) - V^{\hat{\pi}}(s)$. Recall that the Bellman optimality equation gives

$$V^*(s) = r(s, \pi^*(s)) + \gamma P_{s, \pi^*(s)} V^*. \quad (6.37)$$

We first record the following lower bound on $\hat{V}$. For every $s \in \mathcal{S}$,

$$\begin{aligned} \hat{V}(s) &= \max_a \hat{Q}(s, a) \geq \hat{Q}(s, \pi^*(s)) \stackrel{(i)}{\geq} \hat{Q}_{\text{pe}}^*(s, \pi^*(s)) - \frac{1}{N} \\ &\stackrel{(ii)}{\geq} r(s, \pi^*(s)) - b(s, \pi^*(s); \hat{V}_{\text{pe}}^*) + \gamma \hat{P}_{s, \pi^*(s)} \hat{V}_{\text{pe}}^* - \frac{1}{N} \\ &= r(s, \pi^*(s)) - b(s, \pi^*(s); \hat{V}_{\text{pe}}^*) + \gamma \hat{P}_{s, \pi^*(s)} \hat{V} - \frac{1}{N} - \gamma \hat{P}_{s, \pi^*(s)} (\hat{V} - \hat{V}_{\text{pe}}^*) \\ &\stackrel{(iii)}{\geq} r(s, \pi^*(s)) - b(s, \pi^*(s); \hat{V}_{\text{pe}}^*) + \gamma \hat{P}_{s, \pi^*(s)} \hat{V} - \frac{2}{N} \\ &\geq r(s, \pi^*(s)) - b(s, \pi^*(s); \hat{V}) + \gamma P_{s, \pi^*(s)} \hat{V} - \frac{2}{N} \\ &\quad - \gamma |(\hat{P}_{s, \pi^*(s)} - P_{s, \pi^*(s)}) \hat{V}| - |b(s, \pi^*(s); \hat{V}_{\text{pe}}^*) - b(s, \pi^*(s); \hat{V})| \end{aligned}$$

$$\stackrel{(iv)}{\geq} r(s, \pi^*(s)) - 2b(s, \pi^*(s); \hat{V}) + \gamma P_{s, \pi^*(s)} \hat{V}. \quad (6.38)$$

Here, (i) follows from (6.25); (ii) utilizes the fixed point property (6.31); (iii) holds because

$$\hat{P}_{s, \pi^*(s)}(\hat{V} - \hat{V}_{pe}^*) \leq \|\hat{P}_{s, \pi^*(s)}\|_1 \|\hat{V} - \hat{V}_{pe}^*\|_\infty \leq \frac{1}{N};$$

and (iv) follows from Lemma 6.4 and (6.30). Combining (6.37) and (6.38), we obtain

$$\begin{aligned} \rho^\top (V^* - \hat{V}) &= \sum_{s \in \mathcal{S}} \rho(s) (V^*(s) - \hat{V}(s)) \\ &\leq \sum_{s \in \mathcal{S}} \rho(s) \left\{ \gamma P_{s, \pi^*(s)} (V^* - \hat{V}) + 2b(s, \pi^*(s); \hat{V}) \right\} \\ &= \gamma \sum_{s \in \mathcal{S}} \rho(s) P_{s, \pi^*(s)} (V^* - \hat{V}) + 2 \sum_{s \in \mathcal{S}} \rho(s) b(s, \pi^*(s); \hat{V}) \\ &= \gamma \rho^\top P^* (V^* - \hat{V}) + 2\rho^\top b^*, \end{aligned} \quad (6.39)$$

where we introduce $P^* \in \mathbb{R}^{S \times S}$ and $b^* \in \mathbb{R}^{S \times 1}$ whose s -th row and s -th entry are given respectively by

$$[P^*]_{s, \cdot} := P_{s, \pi^*(s)} \quad \text{and} \quad b^*(s) := b(s, \pi^*(s); \hat{V}) \quad (6.40)$$

for all $s \in \mathcal{S}$. Since this relation holds for any state distribution $\rho \in \Delta(\mathcal{S})$, applying it recursively yields

$$\begin{aligned} \rho^\top (V^* - \hat{V}) &\leq (\gamma \rho^\top P^*) (V^* - \hat{V}) + 2\rho^\top b^* \\ &\leq \gamma (\gamma \rho^\top P^*) P^* (V^* - \hat{V}) + 2(\gamma \rho^\top P^*) b^* + 2\rho^\top b^* \\ &= \gamma^2 \rho^\top (P^*)^2 (V^* - \hat{V}) + 2\gamma \rho^\top P^* b^* + 2\rho^\top b^* \\ &\leq \dots \leq \lim_{i \rightarrow \infty} \gamma^i \rho^\top (P^*)^i (V^* - \hat{V}) + 2\rho^\top \left\{ \sum_{i=0}^{\infty} \gamma^i (P^*)^i \right\} b^* \\ &\stackrel{(i)}{=} 2\rho^\top \left\{ \sum_{i=0}^{\infty} \gamma^i (P^*)^i \right\} b^* = 2\rho^\top (I - \gamma P^*)^{-1} b^* \\ &\stackrel{(ii)}{=} \frac{2}{1 - \gamma} \langle d^*, b^* \rangle. \end{aligned} \quad (6.41)$$

Here (i) holds because $\lim_{i \rightarrow \infty} \gamma^i \rho^\top (P^*)^i (V^* - \hat{V}) = 0$ since $\gamma^i \rightarrow 0$ and $\|\rho^\top (P^*)^i\|_1 = 1$ for all $i \geq 0$; while (ii) follows from the definition of the discounted occupancy measure d^* in (2.9b), equivalently expressed in matrix-vector form as

$$(d^*)^\top = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \rho^\top (P^*)^t = (1 - \gamma) \rho^\top (I - \gamma P^*)^{-1}. \quad (6.42)$$

Combining (6.41) with (6.36) gives

$$\langle \rho, V^* - \hat{V} \rangle \leq \langle \rho, V^* - \hat{V} \rangle \leq \frac{2 \langle d^*, b^* \rangle}{1 - \gamma}. \quad (6.43)$$

Step 4: using concentrability to control $\langle d^\star, b^\star \rangle$. We find it helpful to partition the state space $\mathcal{S}$ into the following two disjoint subsets:

$$\mathcal{S}^{\text{small}} := \left\{ s \in \mathcal{S} \mid Nd^b(s, \pi^\star(s)) \leq 8 \log \frac{NS}{(1-\gamma)\delta} \right\}; \quad (6.44a)$$

$$\mathcal{S}^{\text{large}} := \left\{ s \in \mathcal{S} \mid Nd^b(s, \pi^\star(s)) > 8 \log \frac{NS}{(1-\gamma)\delta} \right\}. \quad (6.44b)$$

- First, consider any state $s \in \mathcal{S}^{\text{small}}$. By the definition of $C_\rho^\star$ in (6.3) and the definition of $\mathcal{S}^{\text{small}}$,

$$d^\star(s) \leq C_\rho^\star d^b(s, \pi^\star(s)) \leq \frac{8C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{N}.$$

Combining this bound with the fact that

$$b^\star(s) = b(s, \pi^\star(s); \hat{V}) \leq \frac{1}{1-\gamma} + \frac{5}{N}, \quad (6.45)$$

which is a direct consequence of the definition (6.19), we obtain

$$\begin{aligned} \sum_{s \in \mathcal{S}^{\text{small}}} d^\star(s) b^\star(s) &\leq \sum_{s \in \mathcal{S}^{\text{small}}} \left(\frac{8C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + d^\star(s) \frac{5}{N} \right) \\ &\leq \frac{8SC_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \frac{5}{N}. \end{aligned} \quad (6.46)$$

- Next, we consider any state $s \in \mathcal{S}^{\text{large}}$. By the definition (6.19) of $b(s, a; V)$, we have

$$\begin{aligned} b^\star(s) &\leq \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s, \pi^\star(s)}} \text{Var}_{\hat{P}_{s, \pi^\star(s)}}(\hat{V})} + \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s, \pi^\star(s)}} + \frac{5}{N} \\ &\leq \sqrt{\frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s, \pi^\star(s)}} \text{Var}_{P_{s, \pi^\star(s)}}(\hat{V})} + \frac{4c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s, \pi^\star(s)}}, \end{aligned} \quad (6.47)$$

provided that c_b is sufficiently large. The last inequality follows from Lemmas 6.4 and 6.6, which give

$$\text{Var}_{\hat{P}_{s, \pi^\star(s)}}(\hat{V}) \leq 2\text{Var}_{P_{s, \pi^\star(s)}}(\hat{V}) + \frac{41 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N_{s, \pi^\star(s)}},$$

together with the fact that $N \geq N_{s, a}$. Moreover,

$$\frac{1}{N_{s, \pi^\star(s)}} \stackrel{(i)}{\leq} \frac{12}{Nd^b(s, \pi^\star(s))} \stackrel{(ii)}{\leq} \frac{12C_\rho^\star}{Nd^\star(s)} \quad (6.48)$$

where (i) follows from (6.28) and the definition of $\mathcal{S}^{\text{large}}$, and (ii) follows from the definition of $C_\rho^\star$ in (6.3). Substituting this bound into (6.47) gives

$$b^\star(s) \leq \sqrt{\frac{24c_b C_\rho^\star \log \frac{N}{(1-\gamma)\delta}}{N}} \alpha_1(s) + \frac{48c_b C_\rho^\star \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N} \alpha_2(s) \quad (6.49)$$

where

$$\alpha_1(s) := \frac{1}{\sqrt{d^\star(s)}} \sqrt{\text{Var}_{P_{s, \pi^\star(s)}}(\hat{V})}, \quad \alpha_2(s) := \frac{1}{d^\star(s)}.$$

We now control the weighted sums of α_1 and α_2 . For α_1 , the Cauchy–Schwarz inequality and the identity $\sum_s d^\star(s) = 1$ imply

$$\begin{aligned} \sum_{s \in \mathcal{S}^{\text{large}}} d^\star(s) \alpha_1(s) &= \sum_{s \in \mathcal{S}^{\text{large}}} \sqrt{d^\star(s) \text{Var}_{P_{s, \pi^\star(s)}}(\hat{V})} \\ &\leq \sqrt{S} \cdot \sqrt{\sum_{s \in \mathcal{S}^{\text{large}}} d^\star(s) \text{Var}_{P_{s, \pi^\star(s)}}(\hat{V})}. \end{aligned} \quad (6.50)$$

To further bound the right-hand side of (6.50), we use the following lemma. The proof is provided in Section 6.7.5.

Lemma 6.7. We have

$$\sum_{s \in \mathcal{S}} d^\star(s) \text{Var}_{P_{s, \pi^\star(s)}}(\hat{V}) \leq \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)} \langle d^\star, b^\star \rangle. \quad (6.51)$$

Together, Lemma 6.7 and (6.50) yield

$$\sum_{s \in \mathcal{S}^{\text{large}}} d^\star(s) \alpha_1(s) \leq 2\sqrt{\frac{2S}{1-\gamma}} + 4\sqrt{\frac{S}{1-\gamma}} \langle d^\star, b^\star \rangle, \quad (6.52)$$

where we used $\gamma \geq 1/2$. In addition, the weighted sum of α_2 satisfies

$$\sum_{s \in \mathcal{S}^{\text{large}}} d^\star(s) \alpha_2(s) \leq S, \quad (6.53)$$

again using $\sum_s d^\star(s) = 1$. Combining (6.52) and (6.53) with (6.49), we obtain

$$\sum_{s \in \mathcal{S}^{\text{large}}} d^\star(s) b^\star(s) \leq \sqrt{\frac{384c_b S C_\rho^\star \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}} \langle d^\star, b^\star \rangle + 15\sqrt{\frac{c_b S C_\rho^\star \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}}, \quad (6.54)$$

where we used the consequence of (6.23) and $\varepsilon \leq 1/(1-\gamma)$ that

$$N \geq \frac{c_1 S C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{1-\gamma}, \quad (6.55)$$

provided that c_1 is large enough.

Combining the bounds for $\mathcal{S}^{\text{small}}$ and $\mathcal{S}^{\text{large}}$ in (6.46) and (6.54), respectively, gives

$$\begin{aligned} \langle d^\star, b^\star \rangle &= \sum_{s \in \mathcal{S}^{\text{large}}} d^\star(s) b^\star(s) + \sum_{s \in \mathcal{S}^{\text{small}}} d^\star(s) b^\star(s) \\ &\stackrel{(i)}{\leq} \sqrt{\frac{384c_b S C_\rho^\star \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}} \langle d^\star, b^\star \rangle + 25\sqrt{\frac{c_b S C_\rho^\star \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}} \\ &\stackrel{(ii)}{\leq} \frac{1}{2} \langle d^\star, b^\star \rangle + 192 \frac{c_b S C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + 25\sqrt{\frac{c_b S C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} \\ &\stackrel{(iii)}{\leq} \frac{1}{2} \langle d^\star, b^\star \rangle + 30\sqrt{\frac{c_b S C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}}. \end{aligned}$$

Here, (i) and (iii) follow from the sample size condition (6.55), provided that c_b and c_1 are sufficiently large. Step (i) also uses the fact that $C_\rho^\star \geq 1$ (see Remark 6.1). Step (ii) follows from the AM–GM inequality. Rearranging terms yields

$$\langle d^\star, b^\star \rangle \leq 60 \sqrt{\frac{c_b S C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}}.$$

Combining this with (6.43) gives

$$\langle \rho, V^\star - V^{\hat{\pi}} \rangle \leq \frac{2\langle d^\star, b^\star \rangle}{1-\gamma} \leq 120 \sqrt{\frac{c_b S C_\rho^\star \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3 N}}. \quad (6.56)$$

The desired claim follows from the sample size condition (6.23), after choosing c_1 sufficiently large.

6.4 Model-free offline RL

We have now demonstrated the effectiveness of the pessimism principle amid uncertainty in model-based offline RL in Section 6.3. Similar ideas can also be incorporated into model-free methods to address the challenges arising in the offline setting. Since such model-free approaches are generally statistically suboptimal, we provide only a brief introduction here, focusing primarily on the key ideas and underlying intuition rather than pursuing a treatment as detailed as that of the model-based counterpart. As before, for ease of presentation, our discussion focuses on the independent data model in (6.1); the extension to the Markovian data setting can be found in Yan *et al.* (2023).

6.4.1 Algorithm

Let us now describe how the pessimism principle can be incorporated into Q-learning. Similar to the model-based counterpart, a central idea is to incorporate lower confidence bounds into the iterative value estimation procedure so as to account for uncertainty caused by limited data coverage. Specifically, the algorithm maintains (pessimistic) estimates Q^{lcb} and V^{lcb} for the optimal Q-function and value function throughout its execution. At iteration t , the learner accesses the t -th sample (s_t, a_t, s'_t) from the offline dataset $\mathcal{D}$, and modifies the vanilla Q-learning update rule by inserting a penalty term b_n as follows:

$$Q^{\text{lcb}}(s_t, a_t) \leftarrow (1 - \eta_n) Q^{\text{lcb}}(s_t, a_t) + \eta_n \{r(s_t, a_t) + \gamma V^{\text{lcb}}(s'_t) - b_n\},$$

The corresponding value estimate is then updated according to

$$V^{\text{lcb}}(s_t) \leftarrow \max \left\{ V^{\text{lcb}}(s_t), \max_{a \in \mathcal{A}} Q^{\text{lcb}}(s_t, a) \right\},$$

where the additional maximization step is introduced to preserve the monotonicity of the value updates. Here, η_n and b_n stand for the corresponding learning rate and penalty term used at the current iteration, respectively, where $n := n(s_t, a_t)$ represents the number of visits to the state-action pair (s_t, a_t) up to and including time t .

Clearly, the choices of the learning rates and the penalty terms play a pivotal role in determining both the convergence behavior and the statistical efficiency of this pessimistic variant of Q-learning.

Algorithm 6.3: Pessimistic Q-learning (LCB-Q) for discounted infinite-horizon MDPs

```

1 input: offline dataset  $\mathcal{D}$ , discount factor  $\gamma$ , reward function  $r$ , learning rates  $\{\eta_t\}$ , max
   iteration number  $T$ .
2 initialize  $Q^{\text{lcb}}(s, a) \leftarrow 0$ ,  $V^{\text{lcb}}(s) \leftarrow 0$ ,  $n(s, a) \leftarrow 0$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .
3 for  $t = 1, 2, \dots, T$  do
    /* draw the  $t$ -th sample from  $\mathcal{D}$ . */
4     sample  $(s_t, a_t) \sim d^{\text{b}}$  and  $s'_t \sim P(\cdot | s_t, a_t)$ .
5     set  $n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$ , and  $n \leftarrow n(s_t, a_t)$ .
6     compute  $\eta_n$  and  $b_n$  according to (6.57).
    /* pessimistic Q-learning updates. */
7     update  $Q$ -estimate according to
        
$$Q^{\text{lcb}}(s_t, a_t) \leftarrow (1 - \eta_n) Q^{\text{lcb}}(s_t, a_t) + \eta_n \{r(s_t, a_t) + \gamma V^{\text{lcb}}(s'_t) - b_n\},$$

        and update the value function estimate according to
        
$$V^{\text{lcb}}(s_t) \leftarrow \max \{V^{\text{lcb}}(s_t), \max_{a \in \mathcal{A}} Q^{\text{lcb}}(s_t, a)\}.$$

8 output:  $\hat{\pi}$  s.t.  $\hat{\pi}(s) = \arg \max_{a \in \mathcal{A}} Q^{\text{lcb}}(s, a)$  for all  $s \in \mathcal{S}$ .

```

For concreteness, we focus on the following choices, which closely parallel those adopted for Q-UCB in the online setting (cf. Section 5.5). Specifically, given a confidence level $\delta \in (0, 1)$, by taking

$$H = \left\lceil \frac{4}{1 - \gamma} \log \frac{ST}{\delta} \right\rceil, \quad (6.57a)$$

we choose the learning rate

$$\eta_n = \frac{H + 1}{H + n}, \quad (6.57b)$$

together with the Hoeffding-style penalty term

$$b_n = C_{\text{b}} \sqrt{\frac{H \log(ST/\delta)}{n(1 - \gamma)^2}}, \quad (6.57c)$$

where $C_{\text{b}} > 0$ is some universal constant. Here, H can be viewed as the effective horizon inflated by a logarithmic factor.

The complete procedure, developed by Shi *et al.* (2022c) and Yan *et al.* (2023) under the name *pessimistic Q-learning* (or simply *LCB-Q*), is summarized in Algorithm 6.3.

6.4.2 Theoretical guarantees

The following theorem, due to Yan *et al.* (2023), provides the theoretical performance guarantee for the LCB-Q algorithm described in Algorithm 6.3.

Theorem 6.8 (Yan et al. (2022)). Consider any $0 < \delta < 1$. Let $\rho \in \Delta(\mathcal{S})$ be any given test distribution over the state space. With probability exceeding $1 - \delta$, the policy $\hat{\pi}$ returned by the

LCB-Q algorithm in Algorithm 6.3 with learning rates and penalty in (6.57) satisfies

$$V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \tilde{O}\left(\sqrt{\frac{SC_\rho^*}{T(1-\gamma)^5}}\right),$$

where C_ρ^* is the single-policy concentrability coefficient in Definition 6.1.

According to Theorem 6.8, for any target accuracy level $\varepsilon \in (0, \frac{1}{1-\gamma})$, the LCB-Q algorithm provably yields ε -accuracy with high probability provided that the total number of iterations T exceeds the order of (ignoring logarithmic factors)

$$\frac{SC_\rho^*}{(1-\gamma)^5 \varepsilon^2}.$$

Since each iteration of LCB-Q accesses only a single sample from the offline dataset, this quantity also corresponds to the total number of samples used by the algorithm. Moreover, if a Bernstein-style penalty is employed in place of the Hoeffding-style penalty, the sample complexity of LCB-Q can be further improved to

$$\tilde{O}\left(\frac{SC_\rho^*}{(1-\gamma)^4 \varepsilon^2}\right). \quad (6.58)$$

for sufficiently small ε . When the offline data is near-expert, i.e., $C_\rho^* \leq \tilde{O}(1)$, the resulting sample complexity of LCB-Q with Bernstein-style penalty reduces to

$$\tilde{O}\left(\frac{S}{(1-\gamma)^4 \varepsilon^2}\right),$$

which is, in general, substantially smaller than the sample complexity guarantees established for Q-learning under a generative model in Chapter 4. It is important to note, however, that this improvement pertains only to the suboptimality gap with respect to the initial distribution ρ , namely $V^*(\rho) - V^{\hat{\pi}}(\rho)$, rather than holding uniformly over all entries of $V^* - V^{\hat{\pi}}$.

Finally, we remark that the above statistical guarantees can be further improved through the use of variance reduction techniques. In particular, variance-reduced model-free algorithms can attain a sample complexity of

$$\tilde{O}\left(\frac{SC_\rho^*}{(1-\gamma)^3 \varepsilon^2}\right)$$

as $\varepsilon \rightarrow 0$, thereby achieving asymptotically optimal sample complexity. This improvement, however, comes at the expense of a large burn-in cost. We refer the interested reader to Shi *et al.* (2022c) and Yan *et al.* (2023) for details on such variance-reduced model-free offline algorithms.

6.4.3 Analysis: proof of Theorem 6.8

We adopt the following notation explicitly indexed by the iteration count t , detailed below.

- (s_t, a_t, s'_t) : the state-action-state transition tuple at the t -th iteration;
- $n_t(s, a)$: the visitation count for the state-action pair (s, a) at the t -th iteration;
- $Q_t(s, a)$ and $V_t(s)$ denote respectively $Q^{\text{lcb}}(s, a)$ and $V^{\text{lcb}}(s)$ at the end of the t -th iteration.

For the purpose of analysis, we introduce a sequence of deterministic policies $\{\pi_t\}_{t \geq 0}$, where each $\pi_t : \mathcal{S} \rightarrow \mathcal{A}$ is defined recursively by

$$\pi_t(s) = \begin{cases} \arg \max_{a \in \mathcal{A}} Q_t(s, a), & \text{if } s = s_t \text{ and } V_t(s) > V_{t-1}(s), \\ \pi_{t-1}(s), & \text{otherwise,} \end{cases} \quad (6.59)$$

with an arbitrary initialization π_0 . In other words, $\pi_t(s)$ is updated as the greedy policy w.r.t. the latest Q_t only for the visited state s_t when its value function is strictly increased. In the case of multiple maximizers, ties are broken arbitrarily but deterministically. We emphasize that the policy sequence $\{\pi_t\}$ is introduced solely as an analytical device and is not used by the algorithm itself; its purpose is to help certify the pessimism properties of the monotone value estimates.

Recall the notation and properties of the learning rate as established in Lemma 5.17, which are still valid here. It is easy to check, by induction, that

$$0 \leq V_t(s) \leq \frac{1}{1-\gamma}, \quad \forall s \in \mathcal{S}, \quad 0 \leq t \leq T. \quad (6.60)$$

We first establish the pessimistic property of the Q-estimate and value estimate, given by the following lemma. The proof is deferred to Chapter 6.8.1.

Lemma 6.9. With probability at least $1 - \delta$, for all $s \in \mathcal{S}$ and $t \in [T]$, the optimal-action Q-error obeys

$$Q^*(s, \pi^*(s)) - Q_t(s, \pi^*(s)) \leq \gamma \sum_{i=1}^n \eta_i^n P_{s, \pi^*(s)}(V^* - V_{k_i}) + \beta_n, \quad (6.61)$$

where k_i , $1 \leq i \leq n = n(s, \pi^*(s))$ are the time indices that $(s, \pi^*(s))$ is visited,

$$\beta_n := 3C_b \sqrt{\frac{H\iota}{n(1-\gamma)^2}}. \quad (6.62)$$

In addition, we have

$$Q_t(s, a) \leq Q^*(s, a), \quad V_t(s) \leq V^{\pi_t}(s) \leq V^*(s). \quad (6.63)$$

Step 1: error decomposition. By the monotonicity of V_t and property of $\hat{\pi}$,

$$V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \langle \rho, V^* - V_T \rangle \leq \frac{1}{T} \sum_{t=0}^{T-1} \langle \rho, V^* - V_t \rangle. \quad (6.64)$$

By Lemma 6.9, it follows that

$$\begin{aligned} \sum_{t=0}^{T-1} \langle \rho, V^* - V_t \rangle &\leq \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} \rho(s) \min \left\{ Q^*(s, \pi^*(s)) - V_t(s), \frac{1}{1-\gamma} \right\} \\ &\leq \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} \rho(s) \min \left\{ Q^*(s, \pi^*(s)) - Q_t(s, \pi^*(s)), \frac{1}{1-\gamma} \right\} \\ &\leq \underbrace{\gamma \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} \rho(s) \sum_{i=1}^{n_t(s, \pi^*(s))} \eta_i^{n_t(s, \pi^*(s))} P_{s, \pi^*(s)}(V^* - V_{k_i})}_{=:\zeta} \end{aligned}$$

$$+ \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} \rho(s) \min \left\{ \beta_{n_t(s, \pi^*(s))}, \frac{1}{1-\gamma} \right\}. \quad (6.65)$$

Here, the first inequality follows from $V^*(s) = Q^*(s, \pi^*(s))$, and $0 \leq V^* - V_t \leq \frac{1}{1-\gamma}$ due to (6.60); the second inequality follows from $V_t(s) \geq \max_a Q_t(s, a) \geq Q_t(s, \pi^*(s))$, and the last line uses Lemma 6.9. It then boils down to bounding the two terms.

We introduce the following quantities. Let

$$\Delta_t := V^* - V_t, \quad (6.66)$$

and, for $j \geq 0$, write

$$d_j(s, a) := [\rho(P_{\pi^*})^j](s) \mathbb{1}\{a = \pi^*(s)\}. \quad (6.67)$$

For every $j \geq 0$, set

$$\alpha_j := \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{t=0}^{T-1} \langle \rho(P_{\pi^*})^j, \Delta_t \rangle, \quad (6.68a)$$

$$\theta_j := \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} [\rho(P_{\pi^*})^j](s) \min \left\{ \beta_{n_t(s, \pi^*(s))}, \frac{1}{1-\gamma} \right\}, \quad (6.68b)$$

$$\xi_j := \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^{j+1} \langle \rho(P_{\pi^*})^{j+1}, \Delta_0 \rangle, \quad (6.68c)$$

$$\begin{aligned} \psi_j := & \gamma \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{t=0}^{T-1} \left\{ \sum_{s \in \mathcal{S}, a \in \mathcal{A}} d_j(s, a) \sum_{i=1}^{n_t(s, a)} \eta_i^{n_t(s, a)} P_{s, a} \Delta_{k_i(s, a)} \right. \\ & \left. - \left(1 + \frac{1}{H} \right) \frac{d_j(s_t, a_t)}{d^b(s_t, a_t)} \sum_{i=1}^{n_t(s_t, a_t)} \eta_i^{n_t(s_t, a_t)} P_{s_t, a_t} \Delta_{k_i(s_t, a_t)} \right\}, \end{aligned} \quad (6.68d)$$

$$\phi_j := \gamma \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \left(1 + \frac{1}{H} \right)^2 \sum_{t=0}^{T-1} \left\{ \frac{d_j(s_t, a_t)}{d^b(s_t, a_t)} P_{s_t, a_t} \Delta_t - \left(1 + \frac{1}{H} \right) \sum_{s \in \mathcal{S}, a \in \mathcal{A}} d_j(s, a) P_{s, a} \Delta_t \right\}. \quad (6.68e)$$

With this notation, (6.65) is

$$\alpha_0 \leq \zeta + \theta_0. \quad (6.69)$$

Step 2: recursion along the optimal dynamics. We first recurse on the concrete term ζ appearing in (6.65). Since $d_0(s, a) = \rho(s) \mathbb{1}\{a = \pi^*(s)\}$, the definition of ψ_0 gives

$$\begin{aligned} \zeta &= \gamma \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} d_0(s, a) \sum_{i=1}^{n_t(s, a)} \eta_i^{n_t(s, a)} P_{s, a} \Delta_{k_i(s, a)} \\ &= \gamma \left(1 + \frac{1}{H} \right) \sum_{t=0}^{T-1} \frac{d_0(s_t, a_t)}{d^b(s_t, a_t)} \sum_{i=1}^{n_t(s_t, a_t)} \eta_i^{n_t(s_t, a_t)} P_{s_t, a_t} \Delta_{k_i(s_t, a_t)} + \psi_0. \end{aligned} \quad (6.70)$$

For a fixed state-action pair, group the summands in (6.70) according to the visit that generated the term $\Delta_{k_i(s, a)}$. The learning-rate identity $\sum_{m=i}^{\infty} \eta_i^m = 1 + 1/H$ from Lemma 5.17, the monotonicity of V_t , and hence the monotonicity of Δ_t , imply

$$\sum_{t=0}^{T-1} \frac{d_0(s_t, a_t)}{d^b(s_t, a_t)} \sum_{i=1}^{n_t(s_t, a_t)} \eta_i^{n_t(s_t, a_t)} P_{s_t, a_t} \Delta_{k_i(s_t, a_t)}$$

$$\leq \left(1 + \frac{1}{H}\right) \sum_{t=0}^{T-1} \frac{d_0(s_t, a_t)}{d^b(s_t, a_t)} P_{s_t, a_t} \Delta_t + \left(1 + \frac{1}{H}\right) \langle \rho P_{\pi^*}, \Delta_0 \rangle. \quad (6.71)$$

Combining (6.70), (6.71), and the definition of ϕ_0 , we obtain

$$\begin{aligned} \zeta &\leq \gamma \left(1 + \frac{1}{H}\right)^2 \sum_{t=0}^{T-1} \frac{d_0(s_t, a_t)}{d^b(s_t, a_t)} P_{s_t, a_t} \Delta_t + \psi_0 + \gamma \left(1 + \frac{1}{H}\right)^2 \langle \rho P_{\pi^*}, \Delta_0 \rangle \\ &\leq \gamma \left(1 + \frac{1}{H}\right)^3 \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} d_0(s, a) P_{s, a} \Delta_t + \psi_0 + \phi_0 + \xi_0 \\ &= \alpha_1 + \psi_0 + \phi_0 + \xi_0. \end{aligned} \quad (6.72)$$

Together with (6.69), this yields

$$\alpha_0 \leq \alpha_1 + \xi_0 + \theta_0 + \psi_0 + \phi_0. \quad (6.73)$$

Repeating the same calculation with $\rho(P_{\pi^*})^j$ in place of ρ gives, for all $j \geq 0$,

$$\alpha_j \leq \alpha_{j+1} + \xi_j + \theta_j + \psi_j + \phi_j. \quad (6.74)$$

Iterating (6.74) leads to

$$\alpha_0 \leq \alpha + \xi + \theta + \psi + \phi, \quad (6.75)$$

where

$$\alpha := \limsup_{j \rightarrow \infty} \alpha_j, \quad \xi := \sum_{j=0}^{\infty} \xi_j, \quad \theta := \sum_{j=0}^{\infty} \theta_j, \quad \psi := \sum_{j=0}^{\infty} \psi_j, \quad \phi := \sum_{j=0}^{\infty} \phi_j. \quad (6.76)$$

Step 3: bounding the terms separately. We make note of a useful lemma before delving into bounding the terms separately.

Lemma 6.10. Recall that $H = \left\lceil \frac{4}{1-\gamma} \log \frac{ST}{\delta} \right\rceil$ for some $0 < \delta < 1$. For any vector with non-negative entries $V \in \mathbb{R}^d$, we have

$$\sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H}\right)^3 \right]^j \langle \rho(P_{\pi^*})^j, V \rangle \lesssim \frac{1}{1-\gamma} \langle d_{\rho}^*, V \rangle + \frac{\delta}{ST^4(1-\gamma)} \|V\|_{\infty}. \quad (6.77)$$

Bounding α . First, since $0 \leq \Delta_t \leq (1-\gamma)^{-1}$,

$$\alpha = \limsup_{j \rightarrow \infty} \left[\gamma \left(1 + \frac{1}{H}\right)^3 \right]^j \sum_{t=0}^{T-1} \langle \rho(P_{\pi^*})^j, \Delta_t \rangle \leq \frac{T}{1-\gamma} \limsup_{j \rightarrow \infty} \left[\gamma \left(1 + \frac{1}{H}\right)^3 \right]^j = 0. \quad (6.78)$$

Bounding ξ . Lemma 6.10 gives

$$\begin{aligned} \xi &= \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H}\right)^3 \right]^{j+1} \langle \rho(P_{\pi^*})^{j+1}, \Delta_0 \rangle \\ &\leq \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H}\right)^3 \right]^j \langle \rho(P_{\pi^*})^j, \Delta_0 \rangle \\ &\lesssim \frac{1}{1-\gamma} \langle d_{\rho}^*, \Delta_0 \rangle + \frac{\delta}{ST^4(1-\gamma)} \|\Delta_0\|_{\infty} \lesssim \frac{1}{(1-\gamma)^2}. \end{aligned} \quad (6.79)$$

Bounding θ . When it comes to θ , we can deduce that

$$\begin{aligned}
\theta &= \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{t=1}^T \sum_{s \in \mathcal{S}} [\rho(P_{\pi^*})^j](s) \min \left\{ \beta_{n_t(s, \pi^*(s))}, \frac{1}{1-\gamma} \right\} \\
&= \sum_{t=1}^T \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{s \in \mathcal{S}} [\rho(P_{\pi^*})^j](s) \min \left\{ \beta_{n_t(s, \pi^*(s))}, \frac{1}{1-\gamma} \right\} \\
&\stackrel{(i)}{\lesssim} \frac{1}{1-\gamma} \sum_{t=1}^T \sum_{s \in \mathcal{S}} d_{\rho}^*(s) \min \left\{ \beta_{n_t(s, \pi^*(s))}, \frac{1}{1-\gamma} \right\} + \frac{1}{ST^4(1-\gamma)} \frac{T}{1-\gamma} \\
&\lesssim \sum_{s \in \mathcal{S}} \sum_{t=1}^T d_{\rho}^*(s) \sqrt{\frac{H\iota}{n_t(s, \pi^*(s)) (1-\gamma)^4}} + \frac{1}{T^3(1-\gamma)^2} \\
&\stackrel{(ii)}{\lesssim} \sum_{s \in \mathcal{S}} \sum_{t=1}^T d_{\rho}^*(s) \sqrt{\frac{H\iota}{td^b(s, \pi^*(s)) (1-\gamma)^4}} + \frac{1}{T^3(1-\gamma)^2} \\
&\stackrel{(iii)}{\lesssim} \sum_{s \in \mathcal{S}} d_{\rho}^*(s, \pi^*(s)) \sqrt{\frac{HT\iota}{d^b(s, \pi^*(s)) (1-\gamma)^4}} + \frac{1}{T^3(1-\gamma)^2} \\
&\stackrel{(iv)}{\lesssim} \sqrt{\frac{C_{\rho}^* HT\iota}{(1-\gamma)^4}} \sum_{s \in \mathcal{S}} \sqrt{d_{\rho}^*(s, \pi^*(s))} \stackrel{(v)}{\lesssim} \sqrt{\frac{C_{\rho}^* ST\iota^2}{(1-\gamma)^5}}, \tag{6.80}
\end{aligned}$$

Here, (i) relies on (6.77); (ii) utilizes Li *et al.* (2022b, Lemma 8); (iii) follows from the fact that

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 1 + \int_1^T \frac{1}{\sqrt{x}} dx = 1 + 2(\sqrt{T} - 1) \leq 2\sqrt{T}; \tag{6.81}$$

(iv) uses Definition 6.1; and (v) invokes the Cauchy-Schwarz inequality and the fact that $\sum_s d_{\rho}^*(s, \pi^*(s)) = 1$.

Bound ψ and ϕ . It remains to bound the empirical-process terms ψ and ϕ . Define the weighted occupancy measure

$$\tilde{d}(s, a) := \gamma \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j d_j(s, a). \tag{6.82}$$

Summing (6.68d) over j yields

$$\psi = \sum_{t=0}^{T-1} \left\{ \mathbb{E}_{(s,a) \sim d^b} [f_t(s, a)] - \left(1 + \frac{1}{H} \right) f_t(s_t, a_t) \right\}, \tag{6.83}$$

where the same zero-ratio convention is used in the definition of f_t :

$$f_t(s, a) := \frac{\tilde{d}(s, a)}{d^b(s, a)} \sum_{i=1}^{n_t(s, a)} \eta_i^{n_t(s, a)} P_{s, a} \Delta_{k_i(s, a)}. \tag{6.84}$$

Similarly,

$$\phi = \left(1 + \frac{1}{H} \right)^2 \sum_{t=0}^{T-1} \left\{ g_t(s_t, a_t) - \left(1 + \frac{1}{H} \right) \mathbb{E}_{(s,a) \sim d^b} [g_t(s, a)] \right\}, \tag{6.85}$$

with the same convention,

$$g_t(s, a) := \frac{\tilde{d}(s, a)}{d^b(s, a)} P_{s,a} \Delta_t. \quad (6.86)$$

The functions f_t and g_t are predictable with respect to the history before the sample (s_t, a_t) is drawn, are supported only on actions $a = \pi^*(s)$, and are non-increasing in t for each fixed state-action pair. Moreover, Lemma 6.10 applied to indicator vectors gives

$$\tilde{d}(s, \pi^*(s)) \lesssim \frac{d_\rho^*(s)}{1 - \gamma} + \frac{\delta}{ST^4(1 - \gamma)}, \quad (6.87)$$

and therefore

$$\max\{f_t(s, \pi^*(s)), g_t(s, \pi^*(s))\} \lesssim \frac{d_\rho^*(s)}{(1 - \gamma)^2 d^b(s, \pi^*(s))} + \frac{\delta}{ST^4(1 - \gamma)^2 d^b(s, \pi^*(s))}. \quad (6.88)$$

By the argument in Yan *et al.* (2023), for any nonnegative predictable sequence $h_t(s, a)$ that is non-increasing in t and obeys the envelope in (6.88),

$$\begin{aligned} \sum_{t=0}^{T-1} \left\{ \mathbb{E}_{d^b} h_t - \left(1 + \frac{1}{H}\right) h_t(s_t, a_t) \right\} &\lesssim \frac{C^* H \iota}{(1 - \gamma)^2} + \iota \sum_{s \in \mathcal{S}} \frac{d_\rho^*(s)}{(1 - \gamma)^2 d^b(s, \pi^*(s))}, \\ \sum_{t=0}^{T-1} \left\{ h_t(s_t, a_t) - \left(1 + \frac{1}{H}\right) \mathbb{E}_{d^b} h_t \right\} &\lesssim \frac{C^* H \iota}{(1 - \gamma)^2} + \iota \sum_{s \in \mathcal{S}} \frac{d_\rho^*(s)}{(1 - \gamma)^2 d^b(s, \pi^*(s))}. \end{aligned} \quad (6.89)$$

Applying (6.89) to f_t and g_t , and using $H \asymp \iota/(1 - \gamma)$ and $d_\rho^*(s) \leq C^* d^b(s, \pi^*(s))$, yields

$$\psi + \phi \lesssim \frac{C^* \iota^2}{(1 - \gamma)^3} + \frac{C^* S \iota}{(1 - \gamma)^2}. \quad (6.90)$$

Step 4: putting everything together. Putting (6.78), (6.79), (6.80), and (6.90) together with (6.75) gives

$$\alpha_0 \lesssim \sqrt{\frac{C^* S T \iota^2}{(1 - \gamma)^5}} + \frac{C^* S \iota}{(1 - \gamma)^2} + \frac{C^* \iota^2}{(1 - \gamma)^3} + \frac{1}{(1 - \gamma)^2}. \quad (6.91)$$

In the nontrivial regime where the leading term in (6.91) dominates the displayed lower-order terms, we obtain

$$\alpha_0 \lesssim \sqrt{\frac{C^* S T \iota^2}{(1 - \gamma)^5}}. \quad (6.92)$$

In the complementary regime, the target bound in Theorem 6.8 is at least a constant multiple of the trivial bound $(1 - \gamma)^{-1}$. Thus, substituting (6.92) into (6.64) establishes the advertised bound in the theorem.

6.5 Minimax lower bounds

While we have previously highlighted the sample optimality of the VI-LCB algorithm, a matching lower bound is still needed in order to rigorously establish its minimax optimality. We present such a result below, which serves as a benchmark for assessing the optimality or suboptimality of the offline RL algorithms developed in this chapter and more broadly in the literature.

Theorem 6.11. For any $(\gamma, S, C_\rho^*, \varepsilon)$ obeying $\gamma \in [\frac{2}{3}, 1)$, $S > 12$, $C_\rho^* \geq 3$, and $\varepsilon \leq \frac{1}{672(1-\gamma)}$, one can construct a class of MDPs $\{\mathcal{M}_\theta\}$, an initial state distribution ρ , and a batch dataset with N independent samples and single-policy concentrability coefficient C_ρ^* such that

$$\inf_{\hat{\pi}} \max_{\theta} \mathbb{P}_{\mathcal{M}_\theta} (V^*(\rho) - V^{\hat{\pi}}(\rho) > \varepsilon) \geq \frac{1}{4},$$

provided that

$$N \leq \frac{c_2 S C_\rho^*}{(1-\gamma)^3 \varepsilon^2}$$

for some numerical constant $c_2 > 0$. Here, the infimum is over all estimators $\hat{\pi}$, while $\mathbb{P}_{\mathcal{M}_\theta}$ denotes the probability corresponding to the MDP $\mathcal{M}_\theta$.

In words, Theorem 6.11 asserts that no algorithm can reliably identify an ε -optimal policy unless the sample size is at least on the order of

$$\Omega \left(\frac{S C_\rho^*}{(1-\gamma)^3 \varepsilon^2} \right).$$

Taken together with Theorem 6.3, this lower bound confirms the minimax optimality of model-based offline RL algorithms in terms of sample complexity, up to logarithmic factors. The proof is deferred to Section 6.9.

6.6 Extension: finite-horizon nonstationary MDPs

We now provide a brief discussion of how the pessimism principle can be applied to finite-horizon nonstationary MDPs, covering both model-based and model-free approaches. Since the key ideas largely parallel those developed for the infinite-horizon setting, our treatment here is intentionally concise. We refer the interested reader to Li *et al.* (2024b) and Shi *et al.* (2022c) for more comprehensive treatments and detailed analyses of the model-based and model-free approaches, respectively.

6.6.1 Problem setting

We begin by describing the problem setup.

Offline dataset. Suppose that we are given access to a batch dataset $\mathcal{D}$, containing a collection of K independent sample trajectories generated by a behavior policy $\pi^b = \{\pi_h^b\}_{1 \leq h \leq H}$ starting from an initial state distribution ρ^b . More specifically, the k -th sample trajectory ($1 \leq k \leq K$) is a sequence

$$(s_1^k, a_1^k, s_2^k, a_2^k, \dots, s_H^k, a_H^k, s_{H+1}^k), \quad (6.93)$$

generated by executing the behavior policy π^b in the underlying MDP according to

$$s_1^k \sim \rho^b, \quad a_h^k \sim \pi_h^b(\cdot | s_h^k), \quad \text{and} \quad s_{h+1}^k \sim P_h(\cdot | s_h^k, a_h^k) \quad (6.94)$$

for every $1 \leq h \leq H$. In addition to the dataset $\mathcal{D}$, the learner is assumed to have full access to the reward function.

For notational convenience, we introduce the shorthand notation for the occupancy measures induced by the behavior policy π^b : $\forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$d_h^b(s) := d_h^{\pi^b}(s; \rho^b) \quad \text{and} \quad d_h^b(s, a) := d_h^{\pi^b}(s, a; \rho^b). \quad (6.95)$$

Note that the initial state distribution ρ^b underlying the offline data set does not necessarily coincide with the test distribution ρ used for performance evaluation.

Key parameter. Akin to Definition 6.1, we introduce below the notion of single-policy concentrability tailored to finite-horizon MDPs.

Definition 6.2 (Single-policy concentrability for finite-horizon MDPs). For any given initial state distribution $\rho \in \Delta(\mathcal{S})$, the single-policy concentrability coefficient underlying the batch dataset $\mathcal{D}$ is defined as

$$C_\rho^* := \max_{(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]} \frac{d_h^*(s, a; \rho)}{d_h^b(s, a)}, \quad (6.96)$$

where the convention $0/0 = 0$ is adopted.

Here, C_ρ^* quantifies the distribution shift between the target distribution and the distribution induced by the behavior policy through the maximum density ratio between their corresponding occupancy distributions. As before, this condition does not require uniform coverage of the entire state-action space; it often suffices for the dataset to adequately cover the regions reachable under the optimal policy π^* .

