Taming "data-hungry" reinforcement learning? Stability in continuous state-action spaces

Yaqi Duan
Department of Technology, Operations, and Statistics
Stern School of Business, New York University
New York, NY 10012
yaqi.duan@stern.nyu.edu

Martin J. Wainwright
Laboratory for Information and Decision Systems, Statistics and Data Science Center
Department of Electrical Engineering and Computer Science, and Department of Mathematics
Massachusetts Institute of Technology Cambridge, MA 02139
wainwrigwork@gmail.com

Abstract

We introduce a novel framework for analyzing reinforcement learning (RL) in continuous state-action spaces, and use it to prove fast rates of convergence in both off-line and on-line settings. Our analysis highlights two key stability properties, relating to how changes in value functions and/or policies affect the Bellman operator and occupation measures. We argue that these properties are satisfied in many continuous state-action Markov decision processes. Our analysis also offers fresh perspectives on the roles of pessimism and optimism in off-line and on-line RL.

1 Introduction

Many domains of science and engineering involve making a sequence of decisions over time, with previous decisions influencing the future in uncertain ways [1, 13, 22, 31, 32]. For instance, clinicians managing diabetes [36] or engineers optimizing plasma control in tokamak systems [5] must develop policies that adapt based on evolving conditions and lead to desirable outcomes over a longer period. Markov decision processes (MDPs) and reinforcement learning (RL) provide frameworks and methods for estimating effective policies for such sequential problems. While RL excels in data-rich scenarios such as competitive gaming (e.g., AlphaGo and its extensions [30]), its application in data-scarce areas like healthcare [36] and finance [27] remains challenging due to lack of history, or underlying non-stationarity. With limited data, characterizing and improving the *sample complexity* of RL methods becomes critical.

Considerable research effort has been devoted to studying RL sample complexity in many settings. Existing studies for either the generative or the off-line settings (e.g., [21, 37, 35]) give procedures that, when applied to a dataset of size n, yield a value gap that decays at the rate $1/\sqrt{n}$. In the on-line setting, there are various procedures that yield cumulative regret that grows at the rate $\sqrt{T}$ (e.g., [18, 20, 19, 6]). In contrast, the main result of this paper is to formalize conditions, suitable for RL in continuous domains, under which *much faster rates can be obtained using the same dataset*, achieving a value gap decay of $1/n$ and reducing regret growth to $\log T$.

As revealed by our analysis, these accelerated rates depend on certain *stability properties*, ones that—as we argue—are naturally satisfied in many control problems with continuous state-action

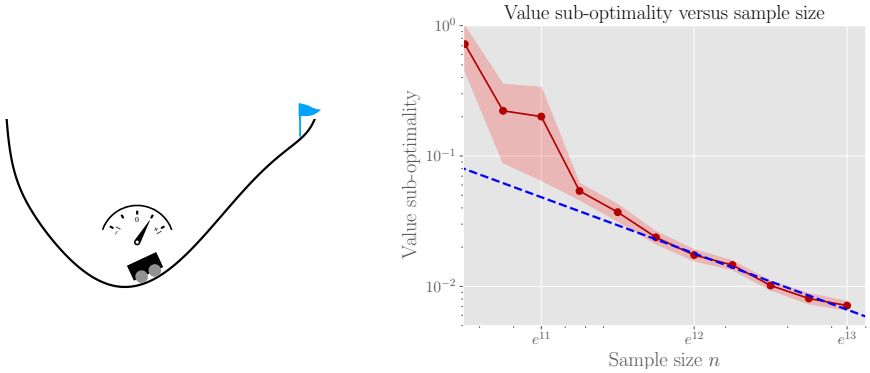
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-2279

spaces. Roughly speaking, these conditions ensure that the evolution of the dynamic system depends in a "smooth" way on the influence of decision policy. Such notions of stability should be expected in various controlled systems with continuous state-action spaces. In robotics, for example, a minor torque or motion perturbation that occurs during a single step should not cause a notable deviation from the intended trajectory. Similarly, in clinical treatment, slight deviations in medication dosage should not significantly compromise effectiveness or safety.

1.1 A simple illustrative example: Mountain Car

The "Mountain Car" problem, a benchmark continuous control task, illustrates the acceleration phenomenon and underlying stability. In this task, as shown in Figure 1(a), a car must reach the top of a hill by adjusting its acceleration within the interval $[-1, 1]$. We employed fitted Q-iteration (FQI) with carefully selected linear basis functions to derive near-optimal policies with off-line data. This learning procedure exhibits a value sub-optimality decay at a rate of $1/n$, a significant improvement over the classical rate of $1/\sqrt{n}$, as detailed in Figure 1(b). (See Appendix D for further explanation. The experiment ran for 3 days on two laptops, each equipped with an Apple M2 Pro CPU and 16 GB RAM.) In this example, slight perturbations in the driving policy lead to only modest changes in future trajectories, which shows the stability. Our theoretical analysis confirms that fast rates are achievable in this and similar continuous control tasks when such stability properties are present.



(a) (b)

Figure 1: Illustration of the "fast rate" phenomenon using FQI on the Mountain Car problem. Each **red point** in the plot represents the average value sub-optimality $J(\pi^\star) - J(\widehat{\pi}_n)$ from $T = 80$ Monte Carlo trials, with the shaded area showing twice the standard errors. The **blue dashed line** is a least-squares fit to the last 6 data points, yielding a 95% confidence interval of $(-1.084, -0.905)$ for the slope, significantly faster than the typical -0.5 "slow rate".

1.2 Contributions of this paper

With this high-level perspective in mind, let us summarize the key contributions of this paper.

Fast rate of convergence: We develop a framework for analyzing RL in continuous state-action spaces, and use it to prove a general result (Theorem 1) under which fast rates can be obtained. The key insight is that stability conditions lead to upper bounds on the value sub-optimality that are proportional to the *squared* norm of Bellman residuals. In the off-line setting, this quadratic scaling improves convergence from a rate of $n^{-\frac{1}{2}}$ to n^{-1}, while in on-line learning, it enhances the regret bound from $\sqrt{T}$ to $\log T$.

Reconsidering pessimism and optimism principles: Our framework provides a novel perspective on the roles of pessimism [21, 4] and optimism [18, 20, 19, 6, 12] in off-line and on-line RL. Our theory reveals that there are settings in which *neither pessimism nor optimism* are required for effective policy optimization—in particular, they are not required as long as one has a sufficiently accurate pilot estimate policy. Moreover, our analysis shows that some procedures based on certainty

equivalence can achieve fast-rate convergence, showing that the benefits gained from incorporating additional pessimism or optimism measures may be limited in this context.

1.3 Related work

In this section, we discuss related work having to do with fast rates in optimization and statistics.

Fast rates in stochastic optimization and risk minimization: For many statistical estimators (e.g., likelihood methods, empirical risk minimization), it is well-understood that the local geometry around the optimum determines whether fast rates can be obtained. For instance, when the loss function exhibits some form of strong convexity (such as exp-concave loss) or strict saddle properties, it can lead to significant reductions in additive regret from $\mathcal{O}(\sqrt{T})$ to just $\mathcal{O}(\log T)$ in stochastic approximation (e.g., [15]), or a decrease in the error rate from $n^{-\frac{1}{2}}$ to n^{-1} in empirical risk minimization [23, 14]. These fast rate phenomena rely on a form of stability, one which relates the similarity of functions to the closeness of their optima. Our work develops a new framework for analyzing value-based RL methods, focusing on identifying specific stability conditions and inherent curvature properties that promote fast rate convergence in RL, similar to the role of stability analysis in statistical learning.

Fast rates in reinforcement learning: In the RL literature, there are various lines of work related to fast rates, but the underlying mechanisms are typically different from those considered here. For problems with discrete state-action spaces, there is a line of recent work [17, 16, 33, 25] that performs gap/marginal-dependent analyses of RL algorithms. However, such separation assumptions are not helpful for continuous action spaces. Other work for discrete state-action spaces [28] has shown convergence rates in off-line RL are influenced by data quality, with a nearly-expert dataset enabling faster rate. In contrast, our analysis reveals that for off-line RL in continuous domains, fast convergence can occur whether or not the dataset has good coverage properties.

An important sub-class of continuous state-action problems are those with linear dynamics and quadratic reward functions (LQR for short). For such problems, it has been shown [24, 29] that value sub-optimality can be connected with the squared error in system identification. Our general theory can also be used to derive guarantees for LQR problems, as we explore in more detail in a follow-up paper [8]. Stability also arises in the analysis of (deterministic) policy optimization and Newton-type algorithms [26, 3], where it is possible to show superlinear convergence in a local neighborhood. This accelerated rate stems from the smoothness of the on-policy transition operator $\mathcal{P}^{\pi_f}$ with respect to changes in the value function f; for instance, see condition (10) in Puterman and Brumelle [26]. Our framework exploits related notions of smoothness, but is tailored to the stochastic setting of reinforcement learning, in which understanding the effect of function approximation and finite sample sizes is essential.

2 Fast rates for value-based reinforcement learning

Let us now set up and state the main result of this paper. We begin in Section 2.1 with background on Markov decision processes (MDPs) and value-based methods, before turning to the statement of our main result in Section 2.2. In Section 2.3, we provide intuition for why stability leads to faster rates, and discuss consequences for both the off-line and on-line settings of RL.

2.1 Markov decision processes and value-based methods

Basic set-up: We consider an episodic Markov decision process (MDP) defined by a quadruple $\big(\mathcal{S}, \mathcal{A}, \boldsymbol{\mathcal{P}} = \{\mathcal{P}_h\}_{h=1}^{H-1}, \{r_h\}_{h=1}^{H}\big)$. We assume that the rewards $r_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are known; however, this condition can be relaxed. A policy π_h at time h is a mapping from any state s to a distribution $\pi_h(\cdot \mid s)$ over the action space $\mathcal{A}$. If the support of $\pi_h(\cdot \mid s)$ is a singleton, we also let $\pi_h(s) \in \mathcal{A}$ denote the single action to be chosen at state s. Given an initial distribution ξ_1 over the states at time $h = 1$, the *expected reward* obtained by choosing actions according to a policy sequence $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_H)$ is given by $J(\boldsymbol{\pi}) \equiv J(\boldsymbol{\pi}; \xi_1) := \mathbb{E}_{\xi_1, \boldsymbol{\pi}}\big[\sum_{h=1}^{H} r_h(S_h, A_h)\big]$, where $S_1 \sim \xi_1$, $S_{h+1} \sim \mathcal{P}_h(\cdot \mid S_h, A_h)$ and $A_h \sim \pi_h(\cdot \mid S_h)$ for $h = 1, 2, \ldots, H$. Our goal is to estimate an *optimal policy* $\boldsymbol{\pi}^\star \in \arg\max_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$.

 https://doi.org/10.52202/079017-2279

Value functions and Bellman operators: Starting from a given state-action pair (s, a) at stage h, the expected return over subsequent stages defines the *state-action value function* $Q_h^{\boldsymbol{\pi}}(s,a) := \mathbb{E}_{\boldsymbol{\pi}}\big[\sum_{h'=h}^{H} r(S_{h'}, A_{h'}) \,\big|\, S_h = s, A_h = a\big]$. The sequence of functions $\boldsymbol{Q}^{\boldsymbol{\pi}} = (Q_1^{\boldsymbol{\pi}}, \ldots, Q_H^{\boldsymbol{\pi}})$ known as the *Q-functions* associated with $\boldsymbol{\pi}$.