Goal. Given the offline dataset $\mathcal{D}$, the objective is to efficiently learn a policy $\hat{\pi}$ that attains near-optimal statistical performance. More specifically, for a prescribed test distribution $\rho \in \Delta(\mathcal{S})$ and a target accuracy level $\varepsilon \in (0, H]$, we aim to design a learning algorithm that guarantees, with high probability,

$$V_1^*(\rho) - V_1^{\hat{\pi}}(\rho) \leq \varepsilon,$$

while requiring as few samples from the dataset as possible.

6.6.2 Model-based algorithm

As in the discounted infinite-horizon setting, minimax-optimal sample complexity in the offline data setting (6.94) is achieved by model-based methods. In the sequel, we briefly describe a simple pessimistic variant of the model-based algorithm.

The VI-LCB algorithm for finite-horizon MDPs. The algorithm first constructs an empirical transition kernel $\hat{P}$ from the offline dataset $\mathcal{D}$. Based on this estimated model, the VI-LCB algorithm maintains estimates of the value functions $\{\hat{V}_h^{\text{lc}b}\}_{1 \leq h \leq H}$ and Q-functions $\{\hat{Q}_h^{\text{lc}b}\}_{1 \leq h \leq H}$, and proceeds backward from $h = H$ to $h = 1$ in the spirit of classical dynamic programming, with the terminal value $\hat{V}_{H+1}^{\text{lc}b} = 0$ (Xie *et al.*, 2021b; Jin *et al.*, 2021b). Specifically, the updates take the form:

$$Q_h^{\text{lc}b}(s, a) = \max \left\{ r_h(s, a) + \hat{P}_{s, a, h} V_{h+1}^{\text{lc}b} - b_h(s, a; V_{h+1}^{\text{lc}b}), 0 \right\}, \quad (6.97a)$$

where $\hat{P}_{s,a,h}$ denotes the empirical estimate of $P_{s,a,h}$ (cf. (2.29)),

$$V_{h+1}^{\text{lb}}(s) = \max_{a \in \mathcal{A}} Q_{h+1}^{\text{lb}}(s, a), \quad (6.97b)$$

and $b_h(s, a; V) \geq 0$ denotes some penalty function that is decreasing in $N_h(s, a)$, the number of times the tuple (s, a, h) is sampled. The output policy $\hat{\pi}$ is then chosen greedily with respect to the estimated Q-functions:

$$\forall (s, h) \in \mathcal{S} \times [H] : \quad \hat{\pi}_h(s) \in \arg \max_a Q_h^{\text{lb}}(s, a). \quad (6.98)$$

In essence, the VI-LCB algorithm performs value iteration on the empirical model $\hat{P}$, while incorporating lower-confidence penalty terms $b_h(s, a; V)$ into each Bellman update to enforce pessimism.

To more effectively capture the variance structure of the underlying MDP, it is, as before, advantageous to adopt a Bernstein-style penalty. Concretely, for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$b_h(s, a; V) = \tilde{\Theta} \left(\sqrt{\frac{\text{Var}_{\hat{P}_{s,a,h}}(V)}{N_h(s, a)}} + \frac{H}{N_h(s, a)} \right).$$

Here, $\text{Var}_{\hat{P}_{s,a,h}}(V)$ denotes the variance of the vector $V \in \mathbb{R}^S$ under the distribution $\hat{P}_{s,a,h}$.

It is worth noting that, in order to establish provable sample optimality, Li *et al.* (2024b) further developed a refined variant of the above model-based offline RL method, based on a two-fold sample splitting procedure that helps decouple complicated statistical dependency arising along each sample trajectory. We omit these technical details for brevity and refer interested readers to Li *et al.* (2024b) for in-depth discussion.

Theoretical guarantees. The following theorem establishes the sample complexity guarantee for a variant of the aforementioned VI-LCB algorithm for finite-horizon MDPs, paralleling the guarantee in Theorem 6.3 for discounted infinite-horizon MDPs.

Theorem 6.12 (Li *et al.* (2024b)). Consider any $\varepsilon \in (0, H]$ and any $0 < \delta < 1$. Let $\rho \in \Delta(\mathcal{S})$ be any given state distribution. Then, there exists a variant of the above VI-LCB algorithm such that, with probability at least $1 - \delta$, the policy $\hat{\pi}$ returned by the algorithm satisfies

$$V_1^*(\rho) - V_1^{\hat{\pi}}(\rho) \leq \varepsilon \quad (6.99)$$

as long as the total number of sample trajectories in $\mathcal{D}$ exceeds

$$K \geq \frac{c_k H^3 S C_\rho^* \log \frac{KH}{\delta}}{\varepsilon^2} \quad (6.100)$$

for some sufficiently large numerical constant $c_k > 0$. Here, C_ρ^* is introduced in Definition 6.2.

Given that each sample trajectory has length H , the total number of samples (namely, KH) required by model-based offline RL algorithms to achieve ε -accuracy can be as low as

$$\tilde{O} \left(\frac{H^4 S C_\rho^*}{\varepsilon^2} \right). \quad (6.101)$$

This sample complexity can be proven to be minimax-optimal up to logarithmic factors (Li *et al.*, 2024b). Importantly, the bound in (6.100) is valid for any $\varepsilon \in (0, H]$, indicating that the algorithm achieves sample-optimal performance without imposing any burn-in sample size requirement.

6.6.3 Model-free algorithm

Next, we briefly describe how the LCB-Q algorithm introduced in Section 6.4 can be adapted to the finite-horizon setting.

The LCB-Q algorithm for finite-horizon MDPs. For each trajectory $k = 1, \dots, K$, the LCB-Q algorithm updates a single entry in the Q-estimate at each step h using the observed transition tuple $(s_h^k, a_h^k, s_{h+1}^k)$. In particular, it employs a modified Q-learning update equipped with a pessimistic penalty term as follows:

$$Q_h^{\text{lb}}(s_h^k, a_h^k) \leftarrow (1 - \eta_n)Q_h^{\text{lb}}(s_h^k, a_h^k) + \eta_n \{r_h(s_h^k, a_h^k) + V_{h+1}^{\text{lb}}(s_{h+1}^k) - b_n\}.$$

Here, η_n denotes the learning rate, depending on the number of times n that the state-action pair (s_h^k, a_h^k) has been visited at step h , with a common choice being $\eta_n = \frac{H+1}{H+n}$. The penalty term $b_n > 0$ is chosen to measure the uncertainty of the corresponding Q-estimate and implement pessimism in the face of uncertainty; for instance, a Hoeffding-style penalty term takes the form

$$b_n = \tilde{\Theta} \left(\sqrt{\frac{H^3}{n}} \right). \quad (6.102)$$

The algorithm then outputs a policy estimate $\hat{\pi}$ obtained by acting greedily with respect to the Q-function estimate Q^{lb} .

The resulting algorithm is a *single-pass* procedure, requiring only one sweep through the offline dataset. It enjoys the following statistical guarantees.

Theorem 6.13 (Shi *et al.* (2022c)). Consider any $\delta \in (0, 1)$ and any $\varepsilon \in (0, H]$. Let $\rho \in \Delta(\mathcal{S})$ be any given state distribution. With probability at least $1 - \delta$, the policy $\hat{\pi}$ returned by the LCB-Q algorithm described above with the Hoeffding-style penalty (6.102) obeys

$$V_1^*(\rho) - V_1^{\hat{\pi}}(\rho) \leq \varepsilon, \quad (6.103)$$

with the proviso that the total number of sample trajectories exceeds

$$K \geq c_a \frac{H^5 S C_\rho^* \log^3 \frac{SAK}{\delta}}{\varepsilon^2} \quad (6.104)$$

for some sufficiently large numerical constant $c_a > 0$. Here, C_ρ^* is introduced in Definition 6.2.

Compared with its model-based counterpart, the LCB-Q algorithm equipped with a Hoeffding-style penalty may incur an additional factor of H^2 in sample complexity, rendering it statistically suboptimal. Comprehensive discussions of this pessimistic model-free approach can be found in Shi *et al.* (2022c).

6.7 Appendix A: proof of auxiliary lemmas in Section 6.3

6.7.1 Proof of Lemma 6.2

For any $Q, Q_1, Q_2 \in \mathbb{R}^{SA}$, write

$$V(s) := \max_a Q(s, a), \quad V_1(s) := \max_a Q_1(s, a), \quad V_2(s) := \max_a Q_2(s, a)$$

for all $s \in \mathcal{S}$. Unless otherwise noted, we assume that

$$Q(s, a), Q_1(s, a), Q_2(s, a) \in \left[0, \frac{1}{1-\gamma}\right] \quad \text{for all } (s, a) \in \mathcal{S} \times \mathcal{A}$$

throughout this subsection.

We first establish two useful properties of the operator $\hat{\mathcal{T}}_{\text{pe}}$. Define

$$\tilde{\mathcal{T}}_{\text{pe}}(Q)(s, a) := r(s, a) - b(s, a; V) + \gamma \hat{P}_{s,a} V,$$

so that

$$\hat{\mathcal{T}}_{\text{pe}}(Q)(s, a) = \max \{ \tilde{\mathcal{T}}_{\text{pe}}(Q)(s, a), 0 \}.$$

Proof of monotonicity. We begin by showing that $\tilde{\mathcal{T}}_{\text{pe}}$ is monotone on the relevant domain:

$$Q \leq \tilde{Q} \quad \implies \quad \tilde{\mathcal{T}}_{\text{pe}}(Q) \leq \tilde{\mathcal{T}}_{\text{pe}}(\tilde{Q}). \quad (6.105)$$

Since $x \mapsto \max\{x, 0\}$ is nondecreasing, this also implies the monotonicity of $\hat{\mathcal{T}}_{\text{pe}}$.

To prove (6.105), fix $(s, a) \in \mathcal{S} \times \mathcal{A}$ and regard $\tilde{\mathcal{T}}_{\text{pe}}(Q)(s, a)$ as a function of V . It suffices to show that this function is coordinate-wise nondecreasing in V . Recall the definition of $b(s, a; V)$ in (6.19), namely

$$b(s, a; V) := \min \left\{ \max\{b_1, b_2\}, \frac{1}{1-\gamma} \right\} + \frac{5}{N}$$

where

$$b_1 := \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(V), \quad b_2 := \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}.$$

For any (s, a) , there are three possible regimes:

- Case 1: the outer clipping is active, namely $\max\{b_1, b_2\} \geq \frac{1}{1-\gamma}$. In this case, the penalty term

$$b(s, a; V) = \frac{1}{1-\gamma} + \frac{5}{N}$$

is independent of V , and hence, for any $s' \in \mathcal{S}$,

$$\frac{\partial \tilde{\mathcal{T}}_{\text{pe}}(Q)(s, a)}{\partial V(s')} = \gamma \hat{P}(s' | s, a) \geq 0.$$

- Case 2: $b_1 < b_2 < \frac{1}{1-\gamma}$. In this case, we have

$$b(s, a; V) = \frac{c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} + \frac{5}{N},$$

which is again independent of V . Thus the same calculation produces the same result as Case 1.

- Case 3: $b_2 \leq b_1 < \frac{1}{1-\gamma}$. As an immediate consequence of $b_2 \leq b_1$,

$$\text{Var}_{\hat{P}_{s,a}}(V) > \frac{4c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N_{s,a}}. \quad (6.106)$$

In this case, the penalty term depends on V :

$$b(s, a; V) = \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(V) + \frac{5}{N}.$$

For any $s' \in \mathcal{S}$, differentiating the penalty term gives

$$\begin{aligned} \frac{\partial b(s, a; V)}{\partial V(s')} &= \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \frac{\hat{P}(s' | s, a)(V(s') - \hat{P}_{s,a}V)}{\sqrt{\text{Var}_{\hat{P}_{s,a}}(V)}} \\ &\stackrel{(i)}{\leq} \frac{1-\gamma}{2} \hat{P}(s' | s, a)(V(s') - \hat{P}_{s,a}V)_+ \\ &\stackrel{(ii)}{\leq} \frac{1}{2} \hat{P}(s' | s, a) \stackrel{(iii)}{\leq} \gamma \hat{P}(s' | s, a). \end{aligned}$$

Here (i) follows from (6.106), (ii) uses $(V(s') - \hat{P}_{s,a}V)_+ \leq \frac{1}{1-\gamma}$, and (iii) follows from $\gamma \geq 1/2$. Consequently,

$$\frac{\partial \tilde{\mathcal{T}}_{\text{pe}}(Q)(s, a)}{\partial V(s')} = \gamma \hat{P}(s' | s, a) - \frac{\partial b(s, a; V)}{\partial V(s')} \geq 0.$$

Combining the three cases, we conclude that $\tilde{\mathcal{T}}_{\text{pe}}(Q)(s, a)$ is coordinate-wise nondecreasing in V at all differentiability points. Since $\tilde{\mathcal{T}}_{\text{pe}}$ is continuous and piecewise differentiable, the monotonicity extends to all points. Finally, $Q \leq \tilde{Q}$ implies

$$V(s) = \max_a Q(s, a) \leq \max_a \tilde{Q}(s, a) = \tilde{V}(s)$$

for any $s \in \mathcal{S}$, which proves (6.105).

Proof of γ -contraction. We next prove that $\hat{\mathcal{T}}_{\text{pe}}$ is a γ -contraction. Consider any $Q_1, Q_2 \in \mathbb{R}^{SA}$, and set $c = \|Q_1 - Q_2\|_\infty$. Then we have

$$\begin{aligned} \tilde{\mathcal{T}}_{\text{pe}}(Q_1) - \tilde{\mathcal{T}}_{\text{pe}}(Q_2) &\stackrel{(i)}{\leq} \tilde{\mathcal{T}}_{\text{pe}}(Q_2 + c1_{SA}) - \tilde{\mathcal{T}}_{\text{pe}}(Q_2) \\ &\stackrel{(ii)}{=} \gamma c \hat{P}1_S = \gamma \|Q_1 - Q_2\|_\infty 1_{SA}. \end{aligned}$$

Here (i) follows from $Q_1 \leq Q_2 + c1_{SA}$ and the monotonicity of $\tilde{\mathcal{T}}_{\text{pe}}$, while (ii) holds since the penalty $b(s, a; V)$ is invariant under constant shifts of the value function V , as $\text{Var}_{\hat{P}_{s,a}}(V) = \text{Var}_{\hat{P}_{s,a}}(V + c \cdot 1_S)$. Swapping the roles of Q_1 and Q_2 yields

$$\|\tilde{\mathcal{T}}_{\text{pe}}(Q_1) - \tilde{\mathcal{T}}_{\text{pe}}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty.$$

Since $x \mapsto \max\{x, 0\}$ is 1-Lipschitz, we also have

$$\|\hat{\mathcal{T}}_{\text{pe}}(Q_1) - \hat{\mathcal{T}}_{\text{pe}}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty. \quad (6.107)$$

Fixed point of $\hat{\mathcal{T}}_{\text{pe}}$. We now discuss the existence and uniqueness of the fixed point of $\hat{\mathcal{T}}_{\text{pe}}$. For any $0 \leq Q \leq \frac{1}{1-\gamma}1_{SA}$, the definition of $\hat{\mathcal{T}}_{\text{pe}}$ ensures that $0 \leq \hat{\mathcal{T}}_{\text{pe}}(Q) \leq \frac{1}{1-\gamma}1_{SA}$. Define the following sequence recursively:

$$Q^{(0)} = 0 \quad \text{and} \quad Q^{(t+1)} = \hat{\mathcal{T}}_{\text{pe}}(Q^{(t)}) \quad t \geq 0.$$

By the contraction property (6.107), the standard proof for the Banach fixed-point theorem (e.g., Agarwal *et al.* (2001, Theorem 1)) guarantees that $Q^{(t)}$ converges to some point $Q^{(\infty)}$ as $t \rightarrow \infty$. Clearly, $Q^{(\infty)}$ is a fixed point of $\hat{\mathcal{T}}_{\text{pe}}$ obeying $0 \leq Q^{(\infty)} \leq \frac{1}{1-\gamma} 1_{SA}$.

Moreover, suppose that there exists another point $\tilde{Q}$ obeying $\tilde{Q} = \hat{\mathcal{T}}_{\text{pe}}(\tilde{Q})$, which clearly satisfies $\tilde{Q} \geq 0$. If $\|\tilde{Q}\|_\infty > \frac{1}{1-\gamma}$, then

$$\begin{aligned} \|\tilde{Q}\|_\infty &= \|\hat{\mathcal{T}}_{\text{pe}}(\tilde{Q})\|_\infty \leq \|r\|_\infty + \gamma \|\hat{P}\|_1 \|\tilde{Q}\|_\infty \leq 1 + \gamma \|\tilde{Q}\|_\infty \\ &< (1 - \gamma) \|\tilde{Q}\|_\infty + \gamma \|\tilde{Q}\|_\infty = \|\tilde{Q}\|_\infty, \end{aligned}$$

resulting in contradiction. Hence every fixed point satisfies $0 \leq \tilde{Q} \leq \frac{1}{1-\gamma} 1_{SA}$. The γ -contraction property (6.107) implies that

$$\|\tilde{Q} - Q^{(\infty)}\|_\infty = \|\hat{\mathcal{T}}_{\text{pe}}(\tilde{Q}) - \hat{\mathcal{T}}_{\text{pe}}(Q^{(\infty)})\|_\infty \leq \gamma \|\tilde{Q} - Q^{(\infty)}\|_\infty.$$

Given that $\gamma < 1$, this immediately implies that $\tilde{Q} = Q^{(\infty)}$, confirming the uniqueness of the fixed point. Denote this unique fixed point by $\hat{Q}_{\text{pe}}^*$.

We finally prove the desired claims for the iterates. Since $Q_0 = 0 \leq \hat{Q}_{\text{pe}}^*$, the monotonicity of $\hat{\mathcal{T}}_{\text{pe}}$ gives

$$Q_1 = \hat{\mathcal{T}}_{\text{pe}}(Q_0) \leq \hat{\mathcal{T}}_{\text{pe}}(\hat{Q}_{\text{pe}}^*) = \hat{Q}_{\text{pe}}^*.$$

Repeating the same argument inductively yields

$$Q_\tau \leq \hat{Q}_{\text{pe}}^* \quad \text{for all } \tau \geq 0.$$

Furthermore, applying the contraction property (6.107) recursively gives

$$\begin{aligned} \|Q_\tau - \hat{Q}_{\text{pe}}^*\|_\infty &= \|\hat{\mathcal{T}}_{\text{pe}}(Q_{\tau-1}) - \hat{\mathcal{T}}_{\text{pe}}(\hat{Q}_{\text{pe}}^*)\|_\infty \leq \gamma \|Q_{\tau-1} - \hat{Q}_{\text{pe}}^*\|_\infty \\ &\leq \cdots \leq \gamma^\tau \|Q_0 - \hat{Q}_{\text{pe}}^*\|_\infty \leq \frac{\gamma^\tau}{1-\gamma}, \end{aligned} \tag{6.108}$$

where the last inequality follows from $Q_0 = 0$ and $\|\hat{Q}_{\text{pe}}^*\|_\infty \leq 1/(1-\gamma)$.

Finally, the other claim (6.21) follows immediately by taking the right-hand side of (6.108) to be no larger than $1/N$. This completes the proof.

6.7.2 Proof of Lemma 6.4

Starting from (6.25), it is straightforward to achieve (6.26a):

$$\|\hat{V} - \hat{V}_{\text{pe}}^*\|_\infty = \max_s \left| \max_a \hat{Q}(s, a) - \max_a \hat{Q}_{\text{pe}}^*(s, a) \right| \leq \|\hat{Q} - \hat{Q}_{\text{pe}}^*\|_\infty \leq \frac{1}{N}.$$

Given the proximity of $\hat{V}$ and $\hat{V}_{\text{pe}}^*$, we can use Lemma 2.2 to bound the difference of the corresponding variance terms as follows:

$$|\text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*) - \text{Var}_{\hat{P}_{s,a}}(\hat{V})| \leq \frac{4}{1-\gamma} \|\hat{V} - \hat{V}_{\text{pe}}^*\|_\infty \leq \frac{4}{(1-\gamma)N}. \tag{6.109}$$

By examining the definition of $b(s, a; V)$ in (6.19), we know that if at least one of the following variance terms is not too small:

$$\max \left\{ \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*), \text{Var}_{\hat{P}_{s,a}}(\hat{V}) \right\} \geq \frac{4c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N_{s,a}}, \tag{6.110}$$

then we have

$$\begin{aligned}
|b(s, a; \hat{V}_{\text{pe}}^*) - b(s, a; \hat{V})| &\leq \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \left| \sqrt{\text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*)} - \sqrt{\text{Var}_{\hat{P}_{s,a}}(\hat{V})} \right| \\
&= \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \frac{|\text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*) - \text{Var}_{\hat{P}_{s,a}}(\hat{V})|}{\sqrt{\text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*)} + \sqrt{\text{Var}_{\hat{P}_{s,a}}(\hat{V})}} \\
&\stackrel{(i)}{\leq} \frac{1-\gamma}{2} |\text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*) - \text{Var}_{\hat{P}_{s,a}}(\hat{V})| \stackrel{(ii)}{\leq} \frac{3}{N},
\end{aligned}$$

where (i) results from (6.110), and (ii) holds due to (6.109). On the other hand, if (6.110) is not satisfied, then one clearly has $b(s, a; \hat{V}_{\text{pe}}^*) = b(s, a; \hat{V})$. In conclusion, in all cases (6.26b) holds true.

6.7.3 Proof of Lemma 6.5

For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, if

$$Nd^b(s, a) < 8 \log \frac{SN}{\delta},$$

then (6.27) holds trivially. It therefore suffices to consider the collection of state-action pairs

$$\mathcal{N}_{\text{large}} := \left\{ (s, a) \mid d^b(s, a) \geq \frac{8 \log \frac{SN}{\delta}}{N} \right\}. \quad (6.111)$$

We first bound the cardinality of $\mathcal{N}_{\text{large}}$. Since

$$|\mathcal{N}_{\text{large}}| \cdot \frac{8 \log \frac{SN}{\delta}}{N} \leq \sum_{(s,a) \in \mathcal{N}_{\text{large}}} d^b(s, a) \leq \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d^b(s, a) \leq 1,$$

we have the crude bound

$$|\mathcal{N}_{\text{large}}| \leq \frac{N}{8 \log \frac{SN}{\delta}} \leq \frac{N}{8}. \quad (6.112)$$

Now fix any $(s, a) \in \mathcal{N}_{\text{large}}$. Since $N_{s,a}$ can be viewed as the sum of N independent Bernoulli random variables, each with mean $d^b(s, a)$, Bernstein's inequality gives

$$\mathbb{P} \left\{ |N_{s,a} - Nd^b(s, a)| \geq \tau \right\} \leq 2 \exp \left(-\frac{\tau^2/2}{v_{s,a} + \tau/3} \right)$$

for any $\tau \geq 0$, where

$$v_{s,a} := N \text{Var} \left(\mathbb{1} \{ (s_i, a_i) = (s, a) \} \right) \leq Nd^b(s, a).$$

A standard rearrangement of the Bernstein bound implies that, with probability at least $1 - \delta$,

$$\begin{aligned}
|N_{s,a} - Nd^b(s, a)| &\leq \sqrt{4v_{s,a} \log \frac{2}{\delta}} + \frac{2}{3} \log \frac{2}{\delta} \\
&\leq \sqrt{4Nd^b(s, a) \log \frac{2}{\delta}} + \log \frac{2}{\delta}.
\end{aligned}$$

Combining this bound with a union bound over $(s, a) \in \mathcal{N}_{\text{large}}$, and using (6.112), we obtain that, with probability at least $1 - \delta$,

$$|N_{s,a} - Nd^b(s, a)| \leq \sqrt{4Nd^b(s, a) \log \frac{N}{\delta}} + \log \frac{N}{\delta} \quad (6.113)$$

holds simultaneously for all $(s, a) \in \mathcal{N}_{\text{large}}$. Finally, for any $(s, a) \in \mathcal{N}_{\text{large}}$, we have $Nd^b(s, a) \geq 8 \log \frac{SN}{\delta}$ by definition. Hence (6.113) implies

$$N_{s,a} \geq Nd^b(s, a) - \left(\sqrt{4Nd^b(s, a) \log \frac{N}{\delta}} + \log \frac{N}{\delta} \right) \geq \frac{Nd^b(s, a)}{12}.$$

Thus (6.27) holds for every $(s, a) \in \mathcal{N}_{\text{large}}$ as well. This completes the proof.

6.7.4 Proof of Lemma 6.6

If $N_{s,a} = 0$, then the desired inequalities hold trivially. It therefore suffices to consider the case where $N_{s,a} > 0$. We begin by recording a key Bernstein-style result, whose proof is deferred to Section 6.7.6.

Lemma 6.14. Consider any fixed pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ with $N_{s,a} > 0$. Let $V \in \mathbb{R}^S$ be any vector independent of $\hat{P}_{s,a}$ satisfying $\|V\|_\infty \leq \frac{1}{1-\gamma}$. With probability at least $1 - 4\delta$, one has

$$|(\hat{P}_{s,a} - P_{s,a})V| \leq \sqrt{\frac{48\text{Var}_{\hat{P}_{s,a}}(V) \log \frac{N}{\delta}}{N_{s,a}}} + \frac{48 \log \frac{N}{\delta}}{(1-\gamma)N_{s,a}} \quad (6.114a)$$

$$\text{Var}_{\hat{P}_{s,a}}(V) \leq 2\text{Var}_{P_{s,a}}(V) + \frac{5 \log \frac{N}{\delta}}{3(1-\gamma)^2 N_{s,a}}. \quad (6.114b)$$

Equipped with this result, we now proceed to the proof of Lemma 6.6. The proof relies on a leave-one-out decoupling argument introduced in Section 3.3.4 and is organized into the following steps.

Step 1: constructing auxiliary state-absorbing MDPs. For each state $s \in \mathcal{S}$ and each scalar $u \geq 0$, we construct an auxiliary MDP $\hat{\mathcal{M}}^{s,u}$ that coincides with the empirical MDP $\hat{\mathcal{M}}$ except at state s . More precisely, the transition kernel of $\hat{\mathcal{M}}^{s,u}$, denoted by $P^{s,u}$, is defined as

$$\begin{aligned} P^{s,u}(\tilde{s} | s, a) &= \mathbf{1}(\tilde{s} = s) && \text{for all } (\tilde{s}, a) \in \mathcal{S} \times \mathcal{A}, \\ P^{s,u}(\cdot | s', a) &= \hat{P}(\cdot | s', a) && \text{for all } (s', a) \in \mathcal{S} \times \mathcal{A} \text{ and } s' \neq s; \end{aligned}$$

and the reward function of $\hat{\mathcal{M}}^{s,u}$, denoted by $r^{s,u}$, is defined as

$$\begin{aligned} r^{s,u}(s, a) &= u && \text{for all } a \in \mathcal{A}, \\ r^{s,u}(s', a) &= r(s', a) && \text{for all } (s', a) \in \mathcal{S} \times \mathcal{A} \text{ and } s' \neq s. \end{aligned}$$

In words, $\hat{\mathcal{M}}^{s,u}$ is obtained from the empirical MDP by removing the randomness in the empirical transition vector $\hat{P}_{s,a}$ for all $a \in \mathcal{A}$ associated with state s , and making s absorbing. We further define the pessimistic Bellman operator $\hat{\mathcal{T}}_{\text{pe}}^{s,u}$ associated with the auxiliary MDP $\hat{\mathcal{M}}^{s,u}$ as

$$\hat{\mathcal{T}}_{\text{pe}}^{s,u}(Q)(s, a) := \max \left\{ r^{s,u}(s, a) + \gamma P_{s,a}^{s,u} V - b^{s,u}(s, a; V), 0 \right\} \quad (6.115)$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, where the penalty term is given by

$$b^{s,u}(s, a; V) := \min \left\{ \max \left\{ \sqrt{\frac{c_b \log \frac{N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{P^{s,u}(\cdot | s, a)}(V), \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} \right\}, \frac{1}{1-\gamma} \right\} + \frac{5}{N}.$$

Step 2: bridging empirical and auxiliary MDPs. Taking

$$u^* = (1 - \gamma) \hat{V}_{\text{pe}}^*(s) + \min \left\{ \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma) \max_a N_{s,a}}, \frac{1}{1-\gamma} \right\} + \frac{5}{N},$$

we claim that there exists a fixed point $\hat{Q}_{s,u^*}^*$ of $\hat{\mathcal{T}}_{\text{pe}}^{s,u^*}$ whose associated value function $\hat{V}_{s,u^*}^*$ coincides with $\hat{V}_{\text{pe}}^*$. It suffices to verify the following two properties.

- Consider any $a \in \mathcal{A}$. Since $P^{s,u}(\cdot | s, a)$ has only one nonzero entry, which equals 1, we have $\text{Var}_{P^{s,u}(\cdot | s, a)}(V) = 0$ for any V and any u . Therefore,

$$b^{s,u}(s, a; V) = \min \left\{ \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}, \frac{1}{1-\gamma} \right\} + \frac{5}{N}. \quad (6.116)$$

Consequently, at state s , we have

$$\begin{aligned} & \max_a \left\{ r^{s,u^*}(s, a) - b^{s,u^*}(s, a; \hat{V}_{\text{pe}}^*) + \gamma \langle P^{s,u^*}(\cdot | s, a), \hat{V}_{\text{pe}}^* \rangle \right\} \\ &= \max_a \left\{ u^* - b^{s,u^*}(s, a; \hat{V}_{\text{pe}}^*) + \gamma \hat{V}_{\text{pe}}^*(s) \right\} \\ &= u^* - \min_a b^{s,u^*}(s, a; \hat{V}_{\text{pe}}^*) + \gamma \hat{V}_{\text{pe}}^*(s) \\ &= (1 - \gamma) \hat{V}_{\text{pe}}^*(s) + \gamma \hat{V}_{\text{pe}}^*(s) = \hat{V}_{\text{pe}}^*(s), \end{aligned} \quad (6.117)$$

where the third identity follows from the choice of u^* together with (6.116).

- Next, consider any $s' \neq s$ and any $a \in \mathcal{A}$. By construction of the auxiliary MDP, we have

$$\begin{aligned} & \max \left\{ r^{s,u^*}(s', a) - b^{s,u^*}(s', a; \hat{V}_{\text{pe}}^*) + \gamma \langle P^{s,u^*}(\cdot | s', a), \hat{V}_{\text{pe}}^* \rangle, 0 \right\} \\ &= \max \left\{ r(s', a) - b(s', a; \hat{V}_{\text{pe}}^*) + \gamma \langle \hat{P}(\cdot | s', a), \hat{V}_{\text{pe}}^* \rangle, 0 \right\} = \hat{Q}_{\text{pe}}^*(s', a), \end{aligned} \quad (6.118)$$

where the last equality uses the fact that $\hat{Q}_{\text{pe}}^*$ is a fixed point of $\hat{\mathcal{T}}_{\text{pe}}$.

Combining (6.117) and (6.118), we see that $\hat{Q}_{s,u^*}^* = \hat{\mathcal{T}}_{\text{pe}}^{s,u^*}(\hat{Q}_{s,u^*}^*)$ by setting

$$\begin{aligned} \max_{a \in \mathcal{A}} \hat{Q}_{s,u^*}^*(s, a) &= \hat{V}_{\text{pe}}^*(s), \\ \hat{Q}_{s,u^*}^*(s', a) &= \hat{Q}_{\text{pe}}^*(s', a) \quad \text{for all } s' \neq s \text{ and } a \in \mathcal{A}. \end{aligned}$$

This establishes the existence of a fixed point of $\hat{\mathcal{T}}_{\text{pe}}^{s,u^*}$ whose associated value function coincides with $\hat{V}_{\text{pe}}^*$.

Step 3: building an ε -net. Consider any $(s, a) \in \mathcal{S} \times \mathcal{A}$ with $N_{s,a} > 0$. Construct the set $\mathcal{U}_{\text{cover}}$ as

$$\mathcal{U}_{\text{cover}} := \left\{ \frac{i}{N} \mid 1 \leq i \leq Nu_{\max} \right\}, \quad (6.119)$$

where

$$u_{\max} = \min \left\{ \frac{2c_b \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}, \frac{1}{1-\gamma} \right\} + \frac{5}{N} + 1.$$

This set can be viewed as an ε -net (Vershynin, 2018) of the interval $[0, u_{\max}] \subseteq [0, \frac{2}{1-\gamma}]$ with $\varepsilon = 1/N$. For each $u \in \mathcal{U}_{\text{cover}}$, construct the auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ as in Step 1. Repeating the argument in the proof of Lemma 6.2, one can show that there exists a unique fixed point $\widehat{Q}_{s,u}^*$ associated with $\widehat{\mathcal{M}}^{s,u}$, which also satisfies $0 \leq \widehat{Q}_{s,u}^* \leq \frac{1}{1-\gamma} \cdot 1_{SA}$. In what follows, we denote by $\widehat{V}_{s,u}^*$ the value function induced by $\widehat{Q}_{s,u}^*$.

Since $\widehat{\mathcal{M}}^{s,u}$ is, by construction, statistically independent of $\widehat{P}_{s,a}$ for any $u \in \mathcal{U}_{\text{cover}}$, we can apply Lemma 6.14 with a union bound over all $u \in \mathcal{U}_{\text{cover}}$ to show that, with probability at least $1 - \delta$,

$$\begin{aligned} |(\widehat{P}_{s,a} - P_{s,a})\widehat{V}_{s,u}^*| &\leq \sqrt{\frac{48 \log \frac{8N^2}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\widehat{P}_{s,a}}(\widehat{V}_{s,u}^*) + \frac{48 \log \frac{8N^2}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}, \\ \text{Var}_{\widehat{P}_{s,a}}(\widehat{V}_{s,u}^*) &\leq 2\text{Var}_{P_{s,a}}(\widehat{V}_{s,u}^*) + \frac{5 \log \frac{8N^2}{(1-\gamma)\delta}}{3(1-\gamma)^2 N_{s,a}} \end{aligned}$$

hold uniformly for all $u \in \mathcal{U}_{\text{cover}}$. Moreover, the number of state-action pairs (s, a) with $N_{s,a} > 0$ is at most N . Therefore, taking union bound over all such pairs yields that, with probability at least $1 - \delta$,

$$|(\widehat{P}_{s,a} - P_{s,a})\widehat{V}_{s,u}^*| \leq \sqrt{\frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\widehat{P}_{s,a}}(\widehat{V}_{s,u}^*) + \frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}, \quad (6.120a)$$

$$\text{Var}_{\widehat{P}_{s,a}}(\widehat{V}_{s,u}^*) \leq 2\text{Var}_{P_{s,a}}(\widehat{V}_{s,u}^*) + \frac{5 \log \frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N_{s,a}} \quad (6.120b)$$

hold uniformly for all $(s, a, u) \in \mathcal{S} \times \mathcal{A} \times \mathcal{U}_{\text{cover}}$ satisfying $N_{s,a} > 0$.

Step 4: a covering argument. In this step, we work on the high-probability event (6.120). Since $\widehat{V}_{\text{pe}}^*$ satisfies $0 \leq \widehat{V}_{\text{pe}}^*(s) \leq \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, there exists some $u_0 \in \mathcal{U}_{\text{cover}}$ such that $|u_0 - u^*| \leq 1/N$. By the construction of the MDP $\widehat{\mathcal{M}}^{s,u}$ and the operator in (6.115), we have

$$\|\widehat{\mathcal{T}}_{\text{pe}}^{s,u_0}(Q) - \widehat{\mathcal{T}}_{\text{pe}}^{s,u^*}(Q)\|_{\infty} \leq |u_0 - u^*| \leq \frac{1}{N}$$

for any $Q \in \mathbb{R}^{SA}$. Consequently, the γ -contraction property of the operator gives

$$\begin{aligned} \|\widehat{Q}_{s,u_0}^* - \widehat{Q}_{s,u^*}^*\|_{\infty} &= \|\widehat{\mathcal{T}}_{\text{pe}}^{s,u_0}(\widehat{Q}_{s,u_0}^*) - \widehat{\mathcal{T}}_{\text{pe}}^{s,u^*}(\widehat{Q}_{s,u^*}^*)\|_{\infty} \\ &\leq \|\widehat{\mathcal{T}}_{\text{pe}}^{s,u^*}(\widehat{Q}_{s,u_0}^*) - \widehat{\mathcal{T}}_{\text{pe}}^{s,u^*}(\widehat{Q}_{s,u^*}^*)\|_{\infty} + \|\widehat{\mathcal{T}}_{\text{pe}}^{s,u_0}(\widehat{Q}_{s,u_0}^*) - \widehat{\mathcal{T}}_{\text{pe}}^{s,u^*}(\widehat{Q}_{s,u_0}^*)\|_{\infty} \\ &\leq \gamma \|\widehat{Q}_{s,u_0}^* - \widehat{Q}_{s,u^*}^*\|_{\infty} + \frac{1}{N}. \end{aligned}$$

It follows that

$$\|\widehat{Q}_{s,u_0}^* - \widehat{Q}_{s,u^*}^*\|_{\infty} \leq \frac{1}{(1-\gamma)N},$$

and hence

$$\|\widehat{V}_{s,u_0}^* - \widehat{V}_{s,u^*}^*\|_{\infty} \leq \|\widehat{Q}_{s,u_0}^* - \widehat{Q}_{s,u^*}^*\|_{\infty} \leq \frac{1}{(1-\gamma)N}.$$

We next use this perturbation bound to compare the corresponding variance terms. Specifically,

$$\text{Var}_{P_{s,a}}(\widehat{V}_{s,u_0}^*) - \text{Var}_{P_{s,a}}(\widehat{V}_{s,u^*}^*)$$

$$\begin{aligned}
&\stackrel{(i)}{\leq} P_{s,a} \left((\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u^*}^*) \circ (\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u^*}^*) \right) - P_{s,a} \left((\hat{V}_{s,u^*}^* - P_{s,a} \hat{V}_{s,u^*}^*) \circ (\hat{V}_{s,u^*}^* - P_{s,a} \hat{V}_{s,u^*}^*) \right) \\
&\leq P_{s,a} \left((\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u^*}^* + \hat{V}_{s,u^*}^* - P_{s,a} \hat{V}_{s,u^*}^*) \circ (\hat{V}_{s,u_0}^* - \hat{V}_{s,u^*}^*) \right) \\
&\leq \frac{2}{1-\gamma} \left| P_{s,a} (\hat{V}_{s,u_0}^* - \hat{V}_{s,u^*}^*) \right| \\
&\leq \frac{2}{1-\gamma} \|\hat{V}_{s,u_0}^* - \hat{V}_{s,u^*}^*\|_\infty \stackrel{(ii)}{\leq} \frac{2}{(1-\gamma)^2 N}.
\end{aligned}$$

Here, (i) follows from a direct use of $\mathbb{E}[X] = \arg \min_c \mathbb{E}[(X - c)^2]$:

$$\begin{aligned}
\text{Var}_{P_{s,a}}(\hat{V}_{s,u_0}^*) &= P_{s,a} \left((\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u_0}^*) \circ (\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u_0}^*) \right) \\
&\leq P_{s,a} \left((\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u^*}^*) \circ (\hat{V}_{s,u_0}^* - P_{s,a} \hat{V}_{s,u^*}^*) \right),
\end{aligned}$$

while (ii) uses the bounds $0 \leq \hat{V}_{s,u_0}^*, \hat{V}_{s,u^*}^* \leq \frac{1}{1-\gamma}$. Swapping $\hat{V}_{s,u_0}^*$ and $\hat{V}_{s,u^*}^*$ gives

$$\left| \text{Var}_{P_{s,a}}(\hat{V}_{s,u_0}^*) - \text{Var}_{P_{s,a}}(\hat{V}_{s,u^*}^*) \right| \leq \frac{2}{(1-\gamma)^2 N}. \quad (6.121)$$

The same argument also applies with $P_{s,a}$ replaced by $\hat{P}_{s,a}$, yielding

$$\left| \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u_0}^*) - \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*) \right| \leq \frac{2}{(1-\gamma)^2 N}. \quad (6.122)$$

With these perturbation bounds in place, we can use the triangle inequality to show that

$$\begin{aligned}
\left| (\hat{P}_{s,a} - P_{s,a}) \hat{V}_{\text{pe}}^* \right| &= \left| (\hat{P}_{s,a} - P_{s,a}) \hat{V}_{s,u^*}^* \right| \\
&\leq \left| (\hat{P}_{s,a} - P_{s,a}) \hat{V}_{s,u_0}^* \right| + \left| (\hat{P}_{s,a} - P_{s,a}) (\hat{V}_{s,u^*}^* - \hat{V}_{s,u_0}^*) \right| \\
&\stackrel{(i)}{\leq} \left| (\hat{P}_{s,a} - P_{s,a}) \hat{V}_{s,u_0}^* \right| + \frac{2}{N(1-\gamma)} \\
&\stackrel{(ii)}{\leq} \sqrt{\frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{N_{s,a}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u_0}^*)} + \frac{60 \log \frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}.
\end{aligned} \quad (6.123)$$

Here, (i) follows from

$$\begin{aligned}
\left| (\hat{P}_{s,a} - P_{s,a}) (\hat{V}_{s,u^*}^* - \hat{V}_{s,u_0}^*) \right| &\leq (\|\hat{P}_{s,a}\|_1 + \|P_{s,a}\|_1) \|\hat{V}_{s,u^*}^* - \hat{V}_{s,u_0}^*\|_\infty \\
&\leq \frac{2}{N(1-\gamma)}.
\end{aligned}$$

Moreover, (ii) follows from (6.122) and (6.120a), which implies

$$\begin{aligned}
\left| (\hat{P}_{s,a} - P_{s,a}) \hat{V}_{s,u_0}^* \right| &\leq \sqrt{\frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{N_{s,a}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u_0}^*)} + \frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} \\
&\leq \sqrt{\frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{N_{s,a}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*)} + \frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} + \sqrt{\frac{96 \log \frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)^2 N \cdot N_{s,a}}} \\
&\leq \sqrt{\frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{N_{s,a}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*)} + \frac{58 \log \frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}},
\end{aligned}$$

where the last step uses $N \geq N_{s,a}$. Similarly, we obtain

$$\begin{aligned}
\text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*) &= \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*) \\
&\leq \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u_0}^*) + \left| \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*) - \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u_0}^*) \right| \\
&\stackrel{(i)}{\leq} 2\text{Var}_{P_{s,a}}(\hat{V}_{s,u_0}^*) + \frac{5 \log \frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N_{s,a}} + \frac{2}{(1-\gamma)^2 N} \\
&\leq 2\text{Var}_{P_{s,a}}(\hat{V}_{s,u^*}^*) + 2 \left| \text{Var}_{P_{s,a}}(\hat{V}_{s,u^*}^*) - \text{Var}_{P_{s,a}}(\hat{V}_{s,u_0}^*) \right| + \frac{5 \log \frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N_{s,a}} + \frac{2}{(1-\gamma)^2 N} \\
&\stackrel{(ii)}{\leq} 2\text{Var}_{P_{s,a}}(\hat{V}_{\text{pe}}^*) + \frac{6}{(1-\gamma)^2 N} + \frac{5 \log \frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N_{s,a}} \\
&\stackrel{(iii)}{\leq} 2\text{Var}_{P_{s,a}}(\hat{V}_{\text{pe}}^*) + \frac{23 \log \frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N_{s,a}}, \tag{6.124}
\end{aligned}$$

where (i) follows from (6.120b) and (6.122), (ii) follows from (6.121), and (iii) uses $N \geq N_{s,a}$.

Step 5: extending the bounds to $\tilde{V}$. Consider any $\tilde{V}$ satisfying $\|\tilde{V} - \hat{V}_{\text{pe}}^*\|_\infty \leq \frac{1}{N}$ and $\|\tilde{V}\|_\infty \leq \frac{1}{1-\gamma}$. Invoking (6.123) and applying the triangle inequality yield

$$\begin{aligned}
|(\hat{P}_{s,a} - P_{s,a})\tilde{V}| &\leq |(\hat{P}_{s,a} - P_{s,a})\hat{V}_{\text{pe}}^*| + |(\hat{P}_{s,a} - P_{s,a})(\hat{V}_{\text{pe}}^* - \tilde{V})| \\
&\stackrel{(i)}{\leq} \sqrt{\frac{48 \log \frac{8N^3}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*) + \frac{60 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} + \frac{2}{N} \\
&\stackrel{(ii)}{\leq} 12 \sqrt{\frac{\log \frac{2N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{s,u^*}^*) + \frac{62 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} \\
&= 12 \sqrt{\frac{\log \frac{2N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*) + \frac{62 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}. \tag{6.125}
\end{aligned}$$

Here, (i) utilizes the bound

$$|(\hat{P}_{s,a} - P_{s,a})(\hat{V}_{\text{pe}}^* - \tilde{V})| \leq (\|\hat{P}_{s,a}\|_1 + \|P_{s,a}\|_1) \|\hat{V}_{\text{pe}}^* - \tilde{V}\|_\infty \leq \frac{2}{N},$$

while (ii) follows from $N \geq N_{s,a}$. Moreover, since $\|\tilde{V} - \hat{V}_{\text{pe}}^*\|_\infty \leq 1/N$, the same perturbation argument used to prove (6.122) gives

$$\left| \text{Var}_{\hat{P}_{s,a}}(\hat{V}_{\text{pe}}^*) - \text{Var}_{\hat{P}_{s,a}}(\tilde{V}) \right| \leq \frac{2}{(1-\gamma)^2 N}.$$

Combining this estimate with (6.125), we obtain

$$\begin{aligned}
|(\hat{P}_{s,a} - P_{s,a})\tilde{V}| &\leq 12 \sqrt{\frac{\log \frac{2N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(\tilde{V}) + 12 \sqrt{\frac{2 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N \cdot N_{s,a}}} + \frac{62 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}} \\
&\leq 12 \sqrt{\frac{\log \frac{2N}{(1-\gamma)\delta}}{N_{s,a}}} \text{Var}_{\hat{P}_{s,a}}(\tilde{V}) + \frac{74 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N_{s,a}}.
\end{aligned}$$

This proves (6.29a). The other desired bound (6.29b) can be established with a similar argument.