The Q-functions $\boldsymbol{Q}^{\boldsymbol{\pi}}$ have an important connection with the *Bellman evaluation operator* for $\boldsymbol{\pi}$. For any policy $\boldsymbol{\pi}$ and stage h, we introduce a linear transition operator $(\mathcal{P}_h^{\boldsymbol{\pi}} f)(s,a) := \int_{\mathcal{S}\times\mathcal{A}} f(s',a')\ \mathcal{P}_h(ds' \mid s,a)\,\pi_{h+1}(da' \mid s')$ for any function $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. With this notation, the *Bellman evaluation operator* at stage h takes the form

$$(\mathcal{T}_h^{\boldsymbol{\pi}} f)(s,a) := r_h(s,a) + (\mathcal{P}_h^{\boldsymbol{\pi}} f)(s,a). \tag{1}$$

From classical dynamic programming, the Q-functions $\boldsymbol{Q}^{\boldsymbol{\pi}}$ must satisfy the Bellman relations $Q_h^{\boldsymbol{\pi}}(s,a) = (\mathcal{T}_h^{\boldsymbol{\pi}} Q_{h+1}^{\boldsymbol{\pi}})(s,a)$ for $h = 1, \ldots, H-1$.

Bellman principle for optimal policies: Under mild regularity conditions, there is at least one policy $\boldsymbol{\pi}^\star$ such that, for any other policy $\boldsymbol{\pi}$, we have $Q_h^{\boldsymbol{\pi}^\star}(s,a) \geq Q_h^{\boldsymbol{\pi}}(s,a)$, for any $h \in [H]$, and uniformly over all state-action pairs (s,a). Any optimal policy $\boldsymbol{\pi}^\star$ must be greedy with respect to the optimal Q-function $\boldsymbol{Q}^\star$. By classical dynamic programming, the optimal Q-function $\boldsymbol{Q}^\star$ is obtained by setting $Q_H^\star = r_H$, and then recursively computing $Q_h^\star = \mathcal{T}_h^\star\, Q_{h+1}^\star$ for $h = H-1, \ldots, 2, 1$, with the *Bellman optimality operator* defined as

$$(\mathcal{T}_h^\star\, f)(s,a) := r_h(s,a) + \mathbb{E}_h\big[\max_{a' \in \mathcal{A}} f(S', a') \,\big|\, s, a\big] \qquad \text{for } S' \sim \mathcal{P}_h(\cdot \mid s, a). \tag{2}$$

Value-based RL methods The main result of this paper applies to a broad class of methods for reinforcement learning. They are known as *value-based*, due to their reliance on the following two step approach for approximating an optimal policy $\boldsymbol{\pi}^\star$: (1) Construct an estimate $\widehat{\boldsymbol{Q}} = (\widehat{f}_1, \ldots, \widehat{f}_H)$ of the optimal value function $\boldsymbol{Q}^\star = (Q_1^\star, \ldots, Q_H^\star)$. (2) Use $\widehat{\boldsymbol{Q}}$ to compute the greedy-optimal policy $\widehat{\pi}_h(s) \in \arg\max_a \widehat{f}_h(s,a)$ for $h = 1, 2, \ldots, H$. It should be noted that there is considerable freedom in the design of a value-based method, since different methods can be used to approximate value functions in Step 1. Rather than applying to a single method, our main result applies to a very broad class of these methods.

Underlying any value-based method is a class $\mathscr{F}$ of functions $(s,a) \mapsto f(s,a)$ used to approximate the state-action value functions.[1] We assume that the function class $\mathscr{F}$ is rich enough—relative to the Bellman evaluation operators—to ensure that for any greedy policy $\boldsymbol{\pi}$ induced by some $\boldsymbol{f} = (f_1, \ldots, f_H) \in \mathscr{F}^H$, we have the inclusion $\mathcal{T}_h^{\boldsymbol{\pi}}\, \mathscr{F} \subseteq \mathscr{F}$ for $h = 1, \ldots, H-1$. We see that this condition depends on the structure of the transition distributions $\mathcal{P}_h(\cdot \mid s,a)$. In many practical examples, the reward function itself has some number of derivatives, and these transition distributions perform some type of smoothing, so that we expect that the output of the Bellman update, given a suitably differentiable function, will remain suitably differentiable.

2.2 Stable problems have fast rates

We now turn the central question in understanding the behavior of any value-based method:

How to translate "closeness" of the Q-function estimate $\widehat{\boldsymbol{Q}}$
to a bound on the value gap $J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}})$?

At a high level, existing theory provides guarantees of the following type: if the Q-function estimates are ε-accurate for some $\varepsilon \in (0,1)$, then the value gap is bounded by a quantity proportional to ε. In contrast, our main result shows that when the MDP is stable in a suitable sense, the value gap can be upper bounded by a quantity proportional to ε^2. This *quadratic as opposed to linear scaling* encapsulates the "fast rate" phenomenon of this paper.

Our analysis isolates two key stability properties required for faster rates; both are Lipschitz conditions with respect to a certain norm. Here we define them with respect to the L^2-norm induced by the

[1] In general, different function classes may be selected at each stage $h = 1, 2, \ldots, H$; here, so as to reduce notational clutter, we assume that the same function class $\mathscr{F}$ is used for each stage.

state-action occupation measure induced by the optimal policy—namely

$$\|f\|_h := \sqrt{\mathbb{E}_{\boldsymbol{\pi}^\star}\left[f^2(S_h, A_h)\right]} \qquad \text{for any } f \in \partial\mathscr{F}^2, \tag{3}$$

and over a neighborhood $\mathcal{N}$ of the optimal Q-value function $\boldsymbol{Q}^\star$.

Bellman stability: The first condition measures the stability of the Bellman optimality operator (2): in particular, we require that there is a scalar $\kappa_h^\star$ such that

$$\left\|\mathcal{T}_h^\star\, f_{h+1} - \mathcal{T}_h^\star\, Q_{h+1}^\star\right\|_h \;\le\; \kappa_h^\star\, \left\|f_{h+1} - Q_{h+1}^\star\right\|_{h+1} \tag{Stb($\mathcal{T}$)}$$

for any $\boldsymbol{f} \in \mathcal{N}$. Moreover, for any pair (h, h') of indices such that $1 \le h < h' \le H-1$, we define

$$\boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star) := \kappa_h^\star\, \kappa_{h+1}^\star \ldots \kappa_{h'-1}^\star\,.$$

Condition (**Stb**($\mathcal{T}$)) is directly linked to the stability of estimating the Q-function $\boldsymbol{Q}^\star$. In typical estimation procedures, such as approximate dynamic programming, the estimation is carried out iteratively in a backward manner, so that it is important to control the propagation of estimation errors across the iterations. Condition (**Stb**($\mathcal{T}$)) captures this property, since it implies that

$$\left\|\mathcal{T}_h^\star\, \mathcal{T}_{h+1}^\star\, \ldots \mathcal{T}_{h'-1}^\star\, f_{h'} - \mathcal{T}_h^\star\, \mathcal{T}_{h+1}^\star\, \ldots \mathcal{T}_{h'-1}^\star\, Q_{h'}^\star\right\|_h \;\le\; \boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star) \cdot \left\|f_{h'} - Q_{h'}^\star\right\|_{h'},$$

which shows how the estimation error $\left(f_{h'} - Q_{h'}^\star\right)$ at step h' can be controlled in terms of estimation error at an earlier time step $h \le h'$.

Occupation measure stability: Our second condition is more subtle, and is key in our argument. Let us begin with some intuition. Consider two sequences of policies

$$\left(\pi_1^\star, \ldots, \pi_{h-1}^\star, \pi_h^\star, \pi_{h+1}^\star, \ldots, \pi_{h'}^\star\right) \qquad \text{and} \qquad \left(\pi_1^\star, \ldots, \pi_{h-1}^\star, \pi_h, \pi_{h+1}^\star, \ldots, \pi_{h'}^\star\right)$$

that only differ at the h-th step, where $\pi_h^\star$ has been replaced by π_h. These two policy sequences induce Markov chains whose distributions differ from stage h onwards, and our second condition controls this difference in terms of the difference $\|f_h - Q_h^\star\|_h$ between the two Q-functions f_h and $Q_h^\star$ that induce π_h and $\pi_h^\star$, respectively.

We adopt $\mathcal{P}_h^\star$ as a convenient shorthand for the transition operator $\mathcal{P}_h^{\boldsymbol{\pi}^\star}$, and define the multi-step transition operator $\mathcal{P}_{h,h'}^\star := \mathcal{P}_h^\star\, \mathcal{P}_{h+1}^\star \cdots \mathcal{P}_{h'-1}^\star$. Using this notation, for any $h' \ge h+1$, we require that there is a scalar $\boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)$ such that

$$\sup_{\substack{g \in \partial\mathscr{F} \\ \|g\|_{h'} > 0}} \frac{\left|\mathbb{E}_{\boldsymbol{\pi}^\star}\left[\left(\mathcal{P}_{h,h'}^\star g\right)(S_h, \pi_h^\star(S_h)) - \left(\mathcal{P}_{h,h'}^\star g\right)(S_h, \pi_h(S_h))\right]\right|}{\|g\|_{h'}} \le \boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)\, \frac{\left\|f_h - Q_h^\star\right\|_h}{\left\|Q_h^\star\right\|_h} \tag{Stb(ξ)}$$

for any $\boldsymbol{f} \in \mathcal{N}$. The renormalization in this definition serves to enforce a natural scale invariance.