6.7.5 Proof of Lemma 6.7

To begin, observe that

$$\begin{aligned}
 (\widehat{V} \circ \widehat{V}) - (\gamma P^* \widehat{V}) \circ (\gamma P^* \widehat{V}) &= (\widehat{V} - \gamma P^* \widehat{V}) \circ (\widehat{V} + \gamma P^* \widehat{V}) \\
 &\stackrel{(i)}{\leq} (\widehat{V} - \gamma P^* \widehat{V} + 2b^*) \circ (\widehat{V} + \gamma P^* \widehat{V}) \\
 &\stackrel{(ii)}{\leq} \frac{2}{1-\gamma} (\widehat{V} - \gamma P^* \widehat{V} + 2b^*).
 \end{aligned} \tag{6.126}$$

Here, (i) follows from $b^* \geq 0$ and $\widehat{V} + \gamma P^* \widehat{V} \geq 0$, while (ii) follows from

$$\|\widehat{V} + \gamma P^* \widehat{V}\|_\infty \leq 2\|\widehat{V}\|_\infty \leq \frac{2}{1-\gamma},$$

together with the fact that $\widehat{V} - \gamma P^* \widehat{V} + 2b^* \geq 0$, which has been verified in (6.38). Using this observation, we obtain

$$\begin{aligned}
 \sum_s d^*(s) \text{Var}_{P_{s, \pi^*(s)}}(\widehat{V}) &\stackrel{(i)}{=} \left\langle d^*, P^*(\widehat{V} \circ \widehat{V}) - (P^* \widehat{V}) \circ (P^* \widehat{V}) \right\rangle \\
 &\stackrel{(ii)}{\leq} \left\langle d^*, P^*(\widehat{V} \circ \widehat{V}) - \frac{1}{\gamma^2} \widehat{V} \circ \widehat{V} + \frac{2}{\gamma^2(1-\gamma)} (\widehat{V} - \gamma P^* \widehat{V} + 2b^*) \right\rangle \\
 &\stackrel{(iii)}{\leq} \left\langle d^*, P^*(\widehat{V} \circ \widehat{V}) - \frac{1}{\gamma} \widehat{V} \circ \widehat{V} + \frac{2}{\gamma^2(1-\gamma)} ((I - \gamma P^*) \widehat{V} + 2b^*) \right\rangle \\
 &= d^{*\top} (I - \gamma P^*) \left\{ -\frac{1}{\gamma} \widehat{V} \circ \widehat{V} + \frac{2}{\gamma^2(1-\gamma)} \widehat{V} \right\} + \frac{4}{\gamma^2(1-\gamma)} \langle d^*, b^* \rangle \\
 &\stackrel{(iv)}{\leq} (1-\gamma) \rho^\top \left\{ -\frac{1}{\gamma} \widehat{V} \circ \widehat{V} + \frac{2}{\gamma^2(1-\gamma)} \widehat{V} \right\} + \frac{4}{\gamma^2(1-\gamma)} \langle d^*, b^* \rangle \\
 &\leq \frac{2}{\gamma^2} \rho^\top \widehat{V} + \frac{4}{\gamma^2(1-\gamma)} \langle d^*, b^* \rangle \stackrel{(v)}{\leq} \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)} \langle d^*, b^* \rangle.
 \end{aligned}$$

Here, (i) follows from the variance identity (2.22); (ii) follows from (6.126); (iii) uses $\gamma < 1$; (iv) is a direct consequence of the matrix-vector representation (6.42); and (v) follows from $\|\rho^\top\|_1 = 1$ and $\|\widehat{V}\|_\infty \leq 1/(1-\gamma)$.

6.7.6 Proof of Lemma 6.14

In this section, we will use concentration inequalities from Section 2.4, including the Bernstein inequality and its empirical variant, to establish Lemma 6.14.

Proof of inequality (6.114a). If $0 < N_{s,a} < 48 \log \frac{N}{\delta}$, then we can immediately see that

$$\left| (\widehat{P}_{s,a} - P_{s,a})V \right| \leq \|V\|_\infty \leq \frac{1}{1-\gamma} \leq \frac{48 \log \frac{N}{\delta}}{(1-\gamma)N_{s,a}},$$

and hence the claim (6.114a) is valid. As a result, it suffices to focus on the case where

$$N_{s,a} \geq 48 \log \frac{N}{\delta}. \tag{6.127}$$

We can apply empirical Bernstein inequality (cf. Lemma 2.6) to show that, with probability exceeding $1 - \delta$,

$$\begin{aligned} |(\hat{P}_{s,a} - P_{s,a})V| &\leq \sqrt{\frac{2 \log\left(\frac{4N}{\delta}\right)}{N_{s,a} - 1}} \sqrt{\text{Var}_{\hat{P}_{s,a}}(V)} + \frac{7\|V\|_\infty \log\left(\frac{4N}{\delta}\right)}{3(N_{s,a} - 1)} \\ &\leq \sqrt{\frac{48 \text{Var}_{\hat{P}_{s,a}}(V) \log \frac{N}{\delta}}{N_{s,a}}} + \frac{48 \log \frac{N}{\delta}}{(1 - \gamma)N_{s,a}} \end{aligned}$$

holds simultaneously for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ satisfying (6.127).

Proof of inequality (6.114b). Define $\bar{V} = V - (P_{s,a}V)1_{\mathcal{S}}$. We can invoke the Bernstein inequality (cf. Lemma 2.5) to show that, with probability at least $1 - \delta$,

$$\begin{aligned} |(P_{s,a} - \hat{P}_{s,a})(\bar{V} \circ \bar{V})| &\leq \sqrt{\frac{4 \text{Var}_{P_{s,a}}(\bar{V} \circ \bar{V}) \log \frac{N}{\delta}}{N_{s,a}}} + \frac{2\|\bar{V} \circ \bar{V}\|_\infty \log \frac{N}{\delta}}{3N_{s,a}} \\ &\leq \sqrt{\frac{4 \text{Var}_{P_{s,a}}(V) \log \frac{N}{\delta}}{(1 - \gamma)^2 N_{s,a}}} + \frac{2 \log \frac{N}{\delta}}{3(1 - \gamma)^2 N_{s,a}} \end{aligned}$$

holds simultaneously for any (s, a) obeying $N_{s,a} > 0$, where the second inequality follows from $P_{s,a}(\bar{V} \circ \bar{V}) = \text{Var}_{P_{s,a}}(V)$. In addition, we also have

$$\begin{aligned} \text{Var}_{s,a}(V) &= P_{s,a}(\bar{V} \circ \bar{V}) = (P_{s,a} - \hat{P}_{s,a})(\bar{V} \circ \bar{V}) + \hat{P}_{s,a}(\bar{V} \circ \bar{V}) \\ &= (P_{s,a} - \hat{P}_{s,a})(\bar{V} \circ \bar{V}) + |(\hat{P}_{s,a} - P_{s,a})V|^2 + \text{Var}_{\hat{P}_{s,a}}(V). \end{aligned}$$

Taking the above two equations collectively yields

$$\begin{aligned} \text{Var}_{P_{s,a}}(V) &\geq -|(P_{s,a} - \hat{P}_{s,a})(\bar{V} \circ \bar{V})| + \text{Var}_{\hat{P}_{s,a}}(V) \\ &\geq -\sqrt{\frac{4 \text{Var}_{P_{s,a}}(V) \log \frac{N}{\delta}}{(1 - \gamma)^2 N_{s,a}}} - \frac{2 \log \frac{N}{\delta}}{3(1 - \gamma)^2 N_{s,a}} + \text{Var}_{\hat{P}_{s,a}}(V). \end{aligned}$$

By rearranging terms and applying the AM-GM inequality, we have

$$\begin{aligned} \text{Var}_{\hat{P}_{s,a}}(V) &\leq \text{Var}_{P_{s,a}}(V) + 2\sqrt{\frac{\text{Var}_{P_{s,a}}(V) \log \frac{N}{\delta}}{(1 - \gamma)^2 N_{s,a}}} + \frac{2 \log \frac{N}{\delta}}{3(1 - \gamma)^2 N_{s,a}} \\ &\leq 2\text{Var}_{P_{s,a}}(V) + \frac{5 \log \frac{N}{\delta}}{3(1 - \gamma)^2 N_{s,a}}. \end{aligned}$$

6.8 Appendix B: proof of auxiliary lemmas in Section 6.4.3

6.8.1 Proof of Lemma 6.9

We start by unrolling the update rule of the Q-function as below. For any state-action pair (s, a) , let $n = n_t(s, a)$ be the number of its visits up to t . Unrolling the update gives

$$Q_t(s, a) = \sum_{i=1}^n \eta_i^n \left\{ r(s, a) + \gamma V_{k_i}(s_{k_i+1}) - b_i \right\}, \quad (6.128)$$

with the convention that the sum is zero if $n = 0$. Here, we make use of the pessimistic initialization $Q_0 = 0$ and s_{k_i+1} is the next state observed at the i -th visit to (s, a) . Combining with the Bellman's optimality equation, we obtain

$$\begin{aligned} (Q^* - Q_t)(s, a) &= \gamma P_{s,a} V^* - \sum_{i=1}^n \eta_i^n \left\{ \gamma V_{k_i}(s_{k_i+1}) - b_i \right\} \\ &= \gamma \sum_{i=1}^n \eta_i^n P_{s,a} (V^* - V_{k_i}) + \gamma \sum_{i=1}^n \eta_i^n [(P_{s,a} - P_{k_i}) V_{k_i}](s, a) + \sum_{i=1}^n \eta_i^n b_i, \end{aligned} \quad (6.129)$$

where $P_{k_i} \in \{0, 1\}^S$ is an indicator vector corresponding to the empirical transition at the i -th visit. By the Azuma–Hoeffding inequality, and applying the union bound, we have that with probability at least $1 - \delta$, the following bound holds simultaneously for every (s, a) and every $1 \leq n \leq T$:

$$\left| \gamma \sum_{i=1}^n \eta_i^n [(P_{s,a} - P_{k_i}) V_{k_i}](s, a) \right| \leq c_0 \sqrt{\frac{H\iota}{n(1-\gamma)^2}}. \quad (6.130)$$

Moreover, by a similar argument as the online case (see Lemma 5.18 and its proof), we have

$$C_b \sqrt{\frac{H\iota}{n(1-\gamma)^2}} \leq \sum_{i=1}^n \eta_i^n b_i(s, a) \leq 2C_b \sqrt{\frac{H\iota}{n(1-\gamma)^2}}. \quad (6.131)$$

Plugging the above into (6.129), we have

$$(Q^* - Q_t)(s, a) \leq \gamma \sum_{i=1}^n \eta_i^n P_{s,a} (V^* - V_{k_i}) + \beta_n. \quad (6.132)$$

Setting C_b sufficiently large and $a = \pi^*(s)$ leads to (6.61).

We next establish the pessimism property (6.63). The lower counterpart of (6.129) gives

$$(Q^* - Q_t)(s, a) \geq \gamma \sum_{i=1}^n \eta_i^n P_{s,a} (V^* - V_{k_i}) - c_0 \sqrt{\frac{H\iota}{n(1-\gamma)^2}} + \sum_{i=1}^n \eta_i^n b_i. \quad (6.133)$$

Taking C_b sufficiently large and using (6.131), the last two terms in (6.133) are nonnegative. We may therefore prove $Q_t \leq Q^*$ and $V_t \leq V^*$ by induction over t . The claim is trivial at $t = 0$. If it holds for all earlier iterates, then $V_{k_i} \leq V^*$ for every i , and hence (6.133) implies $Q_t(s, a) \leq Q^*(s, a)$ for the updated pair; all other pairs are unchanged. The monotone value update then gives

$$V_t(s) \leq \max\{V_{t-1}(s), \max_a Q^*(s, a)\} = V^*(s),$$

which closes the induction.

It remains to justify the policy-value part of (6.63). Fix t and the policy π_t defined by (6.59). For each state s , let $\tau_s \leq t$ be the last time at which the value $V_t(s)$ was attained. Then $V_t(s) = Q_{\tau_s}(s, \pi_t(s))$ unless $V_t(s) = 0$, in which case the desired inequality is immediate. Applying the same unrolling argument as above with Q^{π_t} in place of Q^* gives

$$Q^{\pi_t}(s, \pi_t(s)) - Q_{\tau_s}(s, \pi_t(s)) \geq \gamma \sum_i \eta_i^{n_{\tau_s}} P_{s, \pi_t(s)} (V^{\pi_t} - V_{k_i}), \quad (6.134)$$

where the martingale fluctuation is again dominated by the accumulated penalty. Since $V_{k_i} \leq V_t$ by monotonicity, if

$$m := \min_{s \in S} \{V^{\pi_t}(s) - V_t(s)\},$$

then (6.134) implies

$$V^{\pi_t}(s) - V_t(s) \geq \gamma m, \quad \forall s \in \mathcal{S}.$$

Taking the minimum over s yields $m \geq \gamma m$, and since $\gamma < 1$, this forces $m \geq 0$. Thus $V_t \leq V^{\pi_t}$. Finally, $V^{\pi_t} \leq V^*$ by optimality of V^* , completing the proof of Lemma 6.9.

6.8.2 Proof of Lemma 6.10

For any given integer $K > 0$, one can decompose

$$\begin{aligned} & \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \langle \rho(P_{\pi^*})^j, V \rangle \\ &= \sum_{j=0}^{K-1} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \langle \rho(P_{\pi^*})^j, V \rangle + \sum_{j=K}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \langle \rho(P_{\pi^*})^j, V \rangle \\ &\leq \left(1 + \frac{1}{H} \right)^{3K} \sum_{j=0}^{K-1} \gamma^j \langle \rho(P_{\pi^*})^j, V \rangle + \sum_{j=K}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \|V\|_{\infty} \\ &\leq \underbrace{\left(1 + \frac{1}{H} \right)^{3K} \frac{1}{1-\gamma} \langle d_{\rho}^*, V \rangle}_{=: \alpha_1} + \underbrace{\gamma^K \left(1 + \frac{1}{H} \right)^{3K} \frac{1}{1-\gamma \left(1 + \frac{1}{H} \right)^3} \|V\|_{\infty}}_{=: \alpha_2}. \end{aligned}$$

Here, the last inequality holds since $d_{\rho}^* = (1-\gamma) \sum_{j=0}^{\infty} \gamma^j \rho(P_{\pi^*})^j$.

By taking

$$K = H = \left\lceil \frac{4}{1-\gamma} \log \frac{ST}{\delta} \right\rceil,$$

we can derive

$$\left(1 + \frac{1}{H} \right)^{3K} = \left(1 + \frac{1}{H} \right)^{3H} \stackrel{(i)}{\leq} e^3 = O(1)$$

and

$$\gamma^K = e^{K \log[1-(1-\gamma)]} \stackrel{(ii)}{\leq} e^{-K(1-\gamma)} = \frac{\delta}{ST^4}.$$

Here, (i) holds since $(1 + 1/x)^x \leq e$ for all $x > 0$; (ii) is valid since $\log(1-x) \leq -x$ for all $x \in (0, 1)$. It is also worth noting that

$$\begin{aligned} \frac{1}{1-\gamma \left(1 + \frac{1}{H} \right)^3} &\leq \frac{1}{1-\gamma \left(1 + \frac{1-\gamma}{4} \right)^3} \stackrel{(iii)}{\leq} \frac{1}{1-\gamma \left[1 + \frac{61}{64} (1-\gamma) \right]} \\ &= \frac{1}{(1-\gamma) \left(1 - \frac{61}{64} \gamma \right)} \lesssim \frac{1}{1-\gamma}, \end{aligned} \tag{6.135}$$

where (iii) holds since $(1+x)^3 \leq 1 + 61x/16$ for all $0 < x \leq 1/4$. We then immediately arrive at

$$\alpha_1 \lesssim \frac{1}{1-\gamma} \langle d_{\rho}^*, V \rangle$$

and

$$\alpha_2 \lesssim \frac{\delta}{ST^4 (1-\gamma)} \|V\|_{\infty}.$$

Taking the upper bounds on α_1 and α_2 collectively leads to the advertised inequality

$$\sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \langle \rho(P_{\pi^*})^j, V \rangle \lesssim \frac{1}{1-\gamma} \langle d_{\rho}^*, V \rangle + \frac{\delta}{ST^4(1-\gamma)} \|V\|_{\infty}.$$

6.9 Appendix C: proof of the lower bound in Theorem 6.11

We construct a collection of hard problem instances and use them to establish the minimax lower bounds asserted in Theorem 6.11. Throughout this subsection, we assume that

$$\frac{2}{3} \leq \gamma < 1 \quad \text{and} \quad \frac{224(1-\gamma)\varepsilon}{\gamma} \leq \frac{1}{2}. \quad (6.136)$$

6.9.1 Construction of hard problem instances

Construction of the hard MDPs. We introduce a class of MDPs

$$\left\{ \mathcal{M}_{\theta} = (\mathcal{S}, \mathcal{A}, P_{\theta}, r, \gamma) \mid \theta \in \Theta \subseteq \{0, 1\}^{S-1} \right\}$$

parameterized by $\theta \in \Theta$, each consisting of S states and two actions:

$$\mathcal{S} = \{0, 1, \dots, S-1\} \quad \text{and} \quad \mathcal{A} = \{0, 1\}.$$

We single out the following state distribution

$$\mu(s) = \frac{1}{C(S-1)} \mathbb{1}\{s > 0\} + \left(1 - \frac{1}{C}\right) \mathbb{1}\{s = 0\} \quad (6.137)$$

for some quantity $C > \frac{3}{2}$. With this distribution in place, we define the transition kernel P_{θ} of the MDP $\mathcal{M}_{\theta}$ as follows:

$$P_{\theta}(s' \mid s, a) = \begin{cases} p \mathbb{1}\{s' = s\} + (1-p)\mu(s') & \text{if } s > 0, a = \theta_s, \\ q \mathbb{1}\{s' = s\} + (1-q)\mu(s') & \text{if } s > 0, a = 1 - \theta_s, \\ \mathbb{1}\{s' = 0\} & \text{if } (s, a) = (0, 0), \\ (2\gamma - 1) \mathbb{1}\{s' = 0\} + 2(1-\gamma)\mu(s') & \text{if } (s, a) = (0, 1), \end{cases} \quad (6.138)$$

where p and q are chosen as

$$p = \gamma + \frac{224(1-\gamma)^2\varepsilon}{\gamma}, \quad q = \gamma - \frac{224(1-\gamma)^2\varepsilon}{\gamma}. \quad (6.139)$$

In view of the assumption (6.136), we have

$$p > q \geq \gamma - \frac{1-\gamma}{2} \geq \frac{1}{2}. \quad (6.140)$$

The reward function for each MDP $\mathcal{M}_{\theta}$ is defined as

$$r(s, a) = \begin{cases} 1 & \text{if } s > 0, \\ \frac{4}{7} & \text{if } (s, a) = (0, 0), \\ 0 & \text{if } (s, a) = (0, 1). \end{cases} \quad (6.141)$$

Finally, let us choose the set $\Theta \subseteq \{0, 1\}^{S-1}$. By virtue of the Gilbert-Varshamov lemma (Gilbert, 1952), one can construct $\Theta \subseteq \{0, 1\}^{S-1}$ in a way that

$$|\Theta| \geq e^{(S-1)/8} \quad \text{and} \quad \|\theta - \tilde{\theta}\|_1 \geq \frac{S-1}{8} \quad (6.142)$$

for any $\theta, \tilde{\theta} \in \Theta$ obeying $\theta \neq \tilde{\theta}$. In other words, the set Θ we construct contains an exponentially large number of vectors that are sufficiently separated. This property plays an important role in the ensuing analysis.

Value functions and optimal policies. We next compute the value functions of the constructed MDPs and identify their optimal policies. For notational clarity, for the MDP $\mathcal{M}_\theta$ with $\theta \in \{0, 1\}^{S-1}$, we denote by π_θ^* its optimal policy, and let V_θ^π (resp. V_θ^*) denote the value function of a policy π (resp. π_θ^*). The following lemma collects several useful properties of these value functions and optimal policies; the proof is deferred to Section 6.9.3.

Lemma 6.15. Consider any $\theta \in \{0, 1\}^{S-1}$ and any policy π . One has

$$V_\theta^\pi(s) \leq \frac{1 + \gamma(1 - x_{\pi, \theta}(s))(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0))}{1 - \gamma x_{\pi, \theta}(s)}, \quad (6.143)$$

where $x_{\pi, \theta}(s) := p\pi(\theta_s | s) + q\pi(1 - \theta_s | s)$, and the inequality holds when $\pi = \pi^*$. In addition, the optimal policy π_θ^* and the optimal value function satisfy, for $s > 0$

$$\pi_\theta^*(0 | 0) = 1, \quad \pi_\theta^*(\theta_s | s) = 1, \quad (6.144a)$$

and

$$V_\theta^*(0) = \frac{4}{7(1 - \gamma)} \quad \text{and} \quad V_\theta^*(s) = \frac{1 + \gamma(1 - p)(1 - C^{-1})V_\theta^*(0)}{1 - \gamma(p + (1 - p)C^{-1})}. \quad (6.144b)$$

Construction of the batch dataset. Given any constructed MDP $\mathcal{M}_\theta$, we generate a dataset consisting of N i.i.d. samples $\{(s_i, a_i, s'_i)\}_{1 \leq i \leq N}$ according to (6.1), where the initial state distribution ρ^b and the behavior policy π^b are chosen as

$$\rho^b(s) = \mu(s), \quad \pi^b(a | s) = 1/2, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A},$$

with μ defined in (6.137). For this dataset, the occupancy state distribution coincides with μ , namely,

$$d^b(s) = \mu(s), \quad d^b(s, a) = \mu(s)/2, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (6.145)$$

Moreover, choose the test distribution ρ as

$$\rho(s) = \begin{cases} \frac{1}{S-1}, & \text{if } s > 0, \\ 0, & \text{if } s = 0. \end{cases} \quad (6.146)$$

Then the single-policy concentrability coefficient C_ρ^* of the dataset with respect to the constructed MDP $\mathcal{M}_\theta$ is given by

$$C < C_\rho^* < 2C. \quad (6.147)$$

The proof of the claims in (6.145) and (6.147) can be found in Section 6.9.4.

6.9.2 Establishing the minimax lower bound

Equipped with the above construction, we are ready to establish the desired lower bounds. In view of the choice of the test distribution ρ in (6.146), it suffices to control

$$\langle \rho, V_\theta^* - V_{\hat{\pi}}^\pi \rangle = \frac{1}{S-1} \sum_{s=1}^{S-1} (V_\theta^*(s) - V_{\hat{\pi}}^\pi(s)),$$

where $\hat{\pi}$ denotes a policy estimate computed from the batch dataset.

Step 1: converting $\hat{\pi}$ into an estimate $\hat{\theta}$ of θ . Consider first an arbitrary policy π . Together with (6.143), this gives

$$\begin{aligned} \langle \rho, V_\theta^* - V_\pi^\pi \rangle &\geq \frac{1}{S-1} \sum_{s=1}^{S-1} \left[V_\theta^*(s) - \frac{1 + \gamma(1 - x_{\pi, \theta}(s))(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0))}{1 - \gamma x_{\pi, \theta}(s)} \right] \\ &= \frac{1}{S-1} \sum_{s=1}^{S-1} \frac{1 - (1 - \gamma)(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0))}{(1 - \gamma p)(1 - \gamma x_{\pi, \theta}(s))} \cdot \gamma(p - x_{\pi, \theta}(s)) \\ &\geq \frac{16\varepsilon}{S-1} \sum_{s=1}^{S-1} (1 - \pi(\theta_s | s)). \end{aligned} \quad (6.148)$$

Here, justification for the last line is postponed to Section 6.9.4.

Let $\mathbb{P}_\theta$ denote the probability distribution induced by $\mathcal{M}_\theta$. Suppose for the moment that the policy estimate $\hat{\pi}$ satisfies

$$\mathbb{P}_\theta\{\langle \rho, V_\theta^* - V_{\hat{\pi}}^\pi \rangle \leq \varepsilon\} \geq \frac{3}{4}.$$

Then (6.148) implies that

$$\sum_{s=1}^{S-1} \|\hat{\pi}(\cdot | s) - \pi_\theta^*(\cdot | s)\|_1 \leq \frac{S-1}{8}$$

with probability at least 3/4. If this were the case, we could construct an estimator $\hat{\theta}$ of θ by

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{s=1}^{S-1} \|\hat{\pi}(\cdot | s) - \pi_\theta^*(\cdot | s)\|_1, \quad (6.149)$$

Then for any $\tilde{\theta} \in \Theta$ with $\tilde{\theta} \neq \theta$ one has

$$\begin{aligned} \sum_{s=1}^{S-1} \|\hat{\pi}(\cdot | s) - \pi_{\tilde{\theta}}^*(\cdot | s)\|_1 &\geq \sum_{s=1}^{S-1} \|\pi_\theta^*(\cdot | s) - \pi_{\tilde{\theta}}^*(\cdot | s)\|_1 - \sum_{s=1}^{S-1} \|\hat{\pi}(\cdot | s) - \pi_\theta^*(\cdot | s)\|_1 \\ &= 2\|\theta - \tilde{\theta}\|_1 - \sum_{s=1}^{S-1} \|\hat{\pi}(\cdot | s) - \pi_\theta^*(\cdot | s)\|_1 \\ &> \frac{S-1}{4} - \frac{S-1}{8} = \frac{S-1}{8}, \end{aligned} \quad (6.150)$$

where the first inequality holds by the triangle inequality, the second line arises from the fact $\pi_\theta^*(\theta_s | s) = 1$, and the last line comes from the properties (6.142) about Θ . As a consequence,

$$\mathbb{P}(\hat{\theta} = \theta) \geq \mathbb{P}\left(\sum_{s=1}^{S-1} \|\hat{\pi}(\cdot | s) - \pi_\theta^*(\cdot | s)\|_1 \leq \frac{S-1}{8}\right) \geq \frac{3}{4}. \quad (6.151)$$

We will show next that (6.151) is information-theoretically impossible, namely, one cannot possibly find such a good estimator for θ without sufficiently many samples.

Step 2: probability of error in testing multiple hypotheses. Next, we turn attention to a $|\Theta|$ -ary hypothesis testing problem. For any $\theta \in \Theta$, denote by $\mathbb{P}_\theta$ the probability distribution when the MDP is $\mathcal{M}_\theta$. We will then study the minimax probability of error defined as follows:

$$p_e := \inf_{\psi} \max_{\theta \in \Theta} \mathbb{P}_\theta(\psi \neq \theta), \quad (6.152)$$

where the infimum is taken over all possible tests ψ (constructed based on the batch dataset available). Let $\mu^{\mathbf{b},\theta}$ denote the distribution of a single sample (s_i, a_i, s'_i) under $\mathcal{M}_\theta$. Since the samples are generated independently, we have

$$\begin{aligned} p_e &\stackrel{(i)}{\geq} 1 - \frac{N \max_{\theta, \tilde{\theta} \in \Theta, \theta \neq \tilde{\theta}} \text{KL}(\mu^{\mathbf{b},\theta} \parallel \mu^{\mathbf{b},\tilde{\theta}}) + \log 2}{\log |\Theta|} \\ &\stackrel{(ii)}{\geq} 1 - \frac{8N}{S-1} \max_{\theta, \tilde{\theta} \in \Theta, \theta \neq \tilde{\theta}} \text{KL}(\mu^{\mathbf{b},\theta} \parallel \mu^{\mathbf{b},\tilde{\theta}}) - \frac{8 \log 2}{S-1} \\ &\stackrel{(iii)}{\geq} \frac{1}{2} - \frac{8N}{S-1} \max_{\theta, \tilde{\theta} \in \Theta, \theta \neq \tilde{\theta}} \text{KL}(\mu^{\mathbf{b},\theta} \parallel \mu^{\mathbf{b},\tilde{\theta}}), \end{aligned} \quad (6.153)$$

where (i) arises from Fano's inequality (cf. (Tsybakov, 2009, Corollary 2.6)) and the additivity property of the KL divergence (cf. Tsybakov (2009, Page 85)), (ii) holds since $|\Theta| \geq e^{(S-1)/8}$ (according to our construction (6.142)), and (iii) is valid when $S \geq 1 + 16 \log 2$. Since the samples are generated independently, we have

$$\begin{aligned} \text{KL}(\mu^{\mathbf{b},\theta} \parallel \mu^{\mathbf{b},\tilde{\theta}}) &= \mathbb{E}_{s \sim d^{\mathbf{b}}} \left[\text{KL}(\mu^{\mathbf{b},\theta}(s) \parallel \mu^{\mathbf{b},\tilde{\theta}}(s)) \right] \\ &= \frac{1}{2C(S-1)} \sum_{s=1}^{S-1} \sum_{a \in \{0,1\}} \text{KL}(P_\theta(\cdot | s, a) \parallel P_{\tilde{\theta}}(\cdot | s, a)). \end{aligned}$$

If $\theta_s = \tilde{\theta}_s$, then it is self-evident that

$$\text{KL}(P_{\theta_s}(\cdot | s, 0) \parallel P_{\tilde{\theta}_s}(\cdot | s, 0)) + \text{KL}(P_{\theta_s}(\cdot | s, 1) \parallel P_{\tilde{\theta}_s}(\cdot | s, 1)) = 0. \quad (6.154)$$

Consider now the case that $\theta_s \neq \tilde{\theta}_s$, and suppose without loss of generality that $\theta_s = 0$ and $\tilde{\theta}_s = 1$. Recall that

$$\begin{aligned} P_0(s | s, 0) &= \left(1 - \frac{1}{C(S-1)}\right) p + \frac{1}{C(S-1)}, \\ P_1(s | s, 0) &= \left(1 - \frac{1}{C(S-1)}\right) q + \frac{1}{C(S-1)}. \end{aligned}$$

Since $p > q \geq 1/2$ by (6.140), Lemma 2.12 tells us that

$$\begin{aligned} \text{KL}(P_0(\cdot | s, 0) \parallel P_1(\cdot | s, 0)) &\stackrel{(i)}{\leq} \frac{\left(1 - \frac{1}{C(S-1)}\right)^2 (p-q)^2}{\left(\left(1 - \frac{1}{C(S-1)}\right) p + \frac{1}{C(S-1)}\right) \left(1 - p - (1-p) \frac{1}{C(S-1)}\right)} \\ &\leq \frac{\left(1 - \frac{1}{C(S-1)}\right)^2 (p-q)^2}{p \left(1 - p - \frac{1}{C(S-1)}\right)} \\ &\stackrel{(ii)}{=} \frac{784(1-\gamma)^4 \varepsilon^2}{\gamma^2 \left(\gamma + \frac{14(1-\gamma)^2 \varepsilon}{\gamma}\right) \left(1 - \gamma - \frac{14(1-\gamma)^2 \varepsilon}{\gamma}\right)} \end{aligned}$$

$$\stackrel{\text{(iii)}}{\leq} \frac{1568(1-\gamma)^4 \varepsilon^2}{\gamma^3(1-\gamma)} \stackrel{\text{(iv)}}{\leq} 12544(1-\gamma)^3 \varepsilon^2.$$

Here, (i) follows from Lemma 2.12; (ii) uses the definitions of p and q in (6.139); (iii) holds whenever $\frac{224(1-\gamma)^2 \varepsilon}{\gamma} \leq \frac{1-\gamma}{2}$, which is guaranteed by (6.136); and (iv) follows from $1/2 \leq \gamma < 1$. The same upper bound clearly applies to $\text{KL}(P_0(\cdot | 0, 1) \| P_1(\cdot | 0, 1))$. Substituting these bounds into (6.153) shows that, if the sample size satisfies

$$N \leq \frac{C_\rho^*(S-1)}{802816(1-\gamma)^3 \varepsilon^2} \leq \frac{C(S-1)}{401408(1-\gamma)^3 \varepsilon^2}, \quad (6.155)$$

then it follows that

$$p_e \geq \frac{1}{4}. \quad (6.156)$$

Step 3: putting everything together. To complete the proof, suppose that there exists an estimator $\hat{\pi}$ such that

$$\mathbb{P}_\theta\{\langle \rho, V_\theta^* - V_{\hat{\theta}}^{\hat{\pi}} \rangle \leq \varepsilon\} \geq \frac{3}{4}.$$

Then, by the argument in Step 1, the estimator $\hat{\theta}$ defined in (6.149) must satisfy

$$\mathbb{P}_\theta(\hat{\theta} \neq \theta) < \frac{1}{4}.$$

This is impossible under the sample size condition (6.155), as it contradicts the lower bound in (6.156).

6.9.3 Proof of Lemma 6.15

We begin by deriving the value function at state $s > 0$ under an arbitrary policy π . By the Bellman equation,

$$\begin{aligned} V_\theta^\pi(s) &= \mathbb{E}_{a \sim \pi(\cdot | 0)} \left[r(s, a) + \gamma \sum_{s'} P_\theta(s' | s, a) V_\theta^\pi(s') \right] \\ &\stackrel{\text{(i)}}{\leq} 1 + \gamma \pi(\theta | 0) \left[p V_\theta^\pi(0) + (1-p)(C^{-1} V_\theta^*(s) + (1-C^{-1}) V_\theta^*(0)) \right] \\ &\quad + \gamma \pi(1-\theta | 0) \left[q V_\theta^\pi(0) + (1-q)(C^{-1} V_\theta^*(s) + (1-C^{-1}) V_\theta^*(0)) \right] \\ &\stackrel{\text{(i)}}{=} 1 + \gamma \left[x_{\pi, \theta}(s) V_\theta^\pi(0) + (1-x_{\pi, \theta}(s))(C^{-1} V_\theta^*(s) + (1-C^{-1}) V_\theta^*(0)) \right], \end{aligned} \quad (6.157)$$

where (i) follows from $V_\theta^*(s') = V_\theta^*(s)$ for $s, s' > 0$, and in (ii) we define

$$x_{\pi, \theta} = p\pi(\theta | 0) + q\pi(1-\theta | 0) = q + (p-q)\pi(\theta | 0). \quad (6.158)$$

Rearranging terms in (6.157) gives the desired relation (6.143). Moreover, the inequality holds when we take $\pi = \pi^*$.

We next analyze the value function at state 0. If $\pi(0 | 1) = 1$, then

$$V_\theta^\pi(0) = \frac{4}{7} + \gamma V_\theta^\pi(0) \quad \implies \quad V_\theta^\pi(0) = \frac{4}{7(1-\gamma)}.$$

For any policy π with $\pi(0|1) < 1$, we claim that $V_\theta^\pi(0) < \frac{4}{7(1-\gamma)}$. Otherwise, if $V_\theta^\pi(0) \geq \frac{4}{7(1-\gamma)}$, we have

$$\begin{aligned} V_\theta^\pi(0) &\leq \pi(0|0) \left(\frac{4}{7} + \gamma V_\theta^\pi(0) \right) + \pi(1|0) \gamma \left[2C^{-1} + ((2\gamma - 1) + 2(1 - \gamma)(1 - C^{-1})) V_\theta^\pi(0) \right] \\ &= \pi(0|0) \left(\frac{4}{7} + \gamma V_\theta^\pi(0) \right) + \pi(1|0) \gamma \left[\left(1 - \frac{2(1 - \gamma)}{C} \right) V_\theta^\pi(0) + \frac{2}{C} \right] \\ &\leq \frac{4}{7} \pi(0|0) + \gamma V_\theta^\pi(0) + \frac{6\gamma}{7C} \pi(1|1) \\ &< \frac{4}{7} + \gamma V_\theta^\pi(0), \end{aligned} \tag{6.159}$$

where the first inequality follows from the elementary bound $V_\theta^\pi(s) \leq \frac{1}{1-\gamma}$. Taken together, these two facts show that the optimal policy and the optimal value at state 1 satisfy

$$\pi_\theta^*(0|0) = 1 \quad \text{and} \quad V_\theta^*(1) = \frac{4}{7(1-\gamma)}.$$

It remains to identify $\pi_\theta^*(\cdot|s)$. From (6.143), we have

$$V_\theta^*(s) = C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0) + \frac{1 - (1 - \gamma)(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0))}{1 - \gamma x_{\pi_\theta^*, \theta}}.$$

Recognizing that

$$1 - (1 - \gamma)(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0)) > 1 - (1 - \gamma)\frac{1}{1-\gamma} = 0,$$

we see that the following function is increasing in x :

$$g(x) := C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0) + \frac{1 - (1 - \gamma)(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0))}{1 - \gamma x}. \tag{6.160}$$

In addition, $x_{\pi, \theta}$ in (6.158) is increasing in $\pi(\theta|0)$ since $p > q$. Therefore, the optimal policy satisfies

$$\pi_\theta^*(\theta_s|s) = 1.$$

6.9.4 Proof of auxiliary properties

Proof of (6.145). We first prove (6.145). To this end, with a slight abuse of notation, consider an MDP trajectory $\{(s_t, a_t)\}_{t \geq 0}$ initialized from $s_0 \sim \rho^b = \mu$. Under the behavior policy π^b , we have

$$\begin{aligned} \mathbb{P}\{s_1 = 0\} &= \mu(0) \left\{ \frac{1}{2} P_\theta(0|0, 0) + \frac{1}{2} P_\theta(0|0, 1) \right\} + \sum_{s=1}^{S-1} \mu(s) \left\{ \frac{1}{2} P_\theta(0|s, 0) + \frac{1}{2} P_\theta(0|s, 1) \right\} \\ &= \mu(0) \{ \gamma + (1 - \gamma)\mu(0) \} + \sum_{s=1}^{S-1} \mu(s)(1 - \gamma)\mu(0) \\ &= \mu(0), \end{aligned}$$

where the last identity follows from $\sum_{s=0}^{S-1} \mu(s) = 1$. According to the symmetry, one obtains $\mathbb{P}\{s_1 = s\} = \frac{1 - \mu(0)}{S-1} = \mu(s)$, and hence $s_1 \sim \mu$. Repeating the same argument shows that μ is a stationary distribution of the state process under π^b , so that $s_t \sim \mu$ for all $t \geq 0$. Consequently,

$$d^b(s) = (1 - \gamma) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s | s_0 \sim \rho^b; \pi^b) \right] = \mu(s), \quad \forall s \in \mathcal{S}.$$

In addition, since $\pi^b(a|s) = 1/2$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$d^b(s, a) = d^b(s) \pi^b(a|s) = \mu(s)/2. \tag{6.161}$$

Proof of (6.147). Consider the MDP $\mathcal{M}_\theta$, whose optimal policy π_θ^* satisfies $\pi_\theta^*(\theta_s | s) = 1$ for $s > 0$ by Lemma 6.15. With a slight abuse of notation, let $\{(s_t, a_t)\}_{t \geq 0}$ denote a trajectory generated under the policy π_θ^* , namely, $a_t \sim \pi_\theta^*(\cdot | s_t)$. Then we have

$$\begin{aligned} d^*(0) &= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = 0 | s_0 \sim \rho; \pi_\theta^*) \\ &\leq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t (1 - \gamma^t) < \frac{1}{2}. \end{aligned}$$

Here, the last line makes use of the following relation

$$\begin{aligned} \mathbb{P}(s_t = 0 | s_0 \sim \rho; \pi_\theta^*) &= \mathbb{P}(s_{t-1} = 0 | s_0 \sim \rho; \pi_\theta^*) + (1 - C^{-1})(1 - p) \mathbb{P}(s_{t-1} > 0 | s_0 \sim \rho; \pi_\theta^*) \\ &\leq \gamma \mathbb{P}(s_{t-1} = 0 | s_0 \sim \rho; \pi_\theta^*) + 1 - \gamma \\ &\leq (1 - \gamma) \sum_{k=0}^{t-1} \gamma^k = 1 - \gamma^t. \end{aligned}$$

According to symmetry, combining the above bound with (6.145) yields

$$\frac{d^*(s, \theta_s)}{d^b(s, \theta_s)} = \frac{2(1 - d^*(0))}{(S - 1)\mu(s)} \in (C, 2C), \quad \text{for } s > 0.$$

In addition, one can check that

$$\frac{d^*(0, 0)}{d^b(0, 0)} \leq \frac{1}{\mu(0)} \leq 2C.$$

Putting these bounds together, we establish the advertised result.

Proof of (6.148). The claim follows immediately by combining the relations below:

$$(1 - \gamma)(C^{-1}V_\theta^*(s) + (1 - C^{-1})V_\theta^*(0)) \leq C^{-1} + \frac{4}{7}(1 - C^{-1}) \leq \frac{6}{7},$$

$$(1 - \gamma p)(1 - \gamma x_{\pi, \theta}(s)) \leq (1 - \gamma p)(1 - \gamma q) \leq (1 - \gamma^2)^2,$$

and

$$\gamma(p - x_{\pi, \theta}(s)) = 448(1 - \gamma)^2 \varepsilon (1 - \pi(\theta_s | s)).$$

6.10 Notes

While the core ideas underlying batch RL can be traced back more than two decades (Ernst *et al.*, 2005; Lange *et al.*, 2012), this subfield of RL has witnessed a remarkable resurgence since around 2020, largely under the banner of offline RL (Kumar *et al.*, 2020; Levine *et al.*, 2020; Agarwal *et al.*, 2020b). This renewed interest has catalyzed a substantial body of work across computer science, statistics, operations research, and related disciplines on the statistical and algorithmic foundations of offline RL, motivated by the growing range of applications in which real-time data collection is expensive, safety-critical, or otherwise constrained. We refer the interested reader to Jin *et al.* (2021b), Rashidinejad *et al.* (2021), Xie *et al.* (2021b), Yin and Wang (2021a), Ma *et al.* (2021), Uehara and Sun (2022), Uehara *et al.* (2021), Ren *et al.* (2021), Ma *et al.* (2022), Sonabend *et al.*

(2020), Yin and Wang (2021b), Xie *et al.* (2021a), Yin *et al.* (2021b), Shi *et al.* (2022c), Lyu *et al.* (2022), Yan *et al.* (2023), Zhan *et al.* (2022), Zanette and Wainwright (2022), Zhou *et al.* (2023b), Ye *et al.* (2023), Bhardwaj *et al.* (2023), Li *et al.* (2024b), Yin *et al.* (2022b), Yin *et al.* (2021a), Xu *et al.* (2024), Zhu *et al.* (2024), Zhou *et al.* (2024), Jin *et al.* (2025), and Jiang and Xie (2025) for some examples of recent developments, many of which pursue perspectives that are complementary to the ones presented in this chapter.

A recurring theme in offline RL is to identify the fundamental conditions under which a near-optimal policy can be learned from a historical dataset. Early analyses based on concepts of the concentrability coefficients typically require the offline data distribution to provide adequate coverage of all admissible policies, or equivalently, of all reachable state-action pairs (Munos, 2007; Farahmand *et al.*, 2010; Le *et al.*, 2019; Fan *et al.*, 2019; Chen and Jiang, 2019; Yin *et al.*, 2021b; Yin *et al.*, 2021a). This condition, commonly referred to as the *all-policy concentrability* requirement, is often overly restrictive in practice, particularly in high-dimensional settings where adequate coverage of the entire policy space is difficult to guarantee.