With these notions of stability in hand, we are now equipped to state our main result. Taking as input a value function estimate $\widehat{\boldsymbol{Q}}$, it relates the induced value gap to the *Bellman residuals* $\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h$. Note that these residuals are a way of quantifying proximity to the optimal value function $\boldsymbol{Q}^\star$, which has Bellman residual zero by definition. We assume that $\widehat{\boldsymbol{Q}}$ has Bellman residuals bounded as

$$\left\|\, \mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h \,\right\|_h \;\le\; \varepsilon_h \qquad \text{for } h = 1, 2, \ldots, H-1 \tag{4a}$$

for some sequence $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{H-1}, \varepsilon_H = 0)$ that satisfies the constraint

$$\varepsilon_h \ge \frac{1}{H-h} \sum_{h'=h+1}^{H} \varepsilon_{h'} \qquad \text{for } h = 1, 2, \ldots, H-1. \tag{4b}$$

This last condition means that the Bellman residual ε_h is larger than or equal to the average of the bounds established after step $h+1$. It is natural because estimating at step h is at least as challenging as a stage $h' > h$; indeed, any such state h' occurs earlier in the dynamic programming backward iteration process. As a special case, the bound (4b) holds when $\varepsilon_h = \varepsilon$ for all stages.

With this set-up, we have the following guarantee in terms of the stability coefficients $\boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)$ and $\boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star)$ from conditions (**Stb**(ξ)) and (**Stb**($\mathcal{T}$)).

[2]We let $\partial\mathscr{F}$ be the set of all difference functions of the form $g = f - \widetilde{f}$ for some $f, \widetilde{f} \in \mathscr{F}$.

https://doi.org/10.52202/079017-2279

Theorem 1. *There is a neighborhood of $\boldsymbol{Q}^\star$ such that for any value function estimate $\widehat{\boldsymbol{f}}$ with $\boldsymbol{\varepsilon}$-bounded Bellman residuals* (4a)*, the induced greedy policy $\widehat{\boldsymbol{\pi}}$ has value gap bounded as*

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \;\leq\; 2 \sum_{h=1}^{H-1} \frac{1}{\|Q_h^\star\|_h} \left\{ \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)\, \varepsilon_{h'} \right\} \left\{ \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star)\, \varepsilon_{h'} \right\}. \tag{5}$$

See Appendix A for the proof.

Treating dependence on the stability coefficients as constant, the main take-away is that value suboptimality is bounded above by a quantity proportional to the *squared* norm of the Bellman residuals. Concretely, if the Bellman residuals are uniformly upper bounded by some ε, then equation (5) leads to an upper bound of the form

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \leq c\, H^3\, \varepsilon^2,$$

where c is a universal constant. Due to the quadratic scaling in the Bellman residual error ε, this bound is substantially tighter than the linear in ε rates afforded by a conventional analysis.

2.3 Intuition for fast rates: Smoothness and cancelling terms in the telescope bound

Why does "fast rate" phenomenon formalized in Theorem 1 arise? The fast rates proved in this paper are established by a novel argument, starting from a known telescope bound, which we begin by stating. Given a Q-function estimate $\widehat{\boldsymbol{f}} = \big(\widehat{f}_1, \ldots, \widehat{f}_H\big)$, let $\widehat{\boldsymbol{\pi}}$ denote the induced greedy policy. Then the value gap of $\widehat{\boldsymbol{\pi}}$ with respect to an arbitrary comparator policy $\boldsymbol{\pi}$ is bounded as

$$J(\boldsymbol{\pi}) - J(\widehat{\boldsymbol{\pi}}) \;\leq\; \sum_{h=1}^{H-1} \big(\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big) \big[\big(\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h\big)(S_h, A_h)\big]\,. \tag{6}$$

This result follows by a "telescope" relation induced by the structure of the Bellman updates.[3] For completeness, we provide a proof of the telescope bound in Appendix E.2.

A key feature of inequality (6) is the difference of two expectations $\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}$, corresponding to the occupation measures under $\boldsymbol{\pi}$ versus $\widehat{\boldsymbol{\pi}}$. In standard uses of this inequality, an initial argument is used to guarantee that one of these expectations is negative, and so can be dropped [20, 21].

In contrast, the proof of our Theorem 1 exploits a more refined approach, one that handles the difference of expectations directly. Doing so can be beneficial—and lead to "fast rates"— because various terms in this difference can cancel each other out. Specifically, under the smoothness conditions that underlie Theorem 1, when applying the telescope inequality (6) with comparator $\boldsymbol{\pi} = \boldsymbol{\pi}^\star$, we show that the discrepancy between the occupation measures associated with $\boldsymbol{\pi}^\star$ and $\widehat{\boldsymbol{\pi}}$ is of the *same order* as the Bellman residual associated with $\widehat{\boldsymbol{f}}$. Note that the Bellman residuals of $\widehat{\boldsymbol{f}}$ already appear on the right-hand side of inequality (6), so that this fortuitous cancellation can be exploited—along with a number of auxiliary results laid out in the proof—so as to upper bound the value gap by a quantity proportional to the squared Bellman residual ε^2.

It is worthwhile making an explicit comparison of our cancellation approach with the more standard uses of the telescope relation, which typically consider only one portion of the Bellman residuals (e.g., [18, 20, 21, 19, 6, 12, 35]). We do so in the following two subsections.

2.3.1 Pessimism for off-line RL

In the off-line instantiation of RL, the goal is to learn a "good" policy based on a pre-collected dataset $\mathcal{D}$. Note that no further interaction with the environment is permitted, hence the notion of the learning being off-line. More precisely, an *off-line dataset* $\mathcal{D}$ of size n consists of quadruples

$$\mathcal{D} = \Big\{ \big(s_{h,\,i}, a_{h,\,i}, s'_{h,\,i}, r_{h,\,i}\big) \Big\}_{i=1}^{n},$$

where $s_{h,\,i}$ and $a_{h,\,i}$ represent the i-th state and action at the h-th step in the MDP; $s'_{h,\,i}$ is the successive state; and $r_{h,\,i} = r_h(s_{h,\,i}, a_{h,\,i})$ denotes the scalar reward. Note that while the successive

[3] Results of this type are known; for example, analogous results can be found in past work (e.g., Theorem 2 of the paper [34]; or Lemma 3.2 in the paper [7]).

states are defined by transition dynamics, and the rewards by the reward function, there are no restrictions on how the state-action pairs $(s_{h,i}, a_{h,i})$ are collected. That is, they need not have been generated by any fixed policy, but may have collected from some ensemble of behavioral policies, or even adaptively by human experts. The goal of off-line reinforcement learning is to use the n-sample dataset $\mathcal{D}$ so as to estimate a policy $\widehat{\boldsymbol{\pi}} \equiv \widehat{\boldsymbol{\pi}}_n$ that (approximately) maximizes the expected return $J(\widehat{\boldsymbol{\pi}}_n)$. We expect that—at least for a sensible method for estimating $\widehat{\boldsymbol{\pi}}_n$—the value gap $J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}_n)$ should decay to zero as n increases to infinity, and we are interested in understanding this rate of decay.

The use of pessimism is standard in off-line RL algorithms. Its purpose is to mitigate risks associated with "poor coverage" of the off-line dataset. For instance, the naive approach of simply maximizing Q-function estimates based on an off-line dataset can behave poorly when certain portions of the state-action space are not well covered by the given dataset. The pessimism principle suggests to form a *conservative estimate* of the value function—say with

$$\widehat{f}_h(s,a) \leq \mathcal{T}_h^\star \, \widehat{f}_{h+1}(s,a) \tag{7a}$$

with high probability over state-action pairs (s,a). Thus, the estimated value $\widehat{f}_h(s,a)$ is an underestimate of the Bellman update, a form of conservatism that protects against unrealistically high estimates due to poor coverage. Doing so in the appropriate way ensures that

$$-\mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\big(\mathcal{T}_h^\star \, \widehat{f}_{h+1} - \widehat{f}_h\big)(S_h, A_h)\big] \leq 0. \tag{7b}$$

Applying this upper bound to the inequality (6) yields the sub-optimality bound

$$J(\boldsymbol{\pi}) - J(\widehat{\boldsymbol{\pi}}) \;\leq\; \sum_{h=1}^{H-1} \mathbb{E}_{\boldsymbol{\pi}}\big[\big(\mathcal{T}_h^\star \, \widehat{f}_{h+1} - \widehat{f}_h\big)(S_h, A_h)\big] \, .$$

Upper bounds derived in this manner only contain one portion of the Bellman residual. When the value functions are approximated in a parametric way (e.g., tabular problems, linear function approximation), this line of analysis leads to value sub-optimality decaying at a "slow" $1/\sqrt{n}$ rate in terms of the sample size n (e.g., [21]). In contrast, an application of Theorem 1 can lead to value gaps bounded by $1/n$.

2.3.2 Optimism in on-line RL

In the setting of on-line RL, a learning agent interacts with the environment in a sequential manner, receiving feedback in the form of rewards based on its actions. At the beginning, the learner possesses no prior knowledge of the system's dynamics. In the t-th episode, the agent learns an optimal policy $\widehat{\boldsymbol{\pi}}^{(t)}$ using existing observations, implements the policy and collects data $\big\{\big(s_h^{(t)}, \, a_h^{(t)}, \, r_h^{(t)}\big)\big\}_{h=1}^{H}$ from the new episode. In each round, the system starts at an initial state $s_1^{(t)}$ independently drawn from a fixed distribution ξ_1.

In this on-line setting, it is common to measure the performance of an algorithm by comparing it, over the T rounds of learning, with an oracle that knows and implements an optimal policy. At each round t, we incur the *instantaneous regret* $J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^{(t)})$, where $\boldsymbol{\pi}^\star$ is any optimal policy. Over T rounds, we measure performance in terms of the *cumulative regret*

$$\text{Regret}\big(\{\widehat{\boldsymbol{\pi}}^{(t)}\}_{t=1}^{T}\big) := \max_{\text{policy } \boldsymbol{\pi}} \; \sum_{t=1}^{T} \Big\{ J(\boldsymbol{\pi}) - J(\widehat{\boldsymbol{\pi}}^{(t)}) \Big\} \;=\; \sum_{t=1}^{T} \underbrace{\Big\{ J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^{(t)}) \Big\}}_{\text{Regret at round } t}. \tag{8}$$

In a realistic problem, the cumulative regret of any procedure grows with T, and our goal is to obtain algorithms whose regret grows as slowly as possible.

In contrast to off-line RL, the on-line setting allows for exploring state-action pairs that have been rarely encountered; doing so makes sense since they might be associated with high rewards. Principled exploration of this type can be effected via the *optimism principle*: one constructs function estimates such that

$$\widehat{f}_h(s,a) \geq \mathcal{T}_h^\star \, \widehat{f}_{h+1}(s,a) \tag{9a}$$

https://doi.org/10.52202/079017-2279

with high probability over state-action pairs.[4] Note that $\widehat{f}_h(s,a)$ is optimistic in the sense that it is an over-estimate of the Bellman update $\mathcal{T}_h^\star\, \widehat{f}_{h+1}(s,a)$. In this way, we can ensure that

$$\mathbb{E}_{\boldsymbol{\pi}}\big[\big(\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h\big)(S_h, A_h)\big] \leq 0. \tag{9b}$$

Combining this inequality with the telescope bound (6) allows one to upper bound the regret as

$$\text{Regret}\big(\{\widehat{\boldsymbol{\pi}}^{(t)}\}_{t=1}^T\big) \;=\; \sum_{t=1}^T \big\{J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^{(t)})\big\} \;\leq\; \sum_{t=1}^T \sum_{h=1}^{H-1} \mathbb{E}_{\widehat{\boldsymbol{\pi}}^{(t)}}\big[\big(\widehat{f}_h - \mathcal{T}_h^\star\, \widehat{f}_{h+1}\big)(S_h, A_h)\big]\,.$$

which only includes a single portion of the Bellman residual. In the case of tabular or linear representations of the Q-functions, it results in a regret rate of $\sqrt{T}$ (e.g., see the papers [18, 20]). In contrast, an appropriate use of Theorem 1 leads to regret growing only as $\log(T)$, which corresponds to a much better guarantee.

In summary, then, the fast rates obtained in this paper are based on a different approach than the standard pessimism or optimism principles. Since we deal directly with the difference of expectations in the bound (6), there is no need to nullify either of them through the use of these principles. However, it should be noted that we are assuming smoothness conditions that allow us to control this difference. As we discuss in the sequel, such smoothness conditions rule out certain "hard instances" used in past work on lower bounds (e.g. [18, 20, 21, 37]).

3 Consequences for linear function approximation

In this section, we explore some consequences of our general theory when applied to value-based methods using (finite-dimensional) linear function approximation.

Let $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a given feature map on the state-action space, and consider linear expansions of the form $f_{\boldsymbol{w}}(s,a) = \langle \boldsymbol{\phi}(s,a),\, \boldsymbol{w}\rangle \equiv \sum_{j=1}^d w_j \boldsymbol{\phi}_j(s,a)$ where $\boldsymbol{w} \in \mathbb{R}^d$ is a weight vector. We adopt the conventional assumption that $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$ and $r_h(s,a) \in [0,1]$ for all state-action pairs. Defining the linear function class $\mathscr{F} := \big\{f_{\boldsymbol{w}} \mid \boldsymbol{w} \in \mathbb{R}^d\big\}$, we note that the Minkowski difference class $\partial\mathscr{F}$ is equal to $\mathscr{F}$.

In our analysis of linear approximation, we make use of the norm $\|f\|_h := \sqrt{\mathbb{E}_{\boldsymbol{\pi}^\star}[f^2(S_h, A_h)]}$, corresponding to L^2-norm under the occupation measure induced by the optimal policy $\boldsymbol{\pi}^\star$.

3.1 Consequences for off-line RL

We now turn to some implications of Theorem 1 for off-line reinforcement learning. Let us recall the off-line setting: for each $h = 1, \ldots, H-1$, we are given a dataset $\mathcal{D}_h = \{(s_{h,i}, a_{h,i}, s'_{h,i}, r_{h,i})\}_{i=1}^n$ of quadruples, from which we can compute estimates $\widehat{\boldsymbol{f}} = (\widehat{f}_h)_{h=1}^H$ with certain Bellman residuals $\{\varepsilon_h\}_{h=1}^{H-1}$, which then appear in the bound (5). The remaining factors on the right-hand side of inequality (5) do not depend on the dataset itself (but rather on structural properties of the MDP). Consequently, in terms of statistical understanding, the main challenge is to establish high-probability bounds on the Bellman residuals $\{\varepsilon_h\}_{h=1}^{H-1}$ for a particular estimator.

3.1.1 Fitted Q-iteration (FQI)

As an illustration, let us analyze the use of *fitted Q-iteration* (FQI) for computing estimates of the Q-function. At a given stage $h = 1, \ldots, H-1$, we can use the associated data $\mathcal{D}_h$ to define a regularized objective function

$$\mathcal{L}_h\big(f,\, g\big) := \frac{1}{|\mathcal{D}_h|}\Bigg[\sum_{(s_{h,i}, a_{h,i}, s'_{h,i}, r_{h,i}) \in \mathcal{D}_h} \big\{f(s_{h,\,i},\, a_{h,\,i}) - \big(r_{h,\,i} + \max_{a \in \mathcal{A}} g(s'_{h,\,i}, a)\big)\big\}^2\Bigg] + \Lambda_h^2(f)\,.$$

Here g represents the target function from stage $h+1$, and it defines the targeted responses $y_{h,i}(g) := r_{h,i} + \max_{a\in\mathcal{A}} g(s'_{h,i}, a)$. For a given target g, we obtain a Q-function estimate for

[4] Please refer to, for example, Lemma B.3 in the paper [20] for further details.

stage h by minimizing the functional $f \mapsto \mathcal{L}_h(f, g)$. Given that our objective is defined with a quadratic cost, doing so can be understood as a regression method for estimating the conditional expectation that underlies the Bellman update—viz. $\mathcal{T}_h^\star\, g(s,a) = \mathbb{E}[\, y_{h,\,i}(g) \mid (s_{h,\,i},\, a_{h,\,i}) = (s,a)\,]$. Here $\Lambda_h^2(f) \,=\, \lambda_h \,\|\boldsymbol{w}\|_2^2$ for $f = \langle \boldsymbol{\phi}(\cdot),\, \boldsymbol{w} \rangle$ is a regularizer, with $\lambda_h \,\geq\, 0$ being the regularization weight. Given this set-up, we can generate a Q-function estimate $\widehat{\boldsymbol{f}} \,=\, (\widehat{f}_1, \ldots, \widehat{f}_H)$ by first initializing $\widehat{f}_H \,=\, r_H$, and then recursively computing $\widehat{f}_h \,=\, \arg\min_{f \in \mathscr{F}} \mathcal{L}_h\big(f,\, \widehat{f}_{h+1}\big)$, for $h = H-1, H-2, \ldots, 2, 1$.

3.1.2 Fast rates for FQI-based estimates

In the analysis here, we assume that the dataset consists of i.i.d. tuples (but this can be relaxed as needed). We now state a corollary of Theorem 1, applicable to value function estimates based on FQI with ridge regression.

Corollary 1 (Fast rates for ridge-based FQI)**.** *For FQI based on ridge regression, with a sufficiently large sample size n and with suitable choices of the regularization parameters $\{\lambda_h\}_{h=1}^{H-1}$, the bound* (5) *from Theorem 1 holds with*

$$\varepsilon_h = c\,\sqrt{\{d(H-h)/n\}\,\log(dH/\delta)} \tag{10}$$

with probability at least $1-\delta$.

We omit the proof of Corollary 1, as it follows from standard ridge regression analysis.

Fast rates and comparisons to past work: So as to be able to compare with results from past work, let us consider some consequences of the bound (10) under the following assumptions: (i) $\boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star) = \mathcal{O}(1)$; (ii) $\boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star) = \mathcal{O}(\sqrt{d})$; (iii) $\|Q_h^\star\|_h \asymp H - h + 1$. Then it can be shown that the bound from Corollary 1 takes the form

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \;\leq\; c\, d^{3/2}\, H^3\, n^{-1}\, \log(dH/\delta), \tag{11a}$$

and is valid for a sample size $n \geq cd^2H^3$. Alternatively stated, Corollary 1 guarantees that for FQI using ridge regression with d-dimensional function approximation, the number of samples $n(\epsilon)$ required to obtain ϵ-optimal policy is at most

$$n_{\text{fast}}(\epsilon) \asymp d^{\frac{3}{2}} H^3/\epsilon + d^2 H^3, \tag{11b}$$

where we use $\asymp$ to denote a scaling that ignores constants and logarithmic factors.

Let us compare this guarantee to related work by Zanette et al. [37], who analyzed the use of pessimistic actor-critic methods for linear function classes. When translated into the notation of our paper, their analysis established[5] a sample complexity of the order $n_{\text{Zan}}(\epsilon) \asymp d^2H^3/\epsilon^2$. Consequently, we see that once the target error ϵ is relatively small—$\epsilon \in (0,1)$—then stable MDPs can exhibit a much smaller $(1/\epsilon)$ sample complexity.

It should be noted that past work (e.g., [21, 37]) has established $(1/\epsilon^2)$-lower bounds on the sample complexity of estimating ϵ policies in the off-line setting. However, these lower bounds *do not* contradict our fast rate guarantee (11b), because the "hard instances" used in these lower bound proofs violate the stability condition **(Stb(ξ))**. In particular, even infinitessimally small perturbations in policy lead to occupation measures that are significantly different.

When is pessimism necessary? An interesting aspect of the guarantee from Corollary 1 is that it provides guarantees for off-policy RL (and with fast rates) using a method that does *not* incorporate any form of pessimism. This is a sharp contrast with many other methods for off-policy RL, such as pessimistic forms of Q-learning and actor-critic methods (e.g., [21, 37]).

To be clear, as noted following the bound (11a), the guarantee from Corollary 1 requires the sample size to be lower bounded as $n \geq cd^2H^3$. In contrast, pessimistic schemes only require a sample size sufficiently large to ensure validity of the Bellman residual upper bounds that underlie Corollary 1—meaning that $n \gtrsim d$ up to logarithmic factors. Thus, the pessimism principle can be useful for problems with smaller sample sizes.

[5] See Appendix C.2 for the details of this calculation.

https://doi.org/10.52202/079017-2279

3.2 Consequences for on-line RL

In this section, we explore some consequences of Theorem 1 for on-line reinforcement learning. We begin by describing a two-stage procedure[6] that allows us to convert the risk bounds for FQI from off-line RL into regret in on-line RL:

Phase 1 (*Exploration*) In the initial T_0f episodes, the focus is purely on exploration, resulting in an estimate of Q-function denoted as $\widehat{\boldsymbol{f}}^{(T_0)}$.