Subsequent work sought to relax this stringent requirement by arguing that the offline dataset may only need to provide sufficient coverage of the state-action pairs visited by the target policy (Liu *et al.*, 2020b). Building on this insight, Jin *et al.* (2021b) and Rashidinejad *et al.* (2021) developed offline RL algorithms and analyses based on the notion of *single-policy concentrability*, whereby the coverage requirement is imposed only along the state-action distribution induced by the target policy, as discussed in this chapter. For tabular MDPs, the full-range minimax sample complexity of offline RL under single-policy concentrability was first characterized by Li *et al.* (2024b), who showed the optimality of model-based algorithms for both discounted infinite-horizon MDPs and finite-horizon nonstationary MDPs without imposing any burn-in cost. In addition, Yin and Wang (2021b) and Yin *et al.* (2022b) established instance-dependent statistical guarantees for pessimistic model-based algorithms, revealing how the statistical complexity of offline RL is governed by the variance structure of the problem instance (albeit at the cost of a large burn-in sample size requirement). The more general offline RL setting with function approximation has received much attention as well; see, e.g., Jin *et al.* (2021b), Zanette *et al.* (2021), Yin *et al.* (2022b), Xie *et al.* (2021a), and Di *et al.* (2024). More broadly, the pessimism idea has since found applications in other statistical tasks, such as learning optimal individualized decision rules (Jin *et al.*, 2025) and learning optimal dynamic treatment regimes (Zhou *et al.*, 2023a).

From an algorithmic perspective, the methods discussed in this chapter primarily implement the pessimism principle by penalizing the estimated values of under-covered state-action pairs with the aid of uncertainty quantification. Alternative approaches have also been proposed that enforce pessimism through an adversarial learning perspective. For example, a line of work (Xie *et al.*, 2021a; Uehara and Sun, 2022; Zhan *et al.*, 2022; Wang *et al.*, 2024d; Panaganti *et al.*, 2025; Cen *et al.*, 2025) has suggested searching for the *least favorable model* —that is, the model yielding the lowest value function — among all models that are statistically consistent with the offline data, typically through regularized or constrained optimization formulations.

While the discussion thus far has focused on learning a near-optimal policy from offline data, another major and closely related line of research is *off-policy evaluation* (OPE), whose goal is to estimate the value of a target policy using data generated by a potentially different behavior policy. OPE has evolved into a rich research area in its own right, with deep connections to statistics, causal inference, semiparametric efficiency theory, and related fields. A comprehensive treatment of OPE falls beyond the scope of this chapter; we refer the interested reader to Uehara *et al.* (2022b),

Li *et al.* (2014), Jiang and Li (2016), Thomas and Brunskill (2016), Xie *et al.* (2019), Uehara *et al.* (2020), Yang *et al.* (2020a), Jiang and Huang (2020), Duan *et al.* (2020), Liao *et al.* (2021), Ren *et al.* (2021), Yin and Wang (2020), Yin *et al.* (2021b), Kallus *et al.* (2022), Zhan *et al.* (2021), Xu *et al.* (2022), Hu and Wager (2023), Sonabend-W *et al.* (2023), Mehrabi and Wager (2024), Shi *et al.* (2024a), Bian *et al.* (2025), and Qi *et al.* (2025), as well as the references therein, for detailed discussions.

Extensions

Thus far, we have focused on three canonical — and arguably the most widely studied — formulations of single-agent RL. Many modern applications, however, depart from these classical settings in important ways, giving rise to a rich collection of emerging frameworks that await further investigation. In this chapter, we briefly survey a few representative extensions along these lines, including distributionally robust RL, multi-agent RL, federated RL, statistical inference in RL, and RL with human feedback. Our treatment is necessarily brief and offers only a bird’s-eye view of these rapidly evolving topics. Rather than aiming for an exhaustive account, the purpose is to shed light on how the non-asymptotic perspective developed throughout the previous chapters continues to inform both the statistical principles and the algorithm designs in settings that extend beyond the basic MDP framework.

7.1 Distributionally robust RL

Despite the extensive study of standard RL, its practical deployment is often hindered by the sim-to-real gap and/or various forms of environmental uncertainty (Bertsimas *et al.*, 2019). For instance, a policy learned in an ideal, nominal environment may fail catastrophically when deployed in a setting subject to even small changes in task objectives or adversarial perturbations to the system dynamics (Zhang *et al.*, 2020a; Klopp *et al.*, 2017; Mahmood *et al.*, 2018). As a result, robustness has emerged as a critical objective alongside reward maximization, particularly in high-stakes applications such as robotics, autonomous driving, clinical trials, and financial investments.

To address these challenges, distributionally robust RL (Iyengar, 2005; Nilim and El Ghaoui, 2005; Xu and Mannor, 2012; Bäuerle and Glauner, 2022; Cai *et al.*, 2016) has emerged as a natural yet versatile framework, drawing inspiration and insight from the broader literature on distributionally robust optimization and supervised learning (Rahimian and Mehrotra, 2019; Gao, 2023; Bertsimas *et al.*, 2018; Duchi and Namkoong, 2021; Blanchet and Murthy, 2019; Chen *et al.*, 2019d; Lam, 2019). The objective is to learn a policy that maintains strong performance even when the deployed environment deviates from the nominal one used for training, thereby enhancing robustness to environmental uncertainty and distributional shifts. In what follows, we briefly discuss, through the lens of sample complexity, whether and how the incorporation of distributional robustness affects

the statistical bottlenecks in learning a desirable policy.

7.1.1 Model and preliminaries

In contrast to standard RL, which seeks to learn an optimal policy for a single MDP using samples generated from that MDP, distributionally robust RL aims to learn a policy that remains effective in the presence of model uncertainty. At a high level, using the same collection of samples, it seeks to optimize worst-case performance over a prescribed uncertainty set of transition kernels centered around the nominal kernel. We make this formulation more precise below. For expositional simplicity, we focus on the discounted infinite-horizon setting.

Distributionally robust MDPs (RMDPs). We now introduce the distributionally robust MDP (RMDP) tailored to the discounted infinite-horizon setting. Specifically, an RMDP is formally defined by a tuple

$$\mathcal{M}_{\text{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^\sigma(P^{\text{nom}}), r\},$$

where the state space $\mathcal{S} = \{1, \dots, S\}$, the action space $\mathcal{A} = \{1, \dots, A\}$, the discount factor $0 < \gamma < 1$, and the reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ remain identical to those in a standard MDP. The key distinction lies in the transition kernel:¹ rather than assuming a fixed transition kernel P , the actual transition kernel can deviate arbitrarily within a prescribed uncertainty set $\mathcal{U}^\sigma(P^{\text{nom}})$ centered around a *nominal* kernel $P^{\text{nom}} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$.

Compared with standard MDPs, RMDPs provide a more flexible modeling framework by allowing one to specify both the shape and size of the uncertainty set around the nominal transition kernel. To ensure computational tractability for solving RMDPs, however, the uncertainty set is often chosen to possess additional structure. A common structure in the literature is that the uncertainty set decomposes into a product of independent local uncertainty subsets over states or state-action pairs, known respectively as *s-rectangularity* and *(s, a)-rectangularity* (Zhou *et al.*, 2021c; Wiesemann *et al.*, 2013; Wang *et al.*, 2023b).

For ease of exposition, we focus on the *(s, a)-rectangular* setting. Specifically, let $\rho : \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \rightarrow \mathbb{R}_{\geq 0}$ be a distance-like metric, and let $\sigma > 0$ represent a prescribed radius that controls the level of uncertainty. The uncertainty set is then defined as

$$\mathcal{U}^\sigma(P^{\text{nom}}) := \otimes_{s,a} \mathcal{U}^\sigma(P_{s,a}^{\text{nom}}), \quad (7.1a)$$

where for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, the local uncertainty set is given by

$$\mathcal{U}^\sigma(P_{s,a}^{\text{nom}}) := \left\{ P_{s,a} \in \Delta(\mathcal{S}) : \rho(P_{s,a}, P_{s,a}^{\text{nom}}) \leq \sigma \right\}, \quad (7.1b)$$

with $P_{s,a}$ and $P_{s,a}^{\text{nom}}$ representing the row vectors of P and P^{nom} associated with the state-action pair (s, a) , respectively. In other words, the uncertainty is imposed in a decoupled manner across state-action pairs.

In particular, two popular instances of (7.1) arise when the metric ρ is chosen to be the total-variation (TV) distance, namely $\rho(\cdot, \cdot) = \text{TV}(\cdot, \cdot)$ (see (1.2a)), or the χ^2 divergence, namely $\rho(\cdot, \cdot) = \chi^2(\cdot \| \cdot)$ (see (1.2c)). These two choices will serve as the primary focus of this section. As we shall see shortly, these two choices of divergence metrics have markedly different statistical consequences when it comes to sample complexities.

¹Although uncertainty in the reward function can also be incorporated into the framework, we omit it here for simplicity and focus exclusively on uncertainty in the transition kernel, which constitutes the primary technical challenge.

Robust value functions and Q-functions. In an RMDP, the performance of a policy π is evaluated according to its worst-case performance over all transition kernels in the uncertainty set. This leads to the notions of the *robust value function* $V^{\pi,\sigma}$ and the *robust Q-function* $Q^{\pi,\sigma}$ in $\mathcal{M}_{\text{rob}}$, defined respectively as

$$V^{\pi,\sigma}(s) := \inf_{P \in \mathcal{U}^\sigma(P^{\text{nom}})} V^{\pi,P}(s) \quad (7.2a)$$

$$Q^{\pi,\sigma}(s, a) := \inf_{P \in \mathcal{U}^\sigma(P^{\text{nom}})} Q^{\pi,P}(s, a) \quad (7.2b)$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here, $V^{\pi,P}$ and $Q^{\pi,P}$ denote the value function and Q-function, respectively, of policy π in the standard MDP with transition kernel P . Thus, $V^{\pi,\sigma}$ and $Q^{\pi,\sigma}$ quantify the worst-case discounted cumulative rewards attainable under π when the transition dynamics are allowed to vary within the uncertainty set $\mathcal{U}^\sigma(P^{\text{nom}})$.

Optimal robust policy and robust Bellman operator. As a fundamental generalization of the dynamic programming framework of standard MDPs, there exists at least one deterministic policy that maximizes the robust value function (resp. robust Q-function) simultaneously across all states (resp. state-action pairs) (Iyengar, 2005; Nilim and El Ghaoui, 2005). Accordingly, we denote such a policy (often referred to as an optimal robust policy) by π^* , and let $V^{*,\sigma}$ and $Q^{*,\sigma}$ denote the corresponding *optimal robust value function* and *optimal robust Q-function*, respectively, which satisfy

$$\forall s \in \mathcal{S} : V^{*,\sigma}(s) := V^{\pi^*,\sigma}(s) = \max_{\pi} V^{\pi,\sigma}(s), \quad (7.3a)$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : Q^{*,\sigma}(s, a) := Q^{\pi^*,\sigma}(s, a) = \max_{\pi} Q^{\pi,\sigma}(s, a). \quad (7.3b)$$

A key machinery in RMDPs is the robust counterpart of Bellman's optimality principle, embodied in the following *robust Bellman consistency equation* and *robust Bellman optimality equation*:

$$Q^{\pi,\sigma}(s, a) = r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^{\text{nom}}_{s,a})} \langle P_{s,a}, V^{\pi,\sigma} \rangle, \quad (7.4a)$$

$$Q^{*,\sigma}(s, a) = r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^{\text{nom}}_{s,a})} \langle P_{s,a}, V^{*,\sigma} \rangle \quad (7.4b)$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, where π stands for any policy. Correspondingly, the robust Bellman operator (Iyengar, 2005; Nilim and El Ghaoui, 2005), denoted by $\mathcal{T}^\sigma(\cdot) : \mathbb{R}^{\mathcal{S}\mathcal{A}} \rightarrow \mathbb{R}^{\mathcal{S}\mathcal{A}}$, is defined by

$$\mathcal{T}^\sigma(Q)(s, a) := r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^{\text{nom}}_{s,a})} \langle P_{s,a}, V \rangle \quad (7.5a)$$

for any input vector $Q \in \mathbb{R}^{\mathcal{S}\mathcal{A}}$ and every $(s, a) \in \mathcal{S} \times \mathcal{A}$, where, as usual,

$$\forall s \in \mathcal{S} : V(s) := \max_a Q(s, a). \quad (7.5b)$$

It is well known that $Q^{*,\sigma}$ is the unique fixed point of $\mathcal{T}^\sigma$. Consequently, the optimal robust value function and Q-function can be computed via *distributionally robust value iteration* (DRVI), which is a direct generalization of the classical value iteration algorithm by recursively applying the robust Bellman operator starting from an arbitrary initialization. Analogous to standard value iteration, DRVI converges geometrically fast due to the γ -contraction property of $\mathcal{T}^\sigma$ w.r.t. the ℓ_∞ norm (Iyengar, 2005; Nilim and El Ghaoui, 2005).

Sampling mechanism and goal. To keep the discussion concrete and avoid the complications arising from different data collection schemes, we restrict attention in this section to the generative model as described in Chapter 3, which already captures several key distinctions between the standard MDPs and their distributionally robust counterparts.

Specifically, we draw N independent samples for each state-action pair from the generative model according to the nominal transition kernel P^{nom} . That is, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, we collect

$$s^{(i)}(s, a) \stackrel{\text{ind.}}{\sim} P^{\text{nom}}(\cdot | s, a), \quad i = 1, \dots, N,$$

resulting in a total sample size $N^{\text{total}} := NSA$.

Given these samples, the objective is to learn an ε -optimal robust policy $\hat{\pi}$ satisfying

$$\forall s \in \mathcal{S} : \quad V^{\star, \sigma}(s) - V^{\hat{\pi}, \sigma}(s) \leq \varepsilon. \quad (7.6)$$

Compared with the standard RL setting, the data are collected in exactly the same manner; the key difference lies in the performance criterion, which is now defined in terms of robust value functions.

Having formalized the problem, a natural question is to quantify the impact of robustness on sample complexity. The remainder of this section is devoted to discussing the statistical price of robustness and illustrating how it varies across different choices of uncertainty sets.

7.1.2 A model-based algorithm

In this section, we focus on a model-based approach for RMDPs that attains state-of-the-art statistical guarantees. At a high level, this algorithm first constructs an empirical estimate of the nominal transition kernel using the collected samples and subsequently applies DRVI to compute an optimal robust policy for the resulting empirical model.

Empirical nominal kernel. The empirical nominal transition kernel $\hat{P}^{\text{nom}} \in \mathbb{R}^{SA \times S}$ can be constructed from the empirical transition frequencies. Specifically, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, compute

$$\forall s' \in \mathcal{S} : \quad \hat{P}^{\text{nom}}(s' | s, a) := \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{s' = s^{(i)}(s, a)\}, \quad (7.7)$$

which coincides with the empirical model estimation procedure used in standard RL (see Chapter 3). Replacing the nominal transition kernel P^{nom} with its empirical counterpart $\hat{P}^{\text{nom}}$ yields the empirical RMDP

$$\hat{\mathcal{M}}_{\text{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}(\hat{P}^{\text{nom}}), r\}.$$

As in the population setting, let $\hat{V}^{\pi, \sigma}$ and $\hat{Q}^{\pi, \sigma}$ denote the robust value function and robust Q-function of policy π in $\hat{\mathcal{M}}_{\text{rob}}$, respectively (analogous to (7.3)). In addition, we denote the corresponding optimal robust policy as $\hat{\pi}^{\star}$ and the optimal robust value function (resp. optimal robust Q-function) as $\hat{V}^{\star, \sigma}$ (resp. $\hat{Q}^{\star, \sigma}$) (analogous to (7.4)). The corresponding robust Bellman optimality equation is given by

$$\hat{Q}^{\star, \sigma}(s, a) = r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^{\sigma}(\hat{P}_{s,a}^{\text{nom}})} \langle P_{s,a}, \hat{V}^{\star, \sigma} \rangle \quad (7.8)$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Algorithm 7.1: Distributionally robust value iteration (DRVI) for discounted infinite-horizon RMDPs

```

1 input: discount factor  $\gamma$ , reward function  $r$ , uncertainty level  $\sigma$ , number of iterations  $T$ .
2 initialize  $\hat{Q}_0(s, a) \leftarrow 0$ ,  $\hat{V}_0(s) \leftarrow 0$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .
3 for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
4   draw  $N$  independent samples  $s^{(i)}(s, a) \stackrel{\text{i.i.d.}}{\sim} P^{\text{nom}}(\cdot | s, a)$ .
5   compute  $\hat{P}^{\text{nom}}(\cdot | s, a)$  based on (7.7).
6 for  $t = 1, 2, \dots, T$  do
7   for  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
8     compute  $\hat{Q}_t(s, a)$  according to (7.10);
9   for  $s \in \mathcal{S}$  do
10    compute  $\hat{V}_t(s) = \max_a \hat{Q}_t(s, a)$ ;
11 output: policy  $\hat{\pi}$  obeying  $\hat{\pi}(s) := \arg \max_a \hat{Q}_T(s, a)$ .

```

Equipped with $\hat{P}^{\text{nom}}$, we can define the empirical robust Bellman operator $\hat{\mathcal{T}}^\sigma$ such that: for any input vector $Q \in \mathbb{R}^{\mathcal{S}\mathcal{A}}$ and each $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\hat{\mathcal{T}}^\sigma(Q)(s, a) := r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(\hat{P}_{s,a}^{\text{nom}})} \langle P_{s,a}, V \rangle, \quad (7.9)$$

where $V(s) := \max_a Q(s, a)$ for all $s \in \mathcal{S}$.

Distributionally robust value iteration. To compute the fixed point of $\hat{\mathcal{T}}^\sigma(\cdot)$, we introduce distributionally robust value iteration (DRVI). Starting from an initialization $\hat{Q}_0 = 0$, the update rule at the t -th ($t \geq 1$) iteration is given by

$$\begin{aligned} \hat{Q}_t(s, a) &= \hat{\mathcal{T}}^\sigma(\hat{Q}_{t-1})(s, a) \\ &= r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(\hat{P}_{s,a}^{\text{nom}})} \langle P_{s,a}, \hat{V}_{t-1} \rangle \end{aligned} \quad (7.10)$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\hat{V}_{t-1}(s) = \max_a \hat{Q}_{t-1}(s, a)$ for all $s \in \mathcal{S}$. After running the update for T iterations, we output the greedy policy $\hat{\pi}$ induced by $\hat{Q}_T$, namely,

$$\forall s \in \mathcal{S} : \quad \hat{\pi}(s) = \arg \max_a \hat{Q}_T(s, a). \quad (7.11)$$

The complete procedure is summarized in Algorithm 7.1. As in standard value iteration, the iterates $\{\hat{Q}_t\}_{t \geq 0}$ of DRVI converge linearly to the unique fixed point $\hat{Q}^{*,\sigma}$, owing to the γ -contraction property of the empirical robust Bellman operator $\hat{\mathcal{T}}^\sigma(\cdot)$.

Nevertheless, a direct implementation of (7.10) can be computationally expensive, as it requires solving an optimization problem over an S -dimensional probability simplex for every state-action pair at each iteration. Fortunately, by strong duality (Iyengar, 2005), the inner optimization problem in (7.10) admits an equivalent dual formulation involving only a *scalar* dual variable. Consequently, each update in DRVI can be carried out efficiently by solving a one-dimensional optimization problem. For concreteness, below we illustrate this idea for the two choices of ρ : TV distance and χ^2 divergence.

To this end, for any $V \in \mathbb{R}^S$ and any $\alpha \geq 0$, define the clipped vector $[V]_\alpha$ as

$$\forall s \in \mathcal{S} : \quad [V]_\alpha(s) := \begin{cases} \alpha, & \text{if } V(s) > \alpha, \\ V(s), & \text{otherwise.} \end{cases} \quad (7.12)$$

The lemma below, which follows directly from Iyengar (2005, Lemma 4.3), provides an equivalent dual representation for the inner optimization problem in (7.10).

Lemma 7.1 (Strong duality). Consider any given probability vector $p^{\text{nom}} \in \Delta(\mathcal{S})$, any uncertainty level σ , and the uncertainty set $\mathcal{U}^\sigma(p^{\text{nom}})$. Consider any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, recall the definition of $[V]_\alpha$ in (7.12), and set $V_{\min} = \min_s V(s)$, $V_{\max} = \max_s V(s)$.

- If ρ is taken to be the TV distance, then

$$\inf_{p \in \mathcal{U}^\sigma(p^{\text{nom}})} \langle p, V \rangle = \max_{\alpha \in [V_{\min}, V_{\max}]} \left\{ \langle p^{\text{nom}}, [V]_\alpha \rangle - \sigma \left(\alpha - \min_{s'} [V]_\alpha(s') \right) \right\}. \quad (7.13)$$

- If ρ is taken to be the χ^2 divergence, then

$$\inf_{p \in \mathcal{U}^\sigma(p^{\text{nom}})} \langle p, V \rangle = \max_{\alpha \in [V_{\min}, V_{\max}]} \left\{ \langle p^{\text{nom}}, [V]_\alpha \rangle - \sqrt{\sigma \text{Var}_{p^{\text{nom}}}([V]_\alpha)} \right\}, \quad (7.14)$$

where we recall the definition of $\text{Var}_p(V)$ in (2.22).

7.1.3 Theoretical guarantees: the case of TV distance

We begin with the setting in which the uncertainty set is defined using the total-variation distance. The following theorem develops an upper bound on the sample complexity of DRVI for identifying an ε -optimal robust policy.

Theorem 7.2 (Shi *et al.* (2023) and Shi *et al.* (2026)). Let the metric ρ be the TV distance. Consider any discount factor $\gamma \in [1/4, 1)$, uncertainty level $\sigma \in (0, 1)$, $\delta \in (0, 1)$, and $\varepsilon \in (0, \sqrt{1/\max\{1-\gamma, \sigma\}}]$. Let $\hat{\pi}$ be the policy returned by Algorithm 7.1 after $T = C_1 \log(\frac{N}{1-\gamma})$ iterations. Then, with probability at least $1 - \delta$, one has

$$\forall s \in \mathcal{S} : \quad V^{*,\sigma}(s) - V^{\hat{\pi},\sigma}(s) \leq \varepsilon,$$

provided that the number of samples per state-action pair obeys

$$N \geq \frac{C_2}{(1-\gamma)^2 \max\{1-\gamma, \sigma\} \varepsilon^2} \log \left(\frac{SAN}{(1-\gamma)\delta} \right).$$

Here, $C_1, C_2 > 0$ are some large enough universal constants.

Remark 7.1. Note that Theorem 7.2 is not only valid when invoking Algorithm 7.1. In fact, the theorem holds for any oracle planning algorithm (designed based on the empirical transitions $\hat{P}^{\text{nom}}$) whose output policy $\hat{\pi}$ obeys

$$\|\hat{V}^{*,\sigma} - \hat{V}^{\hat{\pi},\sigma}\|_\infty \leq \tilde{O}\left(\frac{(1-\gamma)^2}{N}\right). \quad (7.15)$$

Thus, the statistical guarantee is largely independent of the specific planning algorithm employed, provided that the empirical planning error is sufficiently small.

Next, we present a matching minimax lower bound. Taken collectively, these upper and lower bounds characterize the sample complexity of learning RMDPs under TV-based uncertainty and establish the statistical optimality of the model-based algorithm DRVI.

Theorem 7.3 (Shi *et al.*, 2023; Shi *et al.*, 2026). Let ρ be the TV distance. Consider any tuple $(S, A, \gamma, \sigma, \varepsilon)$ obeying $\sigma \in (0, 1 - c_0]$, with $0 < c_0 \leq 1/8$ being any small enough positive constant. Suppose further that $\gamma \in [1/2, 1)$ and $\varepsilon \in (0, \frac{c_0}{256(1-\gamma)}]$. Then there exist two γ -discounted infinite-horizon RMDPs $\mathcal{M}_0$ and $\mathcal{M}_1$, an initial state distribution μ , and a dataset with N independent samples for each state-action pair over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that

$$\inf_{\hat{\pi}} \max \left\{ \mathbb{P}_0(V^{*,\sigma}(\mu) - V^{\hat{\pi},\sigma}(\mu) > \varepsilon), \mathbb{P}_1(V^{*,\sigma}(\mu) - V^{\hat{\pi},\sigma}(\mu) > \varepsilon) \right\} \geq \frac{1}{8},$$

provided that

$$N \leq \frac{c_0 \log 2}{8192(1-\gamma)^2 \max\{1-\gamma, \sigma\} \varepsilon^2}.$$

Here, the infimum is over all estimators $\hat{\pi}$, and $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) denotes the probability distribution induced by the RMDP $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

Implications. We next interpret the preceding upper and lower bounds and discuss their implications for the sample complexity of learning RMDPs when the uncertainty set is specified via the TV distance.

- *Minimax-optimal sample complexity.* Theorem 7.2 shows that DRVI (or any oracle planning algorithm satisfying the condition in Remark 7.1) achieves ε -accuracy using

$$\tilde{O} \left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\} \varepsilon^2} \right) \quad (7.16)$$

samples in total. Combined with the lower bound in Theorem 7.3, this establishes the minimax optimality of DRVI (up to logarithmic factors) over a broad range of uncertainty levels σ . Importantly, the sample complexity scales linearly with the dimension of the state-action space and improves as the uncertainty level σ increases in the regime $\sigma \gtrsim 1 - \gamma$.

- *RMDPs can be easier to learn than standard MDPs under TV uncertainty.* Recall from Chapter 3 that learning a standard γ -discounted MDP with access to a generative model requires

$$\tilde{O} \left(\frac{SA}{(1-\gamma)^3 \varepsilon^2} \right) \quad (7.17)$$

samples to achieve ε accuracy. Comparing this with the RMDP sample complexity (7.16), we observe that learning RMDPs under TV uncertainty is at least as easy as — if not easier than — learning standard MDPs. In particular, when $\sigma \lesssim 1 - \gamma$ is small, the sample complexity for learning RMDPs coincides with (7.17) for learning standard MDPs, which is as expected since the RMDP approaches the standard MDP when $\sigma = 0$. In contrast, when $1 - \gamma \lesssim \sigma < 1$, the sample complexity for learning RMDPs simplifies to

$$\tilde{O} \left(\frac{SA}{(1-\gamma)^2 \sigma \varepsilon^2} \right),$$

which improves upon the standard MDP counterpart by a factor of $\sigma/(1-\gamma)$.

7.1.4 Theoretical guarantees: the case of χ^2 divergence

We now turn to the setting in which the uncertainty set is specified by means of the χ^2 divergence. As we shall see, the resulting sample complexity guarantees differ markedly from those obtained under TV-based uncertainty. The following theorem summarizes the corresponding upper bound.

Theorem 7.4 (Shi *et al.*, 2023; Shi *et al.*, 2026). Let the metric ρ be the χ^2 divergence. Consider any uncertainty level $\sigma \in (0, \infty)$, $\gamma \in [1/4, 1)$, $\delta \in (0, 1)$, and $\varepsilon \in (0, \frac{1}{1-\gamma}]$. With probability at least $1 - \delta$, the policy $\hat{\pi}$ returned by Algorithm 7.1 after $T = c_1 \log(\frac{N}{1-\gamma})$ iterations satisfies

$$\forall s \in \mathcal{S} : \quad V^{*,\sigma}(s) - V^{\hat{\pi},\sigma}(s) \leq \varepsilon,$$

provided that the number of samples per state-action pair obeys

$$N \geq \frac{c_2}{(1-\gamma)^3 \varepsilon^2} \left(1 + \frac{\max\{\sqrt{\sigma}, \sigma\}}{1-\gamma} \right) \log \left(\frac{SAN}{\delta} \right).$$

Here, $c_1, c_2 > 0$ are some large enough universal constants.

Remark 7.2. Analogous to Remark 7.1, the sample complexity guarantee in Theorem 7.4 extends beyond DRVI and remains valid for any oracle planning algorithm whose output policy $\hat{\pi}$ satisfies $\|\hat{V}^{*,\sigma} - \hat{V}^{\hat{\pi},\sigma}\|_\infty \leq O\left(\frac{\log(\frac{SAN}{(1-\gamma)\delta})}{N^2}\right)$.

In addition, to gauge the tightness of Theorem 7.4 and pin down the sample complexity limit under χ^2 -based uncertainty, we next present a minimax lower bound.

Theorem 7.5 (Shi *et al.*, 2023; Shi *et al.*, 2026). Let ρ be the χ^2 divergence. Consider any $(S, A, \gamma, \sigma, \varepsilon)$ obeying $\gamma \in [3/4, 1)$, $\sigma \in (0, \infty)$, and $\varepsilon \leq \frac{c_3}{(1-\gamma)}$ for some small universal constant $c_3 > 0$. Then there exist two discounted infinite-horizon RMDPs $\mathcal{M}_0$ and $\mathcal{M}_1$, an initial state distribution μ , and a dataset with N independent samples per (s, a) pair over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that

$$\inf_{\hat{\pi}} \max \left\{ \mathbb{P}_0(V^{*,\sigma}(\mu) - V^{\hat{\pi},\sigma}(\mu) > \varepsilon), \mathbb{P}_1(V^{*,\sigma}(\mu) - V^{\hat{\pi},\sigma}(\mu) > \varepsilon) \right\} \geq \frac{1}{8},$$

provided that the number of samples per state-action pair obeys

$$N \leq \frac{c_4}{(1-\gamma)^3 \varepsilon^2} \left(1 + \frac{\sigma}{1-\gamma} \right)$$

for some universal constant $c_4 > 0$.

Implications. We are now ready to discuss several important implications of the above theorems.

- *Nearly tight sample complexity.* In order to achieve ε -accuracy for RMDPs under the χ^2 divergence, Theorem 7.4 asserts that a total number of samples on the order of

$$\tilde{O} \left(\frac{SA}{(1-\gamma)^3 \varepsilon^2} \left(1 + \frac{\max\{\sqrt{\sigma}, \sigma\}}{1-\gamma} \right) \right) \quad (7.18)$$

is sufficient for DRVI (or any other oracle planning algorithm as discussed in Remark 7.2). Taking this together with the minimax lower bound in Theorem 7.5 confirms that the sample complexity is optimal over a broad range of the uncertainty level σ (when $\sigma \lesssim (1-\gamma)^2$ and $\sigma \gtrsim 1$). In particular,

- when $\sigma \lesssim (1-\gamma)^2$, the sample complexity upper bound $\tilde{O}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$ is sharp and matches the minimax lower bound;
 - when $(1-\gamma)^2 \lesssim \sigma \lesssim 1$, the sample complexity upper bound is on the order of $\tilde{O}\left(\frac{SA\sqrt{\sigma}}{(1-\gamma)^4\epsilon^2}\right)$, leaving a gap of at most $\sqrt{\frac{1}{1-\gamma}}$ from the minimax lower bound;
 - when $\sigma \gtrsim 1$, the sample complexity upper bound $\tilde{O}\left(\frac{SA\sigma}{(1-\gamma)^4\epsilon^2}\right)$ again matches the minimax lower bound.
- *RMDPs are harder to learn than standard MDPs under χ^2 uncertainty.* The minimax lower bound developed in Theorem 7.5 reveals a striking contrast with the TV-distance case. Whereas TV uncertainty can make robust MDPs statistically easier to learn, χ^2 uncertainty never reduces the sample complexity relative to standard MDPs. Indeed, several distinct regimes are worth highlighting:

- when $\sigma \lesssim 1 - \gamma$, the lower bound reduces to

$$\tilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right),$$

which coincides with the sample complexity of standard MDPs;

- when $\sigma \gtrsim 1 - \gamma$, the lower bound becomes

$$\tilde{\Omega}\left(\frac{SA\sigma}{(1-\gamma)^4\epsilon^2}\right),$$

which exceeds the standard MDP sample complexity by a factor of $\sigma/(1-\gamma)$.

7.1.5 Why are the TV and χ^2 cases markedly different?

The preceding results reveal a surprising phenomenon: the sample complexity of robust RL depends critically on both the size and the geometry of the uncertainty set. To gain intuition, consider how estimation errors propagate through a single value-iteration update (cf. (7.10)) from a fixed value function V , when the nominal transition kernel P^{nom} is replaced by its empirical estimate $\hat{P}^{\text{nom}}$. This leads to the following error terms in standard RL and robust RL, respectively:

$$\begin{aligned} \text{standard RL: } \delta_{\text{standard}} &= \left| \langle P_{s,a}^{\text{nom}}, V \rangle - \langle \hat{P}_{s,a}^{\text{nom}}, V \rangle \right|, \\ \text{robust RL: } \delta_{\text{robust}} &= \left| \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^{\text{nom}})} \langle P_{s,a}, V \rangle - \inf_{P_{s,a} \in \mathcal{U}^\sigma(\hat{P}^{\text{nom}})} \langle P_{s,a}, V \rangle \right|. \end{aligned}$$

As a key distinction, the estimation error in standard RL is linear in the model estimation error $P^{\text{nom}} - \hat{P}^{\text{nom}}$, whereas in robust RL, it is governed by an additional inner optimization over the uncertainty set, resulting in a generally nonlinear dependence that often lacks a closed-form characterization. We provide some brief remarks regarding δ_{robust} under both TV distance and χ^2 divergence.

- *TV uncertainty.* For uncertainty sets based on the TV distance, the dual representation in (7.13) reveals that δ_{robust} behaves approximately linearly in $P^{\text{nom}} - \hat{P}^{\text{nom}}$, much like δ_{standard} . Intuitively, this is because TV-based uncertainty sets have a homogeneous shape across all states and exhibit an almost instance-independent structure (w.r.t. uncertainty set center).

Moreover, when $\sigma \gtrsim 1 - \gamma$, the robust value function of a fixed policy exhibits a substantially smaller dynamic range than its standard-RL counterpart. This favorable low-variance statistical property ultimately leads to improved data efficiency and explains why robust RL can require fewer samples than standard RL under TV uncertainty.

- χ^2 *uncertainty*. The situation is fundamentally different for uncertainty sets based on the χ^2 divergence. In this case, δ_{robust} becomes highly nonlinear and can significantly amplify the model estimation error $P^{\text{nom}} - \hat{P}^{\text{nom}}$. This effect stems from the heterogeneous structure of the χ^2 uncertainty sets, which varies across states and depends strongly on the specific center of the set (P^{nom} or $\hat{P}^{\text{nom}}$). As a result, small perturbations in the nominal transition kernel can induce substantially larger perturbations in the robust Bellman update. This amplification effect leads to the increased sample complexity observed under χ^2 uncertainty.

7.2 Competitive multi-agent RL

The field of multi-agent reinforcement learning (MARL) studies how multiple agents interact and make decisions in a shared dynamic environment (Zhang *et al.*, 2021a). Recent breakthroughs in game playing (Vinyals *et al.*, 2019; Brown and Sandholm, 2019) and multi-robot control (Matignon *et al.*, 2012) epitomize its broad applicability. In many applications, however, agents have conflicting objectives and must act strategically to maximize their individual utilities. Such competitive interactions are commonly modeled as *Markov games* (MGs) — also known as *stochastic games* — which provide a mathematically rigorous and versatile framework for studying competitive MARL (Shapley, 1953; Littman, 1994).

In competitive settings, the goal is typically to learn certain equilibrium strategy profiles, most notably the Nash equilibria (Nash Jr, 1950). A key challenge in modern applications is data efficiency: agents often have little prior knowledge of the environment and must learn effective strategies through interaction, often in high-dimensional environments. Despite significant progress, the sample complexity of learning equilibrium strategies in MARL was poorly understood until fairly recently, even in the fundamental setting of two-player zero-sum Markov games.

7.2.1 Markov games: preliminaries

In this section, we introduce the basic framework of Markov games, together with two important solution concepts: Nash equilibrium and coarse correlated equilibrium. For simplicity, we restrict our attention to finite-horizon, non-stationary Markov games.

Markov games. A finite-horizon, non-stationary *multi-agent general-sum Markov game*, denoted by $\mathcal{MG} = \{\mathcal{S}, \{\mathcal{A}_i\}_{1 \leq i \leq m}, H, P, r\}$, models the interaction of m agents in a shared dynamic environment, where agents may have competing objectives. The components of $\mathcal{MG}$ are defined as follows.

- $\mathcal{S} = \{1, \dots, S\}$: the state space of the shared environment, consisting of S different states.
- $\mathcal{A}_i = \{1, \dots, A_i\}$: the action space of the i -th agent ($1 \leq i \leq m$), consisting of A_i actions.
- $\mathcal{A}$ and $\mathcal{A}_{-i}$: the joint action space of all agents and the joint action space excluding agent i , respectively, defined by

$$\mathcal{A} := \mathcal{A}_1 \times \dots \times \mathcal{A}_m; \quad \mathcal{A}_{-i} := \prod_{j:j \neq i} \mathcal{A}_j \quad (1 \leq i \leq m). \quad (7.19)$$

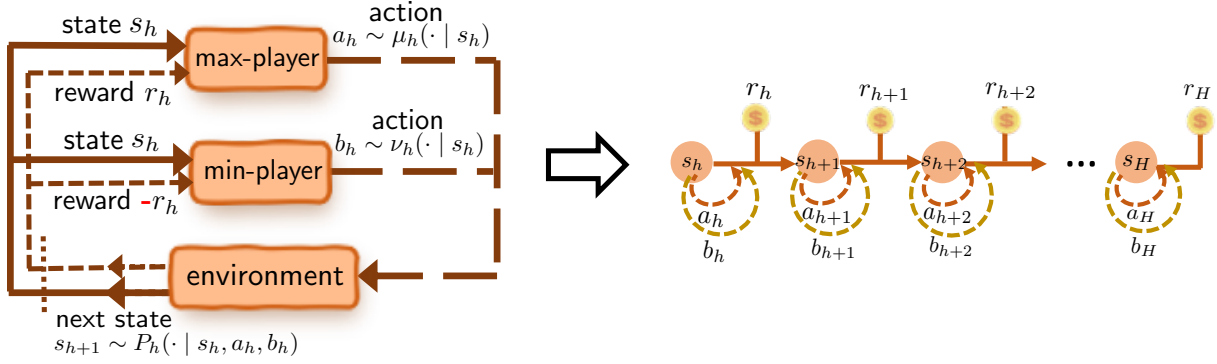


Figure 7.1: Illustration of a two-player zero-sum Markov game.

Throughout, we use $\mathbf{a} \in \mathcal{A}$ (resp. $\mathbf{a}_{-i} \in \mathcal{A}_{-i}$) to denote a joint action profile of all agents (resp. all agents except agent i).

- H : the horizon length.
- $P = \{P_h\}_{1 \leq h \leq H}$, with $P_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$: the probability transition kernel of $\mathcal{MG}$. Specifically, for any $(s, \mathbf{a}, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$, we let $P_h(s' | s, \mathbf{a})$ indicate the probability of transitioning from state s to state s' at step h when the joint action profile $\mathbf{a}$ is executed.
- $r = \{r_{i,h}\}_{1 \leq h \leq H, 1 \leq i \leq m}$, with $r_{i,h} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$: the (deterministic) reward function. Specifically, for any $(s, \mathbf{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $r_{i,h}(s, \mathbf{a})$ stands for the immediate reward received by agent i at step h when the Markov game is in state s and the joint action profile is $\mathbf{a}$. Throughout, rewards are assumed to be normalized so that $r_{i,h}(s, \mathbf{a}) \in [0, 1]$ for any $(s, \mathbf{a}, h, i) \in \mathcal{S} \times \mathcal{A} \times [H] \times [m]$.

An important special case is the *two-player zero-sum Markov game*, which satisfies

$$r_{2,h} = -r_{1,h} \quad \text{for all } h \in [H],$$

denoted by $\mathcal{MG} = \{\mathcal{S}, \{\mathcal{A}_1, \mathcal{A}_2\}, H, P, r\}$. Following standard convention, we assume that $r_{1,h} \geq 0$ for all $h \in [H]$,² and refer to player 1 and player 2 as the *max-player* and *min-player*, respectively. See Figure 7.1 for an illustration.

Markov policies. Having specified the dynamics of a Markov game, we now turn to the policies used by the agents. Throughout this section, we focus on the class of *Markov policies*, under which each agent's action depends only on the current state s and the current step number h , and not on the previously visited states of the game.

Specifically, let $\pi_i = \{\pi_{i,h}\}_{1 \leq h \leq H}$ represent the policy of agent i . Here, $\pi_{i,h}(\cdot | s) \in \Delta(\mathcal{A}_i)$ for any $(s, h) \in \mathcal{S} \times [H]$, where $\pi_{i,h}(a | s)$ denotes the probability that agent i selects action a when the game is in state s at step h . A joint Markov policy can be defined analogously: we let $\pi = (\pi_1, \dots, \pi_m) : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{A})$ represent a joint Markov policy of all agents, where the joint actions of all agents in state s and step h are chosen according to the joint distribution specified by

²Since $r_{2,h} = -r_{1,h} \leq 0$, the reward normalization assumption is violated for player 2. This can be remedied by adding a constant shift to $r_{2,h}$, which typically does not affect the equilibrium strategies or the learning problem.

$\pi_h(\cdot | s) = (\pi_{1,h}, \dots, \pi_{m,h})(\cdot | s) \in \Delta(\mathcal{A})$. For any given joint policy π , we employ π_{-i} to represent the policies of all agents except agent i , and similarly use $\pi_{-i,h}$ to denote the corresponding policies at step h .

Additionally, a joint policy π is said to be a *product policy* if the agents act independently under π ; equivalently, this means the joint action distribution factorizes across agents. In this case, we write $\pi = \pi_1 \times \dots \times \pi_m$ to indicate that π is a product policy.

Value functions. Consider a Markovian trajectory $\{(s_h, \mathbf{a}_h)\}_{1 \leq h \leq H}$, where $s_h \in \mathcal{S}$ is the state and $\mathbf{a}_h \in \mathcal{A}$ is the joint action profile at step h . For any given joint policy π and any step $h \in [H]$, the value function $V_{i,h}^\pi : \mathcal{S} \rightarrow \mathbb{R}$ of agent i is defined as

$$\forall s \in \mathcal{S}, \quad V_{i,h}^\pi(s) := \mathbb{E} \left[\sum_{t=h}^H r_{i,t}(s_t, \mathbf{a}_t) \mid s_h = s \right], \quad (7.20)$$

where the expectation is taken over the Markovian trajectory with the m agents jointly executing policy π (conditional on $s_h = s$). Similarly, we define the Q-function $Q_{i,h}^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ of the i -th agent under policy π as follows

$$\forall (s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}, \quad Q_{i,h}^\pi(s, \mathbf{a}) := \mathbb{E} \left[\sum_{t=h}^H r_{i,t}(s_t, \mathbf{a}_t) \mid s_h = s, \mathbf{a}_h = \mathbf{a} \right],$$

which differs from $V_{i,h}^\pi(s)$ only in that the action at step h is fixed to be $\mathbf{a}$.

In addition, consider the case where all agents except agent i follow the joint policy π_{-i} , while agent i *independently* executes a policy π'_i . We denote by $V_{i,h}^{\pi'_i \times \pi_{-i}}$ the value function resulting from this joint policy $\pi'_i \times \pi_{-i}$. Optimizing over all policies π'_i of agent i , we further define

$$V_{i,h}^{\star, \pi_{-i}}(s) := \max_{\pi'_i : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{A}_i)} V_{i,h}^{\pi'_i \times \pi_{-i}}(s), \quad \forall (s, h, i) \in \mathcal{S} \times [H] \times [m]. \quad (7.21)$$

It is well known that there exists at least one policy, denoted by $\pi_i^\star(\pi_{-i}) : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{A}_i)$, that can simultaneously attain $V_{i,h}^{\star, \pi_{-i}}(s)$ for all $h \in [H]$ and all $s \in \mathcal{S}$. Such a policy is often referred to as a *best-response policy*. Importantly, the best-response policy $\pi_i^\star(\pi_{-i})$ is the best among all policies of the i -th agent executed *independently* of π_{-i} .

Furthermore, if we freeze π_{-i} , then the Bellman optimality condition for agent i can be expressed as (Bertsekas, 2017)

$$V_{i,h}^{\star, \pi_{-i}}(s) = \max_{\mathbf{a}_i \in \mathcal{A}_i} \left\{ \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i,h}(\cdot | s)} \left[r_{i,h}(s, \mathbf{a}) + \langle P_h(\cdot | s, \mathbf{a}), V_{i,h+1}^{\star, \pi_{-i}} \rangle \right] \right\}$$

for all $(s, h, i) \in \mathcal{S} \times [H] \times [m]$, where the joint action profile $\mathbf{a}$ is composed of agent i 's action $\mathbf{a}_i$ and the remaining agents' actions $\mathbf{a}_{-i}$. The terminal value function is set to be $V_{i,H+1}^{\star, \pi_{-i}} = 0$.

Equilibria of Markov games. In a multi-agent Markov game, each agent wishes to maximize its own value function. Since agents may have competing objectives, identifying appropriate equilibrium solutions has become a central theme in the study of Markov games. Two of the most widely studied solution concepts are the Nash equilibrium (NE) (Nash Jr, 1950) and the coarse correlated equilibrium (CCE) (Moulin and Vial, 1978; Aumann, 1987), which we briefly introduce below.

- *Nash equilibrium (NE)*. A product policy $\pi = \pi_1 \times \cdots \times \pi_m$ is said to be a (*mixed-strategy*) *Nash equilibrium* of $\mathcal{MG}$ if the following holds:

$$V_{i,1}^\pi(s) = V_{i,1}^{\star, \pi^{-i}}(s), \quad \forall (s, i) \in \mathcal{S} \times [m]. \quad (7.22)$$

In other words, conditional on the current policies of all other agents and the assumption that all agents take actions *independently*, no agent can improve its value by unilaterally deviating from its current policy. Thus, under a Nash equilibrium, every agent is playing a best response to the policies of the opponents.

- *Coarse correlated equilibrium (CCE)*. A joint policy π is said to be a *coarse correlated equilibrium* of $\mathcal{MG}$ if

$$V_{i,1}^\pi(s) \geq V_{i,1}^{\star, \pi^{-i}}(s), \quad \forall (s, i) \in \mathcal{S} \times [m]. \quad (7.23)$$

Like a Nash equilibrium, a CCE guarantees that no agent can benefit from a unilateral deviation performed independently of the others. The key distinction is that a CCE permits correlation among the agents' actions, whereas a Nash equilibrium requires the joint policy to factorize across agents. Consequently, every Nash equilibrium is also a CCE, although the converse does not generally hold.

For general-sum Markov games, computing a Nash equilibrium is well known to be computationally intractable in general (specifically, PPAD-hard) (Daskalakis, 2013; Daskalakis *et al.*, 2009). This computational barrier motivates the study of alternative equilibrium notions, such as the CCE, which can often be computed efficiently via standard mathematical programming techniques. Another widely studied notion is the correlated equilibrium (CE) (Aumann, 1987; Greenwald *et al.*, 2003; Papadimitriou and Roughgarden, 2008), which sits between NE and CCE; we omit a detailed discussion of the CE in this chapter.

7.2.2 Problem formulation

In many applications, the underlying Markov game is unknown and must be learned from data. As in single-agent RL, MARL has been studied under a variety of data collection mechanisms. To keep the exposition focused and concise, this section restricts attention to the generative-model setting. While idealized, this framework already captures some challenges unique to multi-agent scenarios and provides a useful starting point for understanding other important settings, including online Markov games (Tian *et al.*, 2021; Mao and Başar, 2022; Jin *et al.*, 2024a; Song *et al.*, 2022; Daskalakis *et al.*, 2023; Cui *et al.*, 2023; Feng *et al.*, 2024) and offline Markov games (Cui and Du, 2022b; Cui and Du, 2022a; Yan *et al.*, 2024).

Sampling from a generative model. Throughout this section, we assume access to a generative model. In each query, the learner may choose an arbitrary state-action-step tuple $(s, \mathbf{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and obtain an independent sample

$$s' \stackrel{\text{ind.}}{\sim} P_h(\cdot | s, \mathbf{a}).$$

Here and throughout the section, we use boldface $\mathbf{a}$ to emphasize that it represents a joint action profile of all agents.

While this sampling model resembles its counterpart in single-agent RL, the resulting statistical challenges can be markedly different. In particular, whereas a nearly uniform allocation of samples

across state-action-step tuples is often sufficient in single-agent RL, data-efficient learning in Markov games frequently requires highly adaptive sampling strategies, as we shall illustrate momentarily.

Goal. In practice, it might be challenging to compute an exact equilibrium, and instead one would seek to find approximate solutions. Accordingly, the goal of this section is to learn an ε -approximate equilibrium (either NE or CCE) of $\mathcal{MG} = \{\mathcal{S}, \{\mathcal{A}_i\}_{1 \leq i \leq m}, H, P, r\}$ using as few samples as possible, or equivalently, with a minimal number of calls to the generative model.

To formalize this objective, we need to define the ε -approximate equilibrium. For any policy π , define its suboptimality gap — measured in an ℓ_∞ -based manner — by

$$\text{gap}(\pi) := \max_{s \in \mathcal{S}} \text{gap}(\pi; s), \quad (7.24a)$$

where

$$\text{gap}(\pi; s) := \max_{1 \leq i \leq m} \left\{ V_{i,1}^{\star, \pi-i}(s) - V_{i,1}^\pi(s) \right\}. \quad (7.24b)$$

With this definition in place, a *product* policy $\pi_{\text{NE}} = \pi_1 \times \cdots \times \pi_m$ is said to be an ε -approximate NE — or more concisely, ε -Nash — if the resultant suboptimality gap obeys $\text{gap}(\pi_{\text{NE}}) \leq \varepsilon$. Similarly, a joint (and possibly correlated) policy π_{CCE} is said to be an ε -approximate CCE — or more concisely, ε -CCE — if $\text{gap}(\pi_{\text{CCE}}) \leq \varepsilon$. A key objective is therefore to characterize the sample complexity required to learn such ε -approximate equilibria.

7.2.3 A model-based algorithm

A straightforward and natural model-based approach to equilibrium learning in Markov games is the plug-in estimator (Zhang *et al.*, 2020b; Shi *et al.*, 2024b), which first constructs an empirical estimate of the underlying Markov game from data samples and then computes an equilibrium of the empirical Markov game using classical planning methods. More specifically, the procedure consists of the following two stages.

- 1) *Model estimation.* For each state-joint-action-step tuple $(s, \mathbf{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, draw N independent samples from the generative model:

$$s_h^{(i)}(s, \mathbf{a}) \stackrel{\text{i.i.d.}}{\sim} P_h(\cdot | s, \mathbf{a}), \quad i = 1, \dots, N.$$

Using these samples, construct an empirical transition kernel $\hat{P}$ according to

$$\forall s' \in \mathcal{S}, \quad \hat{P}_h(s' | s, \mathbf{a}) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s' = s_h^{(i)}(s, \mathbf{a})\}$$

for every $(s, \mathbf{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. That is, $\hat{P}_h(s' | s, \mathbf{a})$ records the empirical frequency with which the tuple $(s, \mathbf{a}, h)$ transitions to state s' . This leads to the empirical Markov game $\widehat{\mathcal{MG}} = \{\mathcal{S}, \{\mathcal{A}_i\}_{1 \leq i \leq m}, H, \hat{P}, r\}$.

- 2) *Planning.* Compute either a Nash equilibrium $\hat{\pi}_{\text{NE}}$ or a coarse correlated equilibrium $\hat{\pi}_{\text{CCE}}$ of $\widehat{\mathcal{MG}}$ using an appropriate planning oracle. For instance, NE in two-player zero-sum Markov games can be efficiently computed by Nash value iteration (Shapley, 1953), while CCE in general-sum Markov games can be obtained using linear-programming-based methods (Papadimitriou and Roughgarden, 2008), no-external-regret learning dynamics (Roughgarden, 2016), among others.

The above plug-in estimator provides perhaps the most direct way of extending model-based algorithms for single-agent RL (see Section 3.5) to the problem of equilibrium learning. A natural question is therefore: how many samples are required for this approach to compute an approximate equilibrium? The following theorem provides an answer.

Theorem 7.6 (Zhang *et al.* (2020b) and Shi *et al.* (2024b)). Consider any $\delta \in (0, 1)$. Let $\hat{\pi}_{\text{NE}}$ and $\hat{\pi}_{\text{CCE}}$ denote, respectively, the NE and CCE returned by the planning oracle in the above model-based plug-in algorithm. Then for any $0 < \varepsilon \leq \sqrt{H}$, one has

$$\text{gap}(\hat{\pi}_{\text{NE}}) \leq \varepsilon \quad \text{and} \quad \text{gap}(\hat{\pi}_{\text{CCE}}) \leq \varepsilon$$

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

$$N \geq \frac{C_1 H^3}{\varepsilon^2} \log \left(\frac{SHN \prod_{1 \leq i \leq m} A_i}{\delta} \right)$$

for some large enough universal constant $C_1 > 0$. Here, the metric $\text{gap}(\cdot)$ is defined in (7.24).

Since the Markov game contains a total of $SH \prod_{i=1}^m A_i$ state-joint-action-step combinations, Theorem 7.6 implies that the above model-based plug-in algorithm can learn either an ε -NE or an ε -CCE with a sample complexity of

$$\tilde{O} \left(\frac{SH^4 \prod_{1 \leq i \leq m} A_i}{\varepsilon^2} \right). \quad (7.25)$$

It is worth emphasizing that this theorem concerns only the statistical aspect of the plug-in approach and does not address its computational complexity. Indeed, while the planning oracle can be implemented efficiently in certain settings — such as two-player zero-sum Markov games — it becomes computationally intractable in general-sum Markov games when tasked with computing NE.

While appealing for its simplicity, the model-based plug-in estimator introduced above turns out to be quite expensive and, in general, statistically suboptimal. The main bottleneck stems from its dependence on the size of the joint action space

$$\prod_{1 \leq i \leq m} A_i,$$

which grows exponentially with the number of agents m when the individual action spaces are of comparable size. Consequently, the sample complexity deteriorates rapidly as the number of agents increases, thereby limiting the applicability of this plug-in approach in applications involving many participants. This phenomenon is commonly referred to as the *curse of multi-agency*. A natural question is therefore whether one can circumvent this curse and develop algorithms whose sample complexity exhibits a substantially milder dependence on the size of the joint action space. As we shall see shortly, the answer is often affirmative.

7.2.4 Breaking the curse of multi-agency

Motivated by the goal of overcoming the curse of multi-agency in Markov games, a recent line of work (Mao and Başar, 2022; Song *et al.*, 2022; Jin *et al.*, 2024a; Li *et al.*, 2022a) sought to design equilibrium learning algorithms that can achieve a significantly improved sample complexity than

the plug-in estimator discussed in Section 7.2.3. We discuss one such algorithm, termed **Q-FTRL**, developed by Li *et al.* (2022a) for the generative-model setting.

At a high level, **Q-FTRL** combines two powerful ideas from the RL and online learning literature: optimism in the face of uncertainty, implemented through upper confidence bounds (UCBs) as in online RL (see Chapter 5), and online adversarial learning via the Follow-the-Regularized-Leader (FTRL) framework (Shalev-Shwartz, 2012). A key feature of the algorithm is its use of adaptive sampling, which directs data collection toward statistically informative regions of the state-joint-action space from each agent’s perspective independently, and thereby mitigates the curse of multi-agency. This stands in sharp contrast to the single-agent generative-model setting discussed in Chapter 3, where a uniform, non-adaptive allocation of samples is often sufficient. More broadly, **Q-FTRL** reveals how ideas originating from online RL (see Chapter 5) can play a pivotal role even in the generative-model setting.

Since the **Q-FTRL** algorithm is considerably more sophisticated than the model-based methods introduced in previous chapters, we omit full algorithmic details and refer the interested reader to Li *et al.* (2022a) for a complete description. In the sequel, we present its statistical guarantees and discuss their implications.

Theorem 7.7 (Li *et al.* (2022a)). Consider any $\delta \in (0, 1)$ and any $\varepsilon \in (0, H]$. There exists a model-based algorithm (i.e., **Q-FTRL**) that, with probability at least $1 - \delta$, achieves the following.

- For a two-player zero-sum Markov game, it can find a policy $\hat{\pi}_{\text{NE}}$ satisfying

$$\text{gap}(\hat{\pi}_{\text{NE}}) \leq \varepsilon,$$

with a total sample complexity of

$$\tilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right).$$

- For a general-sum Markov game, it can find a policy $\hat{\pi}_{\text{CCE}}$ satisfying

$$\text{gap}(\hat{\pi}_{\text{CCE}}) \leq \varepsilon$$

with a total sample complexity of

$$\tilde{O}\left(\frac{H^4 S \sum_{i=1}^m A_i}{\varepsilon^2}\right).$$

Theorem 7.7 demonstrates that **Q-FTRL** manages to break the curse of multi-agency. In contrast to the plug-in estimator described in Section 7.2.3, whose sample complexity is proportional to $\prod_{i=1}^m A_i$, the sample complexity of **Q-FTRL** depends only on the total number of individual actions, i.e., $\sum_{i=1}^m A_i$. Consequently, when the action spaces of different agents are of comparable size, **Q-FTRL** reduces an exponential dependence on the number of agents to a linear one, resulting in a dramatic improvement in data efficiency. Another noteworthy feature is that such statistical guarantees hold uniformly over the entire accuracy range $\varepsilon \in (0, H]$, making the algorithm provably appealing even in sample-starved scenarios.

To assess the optimality of the above sample complexity guarantees, we next present a minimax lower bound for learning Markov games (w.r.t. finding either ε -NE or ε -CCE) as follows.

Theorem 7.8. Consider any $m \geq 2$ and any $0 < \varepsilon \leq c_1 H$ for some small enough constant $c_1 > 0$. Then one can construct a collection of m -player zero-sum Markov games $\{\mathcal{MG}_\theta \mid \theta \in \Theta\}$ with S states, horizon H , and A_i actions for the i -th player ($1 \leq i \leq m$) such that

$$\inf_{\hat{\pi}} \max_{\theta \in \Theta} \mathbb{P}_{\mathcal{MG}_\theta} \{ \text{gap}(\hat{\pi}) > \varepsilon \} \geq \frac{1}{4},$$

provided that the total sample size is smaller than

$$\frac{c_2 H^4 S \max_{1 \leq i \leq m} A_i}{\varepsilon^2} \quad (7.26)$$

for some sufficiently small constant $c_2 > 0$. Here, the infimum is over all (joint) policy estimators $\hat{\pi}$, and $\mathbb{P}_{\mathcal{MG}_\theta}$ denotes the probability measure under the Markov game $\mathcal{MG}_\theta$.

Combining Theorem 7.7 with the minimax lower bound in Theorem 7.8 establishes the minimax optimality of Q-FTRL (up to logarithmic factors) when the number m of agents is fixed. Notably, Q-FTRL is provably minimax optimal for learning Nash equilibria in the important special case of two-player zero-sum Markov games.

An intriguing open question is whether the sample complexity of Q-FTRL, or related algorithms, can be further improved to match the lower bound (7.26). Since linear dependence on $\max_{1 \leq i \leq m} A_i$ has already been shown to be attainable in online MARL via V-learning (Song *et al.*, 2022; Jin *et al.*, 2024a) — albeit at the cost of suboptimal dependence on the horizon length — we conjecture the lower bound (7.26) to be tight (up to logarithmic factors) and hence captures the fundamental statistical limit of equilibrium learning in Markov games.

7.3 Federated RL

Collecting, through a single agent, an enormous amount of data required by RL is often prohibitively time-consuming and, in many applications, infeasible. As a result, practical implementations of RL systems often deploy multiple agents to interact with the environment and gather data in parallel. This decentralized data-collection paradigm has motivated the development of distributed and federated RL algorithms, enabling agents to collaboratively learn an optimal policy without actually communicating the collected data to a centralized server. Ideally, such methods would achieve a linear speedup with respect to the number of participating agents.

This federated approach to RL has attracted growing attention, as it significantly reduces bandwidth requirements while offering improved privacy and security guarantees compared to fully centralized approaches. At the same time, the federated setting introduces new statistical and algorithmic challenges that are absent in classical single-agent RL. In this section, we use federated Q-learning as a representative example to illustrate some of these challenges and discuss the fundamental statistical questions arising due to communication constraints.

7.3.1 Federated synchronous Q-learning

As a case study, we consider federated synchronous Q-learning, in which all state-action pairs are updated simultaneously, and each agent is assumed to have access to a generative model (or simulator). This provides a natural framework for studying the interplay between communication constraints and statistical efficiency in federated RL.

Problem setup. Consider a federated RL setting with K agents interacting with the same discounted infinite-horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$. We focus on the *synchronous* setting, where each agent $k \in [K]$ has access to the generative model of the underlying MDP. At iteration t , for every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, each agent k can independently query the generative model and obtain a fresh sample

$$s_t^k(s, a) \stackrel{\text{ind.}}{\sim} P(\cdot | s, a). \quad (7.27)$$

The agents perform local updates using their own samples and may periodically synchronize their estimates through a central server. The goal is to collaboratively learn the optimal Q-function Q^* , while reducing the communication overhead associated with synchronizing the agents' local estimates.

Algorithm description: FedSynQ. We now describe a natural federated synchronous Q-learning algorithm, dubbed **FedSynQ**, which alternates between local Q-learning updates and periodic synchronization across agents. The complete procedure is summarized in Algorithm 7.2.

The **FedSynQ** algorithm initializes, at each agent $k \in [K]$, a local Q-function estimate $Q_0^k = Q_0$. At the beginning of each iteration $t \geq 1$, agent k maintains a local Q-function estimate Q_{t-1}^k together with the corresponding value function estimate V_{t-1}^k :

$$\forall s \in \mathcal{S} : \quad V_{t-1}^k(s) := \max_{a \in \mathcal{A}} Q_{t-1}^k(s, a).$$

Each iteration t performs the following two steps.

1. *Local updates.* Each agent k first updates *all* entries of its local Q-function estimate Q_{t-1}^k independently using the standard synchronous Q-learning update rule, yielding the following *intermediate* Q-function estimate:

$$Q_{t-\frac{1}{2}}^k(s, a) = (1 - \eta)Q_{t-1}^k(s, a) + \eta \left(r(s, a) + \gamma V_{t-1}^k(s_t^k(s, a)) \right) \quad (7.28)$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $s_t^k(s, a)$ is generated according to (7.27), and $0 < \eta < 1$ is the learning rate.

2. *Periodic synchronization and averaging.* Every $\tau \geq 1$ iterations — where τ denotes the synchronization period — the agents synchronize their local Q-function estimates through a central server. At each synchronization round, the server averages the intermediate Q-function estimates $\{Q_{t-\frac{1}{2}}^k\}_{k=1}^K$ and broadcasts the resulting estimate back to all agents. Consequently, the updated Q-function estimate at the end of iteration t can be expressed as

$$Q_t^k(s, a) = \begin{cases} \frac{1}{K} \sum_{k=1}^K Q_{t-\frac{1}{2}}^k(s, a), & \text{if } t \equiv 0 \pmod{\tau} \\ Q_{t-\frac{1}{2}}^k(s, a), & \text{otherwise} \end{cases} \quad (7.29)$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

After T iterations, the **FedSynQ** algorithm outputs the averaged Q-function estimate

$$Q_T = \frac{1}{K} \sum_{k=1}^K Q_T^k.$$

Algorithm 7.2: Federated synchronous Q-learning (FedSynQ)

```

1 inputs: discount factor  $\gamma$ , reward function  $r$ , learning rate  $0 < \eta < 1$ , number of agents  $K$ ,
   synchronization period  $\tau$ , number of iterations  $T$ , initial guess  $Q_0$ .
2 initialize  $Q_0^k \leftarrow Q_0$  for all  $k$ .
3 for  $t = 1, \dots, T$  do
4   for  $k = 1, \dots, K$  do
5     draw  $s_t^k(s, a) \stackrel{\text{ind.}}{\sim} P(\cdot | s, a)$  for every  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .
6     compute  $Q_{t-\frac{1}{2}}^k$  according to (7.28).
7     compute  $Q_t^k$  according to (7.29).
8 return:  $Q_T = \frac{1}{K} \sum_k Q_T^k$ .
```

Without loss of generality, we assume that T is divisible by the synchronization period τ . The resulting communication complexity can be measured by

$$C_{\text{round}} = T/\tau,$$

which corresponds to the total number of communication rounds.

Theoretical guarantees. We are now in a position to establish non-asymptotic performance guarantees for FedSynQ.

Theorem 7.9 (Woo *et al.*, 2023). Consider any given $\delta \in (0, 1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Suppose that the initialization satisfies $0 \leq Q_0 \leq \frac{1}{1-\gamma}$, and the synchronization period τ obeys

$$\tau \leq 1 + \frac{1}{\eta} \min \left\{ \frac{1-\gamma}{8\gamma}, \frac{1}{K} \right\}. \quad (7.30a)$$

There exists some sufficiently large (resp. small) constant $c_T > 0$ (resp. $c_\eta > 0$) such that, with probability at least $1 - \delta$, Algorithm 7.2 achieves $\|Q_T - Q^*\|_\infty \leq \varepsilon$, provided that

$$T \geq \frac{c_T}{K(1-\gamma)^5 \varepsilon^2} \log^3 \left(\frac{SAKT}{\delta \varepsilon (1-\gamma)^2} \right), \quad (7.30b)$$

$$\eta = \frac{c_\eta K (1-\gamma)^4 \varepsilon^2}{\log \frac{SAKT}{\delta}}. \quad (7.30c)$$

Implications. Theorem 7.9 implies that, to obtain an ε -accurate estimate of the optimal Q-function in an ℓ_∞ sense, each agent requires at most

$$\tilde{O} \left(\frac{SA}{K(1-\gamma)^5 \varepsilon^2} \right) \quad (7.31)$$

samples, where each agent draws SA samples at each iteration. This theorem yields several implications as follows.

- *Linear speedup.* The per-agent sample complexity in (7.31) enjoys a linear improvement with respect to the number of agents K . For comparison, recall from Chapter 4 that the sharp sample complexity of single-agent synchronous Q-learning is $\tilde{\Theta} \left(\frac{SA}{(1-\gamma)^4 \varepsilon^2} \right)$ for $\varepsilon \in (0, 1)$. Consequently,

the federated setting provably achieves faster learning whenever $K \gtrsim \frac{1}{1-\gamma}$ up to logarithmic factors, provided that $\varepsilon \in (0, 1)$.

- *Communication efficiency.* A key feature of this federated setting is the use of periodic averaging to reduce communication costs. According to (7.30a), the above theory requires the synchronization period τ to scale inversely with the learning rate η , suggesting that more frequent communication may be needed to compensate for the greater discrepancy of local updates under larger learning rates. To gain further insight, consider the regime where $K \gtrsim \frac{1}{1-\gamma}$ and $\varepsilon \lesssim \frac{1}{K(1-\gamma)^2}$. Substituting the learning rate (7.30c) into the upper bound (7.30a) suggests a feasible synchronization period as

$$\tau = \tilde{\Theta}\left(\frac{1}{K^2(1-\gamma)^4\varepsilon^2}\right).$$

This leads to a communication complexity no larger than

$$C_{\text{round}} = \frac{T}{\tau} \leq \tilde{O}\left(\frac{K}{1-\gamma}\right),$$

which is nearly independent of the target accuracy level ε .

Finally, we note that the communication efficiency of FedSynQ can be further improved by employing more orchestrated communication schedules; see, e.g., Woo *et al.* (2024) and Salgia and Chi (2024).

7.3.2 Trade-offs between statistical and communication efficiency

A central challenge in the design of federated RL algorithms, including federated Q-learning, is to address the natural tension between sample and communication complexities. At one extreme, one may adopt a fully centralized approach that gathers all locally collected data at a central server and then runs a statistically optimal RL algorithm using all samples. While this naïve strategy trivially achieves the optimal sample complexity, it incurs a very large — and often prohibitive — communication cost. In contrast, several recently proposed federated RL algorithms (e.g., Khodadadian *et al.* (2022), Woo *et al.* (2023), and Woo *et al.* (2024)) operate in a more communication-efficient fashion. By reducing the frequency of synchronization between agents and a central server, these methods significantly lower communication costs relative to the fully centralized approach, albeit often at the expense of suboptimal sample complexity.

These results hint at the existence of a fundamental trade-off between sample and communication complexities in federated RL. To fully appreciate the statistical benefits from collaboration, one must understand the price of collaboration, i.e., the minimum amount of communication required to realize such gains. In what follows, we study this trade-off within the class of federated Q-learning algorithms with intermittent communication, which encompasses various communication-efficient schemes.

Intermittent-communication protocols. Consider a popular class of intermittent communication protocols in federated learning. Under such protocols, agents communicate intermittently — namely, only at selected synchronization rounds — sharing their updated estimates of Q^* either with one another or through a central server. Between communication rounds, each agent performs local stochastic-approximation-type updates using only its locally available data, much like a single-agent

Q-learning algorithm. As such, intermittent communication provides a natural framework to extend single-agent Q-learning to the federated setting while reducing communication costs. This class of protocols has been extensively studied in federated learning and distributed optimization and has played a central role in establishing communication lower bounds for distributed stochastic optimization problems (Woodworth *et al.*, 2021).

A generic federated Q-learning algorithm with intermittent communication is outlined in Algorithm 7.3. This generic algorithm is specified by five parameters:

- (i) the total number of updates T ;
- (ii) the number of communication rounds C_{round} ;
- (iii) a step size schedule $\{\eta_t\}_{t=1}^T$;
- (iv) a communication schedule $\mathcal{C} = \{t_r\}_{r=1}^R$;
- (v) a batch size B .

At iteration t , each agent k computes $\{\hat{\mathcal{T}}_b(Q_{t-1}^k)\}_{b=1}^B$, a minibatch of i.i.d. sample Bellman operators $\hat{\mathcal{T}}_b(\cdot)$, $b = 1, \dots, B$, applied to its current estimate Q_{t-1}^k , using B samples from the generative model for each (s, a) pair, and obtains an intermediate local estimate using the Q-learning update rule as follows:

$$Q_{t-\frac{1}{2}}^k = (1 - \eta_t)Q_{t-1}^k + \frac{\eta_t}{B} \sum_{b=1}^B \hat{\mathcal{T}}_b(Q_{t-1}^k). \quad (7.32)$$

Here, $\eta_t \in (0, 1]$ denotes the learning rate at iteration t . The resulting intermediate estimates are synchronized and averaged according to the communication schedule $\mathcal{C} = \{t_r\}_{r=1}^{C_{\text{round}}}$ (which consists of C_{round} synchronization rounds). Specifically,

$$Q_t^k = \begin{cases} \frac{1}{K} \sum_{j=1}^K Q_{t-\frac{1}{2}}^j, & \text{if } t \in \mathcal{C}, \\ Q_{t-\frac{1}{2}}^k, & \text{otherwise.} \end{cases} \quad (7.33)$$

Here, the averaging step may be replaced by any distributed mean estimation routine, including variants that include compression to control the bit-level costs. Without loss of generality, we assume $t_{C_{\text{round}}} = T$, i.e., the final iterates are always synchronized. Evidently, the number of samples taken per agent for each state-action pair is $N = BT$.

Theorem 7.10. Assume that $\gamma \in [5/6, 1)$, $S \geq 4$ and $A \geq 2$. Let $\mathcal{A}$ be a federated Q-learning algorithm with intermittent communication, as described in Algorithm 7.3. Suppose that $\mathcal{A}$ is run for $T \geq \max\{16, \frac{1}{1-\gamma}\}$ iterations with either a linearly rescaled learning rate schedule $\eta_t = \frac{1}{1+c_\eta(1-\gamma)t}$ or a constant learning rate schedule $\eta_t = \eta$ for all $1 \leq t \leq T$. If

$$C_{\text{round}} \leq \frac{c_0}{(1-\gamma) \log^2 N}$$

for some universal constant $c_0 > 0$, then for any choices of communication schedule, batch size B , $c_\eta > 0$ and $\eta \in (0, 1)$, there exists an instance of MDP such that the output of $\mathcal{A}$ obeys

$$\mathbb{E}[\|\hat{Q} - Q^*\|_\infty] \geq \frac{C_\gamma}{\sqrt{N} \log^3 N}$$

for all $K \geq 2$ and $N = BT$. Here, $C_\gamma > 0$ is a quantity that depends only on γ .

Algorithm 7.3: A generic federated Q-learning algorithm $\mathcal{A}$

```

1 input: discount factor  $\gamma$ , reward function  $r$ , number of iterations  $T$ , learning rates  $\{\eta_t\}_{t=1}^T$ ,
   number of synchronization rounds  $R$ , synchronization schedule  $\mathcal{C} = \{t_r\}_{r=1}^R$ , batch size  $B$ .
2 initialize  $Q_0^k \leftarrow 0$  for all  $1 \leq k \leq K$ .
3 for  $t = 1, 2, \dots, T$  do
4   for  $k \in [K]$  do
5     compute  $Q_{t-\frac{1}{2}}^k$  according to (7.32).
6     compute  $Q_t^k$  according to (7.33).
7 return  $\hat{Q} = Q_T$ .
```

The above theorem unveils a fundamental communication requirement for federated Q-learning with intermittent communication. Specifically, for any such algorithm to obtain *any* statistical benefit from collaboration — namely, for its error rate to improve as the number of agents increases — it must perform at least $\tilde{\Omega}(\frac{1}{1-\gamma})$ rounds of communication and transmit $\tilde{\Omega}(\frac{SA}{1-\gamma})$ bits of information per agent, both of which scale linearly with the effective horizon $\frac{1}{1-\gamma}$ of the underlying MDP. The above lower bound on the communication complexity also immediately applies to federated Q-learning algorithms that offer order-optimal sample complexity and thereby a linear speedup with respect to the number of agents. Therefore, this theorem characterizes a fundamental converse for the statistical-communication trade-off in federated Q-learning.

The communication lower bound stems from the bias-variance trade-off underlying the convergence of Q-learning, which we discussed thoroughly in Section 4.3.3. In the fully centralized setting, this trade-off can be achieved through an appropriate choice of step sizes. In the federated setting, however, the communication schedule becomes an additional factor. In particular, the local updates performed between two communication rounds induce a positive estimation bias that depends on the standard deviation of the iterates, a manifestation of the well-documented over-estimation issue in Q-learning (see, e.g., Hasselt (2010) and the lower bound of Q-learning in Theorem 4.6). Since this bias arises solely from local updates, it does not reflect any benefit of collaboration. Communication mitigates this effect by averaging the local iterates across agents, thereby reducing the effective variance of the updates. When communication is too infrequent, however, the accumulated local bias dominates, and averaging of iterates is insufficient to offset the impact of the positive bias induced by the local steps. Consequently, the algorithm may fail to realize any statistical gains from collaboration when communication is too infrequent.

Remark 7.3 (Tightness of the communication lower bound). Salgia and Chi (2024) developed a companion variance-reduced federated Q-learning algorithm that simultaneously achieves order-optimal sample and communication complexities, thereby establishing the tightness of the communication lower bound presented above.

7.4 Statistical inference in RL

While most of this monograph has focused on learning value functions and near-optimal policies, many practical RL applications also require reliable measures of uncertainty to assess the quality of these estimates (Shi *et al.*, 2022b; Zhang *et al.*, 2025c; Shi, 2025). For instance, constructing valid confidence

intervals for value functions is often an important step in supporting downstream decision-making, not to mention its key role in the design of UCB- and LCB-based algorithms discussed in previous chapters. Achieving statistically efficient inference and uncertainty quantification typically requires establishing a distributional theory for the estimators of interest. Motivated by the high-dimensional nature of modern RL problems, a substantial body of recent work has focused on developing non-asymptotic distributional guarantees that remain valid well beyond classical asymptotic regimes. Moreover, since many RL algorithms produce estimates through iterative updates, statistical inference often needs to be developed in a manner that accommodates the online, iterative nature of the learning procedure.

We note that statistical inference for RL is a broad and rapidly growing area; see Shi (2025) for a recent overview. In this section, rather than attempting a comprehensive treatment, we use TD learning as a case study to illustrate some of the key ideas in enabling statistical inference and non-asymptotic distributional theory for RL. Recall that in Section 4.2, we have established non-asymptotic convergence guarantees of synchronous TD learning and showed that its sample complexity in the generative-model setting achieves minimax-optimal rates (Li *et al.*, 2024a); such results, however, typically suppress pre-constants and logarithmic factors, both of which can have a substantial impact on the accuracy of the resulting confidence intervals. A complementary thread of work focused on the asymptotic behavior of TD learning, demonstrating that, when combined with Polyak-Ruppert averaging (Polyak and Juditsky, 1992; Ruppert, 1998), the TD iterates converge to a Gaussian distribution in the asymptotic large-sample limit (e.g., Fort (2015), Mou *et al.* (2020), Borkar *et al.* (2025), and Li *et al.* (2023e)). While informative, such asymptotic results may not directly address finite-sample inference. In this section, we review some recent developments that extend these results to develop non-asymptotic distributional guarantees for the finite-sample regime.

7.4.1 Setting and algorithm

For concreteness, we focus on a simple variant of TD learning for estimating the value function V^π of a fixed policy π in a discounted infinite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Akin to Section 4.2, we consider here an idealized setting in which the agent has access to a sequence of i.i.d. samples; specifically, the data sequence consists of

$$\{(s_t, a_t, s'_t, r_t)\}_{t=1}^T, \quad (7.34a)$$

where, for each $1 \leq t \leq T$, the tuple (s_t, a_t, s'_t, r_t) is generated independently according to

$$s_t \sim \mu, \quad a_t \sim \pi(\cdot | s_t), \quad r_t = r(s_t, a_t), \quad s'_t \sim P(\cdot | s_t, a_t) \quad (7.34b)$$

for some state distribution $\mu \in \Delta(\mathcal{S})$ satisfying $\lambda_0 = \min_{s \in \mathcal{S}} \mu(s) > 0$. In practice, the state-action-reward tuples are often generated sequentially along a Markovian trajectory; here, we adopt the i.i.d. setting to simplify the exposition and avoid additional technical complications. The TD learning variant considered in this section maintains an estimate $V_t : \mathcal{S} \rightarrow \mathbb{R}$ of the target value function and utilizes one sample tuple in each iteration. At iteration t , it updates the value estimate according to

$$V_t(s_t) = (1 - \eta_t)V_{t-1}(s_t) + \eta_t(r_t + \gamma V_{t-1}(s'_t)), \quad (7.35)$$

where $0 < \eta_t < 1$ denotes the learning rate at iteration t .

As discussed in Section 4.2, TD learning can be viewed as a stochastic approximation procedure for finding the fixed point of the Bellman operator. A classical technique for reducing the variability

of the iterates and improving asymptotic statistical efficiency is Polyak-Ruppert averaging (Polyak and Juditsky, 1992; Ruppert, 1998), which forms the final estimate by averaging all iterates (e.g., Mou *et al.* (2020), Mou *et al.* (2023), Durmus *et al.* (2025), Borkar *et al.* (2025), Srikant (2025), and Wu *et al.* (2024)):

$$\bar{V}_T = \frac{1}{T} \sum_{t=1}^T V_t. \quad (7.36)$$

A key step toward statistical inference for TD learning is to characterize the distribution of the value estimation error

$$\bar{\Delta}_T := \bar{V}_T - V^\pi \quad (7.37)$$

of the Polyak-Ruppert averaged TD estimator.

7.4.2 Distributional theory of TD learning

Classical theory in stochastic approximation (Polyak and Juditsky, 1992; Fort, 2015; Borkar *et al.*, 2025) established that, with an appropriate choice of the learning rate schedule, the averaged iterates of a broad class of stochastic approximation algorithms — including TD learning — converge asymptotically to a Gaussian distribution as $T \rightarrow \infty$ with all other problem parameters held fixed. Specializing to the averaged TD estimator, this asymptotic normality says

$$\sqrt{T} \bar{\Delta}_T \xrightarrow{d} \mathcal{N}(0, \Lambda^*), \quad (7.38)$$

where Λ^* is the associated (asymptotic) covariance matrix defined as

$$\Lambda^* := A^{-1} \Gamma A^{-\top}, \quad (7.39)$$

with

$$A := \text{diag}(\mu)(I - \gamma P_\pi), \quad (7.40a)$$

$$\Gamma := \text{diag}(\Gamma_{s,s}), \quad \text{with} \quad \Gamma_{s,s} = \mu(s) \mathbb{E}_{s' \sim P_\pi(\cdot | s)} \left[\left(V^\pi(s) - (r(s) + \gamma V^\pi(s')) \right)^2 \right]. \quad (7.40b)$$

Here, the definition of the matrix P_π can be found in (2.19), $\text{diag}(\mu)$ denotes a diagonal matrix with diagonal entries being $\{\mu(s)\}_{1 \leq s \leq S}$, and Γ is a diagonal matrix whose s -th diagonal entry is given by $\{\Gamma_{s,s}\}_{1 \leq s \leq S}$. Intuitively, $\Gamma_{s,s}$ captures the variance of the TD error at each state, weighted by the state distribution μ .

It remains to strengthen this asymptotic guarantee by establishing a finite-sample, high-dimensional Gaussian approximation for $\bar{\Delta}_T$, which has attracted considerable recent attention (e.g., Srikant (2025), Wu *et al.* (2024), and Samsonov *et al.* (2024a)). To this end, a natural tool is the Berry-Esseen bound, which provides an estimate of the distance between the distribution of an estimator and its Gaussian approximation. In high-dimensional settings, such bounds are particularly useful because they yield uniform guarantees over rich classes of sets (e.g., convex sets) (Raič, 2019; Kloeckner, 2019; Shao and Zhang, 2022; Chen *et al.*, 2010), thereby enabling the construction of confidence regions with finite-sample validity. Throughout this subsection, we measure the discrepancy between two probability distributions P and Q using the convex distance

$$d_{\text{cvx}}(P, Q) := \sup_{\mathcal{A} \in \mathcal{C}_d} |P(\mathcal{A}) - Q(\mathcal{A})|, \quad (7.41)$$

where $\mathcal{C}_d$ is the class of all convex sets in $\mathbb{R}^d$. Recalling the definition of the TV distance $\text{TV}(P, Q)$ in (1.2a), we have $d_{\text{cvx}}(P, Q) \leq \text{TV}(P, Q)$. The following theorem provides a non-asymptotic Berry-Esseen bound (Wu *et al.*, 2024).

Theorem 7.11. Consider the TD algorithm (7.35) with Polyak-Ruppert averaging (7.36), independent samples (7.34), and decaying learning rates $\eta_t = \eta_0 t^{-2/3}$ for some $0 < \eta_0 < 1/4$. Then we have

$$d_{\text{cvx}}(\sqrt{T} \bar{\Delta}_T, \mathcal{N}(0, \Lambda^*)) \lesssim \left(\sqrt{\frac{\eta_0}{(1-\gamma)\lambda_0}} \text{Tr}(\Gamma) \|\Gamma^{-1}\| + \frac{\sqrt{S\kappa(\Gamma)}}{(1-\gamma)\lambda_0\eta_0} \right) T^{-\frac{1}{3}} + \tilde{C} T^{-\frac{1}{2}},$$

where $d_{\text{cvx}}(\cdot, \cdot)$ is defined in (7.41), $\kappa(\Gamma)$ is the condition number of Γ , $\lambda_0 = \min_{s \in S} \mu(s)$. Here $\tilde{C} > 0$ is some problem-related quantity independent of T , and polynomial in salient problem parameters; see details in Wu *et al.*, 2024.

This result establishes that $\sqrt{T} \bar{\Delta}_T$ is uniformly close in distribution, over the class of convex subsets of $\mathbb{R}^d$, to a standard Gaussian random vector with the asymptotic covariance (7.39). In this sense, it provides a finite-sample, high-dimensional analog of the classical central limit theorem and Berry-Esseen bound (e.g., Polyak and Juditsky (1992) and Fort (2015)). We note, however, that this guarantee becomes meaningful only when the sample size T exceeds a polynomial function of the dimension, and the minimal sample size T required to ensure such finite-sample validity remains an open question.

7.4.3 Statistical inference

The distributional theory in Theorem 7.11 shows that the distribution of $\sqrt{T} \bar{\Delta}_T$ is well approximated by a Gaussian distribution with covariance matrix Λ^* in finite samples, which naturally suggests a route to statistical inference through Gaussian confidence regions. The main obstacle is that the covariance matrix Λ^* is unknown and must be estimated from data. We therefore begin by introducing an estimator of Λ^* , before showing how it can be used to construct finite-sample confidence sets.

Covariance matrix estimation. Several approaches have been proposed for estimating covariance matrices in stochastic approximation and RL, such as *plug-in estimation* (Chen *et al.*, 2020a) and *bootstrapping* (White and White, 2010; Hanna *et al.*, 2017; Hao *et al.*, 2021; Ramprasad *et al.*, 2022; Zhu *et al.*, 2023b; Samsonov *et al.*, 2024a). Here, we present a computationally efficient online plug-in estimator for Λ^* , as studied in Wu *et al.*, 2024.

Specifically, denoting by e_s the s -th standard basis vector, we define

$$\begin{aligned} \bar{A}_T &:= \frac{1}{T} \sum_{t=1}^T A_t \quad \text{with } A_t := e_{s_t}(e_{s_t} - \gamma e_{s'_t})^\top, \quad \text{and} \\ \hat{\Gamma}_T &:= \sum_{t=1}^T w_{t,T} (A_t \bar{V}_T - b_t)(A_t \bar{V}_T - b_t)^\top \quad \text{with } b_t := r_t e_{s_t} \end{aligned}$$

as estimators of A and Γ , respectively. Here, the weights $\{w_{t,T}\}_{1 \leq t \leq T}$ in $\hat{\Gamma}_T$ are chosen as

$$w_{t,T} := \frac{t}{\sum_{t=1}^T t} = \frac{2t}{T(T+1)};$$

the rationale is that later iterations tend to be more accurate and therefore provide more reliable information about the covariance matrix. We take

$$\hat{\Lambda}_T = \bar{A}_T^{-1} \hat{\Gamma}_T (\bar{A}_T^{-1})^\top \quad (7.42)$$

to be the final estimator of Λ^* . Crucially, the estimator $\hat{\Lambda}_T$ can be computed and updated in an online fashion.

Confidence sets. Equipped with the covariance estimator $\hat{\Lambda}_T$, we now describe a general procedure for constructing convex confidence sets for V^π . Given a prescribed confidence level $1 - \alpha \in (0, 1)$, it suffices to choose any convex set $\mathcal{C}_\alpha$ that contains at least $1 - \alpha$ probability mass under the Gaussian distribution $\mathcal{N}(\bar{V}_T, \hat{\Lambda}_T/T)$; that is,

$$\mathbb{P}(V \in \mathcal{C}_\alpha \mid \bar{V}_T, \hat{\Lambda}_T) \geq 1 - \alpha \quad (7.43)$$

with $V \sim \mathcal{N}(\bar{V}_T, \hat{\Lambda}_T/T)$. By the Gaussian approximation in Theorem 7.11, such sets serve as approximate confidence regions for V^π . Natural choices of $\mathcal{C}_\alpha$ include, but not limited to, the following.

- (i) *Hyper-rectangles*, yielding simultaneous confidence intervals of the form

$$\mathcal{C}_\alpha = \prod_{s=1}^S [\bar{V}_T(s) - \hat{R}/\sqrt{T}, \bar{V}_T(s) + \hat{R}/\sqrt{T}],$$

where $\hat{R} > 0$ is the (random) quantile of the ℓ_∞ norm of a centered Gaussian random vectors with covariance $\hat{\Lambda}_T/T$. This quantile can be calculated efficiently via simulations.

- (ii) *Ellipsoids* or *balls*, for example

$$\mathcal{C}_\alpha = \left\{ V : T (V - \bar{V}_T)^\top \hat{\Lambda}_T^{-1} (V - \bar{V}_T) \leq R^2 \right\},$$

where R is the $1 - \alpha$ quantile of the $\chi^2(d)$ distribution.

7.5 RL with human feedback

Reinforcement learning from human feedback (RLHF) (Ziegler *et al.*, 2019), a paradigm to fine-tune large language models (LLMs), has been shown to be instrumental in significantly improving the helpfulness, truthfulness and controllability of LLMs (Ouyang *et al.*, 2022). Roughly speaking, there are two critical components of RLHF: (1) *reward modeling*, which maps human preference rankings into a quantitative reward function that can guide policy improvement; and (2) *RL fine-tuning*, which seeks to adjust LLM output to align with human preferences by leveraging the learned reward function, i.e., increasing the probability of preferred answers and decreasing the probability of unfavored answers.

Evidently, the curation of preference data is instrumental in the performance of RLHF, which is commonly modeled as pairwise comparisons from a Bradley-Terry ranking model (Bradley and Terry, 1952). In particular, given a query x , human annotators choose a preferred answer from two candidate answers y_1 and y_2 generated by an LLM. Despite the simple form, collecting large-scale and high-quality preference data can be expensive and time-consuming. Depending on the

availability of preference data, two paradigms of RLHF are considered: (1) *offline* RLHF, where only a pre-collected preference dataset is available, possibly generated from a pre-trained LLM after supervised fine-tuning (SFT); and (2) *online* RLHF, where additional preference data can be collected adaptively to improve alignment.

While there has been significant work on the theoretical underpinnings of RLHF that seeks to uncover algorithmic improvements, significant challenges remain, particularly concerning how to incorporate estimates of reward *uncertainty* without explicitly estimating the reward, given that it is often desirable to only maintain the policy itself.

7.5.1 Background

In RLHF, a language model is described by a policy π , which generates an answer $y \in \mathcal{Y}$ given a prompt $x \in \mathcal{X}$ according to the conditional probability distribution $\pi(\cdot | x)$. The standard RLHF process consists of four stages: supervised fine-tuning (SFT), preference data generation, reward modeling, and RL fine-tuning. In the SFT stage, a language model π_{sft} is obtained by fine-tuning a pre-trained LLM with supervised learning. The remaining stages continue training by leveraging the preference data, which we elaborate below.