Phase 2 (*Fine-tuning*) For $k = 0, 1, \ldots, K-1$ with $K := \lceil \log_2(T/T_0) \rceil$, repeat:

- In the t-th episode, for each $t = T_0\, 2^k + 1, \ldots, T_0\, 2^{k+1}$, execute the greedy policy induced by function $\widehat{\boldsymbol{f}}^{(T_0\, 2^k)}$.
- Update the Q-function estimate $\widehat{\boldsymbol{f}}^{(T_0\, 2^{k+1})}$ using FQI based on observations collected from episodes $T_0\, 2^k + 1, T_0\, 2^k + 2, \ldots, T_0\, 2^{k+1}$.

We assume the burn-in time T_0 is large enough so as to ensure the pilot Q-function estimate $\widehat{\boldsymbol{f}}^{(T_0)}$ obtained in Phase 1 falls within a certain "absorbing" region $\mathcal{N}(\boldsymbol{\rho})$ around $\boldsymbol{Q}^\star$. Under these conditions, we have the following bound on the regret.

Corollary 2. *For FQI based on ridge regression with rewards in* $[0, 1]$*, with a sufficiently large burn-in time* T_0 *and with suitable choices of the regularization parameters* $\{\lambda_h\}_{h=1}^{H-1}$*, the two-phase scheme achieves regret bounded as*

$$\textit{Regret}(T) \;\leq\; c\left\{T_0 \cdot H \;+\; d\sqrt{d}\; H^4\, \log T \;\cdot\; \log(dHK/\delta)\right\}$$

with probability at least $1 - \delta$*.*

See Appendix C.1 for the proof.

Sharper bound on regret: The leading term (as T grows) in the regret bound grows as $\log T$, which is much smaller than the typical $\sqrt{T}$-rate found in past work [18, 20]. The $\sqrt{T}$ rate has been shown to be unimprovable in general, but the worst-case instances[18, 20] that lead to $\sqrt{T}$-regret violate the stability conditions used in our analysis.

When is optimism needed? The use of optimism—by adding bonuses to the current value function estimates so as to encourage exploration—underlies many schemes in on-line RL. An interesting take-away from Corollary 2 is that under the stability conditions highlighted by our theory, it is possible to achieve excellent regret bounds without the use of optimism. In our two-phase scheme, the only exploration occurs in Phase 1. All other data is simply collected using the greedy policy induced by the current Q-function estimate. A well-designed exploration scheme—one that might incorporate the optimism principle—is necessary only during the burn-in Phase 1.

4 Discussion

This paper introduces a novel approach for the analysis of value-based RL methods for continuous state-action spaces. Our analysis highlights two key stability properties of MDPs under which much sharper bounds on value sub-optimality can be guaranteed. Our analysis offers fresh perspectives on the commonly used pessimism and optimism principles, in off-line and on-line settings respectively.

Our study leaves open various questions for future work. First, our main result (Theorem 1) has consequences for linear quadratic control, to be described in an upcoming paper [8]. It provides insight into the role of covariate shift in linear quadratic control, as well as efficient exploration in the on-line setting. Second, our current statistical analysis focused on i.i.d. data with linear function approximation. It is interesting to consider the extensions to dependent data and non-parametric function approximation (e.g. kernels, boosting, and neural networks). Third, while this paper has provided upper bounds, it remains to address the complementary question of lower bounds for policy optimization over the classes of stable MDPs isolated here. Last, to better align our framework with real-world scenarios, we intend to go beyond the idealized completeness condition used in this paper, and treat the role of model mis-specification.

[6] To be clear, the purpose of this scheme is primarily conceptual, rather than practical in nature.

Acknowledgements

This work was partially supported by NSF grant CCF-1955450, ONR grant N00014-21-1-2842, and NSF DMS-2311072 to MJW.

References

[1] A. Altamimi, C. Lagoa, J. G. Borges, M. E. McDill, C. Andriotis, and K. Papakonstantinou. Large-scale wildfire mitigation through deep reinforcement learning. *Frontiers in Forests and Global Change*, 5:734330, 2022.

[2] H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

[3] D. Bertsekas. *Lessons from AlphaZero for optimal, model predictive, and adaptive control*. Athena Scientific, 2022.

[4] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[5] J. Degrave, F. Felici, and J. B. et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602:414–419, 2022.

[6] S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[7] Y. Duan, C. Jin, and Z. Li. Risk bounds and Rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021.

[8] Y. Duan and M. J. Wainwright. Covariate shift in linear quadratic control. *Manuscript*.

[9] Y. Duan and M. J. Wainwright. Policy evaluation from a single path: Multi-step methods, mixing and mis-specification. *arXiv preprint arXiv:2211.03899*, 2022.

[10] Y. Duan and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[11] Y. Duan, M. Wang, and M. J. Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

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

[13] J. Gijsbrechts, R. N. Boute, J. A. Van Mieghem, and D. J. Zhang. Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management*, 24(3):1349–1368, 2022.

[14] A. Gonen and S. Shalev-Shwartz. Fast rates for empirical risk minimization of strict saddle problems. In *Conference on Learning Theory*, pages 1043–1063. PMLR, 2017.

[15] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

[16] J. He, D. Zhou, and Q. Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

[17] Y. Hu, N. Kallus, and M. Uehara. Fast rates for the regret of offline reinforcement learning. *Conference on Learning Theory*, 134:2462–2462, 2021.

[18] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

 https://doi.org/10.52202/079017-2279

[19] C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

[20] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[21] Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? *International Conference on Machine Learning*, pages 5084–5096, 2021.

[22] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

[23] T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. *Advances in Neural Information Processing Systems*, 28, 2015.

[24] H. Mania, S. Tu, and B. Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] T. Nguyen-Tang, M. Yin, S. Gupta, S. Venkatesh, and R. Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. *Association for the Advancement of Artificial Intelligence*, 2023.

[26] M. L. Puterman and S. L. Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research*, 4(1):60–69, 1979.

[27] A. Rao and T. Jelvis. *Foundations of Reinforcement Learning with Applications to Finance*. CRC Press, Boca Raton, FL, 2022.

[28] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[29] B. Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

[30] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

[31] S. Spielberg, A. Tulsyan, N. P. Lawrence, P. D. Loewen, and R. B. Gopaluni. Toward self-driving processes: A deep reinforcement learning approach to control. *Amer. Inst. Chem. Eng. Journal*, 65:e16689, 2022.

[32] L. Tai, G. Paolo, and M. Liu. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 31–36. IEEE, 2017.

[33] X. Wang, Q. Cui, and S. S. Du. On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14865–14877, 2022.

[34] T. Xie and N. Jiang. Q^* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

[35] M. Yin, Y. Duan, M. Wang, and Y.-X. Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

[36] C. Yu, J. Liu, S. Nemati, and G. Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[37] A. Zanette, M. J. Wainwright, and E. Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

A Proofs of Theorem 1

This section is devoted to the proof of Theorem 1, which consists of three main steps. These steps rely on two auxiliary lemmas whose proofs are fairly technical, so that they are deferred to in Appendices B.1 and B.2.

High-level outline: Let us outline the three steps of the proof. In Step 1, we use a one-step expansion of the difference in the occupation measures to reformulate the standard telescope inequality (6). Doing so results in a relation with structure similar to that of the left-hand side of inequality **(Stb(ξ))**. In Step 2, we develop a constraint on the function estimation error $d_h\big(\widehat{f}_h, Q_h^\star\big)$ that ensures the occupation measure produced by policy $\widehat{\boldsymbol{\pi}}$ remains stable and does not deviate too much from the occupation measure associated with the optimal policy $\boldsymbol{\pi}^\star$. In Step 3, we use Bellman stability **(Stb($\mathcal{T}$))** to connect the Q-function error $\widehat{f}_h - Q_h^\star$ with Bellman residuals. With this high-level view in place, we now work through the three steps.

A.1 Step 1: Reformulation of the telescope inequality.

Recall the standard telescope inequality (6). Our proof makes use of an alternative form, which involves the functions

$$\Delta_h(\boldsymbol{\pi};\, s, a\,) \;=\; \sum_{h'=h}^{H-1} \mathcal{P}_{h,h'}^{\boldsymbol{\pi}}\big(\mathcal{T}_{h'}^\star\, \widehat{f}_{h'+1} - \widehat{f}_{h'}\big)(s,a). \tag{12}$$

Lemma 1. *Given a Q-function estimate $\widehat{\boldsymbol{f}} = \big(\widehat{f}_1, \ldots, \widehat{f}_{H-1}, \widehat{f}_H = r_H\big)$ and the associated greedy policy $\widehat{\boldsymbol{\pi}}$, we have the bound*

$$J(\boldsymbol{\pi}) - J(\widehat{\boldsymbol{\pi}}) \;\le\; \sum_{h=1}^{H-1} \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\Delta_h(\boldsymbol{\pi};\, s_h, \pi_h(s_h)) - \Delta_h(\boldsymbol{\pi};\, s_h, \widehat{\pi}_h(s_h))\big] \tag{13}$$

valid for any policy $\boldsymbol{\pi}$.

See Appendix B.1 for the proof.

We apply the bound (13) with $\boldsymbol{\pi} = \boldsymbol{\pi}^\star$. Following some algebra, we find that

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \;\le\; \sum_{h=1}^{H-1} \sum_{h'=h}^{H-1} \widehat{\beta}(h,\, h') \cdot \varepsilon_{h'}\ ,$$

where $\varepsilon_{h'}$ is an upper bound on the Bellman residual $\big\|\mathcal{T}_{h'}^\star\, \widehat{f}_{h'+1} - \widehat{f}_{h'}\big\|_{h'}$ as given in equation (4a). The term $\widehat{\beta}(h,\, h')$ is given by

$$\widehat{\beta}(h,\, h') := \sup_{f \in \partial\mathscr{F}:\, \|f\|_{h'} > 0} \left\{ \frac{1}{\|f\|_{h'}} \Big| \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\Big[\big(\mathcal{P}_{h,h'}^\star\, f\big)(s_h, \pi_h^\star(s_h)) - \big(\mathcal{P}_{h,h'}^\star\, f\big)(s_h, \widehat{\pi}_h(s_h)) \Big] \Big| \right\}. \tag{14a}$$

We note that the left-hand side of inequality **(Stb(ξ))** has a similar form to the term $\widehat{\beta}(h,\, h')$, differing only in that the expectation is taken over the occupation measure of running the optimal policy $\boldsymbol{\pi}^\star$, rather than the estimated policy $\widehat{\boldsymbol{\pi}}$.

A.2 Step 2: Constraint to ensure stability

Our next step is to establish an upper bound on the coefficient $\widehat{\beta}(h,h')$ defined by the estimated policy $\widehat{\boldsymbol{\pi}}$ in terms of the analogous quantity defined by the optimal policy $\boldsymbol{\pi}^\star$—namely, the coefficient

$$\beta(h,h') := \sup_{f \in \partial\mathscr{F}:\, \|f\|_{h'} > 0} \left\{ \frac{1}{\|f\|_{h'}} \Big| \mathbb{E}_{\boldsymbol{\pi}^\star}\Big[\big(\mathcal{P}_{h,h'}^\star\, f\big)(s_h, \pi_h^\star(s_h)) - \big(\mathcal{P}_{h,h'}^\star\, f\big)(s_h, \widehat{\pi}_h(s_h)) \Big] \Big| \right\}. \tag{14b}$$

https://doi.org/10.52202/079017-2279

In order to do so, we demonstrate that a sufficiently small function estimation error $d_h\big(\widehat{f}_h, Q_h^\star\big)$ ensures the inequality

$$\sum_{h=1}^{H-1} \sum_{h'=h}^{H-1} \widehat{\beta}(h,\, h') \cdot \varepsilon_{h'} \;\leq\; 2 \sum_{h=1}^{H-1} \sum_{h'=h}^{H-1} \beta(h,\, h') \cdot \varepsilon_{h'}\,. \tag{15}$$

Once we have established this bound, we can replace the term $\beta(h, h')$ with $\boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star) \cdot \big\|\widehat{f}_h - Q_h^\star\big\|_h / \|Q_h^\star\|_h$, using the inequality **(Stb(ξ))**.

We summarize the result in the following auxiliary lemma:

Lemma 2. *Suppose that the function estimation errors satisfy* $d_h\big(Q_h,\, Q_h^\star\big) \leq \frac{1}{2\, b_{\mathscr{F}}}\, (H - h + 1)^{-1}$ *for* $h = 2, 3, \ldots, H-1$ *and the sequence* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{H-1}, \varepsilon_H = 0)$ *satisfies the regularity condition* (4b)*. Then we have*

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \;\leq\; 2 \sum_{h=1}^{H-1} \frac{\big\|\widehat{f}_h - Q_h^\star\big\|_h}{\|Q_h^\star\|_h} \Big\{ \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)\, \varepsilon_{h'} \Big\}. \tag{16}$$

See Appendix B.2 for the proof.

A.3 Step 3: Connecting Q-function error and Bellman residuals

The remaining piece of the proof is to connect the function difference $\widehat{f}_h - Q_h^\star$ with Bellman residuals $\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h$, using the stability condition **(Stb($\mathcal{T}$))** on the Bellman operator $\boldsymbol{\mathcal{T}}^\star$. This is relatively straightforward: indeed, we claim that

$$\big\|\widehat{f}_h - Q_h^\star\big\|_h \;\leq\; \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star) \cdot \big\|\mathcal{T}_{h'}^\star\, \widehat{f}_{h'+1} - \widehat{f}_{h'}\big\|_{h'}\,. \tag{17}$$

Recall that $Q_h^\star = \mathcal{T}_h^\star\, Q_{h+1}^\star$ for $h = 1, 2, \ldots, H-1$. Therefore, we have

$$\widehat{f}_h - Q_h^\star = \big(\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \mathcal{T}_h^\star\, Q_{h+1}^\star\big) - \big(\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h\big)\,.$$

By employing the triangle inequality and the Bellman stability given in equation **(Stb($\mathcal{T}$))**, we derive that

$$\begin{aligned}
\big\|\widehat{f}_h - Q_h^\star\big\|_h &\leq \big\|\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h\big\|_h + \big\|\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \mathcal{T}_h^\star\, Q_{h+1}^\star\big\|_h \\
&\leq \big\|\mathcal{T}_h^\star\, \widehat{f}_{h+1} - \widehat{f}_h\big\|_h + \kappa_h^\star\, \big\|\widehat{f}_{h+1} - Q_{h+1}^\star\big\|_{h+1}\,.
\end{aligned}$$

Applying this inequality recursively yields the claim (17).

With this piece in place, we can complete the proof of Theorem 1. Indeed, we have

$$\begin{aligned}
J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) &\overset{(a)}{\leq} 2 \sum_{h=1}^{H-1} \frac{\big\|\widehat{f}_h - Q_h^\star\big\|_h}{\|Q_h^\star\|_h} \Big\{ \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)\, \varepsilon_{h'} \Big\} \\
&\overset{(b)}{\leq} 2 \sum_{h=1}^{H-1} \frac{1}{\|Q_h^\star\|_h} \Big\{ \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\mathcal{T}^\star)\, \varepsilon_{h'} \Big\} \Big\{ \sum_{h'=h}^{H-1} \boldsymbol{\kappa}_{h,h'}(\boldsymbol{\pi}^\star)\, \varepsilon_{h'} \Big\}.
\end{aligned}$$

Here step (a) is a restatement of the bound (16) from Lemma 2, whereas step (b) follows from inequality (17). Thus, we have established the claim given in Theorem 1.

B Proof of auxiliary lemmas for Theorem 1

We now turn to proofs of the two auxiliary results used to establish our main theorem, with Lemmas 1 and 2 treated in Appendices B.1 and B.2, respectively.

B.1 Proof of Lemma 1

For any integrable vector function $\boldsymbol{g} = (g_1, \ldots, g_H) \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times H}$, we define

$$D(\boldsymbol{g}) \;=\; \sum_{h=1}^{H} \big(\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big)\big[g_h(S_h, A_h)\big]\,. \tag{18a}$$

We claim that this functional satisfies the recursive relation

$$D(\boldsymbol{g}) = \sum_{h=1}^{H} \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[g_h(S_h, \pi_h(S_h)) - g_h(S_h, \widehat{\pi}_h(S_h))\big] + D(\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}}\boldsymbol{g}), \tag{18b}$$

where we have introduced the shorthand $\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}}\boldsymbol{g} := \big(\mathcal{P}_1^{\boldsymbol{\pi}}\, g_2, \ldots, \mathcal{P}_{H-1}^{\boldsymbol{\pi}}\, g_H, 0\big) \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times H}$.

Taking this claim as given for the moment, let us prove the bound (13) from Lemma 1. First, we set $\boldsymbol{g} := (\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}})^h\, \boldsymbol{g} = \big(\mathcal{P}_{1,1+h}^{\boldsymbol{\pi}}\, g_{1+h}, \ldots, \mathcal{P}_{H-h,H}^{\boldsymbol{\pi}}\, g_H, 0, \ldots, 0\big)$ in equation (18b) for $h = 0, 1, \ldots, H-1$, which yields

$$\begin{aligned} D\big((\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}})^h\, \boldsymbol{g}\big) = \sum_{\substack{1\le h'\le j\le H,\\ j-h'=h}} & \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\{\mathcal{P}_{h',j}^{\boldsymbol{\pi}}\, g_j\}(S_{h'}, \pi_{h'}(S_{h'})) - \{\mathcal{P}_{h',j}^{\boldsymbol{\pi}}\, g_j\}(S_{h'}, \widehat{\pi}_{h'}(S_{h'}))\big] \\ & + D\big((\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}})^{h+1}\, \boldsymbol{g}\big). \end{aligned}$$

Note that $(\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}})^H\, \boldsymbol{g} = 0$, which implies $D\big((\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}})^H\, \boldsymbol{g}\big) = 0$. We then sum the resulting bounds so as to obtain

$$D(\boldsymbol{g}) = \sum_{1\le h\le h'\le H} \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\{\mathcal{P}_{h,h'}^{\boldsymbol{\pi}}\, g_{h'}\}(S_h, \pi_h(S_h)) - \{\mathcal{P}_{h,h'}^{\boldsymbol{\pi}}\, g_{h'}\}(S_h, \widehat{\pi}_h(S_h))\big]. \tag{19}$$

Setting $\boldsymbol{g} = \boldsymbol{\mathcal{T}}^{\star}\widehat{\boldsymbol{f}} - \widehat{\boldsymbol{f}}$, or equivalently $g_h = \mathcal{T}_h^{\star}\, \widehat{f}_{h+1} - \widehat{f}_h$, in equation (19), we find that

$$D\big(\boldsymbol{\mathcal{T}}^{\star}\widehat{\boldsymbol{f}} - \widehat{\boldsymbol{f}}\big) = \sum_{h=1}^{H-1} \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\Delta_h(\boldsymbol{\pi}; S_h, \pi_h(S_h)) - \Delta_h(\boldsymbol{\pi}; S_h, \widehat{\pi}_h(S_h))\big]\,,$$

where we have used the fact (12) that $\Delta_h(\boldsymbol{\pi};\, \cdot) = \sum_{h'=h}^{H} \mathcal{P}_{h,h'}^{\boldsymbol{\pi}}\big(\mathcal{T}_{h'}^{\star}\, \widehat{f}_{h'+1} - \widehat{f}_{h'}\big)$. Thus, we have established the bound (13) stated in Lemma 1.