Reward modeling from preference data. An oracle (e.g., a human labeler or a scoring model) evaluates the quality of two answers y_1 and y_2 given prompt x and reveals its preference. A widely used approach for modeling the probability of pairwise preferences is the Bradley–Terry model (Bradley and Terry, 1952):

$$\begin{aligned} \mathbb{P}(y_1 \succ y_2 | x) &= \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \\ &= \sigma(r^*(x, y_1) - r^*(x, y_2)), \end{aligned} \quad (7.44)$$

where $y_1 \succ y_2$ indicates that y_1 is preferred over y_2 , $r^* : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the ground truth reward function, and $\sigma : \mathbb{R} \rightarrow (0, 1)$ is the logistic function. A preference data sample is denoted by a tuple (x, y_+, y_-) , where y_+ (resp. y_-) is the preferred (resp. unpreferred) answer in the comparison.

Given a preference dataset $\mathcal{D} = \{(x^i, y_+^i, y_-^i)\}$ composed of independent samples, one natural idea to estimate the reward function r is via maximum likelihood estimation (MLE):

$$r_{\text{MLE}} = \arg \min_r \ell(r, \mathcal{D}), \quad (7.45)$$

where $\ell(r, \mathcal{D})$ is the negative log-likelihood of $\mathcal{D}$, given as

$$\ell(r, \mathcal{D}) := - \sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma(r(x^i, y_+^i) - r(x^i, y_-^i)). \quad (7.46)$$

RL fine-tuning. With a reward model r in hand, we seek to fine-tune the policy π to achieve an ideal balance between the expected reward and its distance from an initial reference policy π_{ref} , which is typically set the same as π_{sft} . This is achieved by maximizing the KL-regularized value function $J(r, \pi)$, defined as

$$J(r, \pi) := \mathbb{E}_{x \sim \rho, y \sim \pi(\cdot | x)} [r(x, y)] - \beta \mathbb{E}_{x \sim \rho} [\text{KL}(\pi(\cdot | x) \parallel \pi_{\text{ref}}(\cdot | x))], \quad (7.47)$$

where $\text{KL}(\pi_1 \parallel \pi_2)$ is the KL divergence from π_1 to π_2 , and $\beta > 0$ is a regularization parameter. Consequently, the RL fine-tuned policy π_r with respect to the reward r satisfies

$$\pi_r := \arg \max_{\pi} J(r, \pi), \quad (7.48)$$

which admits a closed-form solution (Rafailov *et al.*, 2023), i.e.,

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}: \quad \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}. \quad (7.49)$$

Here, $Z(r, x)$ is a normalization factor given by

$$Z(r, x) = \sum_{y' \in \mathcal{Y}} \pi_{\text{ref}}(y'|x) \exp(r(x, y')/\beta). \quad (7.50)$$

Direct preference optimization. The closed-form solution (7.49) allows us to write the reward function r in turn as

$$r(x, y) = \beta (\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r, x)). \quad (7.51)$$

Plugging the above equation into the reward MLE (7.45), we obtain the seminal formulation of direct preference optimization (DPO) over the policy space (Rafailov *et al.*, 2023),

$$\pi_{\text{DPO}} = \arg \min_{\pi} - \sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma \left(\beta \left(\log \frac{\pi(y_+^i | x^i)}{\pi_{\text{ref}}(y_+^i | x^i)} - \log \frac{\pi(y_-^i | x^i)}{\pi_{\text{ref}}(y_-^i | x^i)} \right) \right). \quad (7.52)$$

An important advantage of this formulation is that it completely avoids the explicit learning of the reward model, where the normalization factor $Z(r, x)$ cancels out in (7.52), as it is formidable to compute when the number of responses $|\mathcal{Y}|$ is large.

7.5.2 Value-incentivized preference optimization

A major caveat of the standard RLHF framework concerns the lack of accounting for reward uncertainty, which is known to be indispensable in the success of standard RL paradigms in both online and offline settings as seen repetitively in this monograph. This motivates the design of a principled mechanism that can be easily integrated into the RLHF pipeline, while bypassing the difficulties of explicit uncertainty estimation in LLMs.

In view of the suboptimality of naive MLE for reward estimation (Cesa-Bianchi *et al.*, 2017; Rashidinejad *et al.*, 2021), it is tempting to regularize the reward estimate towards those with higher returns (Kumar and Becker, 1982; Liu *et al.*, 2020a)

$$J^*(r) = \max_{\pi} J(r, \pi), \quad (7.53)$$

which measures the resulting value function for the given reward if one acts according to its optimal policy. However, in RLHF, by the definition (7.44), the reward function r^* is only identifiable up to a prompt-dependent global shift. Specifically, suppose that $r_1(x, y)$ and $r_2(x, y)$ are two reward functions that only differ by a prompt-dependent shift $c(x)$, i.e., $r_1(x, y) = r_2(x, y) + c(x)$. It follows that $r_1(x, y_1) - r_1(x, y_2) = r_2(x, y_1) - r_2(x, y_2)$, which means that the two reward functions are not distinguishable in terms of the log-likelihood function since $\ell(r_1, \mathcal{D}) = \ell(r_2, \mathcal{D})$. However, their respective returns are shifted by $J^*(r_1) = J^*(r_2) + \mathbb{E}_{x \sim \rho}[c(x)]$, making $J^*(r)$ improper to be used

directly as a regularizer. To resolve this challenge, we introduce a calibrated value function with respect to a calibration policy π_{cal} , given as

$$J(r, \pi; \pi_{\text{cal}}) = \mathbb{E}_{x \sim \rho, y \sim \pi(\cdot | x)} [r(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot | x)} [r(x, y)] - \beta \mathbb{E}_{x \sim \rho} [\text{KL}(\pi(\cdot | x) \| \pi_{\text{ref}}(\cdot | x))] . \quad (7.54)$$

This calibrated value function is invariant to prompt-dependent shifts of the reward: if $r_1(x, y) = r_2(x, y) + c(x)$, then $J(r_1, \pi; \pi_{\text{cal}}) = J(r_2, \pi; \pi_{\text{cal}})$ for any policy π . We further define the optimal calibrated value

$$J^*(r; \pi_{\text{cal}}) = \max_{\pi} J(r, \pi; \pi_{\text{cal}}), \quad (7.55)$$

which measures the best KL-regularized improvement over the calibration policy under the reward model r .

Cen *et al.* (2025) proposes a regularized MLE of the Bradley-Terry model (7.45) called value-incentivized preference optimization (VPO), which appends the regularization term $J^*(r; \pi_{\text{cal}})$ to the negative log-likelihood

$$r_{\text{VPO}} = \arg \min_r \{ \ell(r, \mathcal{D}) - \text{sign} \cdot \alpha \cdot J^*(r; \pi_{\text{cal}}) \}, \quad (7.56)$$

incentivizing the algorithm to favor (resp. avoid) reward models with higher calibrated value $J^*(r; \pi_{\text{cal}})$ in the online (resp. offline) setting. Here, $\alpha > 0$ is a constant controlling the strength of regularization, and sign is set to 1 in the online setting and -1 in the offline setting, corresponding to the optimism and pessimism for online and offline RL, explained in earlier chapters.

At first glance, the objective function for VPO (7.56) does not immediately imply a computationally-efficient algorithm due to the presence of $J^*(r; \pi_{\text{cal}})$. However, by exploiting the same closed-form solution for the optimal policy given the reward in (7.49), and the reward representation inferred from the policy via (7.51), we can explicitly express $J^*(r; \pi_{\text{cal}})$ as

$$\begin{aligned} J^*(r; \pi_{\text{cal}}) &= \mathbb{E}_{x \sim \rho, y \sim \pi_r(\cdot | x)} [r(x, y) - \beta (\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x))] - \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot | x)} [r(x, y)] \\ &= \beta \mathbb{E}_{x \sim \rho} [\log Z(r, x)] - \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot | x)} [r(x, y)] \\ &\stackrel{(i)}{=} \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot | x)} [\beta \log Z(r, x) - r(x, y)] \\ &\stackrel{(ii)}{=} -\beta \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot | x)} [\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x)], \end{aligned} \quad (7.57)$$

where (i) follows because the bracketed term is independent of y (c.f. (7.49)) and (ii) follows by rearranging (7.49). Given this key ingredient, we can then rewrite (7.56) to directly optimize the policy, in a flavor similar to DP0, as

$$\begin{aligned} \pi_{\text{VPO}} &= \arg \min_{\pi_r} \{ \ell(r, \mathcal{D}) - \text{sign} \cdot \alpha \cdot J^*(r; \pi_{\text{cal}}) \} \\ &= \arg \min_{\pi_r} \left\{ - \sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma \left(\beta \log \frac{\pi_r(y_+^i | x^i)}{\pi_{\text{ref}}(y_+^i | x^i)} - \beta \log \frac{\pi_r(y_-^i | x^i)}{\pi_{\text{ref}}(y_-^i | x^i)} \right) \right. \\ &\quad \left. + \text{sign} \cdot \alpha \beta \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot | x)} [\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x)] \right\} \\ &= \arg \min_{\pi} \left\{ - \sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma \left(\beta \log \frac{\pi(y_+^i | x^i)}{\pi_{\text{ref}}(y_+^i | x^i)} - \beta \log \frac{\pi(y_-^i | x^i)}{\pi_{\text{ref}}(y_-^i | x^i)} \right) \right. \end{aligned}$$

$$+ \text{sign} \cdot \alpha \beta \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{ref}}(y|x)] \}. \quad (7.58)$$

Observing that in the last term $\mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{ref}}(y|x)]$ of (7.58), the reference policy $\pi_{\text{ref}}(y|x)$ does not impact the optimization solution, we can change it to

$$\mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{cal}}(y|x)] = - \mathbb{E}_{x \sim \rho} [\text{KL}(\pi_{\text{cal}}(\cdot|x) \parallel \pi(\cdot|x))],$$

which amounts to adding a KL regularization to the original DPO formulation and offers an interesting interpretation as pushing π against/towards π_{cal} in the online/offline settings, respectively, unveiling the role of reward calibration in RLHF.

In what follows, we elaborate on the development of VPO in the online setting and its possible improvement. For simplicity, we restrict attention to the linear approximation of the reward model.

Assumption 7.1 (Linear Reward). We parameterize the reward model by

$$r_{\theta}(x, y) = \langle \phi(x, y), \theta \rangle, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad (7.59)$$

where $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ is a fixed feature mapping and $\theta \in \mathbb{R}^d$ is the parameter vector. We assume that $\|\phi(x, y)\|_2 \leq 1$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and that $r^*(x, y) = \langle \phi(x, y), \theta^* \rangle$ for some $\theta^* \in \mathbb{R}^d$.

7.5.3 Theory and improvement for online RLHF

The online RLHF procedure extends training by performing reward learning and policy learning iteratively, with a growing preference dataset collected by using the current policy. We use $\pi^{(t)}$ to denote the policy used in the t -th iteration, where the superscript (t) indicates iteration t in the online setting. The t -th iteration of VPO for online RLHF consists of the following steps:

1. **New preference data generation.** We sample a new prompt $x^{(t)} \sim \rho$ and two answers $y_1^{(t)}, y_2^{(t)} \sim \pi^{(t)}(\cdot|x^{(t)})$, query the preference oracle and append $(x^{(t)}, y_+^{(t)}, y_-^{(t)})$ to the preference dataset.
2. **Reward learning.** We train a reward model with preference data $\mathcal{D}^{(t)} := \{(x^{(s)}, y_+^{(s)}, y_-^{(s)})\}_{s=1}^t$ by minimizing the regularized negative log-likelihood, i.e.,

$$r^{(t+1)} = \arg \min_r \{ \ell(r, \mathcal{D}^{(t)}) - \alpha \cdot J^*(r; \pi_{\text{cal}}) \}. \quad (7.60)$$

3. **Policy learning.** This step trains the policy by solving the RL fine-tuning problem:

$$\pi^{(t+1)} = \arg \max_{\pi} J(r^{(t+1)}, \pi). \quad (7.61)$$

In view of the development in Section 7.5.2, the reward learning and policy learning steps can be combined and instead, it is adequate to execute the policy update rule specified in (7.58) using the preference data $\mathcal{D}^{(t)}$ available at the t -th iteration. We summarize the detailed procedure in Algorithm 7.4.

Algorithm 7.4: VP0 for online RLHF

-
- 1 **initialization:** $\pi^{(0)}$.
 - 2 **for** $t = 0, 1, 2, \dots$ **do**
 - 3 Sample $x^{(t)} \sim \rho$, $y_1^{(t)}, y_2^{(t)} \sim \pi^{(t)}(\cdot|x^{(t)})$.
 - 4 Obtain the preference between $(x^{(t)}, y_1^{(t)})$ and $(x^{(t)}, y_2^{(t)})$ from some oracle. Denote the comparison outcome by $(x^{(t)}, y_+^{(t)}, y_-^{(t)})$.
 - 5 Update policy π as
-

$$\pi^{(t+1)} = \arg \min_{\pi} \left\{ - \sum_{s=1}^t \log \sigma \left(\beta \log \frac{\pi(y_+^{(s)}|x^{(s)})}{\pi_{\text{ref}}(y_+^{(s)}|x^{(s)})} - \beta \log \frac{\pi(y_-^{(s)}|x^{(s)})}{\pi_{\text{ref}}(y_-^{(s)}|x^{(s)})} \right) + \alpha \beta \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{ref}}(y|x)] \right\}. \quad (7.62)$$

Theoretical guarantees. Encouragingly, VP0 admits appealing theoretical guarantees under linear function approximation. The next theorem demonstrates that Algorithm 7.4 achieves $\tilde{O}(\sqrt{T})$ cumulative regret under mild assumptions.

Theorem 7.12 (Cen *et al.* (2025)). Under Assumption 7.1, let $r_{\theta^{(t)}}$ denote the corresponding reward model for $\pi^{(t)}$. Assume that $\|\theta^*\|_2 \leq r_{\max}$ and $\|\theta^{(t)}\|_2 \leq r_{\max}$, $\forall t \geq 0$ for some $r_{\max} > 0$. Then, with probability $1 - \delta$, we have

$$\text{Regret} := \sum_{t=1}^T [J^*(r^*; \pi_{\text{cal}}) - J(r^*, \pi^{(t)}; \pi_{\text{cal}})] \leq \tilde{O} \left(\exp(2r_{\max} + r_{\max}/\beta) \sqrt{\kappa d T} \right),$$

with $\alpha = \frac{1}{\exp(2r_{\max} + r_{\max}/\beta)} \sqrt{\frac{T}{\kappa \min\{d \log T, T\}}}$ and $\kappa = \sup_{x,y} \frac{\pi_{\text{cal}}(y|x)}{\pi_{\text{ref}}(y|x)}$.

Theorem 7.12 shows that VP0—without explicit uncertainty quantification—achieves the same $\tilde{O}(\sqrt{T})$ regret for online RLHF as its counterparts in standard contextual bandits with scalar rewards and using UCB for exploration (Lattimore and Szepesvári, 2020).

Adaptive calibration policy. The online procedure above fixes the calibration distribution π_{cal} throughout training. This fixed choice is useful for eliminating the shift ambiguity of the reward function inherent in the Bradley-Terry model, but it also determines which reward differences are emphasized by the optimistic reward estimate and can hinder the exploration. Li and Yan (2025) identifies two opportunities to improve upon VP0:

- Instead of using a fixed π_{cal} , one can adaptively update the calibration policy, in particular, using the current policy $\pi^{(t)}$;
- Instead of drawing both responses following $\pi^{(t)}$, one can sample the responses following $\pi^{(t-1)}$ and $\pi^{(t)}$, respectively.

This leads to a refinement of VPO which in the t -th iteration, adaptively uses the current policy $\pi^{(t)}$ as the calibration policy:

$$\begin{aligned} \pi^{(t+1)} = \operatorname{argmin}_{\pi} \Big\{ & - \sum_{s=1}^t \log \sigma \left(\beta \log \frac{\pi(y_+^{(s)}|x^{(s)})}{\pi_{\text{ref}}(y_+^{(s)}|x^{(s)})} - \beta \log \frac{\pi(y_-^{(s)}|x^{(s)})}{\pi_{\text{ref}}(y_-^{(s)}|x^{(s)})} \right) \\ & + \alpha_t \beta \mathbb{E}_{x \sim \rho, y \sim \pi^{(t)}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{ref}}(y|x)] \Big\}. \end{aligned} \quad (7.63)$$

In the case of multi-armed bandits, this yields a regret bound of order $T^{(\beta+1)/(\beta+2)}$ up to polynomial dependency with other problem parameters. When $\beta = 0$, the rate reduces to the usual $\tilde{O}(\sqrt{T})$ form. In general, the slower exponent in T reflects the trade-off between reward maximization and the KL regularization strength, while avoiding the exponential dependence on quantities such as $r_{\max}/\beta$ that can appear under fixed-calibration sampling protocols such as VPO (cf. Theorem 7.12).

7.6 Notes

Distributionally robust RL. Beyond the discounted infinite-horizon setting, distributionally robust RL has also been studied for finite-horizon MDPs (Xu *et al.*, 2023; Dong *et al.*, 2024; Lu *et al.*, 2024), average-reward MDPs (Chen *et al.*, 2026b), constrained RL (Zhang *et al.*, 2024b), and multi-agent RL (Shi *et al.*, 2025; Shi *et al.*, 2024b; Qu *et al.*, 2026). Several other forms of uncertainty sets have been studied in the literature as well. For instance, Wang and Zou (2021) considered R -contamination uncertain sets and developed a provably robust Q-learning algorithm for the online setting, whereas Clavier *et al.* (2024) studied the ℓ_p -based uncertainty sets under the generative model. Another widely studied class of uncertainty sets is based on KL divergence. In this context, Yang *et al.* (2022), Panaganti and Kalathil (2022), Zhou *et al.* (2021c), Shi and Chi (2024), Xu *et al.* (2023), Wang *et al.* (2023a), Blanchet *et al.* (2023), Wang *et al.* (2024c), Liang *et al.* (2023), and Wang *et al.* (2024b) investigated the statistical performance of various algorithms under several sampling mechanisms.

Last but not least, distributionally robust RL has also been extended beyond the tabular setting. We refer the interested reader to Badrinath and Kalathil (2021), Ramesh *et al.* (2024), Panaganti *et al.* (2022), Wang *et al.* (2024a), Liu and Xu (2024), and Mandal *et al.* (2025) for recent developments involving function approximation.

Markov games. Moving beyond the generative-model setting, a recent line of work has focused on the design and analysis of a class of online MARL algorithms known as *V-learning*, which leverages online adversarial learning techniques and adaptive sampling to mitigate the curse of multi-agents (Mao and Başar, 2022; Song *et al.*, 2022; Jin *et al.*, 2024a). We refer the interested reader to these papers for algorithmic details. Importantly, V-learning was the first framework to reveal the following striking statistical phenomenon: what ultimately matters in the sample complexity is the largest individual action space $\max_{1 \leq i \leq m} A_i$, as opposed to the size of the joint action space $\prod_{i=1}^m A_i$. This insight leads to substantial improvements in data efficiency and helps overcome one of the principal obstacles in multi-agent RL.

Despite this significant progress, the current theoretical guarantees for V-learning remain suboptimal in their horizon dependence. More broadly, the fundamental statistical limits of online learning in Markov games remain unsettled and appear to be considerably more subtle than their

single-agent counterparts. For instance, the optimal horizon dependence is still unknown: it remains unclear whether the minimax sample complexity of equilibrium learning should exhibit the same H^3 dependence (in terms of the number of episodes required) that arises in single-agent online RL, or whether the multi-agent setting gives rise to fundamentally different horizon scaling. Progress on such questions has been hindered both by the challenge of developing sharper upper-bound analyses and by the difficulty of constructing tight lower bounds.

In addition to online Markov games and the generative-model setting, the curse of multi-agency also manifests itself in offline MARL, where it has been addressed in work such as Cui and Du (2022b) and Yan *et al.* (2024). Moreover, while our discussion has focused primarily on competitive Markov games, several other important variants have received considerable attention, including Markov potential games (Song *et al.*, 2022; Ding *et al.*, 2022) and robust Markov games (Shi *et al.*, 2024b; Shi *et al.*, 2025; Jiao and Li, 2024; Blanchet *et al.*, 2023). The sample complexity of MARL under function approximation has likewise been investigated, with particular emphasis on mitigating the curse of multi-agents; see, for example, Dai *et al.* (2024), Cui *et al.* (2023), and Wang *et al.* (2023c). Finally, Yang *et al.* (2025b) and Liu *et al.* (2023b) developed efficient exploration mechanisms for online MARL that avoid explicit uncertainty estimation while still achieving strong performance.

Federated RL. A growing body of recent work has investigated the sample complexity of federated value-based algorithms in various settings, such as online RL (Zheng *et al.*, 2024; Zhang *et al.*, 2025a; Zhang *et al.*, 2026a), offline RL (Woo *et al.*, 2024), asynchronous Q-learning (Woo *et al.*, 2023), average-reward RL (Jiao *et al.*, 2026), and heterogeneous RL (Zhang *et al.*, 2024a; Wang *et al.*, 2025a; Yang *et al.*, 2024b). Federated variants of model-based RL algorithms have also been studied; see, for example, Labbi *et al.* (2025) and Zhou *et al.* (2024). In addition to policy learning, several recent studies (Liu and Olshevsky, 2023; Tian *et al.*, 2024) established that a single round of communication is sufficient to achieve linear speedup of TD learning for *policy evaluation*. Note that these results do not contradict the communication lower bounds established by Salgia and Chi (2024), which concern Q-learning for *policy learning*.

Statistical inference in RL. Moving beyond the i.i.d. data setting (7.34) discussed in this chapter and in Mou *et al.* (2020), Wu *et al.* (2024), Bhandari *et al.* (2018), and Samsonov *et al.* (2024a), there has been substantial recent progress on statistical analysis and inference for Markov-chain-based RL algorithms driven by Markovian sample trajectories; see, e.g., Bhandari *et al.*, 2018; Mou *et al.*, 2020; Fan *et al.*, 2025b; Li *et al.*, 2022b; Ramprasad *et al.*, 2023; Samsonov *et al.*, 2024b; Li *et al.*, 2023d; Samsonov *et al.*, 2024a; Srikant, 2025. The state-of-the-art finite-sample distributional guarantees for Polyak-Ruppert averaged TD learning (e.g., finite-sample Berry-Esseen bounds) have been developed by Wu *et al.*, 2025; Srikant, 2025.

Beyond settings in which samples are generated under a fixed policy, it is also important to study statistical inference in online environments, where data are collected adaptively through sequential interaction with the environment (Zhang *et al.*, 2020c; Chen *et al.*, 2021a; Shen *et al.*, 2024). In such scenarios, adaptive data collection induces intricate dependencies across observations, often invalidating classical inferential techniques developed for independent data. Motivated by these challenges, a recent line of work has investigated statistical inference in bandit problems with adaptively collected data (Kalvit and Zeevi, 2021; Zhang *et al.*, 2020c; Dimakopoulou *et al.*, 2021;

Hadad *et al.*, 2021; Khamaru and Zhang, 2024; Han *et al.*, 2024a; Halder *et al.*, 2025; Guo and Xu, 2025; Fan *et al.*, 2025b; Han *et al.*, 2025; Chang *et al.*, 2026). For instance, one important insight emerging from this literature is that UCB-type algorithms can serve not only as efficient exploration strategies but also as mechanisms for valid statistical inference. Extending these ideas beyond bandits and developing principled inferential paradigms for online RL remains an important direction for future research.

RLHF. RLHF has been attracting considerable attention in recent years, and, in particular, the connection to the classical statistical ranking formulation has been explored (Zhu *et al.*, 2023a). Besides the Bradley-Terry model, other ranking models have also been considered for RLHF, including but not limited to the Mallows model (Chen *et al.*, 2025b). Concurrent with VPO, several papers have proposed similar regularization techniques to impose the principles of optimism and pessimism, respectively, for online and offline RLHF, e.g., Xie *et al.* (2024), Zhang *et al.* (2025d), and Liu *et al.* (2024). Another popular approach to RLHF is via game-theoretic formulations, e.g., Azar *et al.* (2024), which are less prone to reward hacking. Finally, we refer the interested reader to Liu *et al.* (2026) for a recent survey of RLHF from a statistical perspective.

8

Concluding remarks

In this monograph, we have developed a non-asymptotic statistical perspective on reinforcement learning, focusing primarily on tabular MDPs under three canonical sampling mechanisms: access to a simulator, online exploration, and offline data. By emphasizing sharp dependencies of the statistical performance on salient problem parameters — especially the dimension of the state-action space and the planning horizon — our treatment highlights the critical role of the data-collection mechanism in shaping both the fundamental statistical limits and the algorithmic principles needed to attain them.

Several recurring themes have emerged from the non-asymptotic viewpoints developed throughout this monograph, and we highlight a few of them below. First, obtaining sharp finite-sample guarantees requires exploiting the intrinsic Bellman structure of MDPs, rather than relying solely on generic concentration-of-measure arguments. Techniques such as variance-aware analyses, leave-one-out decoupling arguments, and carefully calibrated uncertainty quantification play a pivotal role in overcoming substantial sample-size barriers and enabling statistically optimal RL algorithms. Second, optimism and pessimism arise as complementary guiding principles: optimism drives efficient exploration in online RL, whereas pessimism guards against distribution shift and insufficient coverage in offline RL. In settings beyond these canonical sampling paradigms, the two seemingly opposing principles may even need to be carefully integrated in order to achieve statistical efficiency. Third, the comparison between model-based and model-free approaches reveals a persistent gap in statistical performance. Model-based methods, often coupled with carefully designed uncertainty quantification, remain the only class of approaches known to provably attain minimax-optimal sample complexity across the entire accuracy regime — including the most sample-starved settings — under the three canonical sampling mechanisms studied in this monograph. By contrast, classical model-free methods such as Q-learning may suffer from inherent statistical inefficiency and frequently require sophisticated algorithmic modifications even to achieve asymptotically optimal statistical guarantees.

Taken collectively, the developments covered in this monograph illustrate both the remarkable progress and the substantial challenges that continue to shape modern statistical RL. Ultimately, we hope that this monograph equips new researchers with some of the core conceptual and technical

toolkits of the field, while motivating further theoretical and algorithmic investigations.

Naturally, the scope of this monograph has been selective, and many important directions in modern RL fall outside its coverage. Below, we briefly highlight a few representative examples.

- *Beyond discounted and episodic settings.* This monograph has focused primarily on discounted infinite-horizon MDPs and finite-horizon episodic MDPs. A few other formulations have also received considerable attention in the RL literature. One classic and practically important example is *average-reward RL*, where performance is measured by the long-run average reward rather than by a discounted or finite-horizon return (Puterman, 2014; Bartlett and Tewari, 2009; Wan *et al.*, 2021). Average-reward formulations arise naturally in continuing tasks and long-running control systems and often require techniques that differ substantially from those used in discounted or episodic settings. Significant recent progress has been made in designing efficient algorithms and establishing non-asymptotic performance guarantees in the context of average-reward MDPs (Jin and Sidford, 2021; Liao *et al.*, 2022; Zhang *et al.*, 2021c; Zurek and Chen, 2024; Jin *et al.*, 2024b; Chen and Maguluri, 2026; Jiao *et al.*, 2026), although many fundamental statistical questions remain open.
- *Policy optimization and actor-critic methods.* Policy optimization constitutes another major paradigm in modern RL, which seeks to improve parameterized policies directly through optimization-based procedures (Williams, 1992; Sutton *et al.*, 2000; Kakade, 2002; Peters and Schaal, 2008). Owing to their flexibility in accommodating rich policy parameterizations, policy-gradient methods and their variants (e.g., natural policy gradient methods, trust-region methods, and proximal policy optimization) have played a central role in the successes of deep RL (Sutton *et al.*, 2000; Mnih *et al.*, 2015; Schulman *et al.*, 2015; Silver *et al.*, 2016; Schulman *et al.*, 2017) and have attracted substantial theoretical attention (e.g., Agarwal *et al.* (2019b), Mei *et al.* (2020), Cen *et al.* (2022), Li *et al.* (2023b), Lan (2023), Zhan *et al.* (2023a), and Bhandari and Russo (2024)). Actor-critic methods (Konda and Tsitsiklis, 2000) further combine policy optimization with value estimation, using a critic to evaluate the current policy and guide its improvement. Despite their empirical effectiveness and statistical studies (Xu *et al.*, 2020a; Cai *et al.*, 2020; Yuan *et al.*, 2023; Chen and Zhao, 2023; Tan *et al.*, 2025), the statistical foundations of policy optimization and actor-critic methods remain considerably less developed than those of the model-based and value-based approaches discussed throughout this monograph.
- *RL in continuous time and continuous spaces.* While this monograph has focused on finite-state, finite-action MDPs, many important applications involve continuous state-action spaces and/or continuous-time dynamics (Recht, 2019; Van Hasselt, 2012; Wang *et al.*, 2020a; Hazan *et al.*, 2025; Guo *et al.*, 2026). Such settings are intimately connected to stochastic control and introduce a range of statistical, computational, and modeling challenges. From a statistical perspective, finite-sample guarantees for both sample complexity and regret have been established for a few structured models, most notably linear-quadratic regulators (e.g., Abbasi-Yadkori and Szepesvári (2011), Dean *et al.* (2020), Simchowitz and Foster (2020), Zheng *et al.* (2021), Zhang *et al.* (2021b), and Wang and Janson (2021)). Note that rigorous learning guarantees in continuous domains often rely on additional structural assumptions, such as linearity, kernel structure, or other forms of tractable function approximation.

Beyond the directions highlighted above, a number of other active research areas continue to shape the development of modern RL. We conclude by singling out several directions that have recently attracted significant attention and are likely to play an important role in the future development of the field.

- *Hybrid RL.* While this monograph has addressed generative models, offline RL, and online RL as separate sampling mechanisms, it has omitted important settings in which an agent has simultaneous access to multiple data collection mechanisms. For example, an agent may be allowed to perform a limited amount of online exploration while also leveraging a pre-collected offline dataset, or it may be able to collect samples in real time while making extensive use of a simulator. Such hybrid RL settings have garnered significant recent attention from both theoretical and practical perspectives (e.g., Xie *et al.*, 2023; Song *et al.*, 2023; Li *et al.*, 2024f; Wagenmaker and Pacchiano, 2023; Tan *et al.*, 2024; Panaganti *et al.*, 2024; Qu *et al.*, 2025), offering insights that extend far beyond the studies of online RL, offline RL, and generative models in isolation.
- *Towards more general function approximation.* The tabular theory developed in this monograph should be viewed as a starting point rather than an endpoint for understanding the statistical principles underlying RL. A central objective of modern RL theory is to understand learning and decision-making with rich function approximation (e.g., Foster *et al.* (2021), Fan *et al.* (2020b), Chen *et al.* (2025a), and Yang *et al.* (2025b)). Existing theoretical guarantees beyond the tabular setting, however, typically rely on highly structured function classes, such as linear models, kernel methods, low-rank representations, or other settings that admit tractable statistical and computational analyses. Although these frameworks have yielded important conceptual and technical insights, they remain far from capturing the highly expressive function approximators used in modern practice, most notably deep neural networks. Bridging this gap will require fundamentally new ideas that go well beyond current theories of linear and low-complexity function approximation.
- *Partially observable MDPs (POMDPs).* The MDP framework considered throughout this monograph assumes full observability of the underlying state. In many applications, however, agents must make decisions based only on noisy or partial observations, leading to partially observable MDPs (POMDPs) (Åström, 1965; Kaelbling *et al.*, 1998). The presence of unobserved states introduces substantial statistical and computational challenges, as optimal decisions may depend on the entire observation history rather than the current observation alone. Consequently, both learning, planning, and policy evaluation become significantly more difficult than in fully observable MDPs. From a statistical perspective, obtaining rigorous finite-sample guarantees often requires additional structural assumptions, such as observability, low-rank structure, or suitable mixing conditions. Despite remarkable recent progress (e.g., Liu *et al.* (2022), Liu *et al.* (2023a), Golowich *et al.* (2022), Uehara *et al.* (2022a), Shi *et al.* (2022a), Zhan *et al.* (2023b), and Chen *et al.* (2023)), the statistical foundations of POMDPs remain considerably less mature than those of fully observable MDPs.
- *Transfer learning for RL from heterogeneous data.* Another compelling frontier is the development of RL methods that can effectively leverage heterogeneous data sources. Classical offline RL frameworks often assume that historical data are drawn from a single homogeneous

population. In contrast, real-world applications such as personalized medicine, recommendation systems, and e-commerce routinely involve data collected from diverse subpopulations. Naively pooling these multi-source datasets and treating them as arising from a common distribution often leads to suboptimal statistical performance, as it fails to exploit the latent similarities and structural relationships across sources. Addressing this challenge requires new algorithmic and statistical foundations that can effectively integrate transfer learning into RL (Taylor and Stone, 2009; Chen *et al.*, 2022a; Zhu *et al.*, 2023c; Cai *et al.*, 2024; Chai *et al.*, 2025; Miao *et al.*, 2025). The goal is to enable RL agents to effectively leverage information across related domains, populations, or environments, thereby improving sample efficiency and generalization while maintaining strong statistical guarantees.

- *The interplay between RL and diffusion models.* Diffusion models (Lai *et al.*, 2025; Chen *et al.*, 2024a) have recently emerged as a powerful generative framework for RL and control, with applications ranging from diffusion policies for imitation learning and offline RL (Chi *et al.*, 2025a; Ajay *et al.*, 2023) to diffusion-based planning and policy optimization (Janner *et al.*, 2022). Their ability to represent complex, multimodal action distributions makes them particularly appealing for high-dimensional decision-making problems. Conversely, RL and control have increasingly been employed to improve and fine-tune diffusion models themselves, enabling alignment with downstream objectives that are difficult to specify through conventional likelihood-based training alone (Black *et al.*, 2024; Fan *et al.*, 2023; Han *et al.*, 2024b; Yang *et al.*, 2026b; Zhao *et al.*, 2025a; Jiao *et al.*, 2025; Gao *et al.*, 2025). Despite substantial empirical progress, the statistical and algorithmic foundations of these developments remain significantly underdeveloped, which point to a rich and rapidly growing research agenda at the intersection of RL and generative AI.
- *RL for reasoning in foundation models.* Recent years have witnessed a remarkable expansion of RL beyond its traditional applications in sequential decision-making. In particular, RL has emerged as a cornerstone of the post-training and alignment pipelines for foundation models, where it is increasingly deployed to enhance reasoning, planning, and other high-level cognitive capabilities (Ouyang *et al.*, 2022; Guo *et al.*, 2025). Unlike the settings studied in this monograph, the underlying environments are often generated by large language models themselves, the action spaces consist of long sequences of tokens, and rewards may be sparse, delayed, and derived from automated verifiers and/or human feedback. These introduce a host of new statistical and algorithmic questions concerning exploration, credit assignment, reward design, and the interplay between RL training and model architecture. Developing a principled theoretical framework that explains when and how RL improves reasoning in foundation models has emerged as an important frontier of modern RL (see, e.g., Kim *et al.* (2025), Shang *et al.* (2026), Huang *et al.* (2026), Foster *et al.* (2025), Zhao *et al.* (2025b), Ran-Milo *et al.* (2026), Chen *et al.* (2026a), and Yang *et al.* (2026a)).

Acknowledgements

This monograph will not be possible without the contributions and encouragement of our collaborators and students over the years, including but not limited to Changxiao Cai, Shicong Cen, Chen Cheng, Bo Dai, Hanjun Dai, Simon Du, Wei Fan, Jingchu Gai, Matthieu Geist, Yu Huang, Yuchen Jiao, Gauri Joshi, Jason Lee, Jiadong Liang, Yingbin Liang, Cong Ma, Jincheng Mei, Eric Mazumdar, Alessandro Rinaldo, Sudeep Salgia, Dale Schuurmans, Laixi Shi, Kevin Tan, Bingyan Wang, He Wang, Zixin Wen, Adam Wierman, Jiin Woo, Weichen Wu, Lin Xiao, Tong Yang, Yuepeng Yang, Wenhao Zhan, and Yuchen Zhou.

We would also like to thank the generous support of multiple funding agencies. Y. Chen has been supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-19-1-0030 and FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773, and the Wharton AI & Analytics Initiative's AI Research Fund. Y. Chi has been supported in part by the grants ONR N00014-19-1-2404, AFRL FA8750-20-2-050, NSF CCF-2106778, CNS-2148212, and ECCS-2318441, and in part by funds from federal agencies and industry partners as specified in the NSF Resilient & Intelligent NextG Systems (RINGS) program. J. Fan has been supported in part by the ONR Grants N00014-25-1-2317, ONR N00014-22-1-2340, and N00014-19-1-2120, the NSF grants DMS-2210833, DMS-2053832, and DMS-2052926, and the NIH grant 2R01-GM072611-13. G. Li is supported in part by the Chinese University of Hong Kong Direct Grant for Research and the Hong Kong Research Grants Council ECS 24305724 and GRF 14307525. Y. Wei has been supported in part by the Google Research Scholar Award, the NSF grants CCF-2106778, DMS-2147546/2015447 and CAREER award DMS-2143215, and the Wharton Dean's Research Fund.

References

- Abbasi-Yadkori, Y. and C. Szepesvári. (2011). “Regret Bounds for the Adaptive Control of Linear Quadratic Systems”. In: *Conference on Learning Theory*. Vol. 19. 1–26.
- Abbe, E., J. Fan, K. Wang, and Y. Zhong. (2020). “Entrywise eigenvector analysis of random matrices with low expected rank”. *Annals of statistics*. 48(3): 1452.
- Agarwal, A., N. Jiang, S. M. Kakade, and W. Sun. (2019a). “Reinforcement learning: Theory and algorithms”. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep.* 32: 96.
- Agarwal, A., S. Kakade, and L. F. Yang. (2020a). “Model-based reinforcement learning with a generative model is minimax optimal”. In: *Conference on Learning Theory*. 67–83.
- Agarwal, A., S. M. Kakade, J. D. Lee, and G. Mahajan. (2019b). “Optimality and approximation with policy gradient methods in Markov decision processes”. *arXiv preprint arXiv:1908.00261*: 4.
- Agarwal, R. P., M. Meehan, and D. O’regan. (2001). *Fixed point theory and applications*. Vol. 141. Cambridge university press.
- Agarwal, R., D. Schuurmans, and M. Norouzi. (2020b). “An optimistic perspective on offline reinforcement learning”. In: *International conference on machine learning*. 104–114.
- Agrawal, P., S. Agrawal, and A. Azati. (2026). “Q-learning with Posterior Sampling”. In: *The Fourteenth International Conference on Learning Representations*.
- Agrawal, R. (1995). “Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem”. *Advances in applied probability*. 27(4): 1054–1078.
- Agrawal, S. and R. Jia. (2017). “Optimistic Posterior Sampling for Reinforcement Learning: Worst-Case Regret Bounds”. 30.
- Agterberg, J. and A. R. Zhang. (2025). “Estimating Higher-Order Mixed Memberships via the $\ell_{2,\infty}$ Tensor Perturbation Bound”. *Journal of the American Statistical Association*. 120(550): 1214–1224.
- Ajay, A., Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal. (2023). “Is Conditional Generative Modeling all you need for Decision-Making?” In: *International Conference on Learning Representations*.
- Alon, N. and J. H. Spencer. (2008). *The Probabilistic Method (3rd Edition)*. Wiley.
- Åström, K. J. (1965). “Optimal control of Markov processes with incomplete state information I”. *Journal of mathematical analysis and applications*. 10: 174–205.

- Audibert, J.-Y. and S. Bubeck. (2009). “Minimax policies for adversarial and stochastic bandits”. In: *Colt*. 217–226.
- Auer, P., N. Cesa-Bianchi, and P. Fischer. (2002). “Finite-time analysis of the multiarmed bandit problem”. *Machine learning*. 47(2): 235–256.
- Auer, P. and R. Ortner. (2006). “Logarithmic online regret bounds for undiscounted reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 19.
- Aumann, R. J. (1987). “Correlated equilibrium as an expression of Bayesian rationality”. *Econometrica: Journal of the Econometric Society*: 1–18.
- Ayoub, A., Z. Jia, C. Szepesvári, M. Wang, and L. F. Yang. (2020). “Model-Based Reinforcement Learning with Value-Targeted Regression”. In: *International Conference on Machine Learning*. Vol. 119. 463–474.
- Azar, M. G., H. J. Kappen, M. Ghavamzadeh, and R. Munos. (2011). “Speedy Q-learning”. In: *Advances in Neural Information Processing Systems*. 2411–2419.
- Azar, M. G., Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. (2024). “A general theoretical paradigm to understand learning from human preferences”. In: *International Conference on Artificial Intelligence and Statistics*. 4447–4455.
- Azar, M. G., R. Munos, and H. J. Kappen. (2013). “Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model”. *Machine learning*. 91(3): 325–349.
- Azar, M. G., I. Osband, and R. Munos. (2017). “Minimax regret bounds for reinforcement learning”. In: *International Conference on Machine Learning*. 263–272.
- Azuma, K. (1967). “Weighted sums of certain dependent random variables”. *Tohoku Mathematical Journal, Second Series*. 19(3): 357–367.
- Badrinath, K. P. and D. Kalathil. (2021). “Robust reinforcement learning using least squares policy iteration with provable performance guarantees”. In: *International Conference on Machine Learning*. 511–520.
- Bai, Y., Y. Gao, R. Wan, S. Zhang, and R. Song. (2025). “A review of reinforcement learning in financial applications”. *Annual Review of Statistics and Its Application*. 12(1): 209–232.
- Bai, Y., T. Xie, N. Jiang, and Y.-X. Wang. (2019). “Provably efficient Q-learning with low switching cost”. *Advances in Neural Information Processing Systems*. 32.
- Bartlett, P. L. and A. Tewari. (2009). “REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs”. In: *Uncertainty in Artificial Intelligence*. 35–42.
- Bastani, H., K. Drakopoulos, V. Gupta, I. Vlachogiannis, C. Hadjichristodoulou, P. Lagiou, G. Magiorkinis, D. Paraskevis, and S. Tsiodras. (2021). “Efficient and targeted COVID-19 border testing via reinforcement learning”. *Nature*. 599(7883): 108–113.
- Bäuerle, N. and A. Glauner. (2022). “Distributionally robust Markov decision processes and their connection to risk measures”. *Mathematics of Operations Research*. 47(3): 1757–1780.
- Beck, C. L. and R. Srikant. (2012). “Error bounds for constant step-size Q-learning”. *Systems & control letters*. 61(12): 1203–1208.
- Bellemare, M. G., Y. Naddaf, J. Veness, and M. Bowling. (2013). “The arcade learning environment: An evaluation platform for general agents”. *Journal of artificial intelligence research*. 47: 253–279.
- Bellman, R. (1952). “On the theory of dynamic programming”. *Proceedings of the National Academy of Sciences of the United States of America*. 38(8): 716.
- Bertsekas, D. (2019). *Reinforcement learning and optimal control*. Vol. 1. Athena Scientific.
- Bertsekas, D. P. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific.

- Bertsimas, D., V. Gupta, and N. Kallus. (2018). “Data-driven robust optimization”. *Mathematical Programming*. 167(2): 235–292.
- Bertsimas, D., M. Sim, and M. Zhang. (2019). “Adaptive distributionally robust optimization”. *Management Science*. 65(2): 604–618.
- Bhandari, J. and D. Russo. (2024). “Global optimality guarantees for policy gradient methods”. *Operations Research*. 72(5): 1906–1927.
- Bhandari, J., D. Russo, and R. Singal. (2018). “A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation”. In: *Conference On Learning Theory*. 1691–1692.
- Bhandari, J., D. Russo, and R. Singal. (2021). “A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation”. *Operations Research*. 69(3): 950–973.
- Bhardwaj, M., T. Xie, B. Boots, N. Jiang, and C.-A. Cheng. (2023). “Adversarial model for offline reinforcement learning”. *Advances in Neural Information Processing Systems*. 36: 1245–1269.
- Bhattacharya, S., J. Fan, and J. Hou. (2026). “Inferences on mixing probabilities and ranking in mixed-membership models”. *Journal of the American Statistical Association*. (just-accepted): 1–29.
- Bian, Z., C. Shi, Z. Qi, and L. Wang. (2025). “Off-policy evaluation in doubly inhomogeneous environments”. *Journal of the American Statistical Association*. 120(550): 1102–1114.
- Black, K., M. Janner, Y. Du, I. Kostrikov, and S. Levine. (2024). “Training diffusion models with reinforcement learning”. In: *International Conference on Learning Representations*. Vol. 2024. 4965–4987.
- Blanchet, J., M. Lu, T. Zhang, and H. Zhong. (2023). “Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 66845–66859.
- Blanchet, J. and K. Murthy. (2019). “Quantifying distributional model risk via optimal transport”. *Mathematics of Operations Research*. 44(2): 565–600.
- Borkar, V., S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. (2025). “The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning”. *The Annals of Applied Probability*. 35(2): 936–982.
- Borkar, V. S. and V. S. Borkar. (2008). *Stochastic approximation: a dynamical systems viewpoint*. Vol. 100. Springer.
- Borkar, V. S. and S. P. Meyn. (2000). “The ODE method for convergence of stochastic approximation and reinforcement learning”. *SIAM Journal on Control and Optimization*. 38(2): 447–469.
- Bottou, L. (2010). “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of International Conference on Computational Statistics*. Springer. 177–186.
- Bottou, L., F. E. Curtis, and J. Nocedal. (2018). “Optimization methods for large-scale machine learning”. *SIAM review*. 60(2): 223–311.
- Boucheron, S., G. Lugosi, and P. Massart. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- Bourel, H., O. Maillard, and M. S. Talebi. (2020). “Tightening exploration in upper confidence reinforcement learning”. In: *International Conference on Machine Learning*. 1056–1066.
- Bradley, R. A. and M. E. Terry. (1952). “Rank analysis of incomplete block designs: I. The method of paired comparisons”. *Biometrika*. 39(3/4): 324–345.
- Brantley, K., M. Dudik, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun. (2020). “Constrained episodic reinforcement learning in concave-convex and knapsack settings”. *Advances in Neural Information Processing Systems*. 33: 16315–16326.