It remains to establish the auxiliary claim (18b). Note that the functional D can be decomposed as $D(\boldsymbol{g}) = D_1 + D_2$, where

$$\begin{aligned} D_1 &:= \sum_{h=1}^{H} \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[g_h(S_h, \pi_h(S_h)) - g_h(S_h, \widehat{\pi}_h(S_h))\big] \qquad \text{and} \\ D_2 &:= \sum_{h=1}^{H} \big(\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big)\big[g_h(S_h, \pi_h(S_h))\big]\,. \end{aligned}$$

Applying the tower property of conditional expectation, we find that

$$\begin{aligned} D_2 \;&=\; \sum_{h=1}^{H-1} \big(\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big)\big[\mathbb{E}[g_{h+1}(S_{h+1}, \pi_{h+1}(S_{h+1})) \mid S_h, A_h]\big] \\ &=\; \sum_{h=1}^{H-1} \big(\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big)\big[(\mathcal{P}_h^{\boldsymbol{\pi}}\, g_{h+1})(S_h, A_h)\big] = D\big(\boldsymbol{\mathcal{P}}^{\boldsymbol{\pi}}\boldsymbol{g}\big). \end{aligned}$$

Combining the expressions for D_1 and D_2 above yields the claim (18b).

https://doi.org/10.52202/079017-2279

B.2 Proof of Lemma 2

The key step in proving Lemma 2 is establishing that inequality (15) holds when the function estimation error $d_h\big(\widehat{f}_h, Q_h^\star\big)$ is sufficiently small. In order to do so, we need to establish upper bounds on the term $\widehat{\beta}(h, h')$ by using $\beta(h, h')$. In particular, we will show that for any $1 \le h \le h' \le H-1$,

$$\widehat{\beta}(h,\, h') \;\le\; \beta(h,\, h') \;+\; \sum_{j=1}^{h-1} \widehat{\beta}(j,\, h-1) \;\cdot\; b_{\mathscr{F}} \;\cdot\; d_h\big(\widehat{f}_h,\, Q_h^\star\big)\,. \tag{20}$$

The inequality (20) is derived based on the definitions of metric d_h and parameter $b_{\mathscr{F}} = 1$. After a close examination of the right-hand side of this inequality, it becomes evident that as long as the function estimation error $d_h\big(\widehat{f}_h, Q_h^\star\big)$ is sufficiently small, the terms associated with $d_h\big(\widehat{f}_h, Q_h^\star\big)$ are negligible and are dominated by $\beta(h,\, h')$. Consequently, inequality (15) within the arguments in Appendix A.2 is likely to hold true.

With claim (20) assumed to be valid at this point, we now establish a proper upper bound on the estimation error $d_h\big(\widehat{f}_h, Q_h^\star\big)$ under which inequality (15) is satisfied. By taking linear combinations of inequality (20) using weights $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{H-1}, \varepsilon_H = 0)$, we obtain

$$\begin{aligned}\sum_{h=1}^{H-1}\sum_{h'=h}^{H-1} \widehat{\beta}(h,\, h')\cdot\varepsilon_{h'} \;\le\; &\sum_{h=1}^{H-1}\sum_{h'=h}^{H-1} \beta(h,\, h')\cdot\varepsilon_{h'} \\ &+ \sum_{h=2}^{H-1}\sum_{j=1}^{h-1} \widehat{\beta}(j,\, h-1) \;\cdot\; b_{\mathscr{F}} \;\cdot\; d_h\big(\widehat{f}_h,\, Q_h^\star\big) \sum_{h'=h}^{H-1} \varepsilon_{h'}\,. \end{aligned}\tag{21}$$

When the sequence $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{H-1}, \varepsilon_H = 0)$ is regular in the sense that inequality (4b) holds, the bound (21) reduces to

$$\begin{aligned}\sum_{1\le h\le h'\le H} \widehat{\beta}(h,\, h')\cdot\varepsilon_{h'} \;\le\; &\sum_{1\le h\le h'\le H} \beta(h,\, h')\cdot\varepsilon_{h'} \\ &+ \sum_{1\le h\le h'\le H-2} \widehat{\beta}(h,\, h')\cdot\varepsilon_{h'} \;\cdot\; b_{\mathscr{F}}\,(H-h') \;\cdot\; d_{h'+1}\big(\widehat{f}_{h'+1},\, Q_{h'+1}^\star\big)\,. \end{aligned}$$

Under the condition $d_h\big(\widehat{f}_h,\, Q_h^\star\big) \le \frac{1}{2\,b_{\mathscr{F}}}(H-h+1)^{-1}$ for $2 \le h \le H-1$, the inequality above implies bound (15), which further establishes the bound (16), as stated in Lemma 2.

It remains to prove the relation between $\widehat{\beta}(h,\, h')$ and $\beta(h,\, h')$, as shown in inequality (20).

Proof of bound (20)**:** It is evident that inequality (20) holds for $h = 1$, therefore, we focus on its validation for indices $2 \le h \le H-1$. Recall the definitions of functions $\widehat{\beta}(h, h')$ and $\beta(h, h')$, as given by equations (14a) and (14b). We apply the triangle inequality and derive that

$$\begin{aligned}&\big|\widehat{\beta}(h,\, h') - \beta(h,\, h')\big| \\ &\le \sup_{f\in\partial\mathscr{F}:\,\|f\|_{h'}>0} \left\{ \frac{1}{\|f\|_{h'}} \Big| \big(\mathbb{E}_{\widehat{\boldsymbol{\pi}}} - \mathbb{E}_{\boldsymbol{\pi}^\star}\big) \Big[\big(\mathcal{P}_{h,h'}^\star\, f\big)(S_h, \pi_h^\star(S_h)) - \big(\mathcal{P}_{h,h'}^\star\, f\big)(S_h, \widehat{\pi}_h(S_h)) \Big] \Big| \right\} \\ &= \sup_{f\in\partial\mathscr{F}:\,\|f\|_{h'}>0} \left\{ \frac{1}{\|f\|_{h'}} \Big| \big(\mathbb{E}_{\widehat{\boldsymbol{\pi}}} - \mathbb{E}_{\boldsymbol{\pi}^\star}\big) \Big[\big\{ \big(\mathcal{P}_{h-1}^\star - \mathcal{P}_{h-1}^{\widehat{\boldsymbol{\pi}}}\big)\, \mathcal{P}_{h,h'}^\star\, f \big\}(S_{h-1}, A_{h-1}) \Big] \Big| \right\} \\ &=:\; \Delta\beta(h,\, h')\,. \end{aligned}$$

The term $\Delta\beta(h,\, h')$ involves differences from two sources: (i) the difference in transition kernels $\mathcal{P}_{h-1}^\star - \mathcal{P}_{h-1}^{\widehat{\boldsymbol{\pi}}}$ that captures the divergence between policies $\pi_h^\star$ and $\widehat{\pi}_h$; (ii) the discrepancy of occupation measures at the $(h-1)$-th step reflected by the difference in expectations $\mathbb{E}_{\boldsymbol{\pi}^\star} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}$, which is determined by the policies $(\pi_1^\star, \ldots, \pi_{h-1}^\star)$ and $(\widehat{\pi}_1, \ldots, \widehat{\pi}_{h-1})$ until the $(h-1)$-th step. We treat them separately and write

$$\Delta\beta(h,\, h') \;\le\; \nu_1(h-1,\, h') \;\cdot\; \nu_2(h-1)\,, \tag{22}$$

where the functionals ν_2 and ν_1 are defined as

$$\nu_1(h-1,\, h') \;:=\; \sup_{f\in\partial\mathscr{F}:\, \|f\|_{h'}>0} \left\{ \frac{1}{\|f\|_{h'}} \big\| \big(\mathcal{P}^\star_{h-1} - \mathcal{P}^{\widehat{\boldsymbol{\pi}}}_{h-1}\big)\, \mathcal{P}^\star_{h,h'}\, f \,\big\|_{h-1} \right\},$$

$$\nu_2(h-1) \;:=\; \sup_{f\in\partial\mathscr{F}:\, \|f\|_{h-1}>0} \left\{ \frac{1}{\|f\|_{h-1}} \big| \big(\mathbb{E}_{\widehat{\boldsymbol{\pi}}} - \mathbb{E}_{\boldsymbol{\pi}^\star}\big) \big[f(S_{h-1}, A_{h-1}) \big] \big| \right\}.$$

We first consider the term ν_1. According to the definitions of metric d_h and parameter $b_{\mathscr{F}}$ we find that

$$\big\| \big(\mathcal{P}^\star_{h-1} - \mathcal{P}^{\widehat{\boldsymbol{\pi}}}_{h-1}\big)\, \mathcal{P}^\star_{h,h'}\, f \,\big\|_{h-1} \le d_h\big(\widehat{f}_h,\, Q^\star_h\big) \cdot \big\| \mathcal{P}^\star_{h,h'}\, f \,\big\|_h \overset{(*)}{\le} d_h\big(\widehat{f}_h,\, Q^\star_h\big) \cdot b_{\mathscr{F}}\, \|f\|_{h'}\,,$$

which in turn implies

$$\nu_1(h-1,\, h') \;\le\; b_{\mathscr{F}} \;\cdot\; d_h\big(\widehat{f}_h,\, Q^\star_h\big)\,. \tag{23a}$$

The proof of inequality $(*)$ for $b_{\mathscr{F}} = 1$, as mentioned above, can be found in Appendix E.1.

As for term ν_2, we claim that

$$\nu_2(h-1) \;\le\; \sum_{j=1}^{h-1} \widehat{\beta}(j,\, h-1)\,. \tag{23b}$$

Combining the bound $\widehat{\beta}(h,\, h') \le \beta(h,\, h') + \Delta\beta(h,\, h')$ with inequalities (22), (23a) and (23b), we establish the bound (20), as claimed. It remains to prove the claim (23b).

Proof of inequality (23b)**:** This proof is analogous to that of Lemma 1. We begin by introducing an analogue of the functional $D(\boldsymbol{g})$ from equation (18a); in particular, for any index $h \in [H-1]$ and function $g \in \partial\mathscr{F}$, define

$$D^\star_h(g) := \big(\mathbb{E}_{\boldsymbol{\pi}^\star} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big)\big[g(S_h, A_h)\big]\,.$$

Using the notation of $D^\star_h$, we can rewrite the left-hand side of inequality (23b) as $\nu_2(h-1) = \sup_{f\in\partial\mathscr{F}:\, \|f\|_{h-1}>0} \big\{ |D^\star_{h-1}(f)| / \|f\|_{h-1} \big\}$. Following the same arguments as in the proof of inequality (18b), we can show that

$$D^\star_h(g) = \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[g(S_h, \pi^\star_h(S_h)) - g(S_h, \widehat{\pi}_h(S_h))\big] + D^\star_{h-1}(\mathcal{P}^\star_{h-1}\, g) \qquad \text{for } h = 1, 2, \ldots, H, \tag{24}$$

where we set $D^\star_0 \equiv 0$.