- Brown, N. and T. Sandholm. (2019). “Superhuman AI for multiplayer poker”. *Science*. 365(6456): 885–890.
- Bubeck, S. and N. Cesa-Bianchi. (2012). “Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems”. In: *Foundations and Trends® in Machine Learning*. Vol. 5. No. 1. 1–122.
- Bura, A., A. HasanzadeZonuzi, D. Kalathil, S. Shakkottai, and J.-F. Chamberland. (2022). “DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning”. *Advances in neural information processing systems*. 35: 1047–1059.
- Burnetas, A. N. and M. N. Katehakis. (1996). “Optimal adaptive policies for sequential allocation problems”. *Advances in Applied Mathematics*. 17(2): 122–142.
- Cai, C., T. T. Cai, and H. Li. (2024). “Transfer learning for contextual multi-armed bandits”. *The Annals of Statistics*. 52(1): 207–232.
- Cai, C., G. Li, H. V. Poor, and Y. Chen. (2022). “Nonconvex Low-Rank Tensor Completion from Noisy Data”. *Operations Research*. 70(2): 1219–1237.
- Cai, J.-F., X. Qu, W. Xu, and G.-B. Ye. (2016). “Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction”. *Applied and computational harmonic analysis*. 41(2): 470–490.
- Cai, Q., Z. Yang, C. Jin, and Z. Wang. (2020). “Provably efficient exploration in policy optimization”. In: *International Conference on Machine Learning*. 1283–1294.
- Cai, Q., Z. Yang, J. D. Lee, and Z. Wang. (2019). “Neural temporal-difference learning converges to global optima”. *Advances in Neural Information Processing Systems*. 32.
- Cappé, O., A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. (2013). “Kullback-Leibler upper confidence bounds for optimal sequential allocation”. *The Annals of Statistics*: 1516–1541.
- Cayci, S., S. Satpathi, N. He, and R. Srikant. (2023). “Sample complexity and overparameterization bounds for temporal-difference learning with neural network approximation”. *IEEE Transactions on Automatic Control*. 68(5): 2891–2905.
- Cen, S., C. Cheng, Y. Chen, Y. Wei, and Y. Chi. (2022). “Fast global convergence of natural policy gradient methods with entropy regularization”. *Operations Research*. 70(4): 2563–2578.
- Cen, S., J. Mei, K. Goshvadi, H. Dai, T. Yang, S. Yang, D. Schuurmans, Y. Chi, and B. Dai. (2025). “Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF”. In: *The Thirteenth International Conference on Learning Representations*.
- Cesa-Bianchi, N., C. Gentile, G. Lugosi, and G. Neu. (2017). “Boltzmann exploration done right”. In: *Advances in Neural Information Processing Systems*. Vol. 30.
- Chai, J., E. Chen, and J. Fan. (2025). “Deep Transfer Q-Learning for Offline Non-Stationary Reinforcement Learning”. *arXiv preprint arXiv:2501.04870*.
- Chang, X., X. Chen, Z. Lai, H. Li, Z. Liu, and Y. Zhang. (2026). “Online statistical inference for contextual bandits via stochastic gradient descent”. *Journal of the American Statistical Association*: 1–24.
- Chen, E. Y., M. I. Jordan, and S. Li. (2022a). “Transferred Q-learning”. *arXiv preprint arXiv:2202.04709*.
- Chen, F., Z. Jia, A. Rakhlin, and T. Xie. (2026a). “Outcome-based online reinforcement learning: Algorithms and fundamental limits”. *Advances in Neural Information Processing Systems*. 38: 106095–106133.
- Chen, F., S. Mei, and Y. Bai. (2025a). “Unified algorithms for RL with Decision-Estimation Coefficients: PAC, reward-free, preference-based learning and beyond”. *The Annals of Statistics*. 53(1): 426–456.

- Chen, F., H. Wang, C. Xiong, S. Mei, and Y. Bai. (2023). “Lower bounds for learning in revealing POMDPs”. In: *International Conference on Machine Learning*. 5104–5161.
- Chen, H., H. Zhao, H. Lam, D. Yao, and W. Tang. (2025b). “MallowsPO: Fine-tune your LLM with preference dispersions”. In: *International Conference on Learning Representations*. Vol. 2025. 31387–31419.
- Chen, H., W. Lu, and R. Song. (2021a). “Statistical inference for online decision making: In a contextual bandit setting”. *Journal of the American Statistical Association*. 116(533): 240–255.
- Chen, J. and N. Jiang. (2019). “Information-theoretic considerations in batch reinforcement learning”. In: *International Conference on Machine Learning*. 1042–1051.
- Chen, L., M. Jafarnia-Jahromi, R. Jain, and H. Luo. (2021b). “Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path”. In: *Advances in Neural Information Processing Systems*. Vol. 34.
- Chen, L. H., L. Goldstein, and Q.-M. Shao. (2010). *Normal approximation by Stein’s method*. Springer Science & Business Media.
- Chen, M., S. Mei, J. Fan, and M. Wang. (2024a). “Opportunities and challenges of diffusion models for generative AI”. *National Science Review*. 11(12): nwae348.
- Chen, P., C. Gao, and A. Y. Zhang. (2022b). “Partial recovery for top- k ranking”. *The Annals of Statistics*. 50(3): 1618–1652.
- Chen, X., J. D. Lee, X. T. Tong, and Y. Zhang. (2020a). “Statistical inference for model parameters in stochastic gradient descent”. *Annals of Statistics*. 48(1): 251–273.
- Chen, X., G. Qu, Y. Tang, S. Low, and N. Li. (2022c). “Reinforcement learning for selective key applications in power systems: Recent advances and future challenges”. *IEEE Transactions on Smart Grid*. 13(4): 2935–2958.
- Chen, X. and L. Zhao. (2023). “Finite-time analysis of single-timescale actor-critic”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 7017–7049.
- Chen, Y., Y. Chi, J. Fan, and C. Ma. (2019a). “Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval”. *Mathematical Programming*. 176(1): 5–37.
- Chen, Y., Y. Chi, J. Fan, and C. Ma. (2021c). “Spectral methods for data science: A statistical perspective”. *Foundations and Trends® in Machine Learning*. 14(5): 566–806.
- Chen, Y., Y. Chi, J. Fan, C. Ma, and Y. Yan. (2020b). “Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization”. *SIAM journal on optimization*. 30(4): 3098–3121.
- Chen, Y., J. Fan, C. Ma, and K. Wang. (2019b). “Spectral method and regularized MLE are both optimal for top- K ranking”. *Annals of statistics*. 47(4): 2204.
- Chen, Y., J. Fan, C. Ma, and Y. Yan. (2019c). “Inference and uncertainty quantification for noisy matrix completion”. *Proceedings of the National Academy of Sciences*. 116(46): 22931–22937.
- Chen, Z., S. T. Maguluri, S. Shakkottai, and K. Shanmugam. (2024b). “A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation”. *Operations Research*. 72(4): 1352–1367.
- Chen, Z. and S. T. Maguluri. (2026). “Non-Asymptotic Convergence of Stochastic Iterative Algorithms: A Lyapunov Framework”. *arXiv preprint arXiv:2605.31309*.
- Chen, Z., S. T. Maguluri, S. Shakkottai, and K. Shanmugam. (2020c). “Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes”. *Advances in Neural Information Processing Systems*. 33: 8223–8234.

- Chen, Z., S. T. Maguluri, S. Shakkottai, and K. Shanmugam. (2021d). “Finite-sample analysis of off-policy TD-learning via generalized Bellman operators”. *Advances in Neural Information Processing Systems*. 34: 21440–21452.
- Chen, Z., S. T. Maguluri, and M. Zubeldia. (2025c). “Concentration of contractive stochastic approximation: Additive and multiplicative noise”. *The Annals of Applied Probability*. 35(2): 1298–1352.
- Chen, Z., S. Zhang, T. T. Doan, J.-P. Clarke, and S. T. Maguluri. (2022d). “Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning”. *Automatica*. 146: 110623.
- Chen, Z., S. Zhang, Z. Zhang, S. U. Haque, and S. T. Maguluri. (2025d). “A non-asymptotic theory of seminorm Lyapunov stability: From deterministic to stochastic iterative algorithms”. *arXiv preprint arXiv:2502.14208*.
- Chen, Z., M. Sim, and H. Xu. (2019d). “Distributionally robust optimization with infinitely constrained ambiguity sets”. *Operations Research*. 67(5): 1328–1344.
- Chen, Z., S. Wang, and N. Si. (2026b). “Sample complexity of distributionally robust average-reward reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 38. 85402–85463.
- Chi, C., Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. (2025a). “Diffusion policy: Visuomotor policy learning via action diffusion”. *The International Journal of Robotics Research*. 44(10-11): 1684–1704.
- Chi, Y., Y. Chen, and Y. Wei. (2025b). “Statistical and algorithmic foundations of reinforcement learning”. In: *INFORMS TutORials in Operations Research*. 104–144.
- Clavier, P., L. Shi, E. Le Pennec, E. Mazumdar, A. Wierman, and M. Geist. (2024). “Near-Optimal Distributionally Robust Reinforcement Learning with General L_p Norms”. In: *Advances in Neural Information Processing Systems*. Vol. 37. 1750–1810.
- Cui, Q. and S. S. Du. (2022a). “Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus”. *Advances in Neural Information Processing Systems*. 35: 11739–11751.
- Cui, Q. and S. S. Du. (2022b). “When are offline two-player zero-sum Markov games solvable?”. *Advances in Neural Information Processing Systems*. 35: 25779–25791.
- Cui, Q., K. Zhang, and S. Du. (2023). “Breaking the curse of multiagents in a large state space: RL in Markov games with independent linear function approximation”. In: *The Thirty Sixth Annual Conference on Learning Theory*. 2651–2652.
- Dai, Y., Q. Cui, and S. S. Du. (2024). “Refined sample complexity for Markov games with independent linear function approximation”. In: *Conference on Learning Theory*. 1260–1261.
- Dann, C. and E. Brunskill. (2015). “Sample complexity of episodic fixed-horizon reinforcement learning”. *Advances in Neural Information Processing Systems*. 28.
- Dann, C., T. Lattimore, and E. Brunskill. (2017). “Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning”. *arXiv preprint arXiv:1703.07710*.
- Dann, C., L. Li, W. Wei, and E. Brunskill. (2019). “Policy certificates: Towards accountable reinforcement learning”. In: *International Conference on Machine Learning*. 1507–1516.
- Daskalakis, C. (2013). “On the complexity of approximating a Nash equilibrium”. *ACM Transactions on Algorithms (TALG)*. 9(3): 1–35.
- Daskalakis, C., P. W. Goldberg, and C. H. Papadimitriou. (2009). “The complexity of computing a Nash equilibrium”. *SIAM Journal on Computing*. 39(1): 195–259.

- Daskalakis, C., N. Golowich, and K. Zhang. (2023). “The complexity of markov equilibrium in stochastic games”. In: *Conference on Learning Theory*. 4180–4234.
- Dean, S., H. Mania, N. Matni, B. Recht, and S. Tu. (2020). “On the sample complexity of the linear quadratic regulator”. *Foundations of Computational Mathematics*. 20(4): 633–679.
- Degenne, R. and V. Perchet. (2016). “Anytime optimal algorithms in stochastic multi-armed bandits”. In: *International Conference on Machine Learning*. 1587–1595.
- Devraj, A. M. and S. Meyn. (2017). “Zap Q-learning”. *Advances in Neural Information Processing Systems*. 30.
- Devraj, A. M. and S. P. Meyn. (2021). “Q-learning with uniformly bounded variance”. *IEEE Transactions on Automatic Control*. 67(11): 5948–5963.
- Di, Q., H. Zhao, J. He, and Q. Gu. (2024). “Pessimistic nonlinear least-squares value iteration for offline reinforcement learning”. In: *International Conference on Learning Representations*. Vol. 2024. 4377–4410.
- Diehl, C., T. Sievernich, M. Krüger, F. Hoffmann, and T. Bertran. (2021). “Umbrella: Uncertainty-aware model-based offline reinforcement learning leveraging planning”. *arXiv preprint arXiv:2111.11097*.
- Dimakopoulou, M., Z. Ren, and Z. Zhou. (2021). “Online multi-armed bandits with adaptive inference”. *Advances in Neural Information Processing Systems*. 34: 1939–1951.
- Ding, D., C.-Y. Wei, K. Zhang, and M. Jovanovic. (2022). “Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence”. In: *International Conference on Machine Learning*. 5166–5220.
- Ding, D., X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. (2021). “Provably efficient safe exploration via primal-dual policy optimization”. In: *International conference on artificial intelligence and statistics*. 3304–3312.
- Doan, T., S. Maguluri, and J. Romberg. (2019). “Finite-time analysis of distributed TD (0) with linear function approximation on multi-agent reinforcement learning”. In: *International Conference on Machine Learning*. 1626–1635.
- Domingues, O. D., P. Ménard, E. Kaufmann, and M. Valko. (2021). “Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited”. In: *Algorithmic Learning Theory*. 578–598.
- Dong, J., J. Li, B. Wang, and J. Zhang. (2024). “Online Policy Optimization for Robust Markov Decision Process”. In: *Uncertainty in Artificial Intelligence*. 1146–1175.
- Dong, K., Y. Wang, X. Chen, and L. Wang. (2019). “Q-learning with UCB exploration is sample efficient for infinite-horizon MDP”. In:
- Du, S., S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. (2021). “Bilinear classes: A structural framework for provable generalization in RL”. In: *International Conference on Machine Learning*. 2826–2836.
- Du, S. S., S. M. Kakade, R. Wang, and L. F. Yang. (2020). “Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?” In: *International Conference on Learning Representations*.
- Duan, Y., Z. Jia, and M. Wang. (2020). “Minimax-optimal off-policy evaluation with linear function approximation”. In: *International Conference on Machine Learning*. 2701–2709.
- Dubhashi, D. P. and A. Panconesi. (2009). *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press.
- Duchi, J. C. and H. Namkoong. (2021). “Learning models with uniform performance via distributionally robust optimization”. *The Annals of Statistics*. 49(3): 1378–1406.

- Durmus, A., E. Moulines, A. Naumov, and S. Samsonov. (2025). “Finite-time high-probability bounds for Polyak-Ruppert averaged iterates of linear stochastic approximation”. *Mathematics of Operations Research*. 50(2): 935–964.
- Ebert, F., C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine. (2018). “Visual foresight: Model-based deep reinforcement learning for vision-based robotic control”. *arXiv preprint arXiv:1812.00568*.
- Efroni, Y., S. Mannor, and M. Pirotta. (2020). “Exploration-exploitation in constrained MDPs”. *arXiv preprint arXiv:2003.02189*.
- El Gamal, A. and M. Raginsky. (2026). “Information-theoretic Limits of Learning and Estimation”. *arXiv preprint arXiv:2605.06710*.
- El Karoui, N. (2015). “On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators”. *Probability Theory and Related Fields*: 1–81.
- El Karoui, N. (2018). “On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators”. *Probability Theory and Related Fields*. 170(1): 95–175.
- El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu. (2013). “On robust regression with high-dimensional predictors”. *Proceedings of the National Academy of Sciences*. 110(36): 14557–14562.
- Ernst, D., P. Geurts, and L. Wehenkel. (2005). “Tree-based batch mode reinforcement learning”. *Journal of Machine Learning Research*. 6.
- Even-Dar, E. and Y. Mansour. (2003). “Learning rates for Q-learning”. *Journal of machine learning Research*. 5(Dec): 1–25.
- Fan, J., J. Hou, and M. Yu. (2024). “Uncertainty quantification of MLE for entity ranking with covariates”. *Journal of Machine Learning Research*. 25(358): 1–83.
- Fan, J., R. Li, C.-H. Zhang, and H. Zou. (2020a). *Statistical foundations of data science*. Chapman and Hall/CRC.
- Fan, J., Z. Lou, W. Wang, and M. Yu. (2025a). “Ranking inferences based on the top choice of multiway comparisons”. *Journal of the American Statistical Association*. 120(549): 237–250.
- Fan, J., Z. Lou, W. Wang, and M. Yu. (2026). “Spectral ranking inferences based on general multiway comparisons”. *Operations Research*. 74(1): 161–180.
- Fan, J., Z. Wang, Y. Xie, and Z. Yang. (2019). “A Theoretical Analysis of Deep Q-Learning”. *arXiv e-prints*: arXiv-1901.
- Fan, J., Z. Wang, Y. Xie, and Z. Yang. (2020b). “A theoretical analysis of deep Q-learning”. In: *Learning for dynamics and control*. 486–489.
- Fan, W., K. Tan, and Y. Wei. (2025b). “Statistical inference under adaptive sampling with LinUCB”. *arXiv preprint arXiv:2512.00222*.
- Fan, Y., O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. (2023). “Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models”. *Advances in Neural Information Processing Systems*. 36: 79858–79885.
- Farahmand, A.-m., C. Szepesvári, and R. Munos. (2010). “Error propagation for approximate policy and value iteration”. In: *Advances in Neural Information Processing Systems*. Vol. 23.
- Feng, S., M. Yin, Y. X. Wang, J. Yang, and Y. Liang. (2024). “Improving Sample Efficiency of Model-Free Algorithms for Zero-Sum Markov Games”. *International Conference on Machine Learning*. 235: 13387–13422.

- Filippi, S., O. Cappé, and A. Garivier. (2010). “Optimism in reinforcement learning and Kullback-Leibler divergence”. In: *Annual Allerton Conference on Communication, Control, and Computing*. IEEE. 115–122.
- Fort, G. (2015). “Central limit theorems for stochastic approximation with controlled Markov chain dynamics”. *ESAIM: Probability and Statistics*. 19: 60–80.
- Foster, D. J., S. M. Kakade, J. Qian, and A. Rakhlin. (2021). “The statistical complexity of interactive decision making”. *arXiv preprint arXiv:2112.13487*.
- Foster, D. J., Z. Mhammedi, and D. Rohatgi. (2025). “Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration”. *arXiv preprint arXiv:2503.07453*.
- Foster, D. J. and A. Rakhlin. (2023). “Foundations of reinforcement learning and interactive decision making”. *arXiv preprint arXiv:2312.16730*.
- Freedman, D. A. (1975). “On tail probabilities for martingales”. *the Annals of Probability*: 100–118.
- Gao, R. (2023). “Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality”. *Operations Research*. 71(6): 2291–2306.
- Gao, X., J. Zha, and X. Y. Zhou. (2025). “Reward-directed score-based diffusion models via Q-learning”. *Journal of Machine Learning Research*. 26(302): 1–46.
- Gilbert, E. N. (1952). “A comparison of signalling alphabets”. *The Bell system technical journal*. 31(3): 504–522.
- Golowich, N., A. Moitra, and D. Rohatgi. (2022). “Learning in observable POMDPs, without computationally intractable oracles”. *Advances in neural information processing systems*. 35: 1458–1473.
- Gopalan, A. and S. Mannor. (2015). “Thompson Sampling for Learning Parameterized Markov Decision Processes”. In: *Conference on Learning Theory*. Vol. 40. 861–898.
- Greenwald, A., K. Hall, and R. Serrano. (2003). “Correlated Q-learning”. In: *International Conference on Machine Learning*. Vol. 3. 242–249.
- Guo, D., D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, *et al.* (2025). “DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning”. *Nature*. 645(8081): 633–638.
- Guo, X., A. Hu, and Y. Zhang. (2026). “Introduction to continuous time RL: a stochastic control perspective”.
- Guo, Y. and Z. Xu. (2025). “Statistical inference for misspecified contextual bandits”. *arXiv preprint arXiv:2509.06287*.
- Gupta, H., R. Srikant, and L. Ying. (2019). “Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 32.
- Hadad, V., D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey. (2021). “Confidence intervals for policy evaluation in adaptive experiments”. *Proceedings of the national academy of sciences*. 118(15): e2014602118.
- Halder, B., S. Pan, and K. Khamaru. (2025). “Stable Thompson sampling: valid inference via variance inflation”. *arXiv preprint arXiv:2505.23260*.
- Han, Q., K. Khamaru, and C.-H. Zhang. (2024a). “UCB algorithms for multi-armed bandits: Precise regret and adaptive inference”. *arXiv preprint arXiv:2412.06126*.
- Han, Q., W. W. Sun, and Y. Zhang. (2025). “Online Statistical Inference in Decision-Making with Matrix Context”. *The Annals of Statistics*. 53(5): 1963–1986.

- Han, Y., M. Razaviyayn, and R. Xu. (2024b). “Stochastic control for fine-tuning diffusion models: Optimality, regularity, and convergence”. *arXiv preprint arXiv:2412.18164*.
- Hanna, J., P. Stone, and S. Niekum. (2017). “Bootstrapping with models: Confidence intervals for off-policy evaluation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. No. 1.
- Hao, B., X. Ji, Y. Duan, H. Lu, C. Szepesvari, and M. Wang. (2021). “Bootstrapping fitted Q-evaluation for off-policy inference”. In: *International Conference on Machine Learning*. 4074–4084.
- Hasselt, H. (2010). “Double Q-learning”. In: *Advances in Neural Information Processing Systems*. Vol. 23. Citeseer. 2613–2621.
- Hazan, E., S. S. Shwartz, and N. Srebro. (2025). “Research program: Theory of learning in dynamical systems”. *arXiv preprint arXiv:2512.19410*.
- He, J., H. Zhao, D. Zhou, and Q. Gu. (2023). “Nearly minimax optimal reinforcement learning for linear Markov decision processes”. In: *International Conference on Machine Learning*. 12790–12822.
- Hu, Y. and S. Wager. (2023). “Off-policy evaluation in partially observed Markov decision processes under sequential ignorability”. *The Annals of Statistics*. 51(4): 1561–1585.
- Hu, Y., Q. Da, A. Zeng, Y. Yu, and Y. Xu. (2018). “Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application”. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 368–377.
- Huang, J., J. Chen, L. Zhao, T. Qin, N. Jiang, and T.-Y. Liu. (2022). “Towards Deployment-Efficient Reinforcement Learning: Lower Bound and Optimality”. In: *International Conference on Learning Representations*.
- Huang, Y., Z. Wen, Y. Chi, Y. Wei, A. Singh, Y. Liang, and Y. Chen. (2026). “On the Emergence of Implicit Curriculum in RLVR Learning Dynamics”. In: *International Conference on Machine Learning*.
- Hung, Y.-H., P.-C. Hsieh, X. Liu, and P. Kumar. (2021). “Reward-biased maximum likelihood estimation for linear stochastic bandits”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 9. 7874–7882.
- Iyengar, G. N. (2005). “Robust dynamic programming”. *Mathematics of Operations Research*. 30(2): 257–280.
- Jaakkola, T., M. Jordan, and S. Singh. (1993). “Convergence of stochastic iterative dynamic programming algorithms”. In: *Advances in Neural Information Processing Systems*. Vol. 6.
- Jaech, A., A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, *et al.* (2024). “Openai o1 system card”. *arXiv preprint arXiv:2412.16720*.
- Jaksch, T., R. Ortner, and P. Auer. (2010). “Near-optimal Regret Bounds for Reinforcement Learning”. *Journal of Machine Learning Research*. 11: 1563–1600.
- Janner, M., Y. Du, J. Tenenbaum, and S. Levine. (2022). “Planning with Diffusion for Flexible Behavior Synthesis”. In: *International Conference on Machine Learning*. 9902–9915.
- Javanmard, A. and A. Montanari. (2018). “Debiasing the Lasso: Optimal sample size for Gaussian designs”. *The Annals of Statistics*. 46(6A): 2593–2622.
- Ji, X. and G. Li. (2023). “Regret-optimal model-free reinforcement learning for discounted MDPs with short burn-in time”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 80674–80689.

- Jiang, N. and A. Agarwal. (2018). “Open problem: The dependence of sample complexity lower bounds on planning horizon”. In: *Conference On Learning Theory*. 3395–3398.
- Jiang, N. and J. Huang. (2020). “Minimax value interval for off-policy evaluation and policy optimization”. *Advances in Neural Information Processing Systems*. 33: 2747–2758.
- Jiang, N. and L. Li. (2016). “Doubly robust off-policy value evaluation for reinforcement learning”. In: *International conference on machine learning*. 652–661.
- Jiang, N. and T. Xie. (2025). “Offline reinforcement learning in large state spaces: Algorithms and guarantees”. *Statistical Science*. 40(4): 570–596.
- Jiao, Y., Y. Chen, and G. Li. (2025). “Connections between reinforcement learning with feedback, test-time scaling, and diffusion guidance: An anthology”. *arXiv preprint arXiv:2509.04372*.
- Jiao, Y. and G. Li. (2024). “Minimax-Optimal Multi-Agent Robust Reinforcement Learning”. *arXiv preprint arXiv:2412.19873*.
- Jiao, Y., J. Woo, G. Li, G. Joshi, and Y. Chi. (2026). “Sample Complexity of Average-Reward Q-Learning: From Single-agent to Federated Reinforcement Learning”. *arXiv preprint arXiv:2601.13642*.
- Jin, C., Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. (2018). “Is Q-learning provably efficient?” In: *Advances in Neural Information Processing Systems*. 4863–4873.
- Jin, C., A. Krishnamurthy, M. Simchowitz, and T. Yu. (2020a). “Reward-free exploration for reinforcement learning”. In: *International Conference on Machine Learning*. 4870–4879.
- Jin, C., Q. Liu, and S. Miryoosefi. (2021a). “Bellman Eluder dimension: New rich classes of RL problems, and sample-efficient algorithms”. *Advances in neural information processing systems*. 34: 13406–13418.
- Jin, C., Q. Liu, Y. Wang, and T. Yu. (2024a). “V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning”. *Mathematics of Operations Research*. 49(4): 2295–2322.
- Jin, C., Z. Yang, Z. Wang, and M. I. Jordan. (2020b). “Provably efficient reinforcement learning with linear function approximation”. In: *Conference on Learning Theory*. 2137–2143.
- Jin, Y., R. Gummadi, Z. Zhou, and J. Blanchet. (2024b). “Feasible Q-Learning for Average Reward Reinforcement Learning”. In: *International Conference on Artificial Intelligence and Statistics*. 1630–1638.
- Jin, Y., Z. Ren, Z. Yang, and Z. Wang. (2025). “Policy learning “without” overlap: Pessimism and generalized empirical Bernstein’s inequality”. *The Annals of Statistics*. 53(4): 1483–1512.
- Jin, Y., Z. Yang, and Z. Wang. (2021b). “Is Pessimism Provably Efficient for Offline RL?” In: *International Conference on Machine Learning*. 5084–5096.
- Jin, Y. and A. Sidford. (2021). “Towards tight bounds on the sample complexity of average-reward MDPs”. In: *International Conference on Machine Learning*. 5055–5064.
- Johnson, R. and T. Zhang. (2013). “Accelerating stochastic gradient descent using predictive variance reduction”. In: *Advances in Neural Information Processing Systems*. 315–323.
- Kaelbling, L. P., M. L. Littman, and A. R. Cassandra. (1998). “Planning and acting in partially observable stochastic domains”. *Artificial intelligence*. 101(1-2): 99–134.
- Kakade, S. M. (2002). “A natural policy gradient”. In: *Advances in Neural Information Processing Systems*. 1531–1538.
- Kakade, S. (2003). “On the sample complexity of reinforcement learning”. *PhD thesis*. University of London.
- Kaledin, M., E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai. (2020). “Finite time analysis of linear two-timescale stochastic approximation with Markovian noise”. In: *Conference on Learning Theory*. 2144–2203.

- Kallus, N., X. Mao, K. Wang, and Z. Zhou. (2022). “Doubly robust distributionally robust off-policy evaluation and learning”. In: *International Conference on Machine Learning*. 10598–10632.
- Kalvit, A. and A. Zeevi. (2021). “A closer look at the worst-case behavior of multi-armed bandit algorithms”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 8807–8819.
- Katehakis, M. N. and H. Robbins. (1995). “Sequential choice from several populations”. *Proceedings of the National Academy of Sciences*. 92(19): 8584–8585.
- Ke, Z. T. and J. Wang. (2025). “Optimal network membership estimation under severe degree heterogeneity”. *Journal of the American Statistical Association*. 120(550): 948–962.
- Kearns, M., Y. Mansour, and A. Y. Ng. (2002). “A sparse sampling algorithm for near-optimal planning in large Markov decision processes”. *Machine learning*. 49(2-3): 193–208.
- Kearns, M. J. and S. P. Singh. (1999). “Finite-sample convergence rates for Q-learning and indirect algorithms”. In: *Advances in Neural Information Processing Systems*. 996–1002.
- Khamaru, K., A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan. (2021). “Is temporal difference learning optimal? an instance-dependent analysis”. *SIAM Journal on Mathematics of Data Science*. 3(4): 1013–1040.
- Khamaru, K. and C.-H. Zhang. (2024). “Inference with the upper confidence bound algorithm”. *arXiv preprint arXiv:2408.04595*.
- Khodadadian, S., P. Sharma, G. Joshi, and S. T. Maguluri. (2022). “Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling”. In: *International Conference on Machine Learning*. 10997–11057.
- Kim, J., D. Wu, J. Lee, and T. Suzuki. (2025). “Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, RL and distillation”. *arXiv preprint arXiv:2502.01694*.
- Kiran, B. R., I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez. (2021). “Deep reinforcement learning for autonomous driving: A survey”. *IEEE transactions on intelligent transportation systems*. 23(6): 4909–4926.
- Kloeckner, B. (2019). “Effective Berry–Esseen and concentration bounds for Markov chains with a spectral gap”. *The Annals of Applied Probability*. 29(3): 1778–1807.
- Klopp, O., K. Lounici, and A. B. Tsybakov. (2017). “Robust matrix completion”. *Probability Theory and Related Fields*. 169(1-2): 523–564.
- Konda, V. R. and J. N. Tsitsiklis. (2000). “Actor-critic algorithms”. In: *Advances in Neural Information Processing Systems*. 1008–1014.
- Konda, V. R. and J. N. Tsitsiklis. (2004). “Convergence Rate of Linear Two-Time-Scale Stochastic Approximation”. *Annals of Applied Probability*: 796–819.
- Kumar, A., T. Buckley, J. B. Lanier, Q. Wang, A. Kavelaars, and I. Kuzovkin. (2019). “Offworld gym: open-access physical robotics environment for real-world reinforcement learning benchmark and research”. *arXiv preprint arXiv:1910.08639*.
- Kumar, A., A. Zhou, G. Tucker, and S. Levine. (2020). “Conservative q-learning for offline reinforcement learning”. *Advances in neural information processing systems*. 33: 1179–1191.
- Kumar, P. and A. Becker. (1982). “A new family of optimal adaptive controllers for Markov chains”. *IEEE Transactions on Automatic Control*. 27(1): 137–146.
- Labbi, S., D. Tiapkin, L. Mancini, P. Mangold, and E. Moulines. (2025). “Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents”. In: *International Conference on Artificial Intelligence and Statistics*. 1315–1323.
- Lai, C.-H., Y. Song, D. Kim, Y. Mitsufuji, and S. Ermon. (2025). “The principles of diffusion models”. *arXiv preprint arXiv:2510.21890*.

- Lai, T. L. (1987). “Adaptive treatment allocation and the multi-armed bandit problem”. *The annals of statistics*: 1091–1114.
- Lai, T. L. (2003). “Stochastic approximation”. *The Annals of Statistics*. 31(2): 391–406.
- Lai, T. L. and H. Robbins. (1985). “Asymptotically efficient adaptive allocation rules”. *Advances in Applied Mathematics*. 6(1): 4–22.
- Lakshminarayanan, C. and C. Szepesvari. (2018). “Linear stochastic approximation: How far does constant step-size and iterate averaging go?” In: *International Conference on Artificial Intelligence and Statistics*. 1347–1355.
- Lam, H. (2019). “Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization”. *Operations Research*. 67(4): 1090–1105.
- Lan, G. (2023). “Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes”. *Mathematical programming*. 198(1): 1059–1106.
- Lange, S., T. Gabel, and M. Riedmiller. (2012). “Batch reinforcement learning”. In: *Reinforcement learning*. Springer. 45–73.
- Lattimore, T. and M. Hutter. (2012). “PAC bounds for discounted MDPs”. In: *International Conference on Algorithmic Learning Theory*. Springer. 320–334.
- Lattimore, T. and C. Szepesvári. (2020). *Bandit algorithms*. Cambridge University Press.
- Le, H., C. Voloshin, and Y. Yue. (2019). “Batch policy learning under constraints”. In: *International Conference on Machine Learning*. 3703–3712.
- Levine, S., A. Kumar, G. Tucker, and J. Fu. (2020). “Offline reinforcement learning: Tutorial, review, and perspectives on open problems”. *arXiv preprint arXiv:2005.01643*.
- Li, G., C. Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi. (2021a). “Tightening the dependence on horizon in the sample complexity of Q-learning”. In: *International Conference on Machine Learning*. 6296–6306.
- Li, G., C. Cai, Y. Chen, Y. Wei, and Y. Chi. (2024a). “Is Q-Learning Minimax Optimal? A Tight Sample Complexity Analysis”. *Operations Research*. 72(1): 222–236.
- Li, G., Y. Chi, Y. Wei, and Y. Chen. (2022a). “Minimax-optimal multi-agent RL in Markov games with a generative model”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 15353–15367.
- Li, G., L. Shi, Y. Chen, and Y. Chi. (2023a). “Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning”. *Information and Inference: A Journal of the IMA*. 12(2): 969–1043.
- Li, G., L. Shi, Y. Chen, Y. Chi, and Y. Wei. (2024b). “Settling the sample complexity of model-based offline reinforcement learning”. *The Annals of Statistics*. 52(1): 233–260.
- Li, G., L. Shi, Y. Chen, Y. Gu, and Y. Chi. (2021b). “Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 34.
- Li, G., Y. Wei, Y. Chi, and Y. Chen. (2023b). “Softmax policy gradient methods can take exponential time to converge”. *Mathematical Programming*. 201(1-2): 707–802.
- Li, G., Y. Wei, Y. Chi, and Y. Chen. (2024c). “Breaking the sample size barrier in model-based reinforcement learning with a generative model”. *Operations Research*. 72(1): 203–221.
- Li, G., Y. Wei, Y. Chi, Y. Gu, and Y. Chen. (2020). “Breaking the sample size barrier in model-based reinforcement learning with a generative model”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 12861–12872.

- Li, G., Y. Wei, Y. Chi, Y. Gu, and Y. Chen. (2022b). “Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction”. *IEEE Transactions on Information Theory*. 68(1): 448–473.
- Li, G., W. Wu, Y. Chi, C. Ma, A. Rinaldo, and Y. Wei. (2024d). “High-probability sample complexities for policy evaluation with linear function approximation”. *IEEE Transactions on Information Theory*. 70(8): 5969–5999.
- Li, G. and Y. Yan. (2025). “Towards Efficient Online Exploration for Reinforcement Learning with Human Feedback”. *arXiv preprint arXiv:2509.22633*.
- Li, G., Y. Yan, Y. Chen, and J. Fan. (2024e). “Minimax-optimal reward-agnostic exploration in reinforcement learning”. In: *Conference on Learning Theory*. 3431–3436.
- Li, G., W. Zhan, J. D. Lee, Y. Chi, and Y. Chen. (2024f). “Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 36.
- Li, L., R. Munos, and C. Szepesvári. (2014). “On minimax optimal offline policy evaluation”. *arXiv preprint arXiv:1409.3653*.
- Li, T., G. Lan, and A. Pananjady. (2023c). “Accelerated and instance-optimal policy evaluation with linear function approximation”. *SIAM Journal on Mathematics of Data Science*. 5(1): 174–200.
- Li, X., J. Liang, and Z. Zhang. (2023d). “Online Statistical Inference for Nonlinear Stochastic Approximation with Markovian Data”. *arXiv preprint arXiv:2302.07690*.
- Li, X., W. Yang, J. Liang, Z. Zhang, and M. I. Jordan. (2023e). “A statistical analysis of polyak-ruppert averaged q-learning”. In: *International Conference on Artificial Intelligence and Statistics*. 2207–2261.
- Li, Y., R. Wang, and L. F. Yang. (2022c). “Settling the horizon-dependence of sample complexity in reinforcement learning”. In: *Symposium on Foundations of Computer Science (FOCS)*. IEEE. 965–976.
- Liang, Z., X. Ma, J. Blanchet, J. Zhang, and Z. Zhou. (2023). “Single-Trajectory Distributionally Robust Reinforcement Learning”. *arXiv preprint arXiv:2301.11721*.
- Liao, P., P. Klasnja, and S. Murphy. (2021). “Off-policy estimation of long-term average outcomes with applications to mobile health”. *Journal of the American Statistical Association*. 116(533): 382–391.
- Liao, P., Z. Qi, R. Wan, P. Klasnja, and S. A. Murphy. (2022). “Batch policy learning in average reward markov decision processes”. *Annals of statistics*. 50(6): 3364.
- Ling, S. (2022). “Improved performance guarantees for orthogonal group synchronization via generalized power method”. *SIAM Journal on Optimization*. 32(2): 1018–1048.
- Littman, M. L. (1994). “Markov games as a framework for multi-agent reinforcement learning”. In: *Machine Learning Proceedings*. Elsevier. 157–163.
- Liu, B., J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik. (2015). “Finite-sample analysis of proximal gradient TD algorithms”. In: *Uncertainty in Artificial Intelligence*. 504–513.
- Liu, P., C. Shi, and W. W. Sun. (2026). “Reinforcement learning from human feedback: A statistical perspective”. *arXiv preprint arXiv:2604.02507*.
- Liu, Q., A. Chung, C. Szepesvári, and C. Jin. (2022). “When is partially observable reinforcement learning not scary?” In: *Conference on Learning Theory*. 5175–5220.
- Liu, Q., P. Netrapalli, C. Szepesvari, and C. Jin. (2023a). “Optimistic MLE: A generic model-based algorithm for partially observable sequential decision making”. In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 363–376.

- Liu, R. and A. Olshevsky. (2023). “Distributed TD(0) with almost no communication”. *IEEE Control Systems Letters*. 7: 2892–2897.
- Liu, T., R. Zhou, D. Kalathil, P. Kumar, and C. Tian. (2021). “Learning policies with zero or bounded constraint violation for constrained mdps”. *Advances in Neural Information Processing Systems*. 34: 17183–17193.
- Liu, X., P.-C. Hsieh, Y. H. Hung, A. Bhattacharya, and P. Kumar. (2020a). “Exploration through reward biasing: Reward-biased maximum likelihood estimation for stochastic multi-armed bandits”. In: *International Conference on Machine Learning*. 6248–6258.
- Liu, Y., A. Swaminathan, A. Agarwal, and E. Brunskill. (2020b). “Provably good batch reinforcement learning without great exploration”. *arXiv preprint arXiv:2007.08202*.
- Liu, Z., M. Lu, W. Xiong, H. Zhong, H. Hu, S. Zhang, S. Zheng, Z. Yang, and Z. Wang. (2023b). “Maximize to explore: One objective function fusing estimation, planning, and exploration”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 22151–22165.
- Liu, Z., M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang. (2024). “Provably Mitigating Overoptimization in RLHF: Your SFT Loss is Implicitly an Adversarial Regularizer”. *arXiv preprint arXiv:2405.16436*.
- Liu, Z. and P. Xu. (2024). “Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation”. In: *International Conference on Artificial Intelligence and Statistics*. 2719–2727.
- Lu, M., H. Zhong, T. Zhang, and J. Blanchet. (2024). “Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms”. In: *Advances in Neural Information Processing Systems*. Vol. 37. 12528–12580.
- Lyu, J., X. Ma, X. Li, and Z. Lu. (2022). “Mildly conservative Q-learning for offline reinforcement learning”. *Advances in Neural Information Processing Systems*. 35: 1711–1724.
- Ma, C., K. Wang, Y. Chi, and Y. Chen. (2020). “Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion, and Blind Deconvolution”. *Foundations of Computational Mathematics*. 20(3): 451–633.
- Ma, J. Y., J. Yan, D. Jayaraman, and O. Bastani. (2022). “Offline goal-conditioned reinforcement learning via f -advantage regression”. *Advances in neural information processing systems*. 35: 310–323.
- Ma, Y., D. Jayaraman, and O. Bastani. (2021). “Conservative offline distributional reinforcement learning”. *Advances in neural information processing systems*. 34: 19235–19247.
- Maei, H., C. Szepesvari, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton. (2009). “Convergent temporal-difference learning with arbitrary smooth function approximation”. *Advances in neural information processing systems*. 22.
- Mahmood, A. R., D. Korenkevych, G. Vasan, W. Ma, and J. Bergstra. (2018). “Benchmarking reinforcement learning algorithms on real-world robots”. In: *Conference on robot learning*. 561–591.
- Mandal, S., Y. Murthy, and R. Srikant. (2025). “Finite-Time Bounds for Distributionally Robust TD Learning with Linear Function Approximation”. *arXiv preprint arXiv:2510.01721*.
- Mao, W. and T. Başar. (2022). “Provably efficient reinforcement learning in decentralized general-sum Markov games”. *Dynamic Games and Applications*: 1–22.
- Matignon, L., L. Jeanpierre, and A.-I. Mouaddib. (2012). “Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes”. In: *Twenty-sixth AAAI conference on artificial intelligence*.

- Maurer, A. and M. Pontil. (2009). “Empirical Bernstein bounds and sample variance penalization”. In: *Conference on Learning Theory*.
- Mehrabi, M. and S. Wager. (2024). “Off-policy evaluation in Markov decision processes under weak distributional overlap”. *arXiv preprint arXiv:2402.08201*.
- Mei, J., C. Xiao, C. Szepesvari, and D. Schuurmans. (2020). “On the global convergence rates of softmax policy gradient methods”. In: *International conference on machine learning*. 6820–6829.
- Ménard, P., O. D. Domingues, X. Shang, and M. Valko. (2021). “UCB momentum Q-learning: Correcting the bias without forgetting”. In: *International Conference on Machine Learning*. 7609–7618.
- Mete, A., R. Singh, X. Liu, and P. Kumar. (2021). “Reward biased maximum likelihood estimation for reinforcement learning”. In: *Learning for Dynamics and Control*. 815–827.
- Meyn, S. (2022). *Control systems and reinforcement learning*. Cambridge University Press.
- Meyn, S. (2023). “Stability of Q-learning through design and optimism”. *arXiv preprint arXiv:2307.02632*.
- Miao, R., B. Shahbaba, and A. Qu. (2025). “Reinforcement learning for individual optimal policy from heterogeneous data”. *Annals of statistics*. 53(4): 1513.
- Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski. (2015). “Human-level control through deep reinforcement learning”. *Nature*. 518(7540): 529–533.
- Modi, A., N. Jiang, A. Tewari, and S. Singh. (2020). “Sample Complexity of Reinforcement Learning using Linearly Combined Model Ensembles”. In: *International Conference on Artificial Intelligence and Statistics*. Vol. 108. 2010–2020.
- Mou, W., K. Khamaru, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. (2022). “Optimal variance-reduced stochastic approximation in Banach spaces”. *arXiv preprint arXiv:2201.08518*.
- Mou, W., C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. (2020). “On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration”. In: *Conference on Learning Theory*. 2947–2997.
- Mou, W., A. Pananjady, and M. J. Wainwright. (2023). “Optimal oracle inequalities for projected fixed-point equations, with applications to policy evaluation”. *Mathematics of Operations Research*. 48(4): 2308–2336.
- Mou, W., A. Pananjady, M. J. Wainwright, and P. L. Bartlett. (2024). “Optimal and instance-dependent guarantees for Markovian linear stochastic approximation”. *Mathematical Statistics and Learning*. 7(1): 41–153.
- Moulin, H. and J.-P. Vial. (1978). “Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon”. *International Journal of Game Theory*. 7(3): 201–221.
- Munos, R. (2007). “Performance bounds in L_p -norm for approximate value iteration”. *SIAM journal on control and optimization*. 46(2): 541–561.
- Murphy, S. A. (2005). “A generalization error for Q-learning”. In: *Journal of machine learning research*. 1073–1097.
- Nash Jr, J. F. (1950). “Equilibrium points in n -person games”. *Proceedings of the national academy of sciences*. 36(1): 48–49.
- Nilim, A. and L. El Ghaoui. (2005). “Robust control of Markov decision processes with uncertain transition matrices”. *Operations Research*. 53(5): 780–798.
- Osband, I., D. Russo, and B. Van Roy. (2013). “More Efficient Reinforcement Learning via Posterior Sampling”. In: *Advances in Neural Information Processing Systems*. Vol. 26.

- Osband, I. and B. Van Roy. (2016). “On lower bounds for regret in reinforcement learning”. *arXiv preprint arXiv:1608.02732*.
- Osband, I. and B. Van Roy. (2017). “Why is Posterior Sampling Better than Optimism for Reinforcement Learning?” In: *International Conference on Machine Learning*. Vol. 70. *Proceedings of Machine Learning Research*. 2701–2710.
- Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.* (2022). “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 27730–27744.
- Ouyang, Y., A. Gopalan, A. Nayyar, R. Jain, and B. Ravindran. (2017). “Learning Unknown Markov Decision Processes: A Thompson Sampling Approach”. In: *Advances in Neural Information Processing Systems*. Vol. 30.
- Panaganti, K. and D. Kalathil. (2022). “Sample complexity of robust reinforcement learning with a generative model”. In: *International Conference on Artificial Intelligence and Statistics*. 9582–9602.
- Panaganti, K., A. Wierman, and E. Mazumdar. (2024). “Model-Free Robust φ -Divergence Reinforcement Learning Using Both Offline and Online Data”. In: *International Conference on Machine Learning*. 39324–39363.
- Panaganti, K., Z. Xu, D. Kalathil, and M. Ghavamzadeh. (2022). “Robust reinforcement learning using offline data”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 32211–32224.
- Panaganti, K., Z. Xu, D. Kalathil, and M. Ghavamzadeh. (2025). “Bridging Distributionally Robust Learning and Offline RL: An Approach to Mitigate Distribution Shift and Partial Data Coverage”. In: *7th Annual Learning for Dynamics & Control Conference*. 619–634.
- Pananjady, A. and M. J. Wainwright. (2020). “Instance-dependent ℓ_∞ -bounds for policy evaluation in tabular reinforcement learning”. *IEEE Transactions on Information Theory*. 67(1): 566–585.
- Papadimitriou, C. H. and T. Roughgarden. (2008). “Computing correlated equilibria in multi-player games”. *Journal of the ACM (JACM)*. 55(3): 1–29.
- Patil, G., L. Prashanth, D. Nagaraj, and D. Precup. (2023). “Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation”. In: *International Conference on Artificial Intelligence and Statistics*. 5438–5448.
- Paulin, D. (2015). “Concentration inequalities for Markov chains by Marton couplings and spectral methods”. *Electronic Journal of Probability*. 20: 1–32.
- Peters, J. and S. Schaal. (2008). “Natural actor-critic”. *Neurocomputing*. 71(7-9): 1180–1190.
- Polyak, B. T. and A. B. Juditsky. (1992). “Acceleration of stochastic approximation by averaging”. *SIAM Journal on Control and Optimization*. 30(4): 838–855.
- Polyanskiy, Y. and Y. Wu. (2025). *Information theory: From coding to learning*. Cambridge university press.
- Prudencio, R. F., M. R. Maximo, and E. L. Colombini. (2022). “A survey on offline reinforcement learning: Taxonomy, review, and open problems”. *arXiv preprint arXiv:2203.01387*.
- Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
- Qi, Z., C. Bai, Z. Wang, and L. Wang. (2025). “Distributional off-policy evaluation in reinforcement learning”. *Journal of the American Statistical Association*. 120(551): 1517–1530.
- Qiao, D. and Y.-X. Wang. (2022). “Near-optimal deployment efficiency in reward-free reinforcement learning with linear function approximation”. *arXiv preprint arXiv:2210.00701*.

- Qiao, D., M. Yin, M. Min, and Y.-X. Wang. (2022). “Sample-efficient reinforcement learning with $\log \log(t)$ switching cost”. In: *International Conference on Machine Learning*. 18031–18061.
- Qu, C., L. Shi, K. Panaganti, P. You, and A. Wierman. (2025). “Hybrid Transfer Reinforcement Learning: Provable Sample Efficiency from Shifted-Dynamics Data”. In: *International Conference on Artificial Intelligence and Statistics*. 1054–1062.
- Qu, C., C. Yeh, K. Panaganti, E. Mazumdar, and A. Wierman. (2026). “Distributionally Robust Cooperative Multi-Agent Reinforcement Learning via Robust Value Factorization”. *arXiv preprint arXiv:2602.11437*.
- Qu, G. and A. Wierman. (2020). “Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning”. In: *Conference on Learning Theory*. 3185–3205.
- Rafailov, R., A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. (2023). “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems*. Vol. 36.
- Rahimian, H. and S. Mehrotra. (2019). “Distributionally robust optimization: A review”. *arXiv preprint arXiv:1908.05659*.
- Raič, M. (2019). “A multivariate Berry–Esseen theorem with explicit constants”. *Bernoulli*. 25(4A): 2824–2853.
- Ramesh, S. S., P. G. Sessa, Y. Hu, A. Krause, and I. Bogunovic. (2024). “Distributionally robust model-based reinforcement learning with large state spaces”. In: *International Conference on Artificial Intelligence and Statistics*. 100–108.
- Ramprasad, P., Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng. (2022). “Online bootstrap inference for policy evaluation in reinforcement learning”. *Journal of the American Statistical Association*: 1–14.
- Ramprasad, P., Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng. (2023). “Online bootstrap inference for policy evaluation in reinforcement learning”. *Journal of the American Statistical Association*. 118(544): 2901–2914.
- Ran-Milo, Y., Y. Alexander, S. Mendel, and N. Cohen. (2026). “Outcome-Based RL Provably Leads Transformers to Reason, but Only With the Right Data”. *arXiv preprint arXiv:2601.15158*.
- Rashidinejad, P., B. Zhu, C. Ma, J. Jiao, and S. Russell. (2021). “Bridging offline reinforcement learning and imitation learning: A tale of pessimism”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 11702–11716.
- Recht, B. (2019). “A tour of reinforcement learning: The view from continuous control”. *Annual Review of Control, Robotics, and Autonomous Systems*. 2(1): 253–279.
- Ren, T., J. Li, B. Dai, S. S. Du, and S. Sanghavi. (2021). “Nearly horizon-free offline reinforcement learning”. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. 15621–15634.
- Robbins, H. (1952). “Some aspects of the sequential design of experiments”.
- Robbins, H. and S. Monro. (1951). “A stochastic approximation method”. *The Annals of Mathematical Statistics*: 400–407.
- Roughgarden, T. (2016). *Twenty lectures on algorithmic game theory*. Cambridge University Press.
- Ruppert, D. (1998). “Efficient estimators from a slowly converging robbins-monro process”. *Tech. rep.*
- Russo, D. and B. Van Roy. (2013). “Eluder dimension and the sample complexity of optimistic exploration”. *Advances in Neural Information Processing Systems*. 26.
- Salgia, S. and Y. Chi. (2024). “The Sample-Communication Complexity Trade-off in Federated Q-Learning”. In: *Advances in Neural Information Processing Systems*.

- Samsonov, S., E. Moulines, Q.-M. Shao, Z.-S. Zhang, and A. Naumov. (2024a). “Gaussian approximation and multiplier bootstrap for Polyak-Ruppert averaged linear stochastic approximation with applications to TD learning”. In: *Advances in Neural Information Processing Systems*. Vol. 37. 12408–12460.
- Samsonov, S., D. Tiapkin, A. Naumov, and E. Moulines. (2024b). “Improved high-probability bounds for the temporal difference learning algorithm via exponential stability”. In: *Conference on Learning Theory*. 4511–4547.
- Samworth, R. J. and R. D. Shah. (2026). “Modern Statistical Methods and Theory”.
- Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz. (2015). “Trust region policy optimization”. In: *International Conference on Machine Learning*. 1889–1897.
- Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. (2017). “Proximal policy optimization algorithms”. *arXiv preprint arXiv:1707.06347*.
- Shah, S., D. Dey, C. Lovett, and A. Kapoor. (2018). “Airsim: High-fidelity visual and physical simulation for autonomous vehicles”. In: *Field and Service Robotics: Results of the 11th International Conference*. Springer. 621–635.
- Shalev-Shwartz, S. (2012). “Online learning and online convex optimization”. *Foundations and Trends® in Machine Learning*. 4(2): 107–194.
- Shang, S., H. Strauss, S. Wei, S. Arora, and N. Razin. (2026). “When Errors Can Be Beneficial: A Categorization of Imperfect Rewards for Policy Gradient”. *arXiv preprint arXiv:2604.25872*.
- Shao, Q.-M. and Z.-S. Zhang. (2022). “Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms”. *Bernoulli*. 28(3): 1548–1576.
- Shapley, L. S. (1953). “Stochastic games”. *Proceedings of the National Academy of Sciences*. 39(10): 1095–1100.
- Shen, Y., H. Cai, and R. Song. (2024). “Doubly robust interval estimation for optimal policy evaluation in online learning”. *Journal of the American Statistical Association*. 119(548): 2811–2821.
- Shi, C. (2025). “Statistical inference in reinforcement learning: A selective survey”. *arXiv preprint arXiv:2502.16195*.
- Shi, C., M. Uehara, J. Huang, and N. Jiang. (2022a). “A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes”. In: *International Conference on Machine Learning*. 20057–20094.
- Shi, C., S. Zhang, W. Lu, and R. Song. (2022b). “Statistical inference of the value function for reinforcement learning in infinite-horizon settings”. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 84(3): 765–793.
- Shi, C., J. Zhu, Y. Shen, S. Luo, H. Zhu, and R. Song. (2024a). “Off-policy confidence interval estimation with confounded markov decision process”. *Journal of the American Statistical Association*. 119(545): 273–284.
- Shi, L. and Y. Chi. (2024). “Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity”. *Journal of Machine Learning Research*. 25(200): 1–91.
- Shi, L., J. Gai, E. Mazumdar, Y. Chi, and A. Wierman. (2025). “Breaking the Curse of Multiagency in Robust Multi-Agent Reinforcement Learning”. In: *International Conference on Machine Learning*.

- Shi, L., G. Li, Y. Wei, Y. Chen, and Y. Chi. (2022c). “Pessimistic Q-Learning for Offline Reinforcement Learning: Towards Optimal Sample Complexity”. In: *International Conference on Machine Learning*. 19967–20025.
- Shi, L., G. Li, Y. Wei, Y. Chen, M. Geist, and Y. Chi. (2023). “The curious price of distributional robustness in reinforcement learning with a generative model”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 79903–79917.
- Shi, L., G. Li, Y. Wei, Y. Chen, M. Geist, and Y. Chi. (2026). “The curious price of distributional robustness in reinforcement learning with a generative model”. *Operations Research*.
- Shi, L., E. Mazumdar, Y. Chi, and A. Wierman. (2024b). “Sample-Efficient Robust Multi-Agent Reinforcement Learning in the Face of Environmental Uncertainty”. In: *International Conference on Machine Learning*. 44909–44959.
- Sidford, A., M. Wang, X. Wu, L. Yang, and Y. Ye. (2018a). “Near-optimal time and sample complexities for solving Markov decision processes with a generative model”. In: *Advances in Neural Information Processing Systems*. 5186–5196.
- Sidford, A., M. Wang, X. Wu, and Y. Ye. (2018b). “Variance reduced value iteration and faster algorithms for solving Markov decision processes”. In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 770–787.
- Sidford, A., M. Wang, L. Yang, and Y. Ye. (2020). “Solving discounted stochastic two-player games with near-optimal time and sample complexity”. In: *International Conference on Artificial Intelligence and Statistics*. 2992–3002.
- Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.* (2016). “Mastering the game of Go with deep neural networks and tree search”. *Nature*. 529(7587): 484–489.
- Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.* (2017). “Mastering the game of Go without human knowledge”. *Nature*. 550(7676): 354–359.
- Simchowitz, M. and D. J. Foster. (2020). “Naive Exploration is Optimal for Online LQR”. In: *International Conference on Machine Learning*. Vol. 119. *Proceedings of Machine Learning Research*. 8937–8948.
- Slivkins, A. (2019). “Introduction to multi-armed bandits”. *Foundations and Trends® in Machine Learning*. 12(1-2): 1–286.
- Sonabend, A., J. Lu, L. A. Celi, T. Cai, and P. Szolovits. (2020). “Expert-supervised reinforcement learning for offline policy learning and evaluation”. *Advances in Neural Information Processing Systems*. 33: 18967–18977.
- Sonabend-W, A., N. Laha, A. N. Ananthakrishnan, T. Cai, and R. Mukherjee. (2023). “Semi-supervised off-policy reinforcement learning and value estimation for dynamic treatment regimes”. *Journal of Machine Learning Research*. 24(323): 1–86.
- Song, Y., Y. Zhou, A. Sekhari, D. Bagnell, A. Krishnamurthy, and W. Sun. (2023). “Hybrid RL: Using both offline and online data can make RL efficient”. In: *International Conference on Learning Representations*.
- Song, Z., S. Mei, and Y. Bai. (2022). “When Can We Learn General-Sum Markov Games with a Large Number of Players Sample-Efficiently?” In: *International Conference on Learning Representations*.
- Srikant, R. and L. Ying. (2019). “Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning”. In: *Conference on Learning Theory*. 2803–2830.

- Srikant, R. (2025). “Rates of convergence in the central limit theorem for Markov chains, with an application to TD learning”. *Mathematics of Operations Research*.
- Strehl, A. L., L. Li, E. Wiewiora, J. Langford, and M. L. Littman. (2006). “PAC model-free reinforcement learning”. In: *Proceedings of the 23rd International Conference on Machine Learning*. 881–888.
- Sun, T., Y. Sun, Y. Xu, and W. Yin. (2020). “Markov chain block coordinate descent”. *Computational Optimization and Applications*. 75(1): 35–61.
- Sutton, R. S. (1988). “Learning to predict by the methods of temporal differences”. *Machine learning*. 3(1): 9–44.
- Sutton, R. S. and A. G. Barto. (2018). *Reinforcement learning: An introduction*. MIT press.
- Sutton, R. S., H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. (2009). “Fast gradient-descent methods for temporal-difference learning with linear function approximation”. In: *International conference on machine learning*. 993–1000.
- Sutton, R. S., D. A. McAllester, S. P. Singh, and Y. Mansour. (2000). “Policy gradient methods for reinforcement learning with function approximation”. In: *Advances in Neural Information Processing Systems*. 1057–1063.
- Szepesvári, C. (1997). “The asymptotic convergence-rate of Q-learning”. In: *NIPS*. Vol. 10. Citeseer. 1064–1070.
- Szepesvári, C. (2010). “Algorithms for reinforcement learning”. *Synthesis lectures on artificial intelligence and machine learning*. 4(1): 1–103.
- Szita, I. and C. Szepesvári. (2010). “Model-based reinforcement learning with nearly tight exploration complexity bounds”. In: *International Conference on Machine Learning*. 1031–1038.
- Tan, K., W. Fan, and Y. Wei. (2024). “Hybrid Reinforcement Learning Breaks Sample Size Barriers In Linear MDPs”. In: *Advances in Neural Information Processing Systems*.
- Tan, K., W. Fan, and Y. Wei. (2025). “Actor-critics can achieve optimal sample efficiency”. *arXiv preprint arXiv:2505.03710*.
- Tang, C., B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone. (2025). “Deep reinforcement learning for robotics: A survey of real-world successes”. *Annual Review of Control, Robotics, and Autonomous Systems*. 8(1): 153–188.
- Tang, S. and J. Wiens. (2021). “Model selection for offline reinforcement learning: Practical considerations for healthcare settings”. In: *Machine Learning for Healthcare Conference*. 2–35.
- Taylor, M. E. and P. Stone. (2009). “Transfer learning for reinforcement learning domains: A survey.” *Journal of Machine Learning Research*. 10(7).
- Tesauro, G. et al. (1995). “Temporal difference learning and TD-Gammon”. *Communications of the ACM*. 38(3): 58–68.
- Thomas, P. and E. Brunskill. (2016). “Data-efficient off-policy policy evaluation for reinforcement learning”. In: *International conference on machine learning*. 2139–2148.
- Thompson, W. R. (1933). “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. *Biometrika*. 25(3/4): 285–294.
- Thrun, S. and A. Schwartz. (1993). “Issues in using function approximation for reinforcement learning”. In: *Proceedings of the Fourth Connectionist Models Summer School*.
- Tian, H., I. C. Paschalidis, and A. Olshevsky. (2024). “One-Shot Averaging for Distributed TD(λ) Under Markov Sampling”. *IEEE Control Systems Letters*.
- Tian, Y., Y. Wang, T. Yu, and S. Sra. (2021). “Online learning in unknown Markov games”. In: *International conference on machine learning*. 10279–10288.

- Tiapkin, D., D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, P. Perrault, M. Valko, and P. Ménard. (2023). “Model-free posterior sampling via learning rate randomization”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 73719–73774.
- Tropp, J. (2011). “Freedman’s inequality for matrix martingales”. *Electronic Communications in Probability*. 16: 262–270.
- Tsitsiklis, J. and B. Van Roy. (1997). “An Analysis of Temporal-Difference Learning with Function Approximation”. *IEEE Transactions on Automatic Control*. 42(5): 674–690.
- Tsitsiklis, J. N. (1994). “Asynchronous stochastic approximation and Q-learning”. *Machine Learning*. 16(3): 185–202.
- Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Vol. 11. Springer.
- Uehara, M., J. Huang, and N. Jiang. (2020). “Minimax weight and Q-function learning for off-policy evaluation”. In: *International Conference on Machine Learning*. 9659–9668.
- Uehara, M., M. Imaizumi, N. Jiang, N. Kallus, W. Sun, and T. Xie. (2021). “Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency”. *arXiv preprint arXiv:2102.02981*.
- Uehara, M., A. Sekhari, J. D. Lee, N. Kallus, and W. Sun. (2022a). “Provably efficient reinforcement learning in partially observable dynamical systems”. *Advances in Neural Information Processing Systems*. 35: 578–592.
- Uehara, M., C. Shi, and N. Kallus. (2022b). “A review of off-policy evaluation in reinforcement learning”. *arXiv preprint arXiv:2212.06355*.
- Uehara, M. and W. Sun. (2022). “Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage”. In: *International Conference on Learning Representations*.
- Van Hasselt, H. (2012). “Reinforcement learning in continuous state and action spaces”. In: *Reinforcement learning: State-of-the-art*. Springer. 207–251.
- Van Hasselt, H., A. Guez, and D. Silver. (2016). “Deep reinforcement learning with double Q-learning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. No. 1.
- Vaswani, S., L. Yang, and C. Szepesvári. (2022). “Near-optimal sample complexity bounds for constrained MDPs”. *Advances in Neural Information Processing Systems*. 35: 3110–3122.
- Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.
- Vinyals, O., I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, and P. Georgiev. (2019). “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. *Nature*. 575(7782): 350–354.
- Wagenmaker, A. and A. Pacchiano. (2023). “Leveraging offline data in online reinforcement learning”. In: *International Conference on Machine Learning*. 35300–35338.
- Wagenmaker, A. J., Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. (2022). “Reward-free RL is no harder than reward-aware RL in linear Markov decision processes”. In: *International Conference on Machine Learning*. 22430–22456.
- Wainwright, M. (2019a). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Wainwright, M. J. (2019b). “Stochastic approximation with cone-contractive operators: Sharp ℓ_∞ -bounds for Q-learning”. *arXiv preprint arXiv:1905.06265*.
- Wainwright, M. J. (2019c). “Variance-reduced Q-learning is minimax optimal”. *arXiv preprint arXiv:1906.04697*.

- Wan, Y., A. Naik, and R. S. Sutton. (2021). “Learning and planning in average-reward Markov decision processes”. In: *International Conference on Machine Learning*. 10653–10662.
- Wang, B., Y. Yan, and J. Fan. (2021a). “Sample-efficient reinforcement learning for linearly-parameterized MDPs with a generative model”. In: *Advances in Neural Information Processing Systems*. Vol. 34.
- Wang, F. and L. Janson. (2021). “Exact asymptotics for linear quadratic adaptive control”. *Journal of Machine Learning Research*. 22(265): 1–112.
- Wang, H., T. Zariwopoulou, and X. Y. Zhou. (2020a). “Reinforcement learning in continuous time and space: A stochastic control approach”. *Journal of Machine Learning Research*. 21(198): 1–34.
- Wang, H., L. Shi, and Y. Chi. (2024a). “Sample Complexity of Offline Distributionally Robust Linear Markov Decision Processes”. *Reinforcement Learning Journal*. 3: 1467–1510.
- Wang, H., X. Xu, and Y. Chi. (2026). “Provably Efficient and Agile Randomized Q-Learning”. In: *International Conference on Artificial Intelligence and Statistics*.
- Wang, J., Z. Ren, R. Zhan, and Z. Zhou. (2024b). “Distributionally robust policy learning under concept drifts”. *arXiv preprint arXiv:2412.14297*.
- Wang, M. (2020). “Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time”. *Mathematics of Operations Research*. 45(2): 517–546.
- Wang, M., P. Yang, and L. Su. (2025a). “On the Convergence Rates of Federated Q-Learning across Heterogeneous Environments”. *Transactions on Machine Learning Research*. ISSN: 2835-8856.
- Wang, R., S. S. Du, L. Yang, and R. R. Salakhutdinov. (2020b). “On reward-free reinforcement learning with linear function approximation”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 17816–17826.
- Wang, S., N. Si, J. Blanchet, and Z. Zhou. (2023a). “A finite sample complexity bound for distributionally robust Q-learning”. In: *International Conference on Artificial Intelligence and Statistics*. 3370–3398.
- Wang, S., N. Si, J. Blanchet, and Z. Zhou. (2023b). “On the foundation of distributionally robust reinforcement learning”. *arXiv preprint arXiv:2311.09018*.
- Wang, S., N. Si, J. Blanchet, and Z. Zhou. (2024c). “Sample complexity of variance-reduced distributionally robust Q-learning”. *Journal of Machine Learning Research*. 25(341): 1–77.
- Wang, Y., R. Wang, S. S. Du, and A. Krishnamurthy. (2021b). “Optimism in Reinforcement Learning with Generalized Linear Function Approximation”. In: *International Conference on Learning Representations*.
- Wang, Y., Q. Liu, Y. Bai, and C. Jin. (2023c). “Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation”. In: *Conference on Learning Theory*. 2793–2848.
- Wang, Y., R. Wang, and S. Kakade. (2021c). “An exponential lower bound for linearly realizable MDP with constant suboptimality gap”. *Advances in Neural Information Processing Systems*. 34: 9521–9533.
- Wang, Y., J. Xiong, and S. Zou. (2024d). “Achieving the Asymptotically Minimax Optimal Sample Complexity of Offline Reinforcement Learning: A DRO-Based Approach”. *Transactions on Machine Learning Research*. ISSN: 2835-8856.
- Wang, Y. and S. Zou. (2021). “Online Robust Reinforcement Learning with Model Uncertainty”. In: *Advances in Neural Information Processing Systems*. Vol. 34.

- Wang, Z., D. Zhou, J. Lui, and W. Sun. (2025b). “Model-based RL as a minimalist approach to horizon-free and second-order bounds”. In: *International Conference on Learning Representations*. Vol. 2025. 91356–91386.
- Watkins, C. J. C. H. (1989). “Learning from delayed rewards”. *PhD thesis, King’s College, University of Cambridge*.
- Watkins, C. J. and P. Dayan. (1992). “Q-learning”. *Machine learning*. 8(3-4): 279–292.
- Wei, H., A. Ghosh, N. Shroff, L. Ying, and X. Zhou. (2023). “Provably efficient model-free algorithms for non-stationary CMDPs”. In: *International Conference on Artificial Intelligence and Statistics*. 6527–6570.
- Wei, H., X. Liu, and L. Ying. (2022). “Triple-Q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation”. In: *International Conference on Artificial Intelligence and Statistics*. 3274–3307.
- Weisz, G., P. Amortila, and C. Szepesvári. (2021). “Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions”. In: *Algorithmic Learning Theory*. 1237–1264.
- Weng, B., H. Xiong, L. Zhao, Y. Liang, and W. Zhang. (2020a). “Momentum Q-learning with finite-sample convergence guarantee”. *arXiv preprint arXiv:2007.15418*.
- Weng, W., H. Gupta, N. He, L. Ying, and R. Srikant. (2020b). “The Mean-Squared Error of Double Q-Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 33.
- White, M. and A. White. (2010). “Interval estimation for reinforcement-learning algorithms in continuous-state domains”. In: *Advances in Neural Information Processing Systems*. Vol. 23.
- Wiesemann, W., D. Kuhn, and B. Rustem. (2013). “Robust Markov decision processes”. *Mathematics of Operations Research*. 38(1): 153–183.
- Williams, R. J. (1992). “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. *Machine learning*. 8(3-4): 229–256.
- Woo, J., G. Joshi, and Y. Chi. (2023). “The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond”. In: *International Conference on Machine Learning*. 37157–37216.
- Woo, J., L. Shi, G. Joshi, and Y. Chi. (2024). “Federated Offline Reinforcement Learning: Collaborative Single-Policy Coverage Suffices”. In: *Forty-first International Conference on Machine Learning*.
- Woodworth, B., B. Bullins, O. Shamir, and N. Srebro. (2021). “The Min-max Complexity of Distributed Stochastic Convex Optimization with Intermittent Communication”. In: *Conference on Learning Theory*. 4386–4437.
- Wu, W., G. Li, Y. Wei, and A. Rinaldo. (2024). “Statistical Inference for Temporal Difference Learning with Linear Function Approximation”. *arXiv preprint arXiv:2410.16106*.
- Wu, W., Y. Wei, and A. Rinaldo. (2025). “Uncertainty quantification for Markov chains with application to temporal difference learning”. *arXiv preprint arXiv:2502.13822*.
- Xia, E., K. Khamaru, M. J. Wainwright, and M. I. Jordan. (2024). “Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning”. *IEEE Transactions on Information Theory*.
- Xie, T., C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. (2021a). “Bellman-consistent pessimism for offline reinforcement learning”. In: *Advances in Neural Information Processing Systems*. 6683–6694.
- Xie, T., D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade. (2023). “The Role of Coverage in Online Reinforcement Learning”. In: *International Conference on Learning Representations*.

- Xie, T., D. J. Foster, A. Krishnamurthy, C. Rosset, A. Awadallah, and A. Rakhlin. (2024). “Exploratory Preference Optimization: Harnessing Implicit Q^* -Approximation for Sample-Efficient RLHF”. *arXiv preprint arXiv:2405.21046*.
- Xie, T., N. Jiang, H. Wang, C. Xiong, and Y. Bai. (2021b). “Policy finetuning: Bridging sample-efficient offline and online reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 27395–27407.
- Xie, T., Y. Ma, and Y.-X. Wang. (2019). “Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling”. *Advances in neural information processing systems*. 32.
- Xiong, H., L. Zhao, Y. Liang, and W. Zhang. (2020). “Finite-time analysis for double Q-learning”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 16628–16638.
- Xiong, Z., R. Shen, Q. Cui, M. Fazel, and S. S. Du. (2022). “Near-optimal randomized exploration for tabular markov decision processes”. *Advances in neural information processing systems*. 35: 6358–6371.
- Xu, H. and S. Mannor. (2012). “Distributionally Robust Markov Decision Processes”. *Mathematics of Operations Research*. 37(2): 288–300.
- Xu, P. and Q. Gu. (2020). “A finite-time analysis of Q-learning with neural network function approximation”. In: *International conference on machine learning*. 10555–10565.
- Xu, T. and Y. Liang. (2021). “Sample complexity bounds for two timescale value-based reinforcement learning algorithms”. In: *International conference on artificial intelligence and statistics*. 811–819.
- Xu, T., Y. Wang, S. Zou, and Y. Liang. (2024). “Provably efficient offline reinforcement learning with trajectory-wise reward”. *IEEE Transactions on Information Theory*. 70(9): 6481–6518.
- Xu, T., Z. Wang, and Y. Liang. (2020a). “Improving sample complexity bounds for (natural) actor-critic algorithms”. *Advances in Neural Information Processing Systems*. 33: 4358–4369.
- Xu, T., Z. Wang, Y. Zhou, and Y. Liang. (2020b). “Reanalysis of variance reduced temporal difference learning”. *arXiv preprint arXiv:2001.01898*.
- Xu, T., S. Zou, and Y. Liang. (2019). “Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples”. In: *Advances in Neural Information Processing Systems*. Vol. 32.
- Xu, Y., C. Shi, S. Luo, L. Wang, and R. Song. (2022). “Quantile off-policy evaluation via deep conditional generative learning”. *arXiv preprint arXiv:2212.14466*.
- Xu, Z., K. Panaganti, and D. Kalathil. (2023). “Improved Sample Complexity Bounds for Distributionally Robust Reinforcement Learning”. *arXiv preprint arXiv:2303.02783*.
- Yan, Y., G. Li, Y. Chen, and J. Fan. (2023). “The efficacy of pessimism in asynchronous Q-learning”. *IEEE Transactions on Information Theory*. 69(11): 7185–7219.
- Yan, Y., G. Li, Y. Chen, and J. Fan. (2024). “Model-based reinforcement learning for offline zero-sum Markov games”. *Operations research*. 72(6): 2430–2445.
- Yang, K., L. Yang, and S. Du. (2021). “Q-learning with logarithmic regret”. In: *International Conference on Artificial Intelligence and Statistics*. 1576–1584.
- Yang, L. and M. Wang. (2019). “Sample-optimal parametric Q-learning using linearly additive features”. In: *International Conference on Machine Learning*. 6995–7004.
- Yang, L. and M. Wang. (2020). “Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound”. In: *International Conference on Machine Learning*. 10746–10756.
- Yang, M., O. Nachum, B. Dai, L. Li, and D. Schuurmans. (2020a). “Off-policy evaluation via the regularized lagrangian”. *Advances in Neural Information Processing Systems*. 33: 6551–6561.

- Yang, S., Y. Du, S. K. S. Ghasemipour, J. Tompson, L. P. Kaelbling, D. Schuurmans, and P. Abbeel. (2024a). “Learning Interactive Real-World Simulators”. In: *The Twelfth International Conference on Learning Representations*.
- Yang, T., S. Cen, Y. Wei, Y. Chen, and Y. Chi. (2024b). “Federated Natural Policy Gradient and Actor Critic Methods for Multi-task Reinforcement Learning”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Yang, T., B. Dai, L. Xiao, and Y. Chi. (2025a). “Exploration from a primal-dual lens: Value-incentivized actor-critic methods for sample-efficient online RL”. In: *Advances in Neural Information Processing Systems*. Vol. 38. 40470–40521.
- Yang, T., B. Dai, L. Xiao, and Y. Chi. (2025b). “Incentivize without Bonus: Provably Efficient Model-based Online Multi-agent RL for Markov Games”. In: *International Conference on Machine Learning*. 70704–70744.
- Yang, T., Y. Huang, Y. Liang, and Y. Chi. (2026a). “Agentic Transformers Provably Learn to Search via Reinforcement Learning”. *arXiv preprint arXiv:2606.00183*.
- Yang, T., M. Ryu, C.-W. Hsu, G. Tennenholtz, Y. Chi, C. Boutilier, and B. Dai. (2026b). “Diffusion Controller: Framework, Algorithms and Parameterization”. In: *International Conference on Machine Learning*.
- Yang, W., L. Zhang, and Z. Zhang. (2022). “Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics”. *The Annals of Statistics*. 50(6): 3223–3248.
- Yang, Z., C. Jin, Z. Wang, M. Wang, and M. I. Jordan. (2020b). “On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 8146–8157.
- Ye, C., R. Yang, Q. Gu, and T. Zhang. (2023). “Corruption-robust offline reinforcement learning with general function approximation”. *Advances in Neural Information Processing Systems*. 36: 36208–36221.
- Yin, D., B. Hao, Y. Abbasi-Yadkori, N. Lazić, and C. Szepesvári. (2022a). “Efficient local planning with linear function approximation”. In: *International Conference on Algorithmic Learning Theory*. 1165–1192.
- Yin, M., Y. Bai, and Y.-X. Wang. (2021a). “Near-optimal offline reinforcement learning via double variance reduction”. *Advances in neural information processing systems*. 34: 7677–7688.
- Yin, M., Y. Bai, and Y.-X. Wang. (2021b). “Near-Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning”. In: *International Conference on Artificial Intelligence and Statistics*. 1567–1575.
- Yin, M., Y. Duan, M. Wang, and Y.-X. Wang. (2022b). “Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism”. In: *International Conference on Learning Representations*.
- Yin, M. and Y.-X. Wang. (2020). “Asymptotically efficient off-policy evaluation for tabular reinforcement learning”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 3948–3958.
- Yin, M. and Y.-X. Wang. (2021a). “Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings”. *Advances in neural information processing systems*. 34: 12890–12903.
- Yin, M. and Y.-X. Wang. (2021b). “Towards instance-optimal offline reinforcement learning with pessimism”. In: *Advances in Neural Information Processing Systems*. Vol. 34.

- Yu, C., J. Liu, S. Nemati, and G. Yin. (2021). “Reinforcement learning in healthcare: A survey”. *ACM Computing Surveys (CSUR)*. 55(1): 1–36.
- Yuan, R., S. S. Du, R. M. Gower, A. Lazaric, and L. Xiao. (2023). “Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies”. In: *International Conference on Learning Representations*.
- Zanette, A. (2021). “Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL”. In: *International Conference on Machine Learning*. 12287–12297.
- Zanette, A. and E. Brunskill. (2019). “Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds”. In: *International Conference on Machine Learning*. 7304–7312.
- Zanette, A., A. Lazaric, M. Kochenderfer, and E. Brunskill. (2020a). “Learning near optimal policies with low inherent Bellman error”. In: *International Conference on Machine Learning*. 10978–10989.
- Zanette, A., A. Lazaric, M. J. Kochenderfer, and E. Brunskill. (2020b). “Provably efficient reward-agnostic navigation with linear value iteration”. *Advances in Neural Information Processing Systems*. 33: 11756–11766.
- Zanette, A. and M. J. Wainwright. (2022). “Bellman residual orthogonalization for offline reinforcement learning”. *Advances in Neural Information Processing Systems*. 35: 3137–3151.
- Zanette, A., M. J. Wainwright, and E. Brunskill. (2021). “Provable benefits of actor-critic methods for offline reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 34.
- Zhan, R., V. Hadad, D. A. Hirshberg, and S. Athey. (2021). “Off-policy evaluation via adaptive weighting with data from contextual bandits”. In: *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2125–2135.
- Zhan, W., S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi. (2023a). “Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence”. *SIAM Journal on Optimization*. 33(2): 1061–1091.
- Zhan, W., B. Huang, A. Huang, N. Jiang, and J. Lee. (2022). “Offline reinforcement learning with realizability and single-policy concentrability”. In: *Conference on Learning Theory*. 2730–2775.
- Zhan, W., M. Uehara, W. Sun, and J. D. Lee. (2023b). “PAC reinforcement learning for predictive state representations”. *International Conference on Learning Representations*.
- Zhang, A. Y. and H. Y. Zhou. (2024). “Leave-one-out singular subspace perturbation analysis for spectral clustering”. *The Annals of Statistics*. 52(5): 2004–2033.
- Zhang, C., H. Wang, A. Mitra, and J. Anderson. (2024a). “Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning”. In: *The Twelfth International Conference on Learning Representations*.
- Zhang, H., Z. Zheng, and L. Xue. (2025a). “Gap-Dependent Bounds for Federated Q-Learning”. In: *International Conference on Machine Learning*.
- Zhang, H., Z. Zheng, and L. Xue. (2025b). “Q-Learning with Fine-Grained Gap-Dependent Regret”. *arXiv preprint arXiv:2510.06647*.
- Zhang, H., Z. Zheng, and L. Xue. (2026a). “Regret-optimal Q-learning with low cost for single-agent and federated reinforcement learning”. *Advances in Neural Information Processing Systems*. 38: 128711–128799.

- Zhang, H., H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh. (2020a). “Robust deep reinforcement learning against adversarial perturbations on state observations”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 21024–21037.
- Zhang, K., S. Kakade, T. Basar, and L. Yang. (2020b). “Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 1166–1178.
- Zhang, K., Z. Yang, and T. Başar. (2021a). “Multi-agent reinforcement learning: A selective overview of theories and algorithms”. *Handbook of Reinforcement Learning and Control*: 321–384.
- Zhang, K., L. Janson, and S. Murphy. (2020c). “Inference for batched bandits”. *Advances in neural information processing systems*. 33: 9818–9829.
- Zhang, L., Y. Peng, J. Liang, W. Yang, and Z. Zhang. (2025c). “Estimation and inference in distributional reinforcement learning”. *The Annals of Statistics*. 53(5): 1987–2011.
- Zhang, R., Y. Li, and N. Li. (2021b). “On the regret analysis of online LQR control with predictions”. In: *2021 American Control Conference (ACC)*. IEEE. 697–703.
- Zhang, S., D. Yu, H. Sharma, H. Zhong, Z. Liu, Z. Yang, S. Wang, H. H. Awadalla, and Z. Wang. (2025d). “Self-Exploring Language Models: Active Preference Elicitation for Online Alignment”. *Transactions on Machine Learning Research*. ISSN: 2835-8856.
- Zhang, S., Z. Zhang, and S. T. Maguluri. (2021c). “Finite Sample Analysis of Average-Reward TD Learning and Q-Learning”. *Advances in Neural Information Processing Systems*. 34: 1230–1242.
- Zhang, T. (2022). “Feel-good Thompson sampling for contextual bandits and reinforcement learning”. *SIAM Journal on Mathematics of Data Science*. 4(2): 834–857.
- Zhang, T. (2023). *Mathematical analysis of machine learning algorithms*. Cambridge University Press.
- Zhang, W., D. Zhou, and Q. Gu. (2021d). “Reward-free model-based reinforcement learning with linear function approximation”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 1582–1593.
- Zhang, X., Y. Ma, and A. Singla. (2020d). “Task-agnostic exploration in reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 11734–11743.
- Zhang, Z., K. Panaganti, L. Shi, Y. Sui, A. Wierman, and Y. Yue. (2024b). “Distributionally robust constrained reinforcement learning under strong duality”. *arXiv preprint arXiv:2406.15788*.
- Zhang, Z., Y. Chen, J. Lee, S. Du, L. Yang, and R. Wang. (2026b). “Deployment efficient reward-free exploration with linear function approximation”. *Advances in Neural Information Processing Systems*. 38: 25364–25399.
- Zhang, Z., Y. Chen, J. Lee, and S. S. Du. (2025e). “Settling the sample complexity of online reinforcement learning”. *Journal of the ACM*. 72(3): 1–63.
- Zhang, Z., Y. Chen, J. D. Lee, and S. S. Du. (2024c). “Settling the sample complexity of online reinforcement learning”. In: *Conference on Learning Theory*. 5213–5219.
- Zhang, Z., S. S. Du, and X. Ji. (2020e). “Nearly minimax optimal reward-free reinforcement learning”. *arXiv preprint arXiv:2010.05901*.
- Zhang, Z., X. Ji, and S. Du. (2021e). “Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon”. In: *Conference on Learning Theory*. 4528–4531.
- Zhang, Z., X. Ji, and S. Du. (2022). “Horizon-free reinforcement learning in polynomial time: the power of stationary policies”. In: *Conference on Learning Theory*. 3858–3904.

- Zhang, Z., J. Lee, Y. Chen, and S. Du. (2024d). “Horizon-free regret for linear markov decision processes”. In: *International Conference on Learning Representations*. Vol. 2024. 54205–54216.
- Zhang, Z., Y. Zhou, and X. Ji. (2020f). “Almost Optimal Model-Free Reinforcement Learning via Reference-Advantage Decomposition”. In: *Advances in Neural Information Processing Systems*. Vol. 33.
- Zhang, Z., Y. Zhou, and X. Ji. (2021f). “Model-free reinforcement learning: from clipped pseudo-regret to sample complexity”. In: *International Conference on Machine Learning*. 12653–12662.
- Zhao, H., H. Chen, J. Zhang, D. Yao, and W. Tang. (2025a). “Score as Action: Fine Tuning Diffusion Generative Models by Continuous-time Reinforcement Learning”. In: *International Conference on Machine Learning*. 77371–77389.
- Zhao, H., D. Liang, W. Tang, D. Yao, and N. Kallus. (2025b). “Diffpo: Training diffusion llms to reason fast and furious via reinforcement learning”. *arXiv preprint arXiv:2510.02212*.
- Zhao, H., J. He, and Q. Gu. (2024). “A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation”. *Advances in Neural Information Processing Systems*. 37: 94684–94735.
- Zheng, Y., L. Furieri, M. Kamgarpour, and N. Li. (2021). “Sample complexity of linear quadratic Gaussian (LQG) control for output feedback systems”. In: *Learning for dynamics and control*. 559–570.
- Zheng, Z., F. Gao, L. Xue, and J. Yang. (2024). “Federated Q-learning: Linear regret speedup with low communication cost”. In: *International Conference on Learning Representations*. Vol. 2024. 57399–57443.
- Zhong, Y. and N. Boumal. (2018). “Near-optimal bounds for phase synchronization”. *SIAM Journal on Optimization*. 28(2): 989–1016.
- Zhou, D., Q. Gu, and C. Szepesvari. (2021a). “Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes”. In: *Conference on Learning Theory*. 4532–4576.
- Zhou, D., J. He, and Q. Gu. (2021b). “Provably efficient reinforcement learning for discounted MDPs with feature mapping”. In: *International Conference on Machine Learning*. 12793–12802.
- Zhou, D., Y. Zhang, A. Sonabend-W, Z. Wang, J. Lu, and T. Cai. (2024). “Federated offline reinforcement learning”. *Journal of the American Statistical Association*. 119(548): 3152–3163.
- Zhou, Y., Z. Qi, C. Shi, and L. Li. (2023a). “Optimizing pessimism in dynamic treatment regimes: A Bayesian learning approach”. In: *International Conference on Artificial Intelligence and Statistics*. 6704–6721.
- Zhou, Z., Z. Zhou, Q. Bai, L. Qiu, J. Blanchet, and P. Glynn. (2021c). “Finite-sample regret bound for distributionally robust offline tabular reinforcement learning”. In: *International Conference on Artificial Intelligence and Statistics*. 3331–3339.
- Zhou, Z., S. Athey, and S. Wager. (2023b). “Offline multi-action policy learning: Generalization and optimization”. *Operations Research*. 71(1): 148–183.
- Zhu, B., M. Jordan, and J. Jiao. (2023a). “Principled reinforcement learning with human feedback from pairwise or k-wise comparisons”. In: *International Conference on Machine Learning*. 43037–43067.
- Zhu, J., R. Wan, Z. Qi, S. Luo, and C. Shi. (2024). “Robust offline reinforcement learning with heavy-tailed rewards”. In: *International Conference on Artificial Intelligence and Statistics*. 541–549.
- Zhu, W., X. Chen, and W. B. Wu. (2023b). “Online covariance matrix estimation in stochastic gradient descent”. *Journal of the American Statistical Association*. 118(541): 393–404.

- Zhu, Z., K. Lin, A. K. Jain, and J. Zhou. (2023c). “Transfer learning in deep reinforcement learning: A survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45(11): 13344–13362.
- Ziegler, D. M., N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. (2019). “Fine-tuning language models from human preferences”. *arXiv preprint arXiv:1909.08593*.
- Zou, S., T. Xu, and Y. Liang. (2019). “Finite-sample analysis for sarsa with linear function approximation”. *Advances in neural information processing systems*. 32.
- Zurek, M. and Y. Chen. (2024). “Span-Based Optimal Sample Complexity for Weakly Communicating and General Average Reward MDPs”. In: *Advances in Neural Information Processing Systems*.
- Zurek, M. and Y. Chen. (2025). “The Plug-in Approach for Average-Reward and Discounted MDPs: Optimal Sample Complexity Analysis”. In: *Algorithmic Learning Theory*. 1386–1387.