We consider function $g := \mathcal{P}^\star_{j,h-1}\, f$ for $1 \le j < h \le H-1$. It follows from equation (24) that

$$D^\star_j\big(\mathcal{P}^\star_{j,h-1}\, f\big) = \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\big(\mathcal{P}^\star_{j,h-1} f\big)(S_j, \pi^\star_j(S_j)) - \big(\mathcal{P}^\star_{j,h-1} f\big)(S_j, \widehat{\pi}_j(S_j))\big] + D^\star_{j-1}\big(\mathcal{P}^\star_{j-1,h-1}\, f\big)\,,$$

where we have used the relation $\mathcal{P}^\star_{j-1}\mathcal{P}^\star_{j,h-1} = \mathcal{P}^\star_{j-1,h-1}$. Recalling the definition of $\widehat{\beta}(j,\, h-1)$ in equation (14a), applying the triangle inequality yields

$$\big|D^\star_j\big(\mathcal{P}^\star_{j,h-1}\, f\big)\big| \;\le\; \widehat{\beta}(j,\, h-1) \cdot \|f\|_{h-1} \;+\; \big|D^\star_{j-1}\big(\mathcal{P}^\star_{j-1,h-1}\, f\big)\big|\,.$$

Summing this equation over indices $j = 1, 2, 3, \ldots, h-1$ yields

$$|D^\star_{h-1}(f)| \;\le\; \sum_{j=1}^{h-1} \widehat{\beta}(j,\, h-1) \cdot \|f\|_{h-1}\,,$$

which establishes inequality (23b).

https://doi.org/10.52202/079017-2279

C Proof of corollaries

This section cotains proofs of several corollaries.

C.1 Proof of Corollary 2

We now turn to proving Corollary 2 regarding ridge-based FQI in on-line settings.

In Phase 1 of pure exploration, the cumulative regret is always bounded from above by $T_0 \cdot H$. During Phase 2 of fine-tuning, we let $\widehat{\boldsymbol{\pi}}^k$ be the policy employed in the rounds $T_0 2^k+1$, $T_0 2^k+2,\ldots$, $T_0 2^{k+1}$, which is determined by the estimate $\widehat{\boldsymbol{f}}^{(T_0 2^k)}$ calculated at the end of the $(T_0 2^k)$-th round. To estimate the regret, we consider the decomposition

$$\sum_{t=T_0+1}^{T} \big\{J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^{(t)})\big\} \le \sum_{k=0}^{K-1} \sum_{t=T_0 2^k}^{T_0 2^{k+1}} \big\{J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^{(t)})\big\} = \sum_{k=0}^{K-1} T_0\, 2^k \big\{J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^{k})\big\}.$$

We leverage our bound (11a) for off-line RL in Section 3.1.2 to control the value sub-optimality $J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^k)$. Recall that the policy $\widehat{\boldsymbol{\pi}}^k$ is derived from i.i.d. trajectories collected from the rounds $T_0\, 2^{k-1}+1, T_0\, 2^{k-1}+2, \ldots, T_0\, 2^k$. We divide those $T_0\, 2^{k-1}$ trajectories into $H-1$ equal shares and use each share to conduct estimation in one iteration of the FQI procedure. This subsampling technique ensures the independence of samples used in different iterations. It is primarily adopted for the sake of convenience (to keep the explanations concise) and is not essential in general. It follows from inequality (11a) that the bound

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}^k) \;\le\; c\, \frac{d\sqrt{d}\, H^4}{T_0\, 2^k}\, \log(dHK/\delta)$$

holds uniformly for indices $k = 0, 1, \ldots, K-1$ with a probability exceeding $1-\delta$.

Putting together the pieces, we arrive at

$$\text{Regret}(T) \le T_0 \cdot H + c\, d\sqrt{d}\, H^4 K\, \log(dHK/\delta).$$

We then derive the regret bound in Corollary 2 by noticing that $K = \mathcal{O}(\log T)$.

C.2 Comparing to known off-line bounds

In this section, we derive the sample complexity $n_{\text{Zan}}(\epsilon) \asymp \frac{d^2H^3}{\epsilon^2}$ in Section 3.1.2 based on the results of Zanette et al. [37]; it gives the conventional $1/\sqrt{n}$ slow rate to which we compare. Zanette et al. [37] proved upper bounds on a pessimistic actor-critic scheme based on d-dimensional linear function approximation. Using our notation, Theorem 1 in their paper [37] can be expressed as

$$J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \;\le\; c\left\{\frac{1}{H}\sum_{h=1}^{H-1} \sqrt{\overline{\boldsymbol{\phi}}_h{}^\top (\widehat{\boldsymbol{\Sigma}}_{h,\mathcal{D}} + \lambda_h \boldsymbol{I})^{-1}\, \overline{\boldsymbol{\phi}}_h}\right\} \sqrt{\frac{dH^4}{n}}, \tag{25}$$

where the vector $\overline{\boldsymbol{\phi}}_h$ is given by $\overline{\boldsymbol{\phi}}_h := \mathbb{E}_{\boldsymbol{\pi}^\star}\big[\boldsymbol{\phi}(S_h, A_h)\big]$, the covariance matrix

$$\widehat{\boldsymbol{\Sigma}}_{h,\mathcal{D}} := \frac{1}{|\mathcal{D}_h|} \sum_{(s_{h,i}, a_{h,i}, s'_{h,i}, r_{h,i}) \in \mathcal{D}_h} \boldsymbol{\phi}(s_{h,i},\, a_{h,i}) \boldsymbol{\phi}(s_{h,i},\, a_{h,i})^\top.$$

We now consider the explicit dependence of this upper bound on dimension d, horizon H and sample size n. The divergence term $\overline{\boldsymbol{\phi}}_h{}^\top (\widehat{\boldsymbol{\Sigma}}_{h,\mathcal{D}} + \lambda_h \boldsymbol{I})^{-1}\, \overline{\boldsymbol{\phi}}_h$ measures the conditioning of the regularized covariance matrix $(\widehat{\boldsymbol{\Sigma}}_{h,\mathcal{D}} + \lambda_h \boldsymbol{I})$ along a specific direction of $\overline{\boldsymbol{\phi}}_h$. When the feature mapping $\boldsymbol{\phi}$ operates within a d-dimensional space, it is reasonable to assume that

$$\overline{\boldsymbol{\phi}}_h{}^\top (\widehat{\boldsymbol{\Sigma}}_{h,\mathcal{D}} + \lambda_h \boldsymbol{I})^{-1}\, \overline{\boldsymbol{\phi}}_h \;\le\; c'\, d.$$

The bound (25) then reduces to $J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \;\le\; c\, dH^2/\sqrt{n}$. Regarding the dependence on horizon H, we conjecture that by incorporating the law of total variance in a more refined manner, it may be possible to further reduce the dependence by a factor of $\sqrt{H}$. Under these conditions, the bound takes the form $J(\boldsymbol{\pi}^\star) - J(\widehat{\boldsymbol{\pi}}) \le\; c\, d\sqrt{H^3/n}$.

D Details of the mountain car experiment

In this experiment, a car is situated in a valley between two hills. The car's objective is to overcome the gravitational pull and reach the top of the right hill by efficiently controlling its acceleration.

D.1 Structure of the Markov decision process

The Markov decision process underlying the mountain car problem has a state space $\mathcal{S} \subset \mathbb{R}^2$ and an action space $\mathcal{A} \subset \mathbb{R}$. The state $s = (p, v)$ consists of the current position p and velocity v, whereas the scalar action $a = f$ corresponds to the applied input force. The state variables (p, v) and action f are restricted as

$$\begin{aligned} p &\in [p_{\min}, p_{\max}] = [-1.2, 0.6], \\ v &\in [v_{\min}, v_{\max}] = [-0.07, 0.07] \quad \text{and} \\ f &\in [f_{\min}, f_{\max}] = [-1, 1]\,. \end{aligned}$$

The mountain is described by the function

$$m(p) = \tfrac{1}{3}\sin(3p) + \frac{0.025}{(p_{\max} - p)(p - p_{\min})},$$

over the interval $p \in [p_{\min}, p_{\max}]$.

Let m' be the derivative of the mountain shape function m, which represents the instantaneous slope, and let $(\sigma_v, \sigma_p) = (0.01, 0.0025)$ be a pair of standard deviations that dictate the amount of randomness in the updates. For an interval $[a, b]$, we define the truncation function

$$\Psi_{[a,b]}(u) := \begin{cases} u & \text{if } u \in [a, b], \\ b & \text{if } u > b, \\ a & \text{if } u < a. \end{cases}$$

With this notation, at each discrete time step $h = 0, 1, 2, \ldots$, the position and velocity of the car evolve as

$$\begin{aligned} v_{h+1} &= \Psi_{[v_{\min}, v_{\max}]}\big(\, v_h + 0.0015\, f_h - 0.0025\, m'(p_h) + \sigma_v Z_h\big) \\ p_{h+1} &= \Psi_{[p_{\min}, p_{\max}]}\big(\, p_h + v_{h+1} + \sigma_p Z'_h\big) \end{aligned}$$

where (Z_h, Z'_h) are a pair of independent standard normal variables. Note that the system dynamics are non-linear due to both the presence of the derivative m' and the truncation function Ψ.

The objective of the car is to reach the peak of the mountain, designated by the position $p_{\text{goal}} = 0.45$. The reward at state-action pair (s, a) is given by

$$r(s, a) := -\tfrac{1}{10} f^2 + 100\big[\max\{0,\, p - p_{\text{goal}}\}\big]^2.$$

For any policy π, we define the γ-discounted value function

$$J(\pi) := \mathbb{E}_\pi\Big[\sum_{h=0}^{\infty} \gamma^h\, r(S_h, A_h)\Big],$$

using $\gamma = 0.97$. The initial state $s_0 = (p_0, v_0)$ is generated with p_0 following a uniform distribution over the interval $[-0.6, -0.4]$, and we initialize with velocity $v_0 = 0$.

D.2 Fitted Q-iteration (FQI) with linear function approximation

Here we describe the use of fitted Q-iteration (FQI) with linear function approximation to estimate the optimal Q-function, along with the corresponding greedy policy $\widehat{\pi}$.

Linear function approximation We approximate the the optimal Q-function $(s, a) \mapsto Q^\star(s, a)$ using a d-dimensional linear function class with $d = 3000$ features. We begin by defining the *base*

 https://doi.org/10.52202/079017-2279

feature maps $\boldsymbol{\phi}_p : [\, p_{\min}, p_{\max}\,] \to \mathbb{R}^{50}$ for position, and $\boldsymbol{\phi}_v : [\, v_{\min}, v_{\max}\,] \to \mathbb{R}^{15}$ for velocity, with components given by

$$\begin{cases} \phi_{p,2j+1}(p) := \cos(jp), & \text{for } j = 0, 1, \ldots, 24, \quad \text{and} \\ \phi_{p,2j}(p) := \sin(jp), & \text{for } j = 1, 2, \ldots, 25\,; \end{cases}$$

$$\begin{cases} \phi_{v,2j+1}(v) := \cos(jv), & \text{for } j = 0, 1, \ldots, 7, \quad \text{and} \\ \phi_{v,2j}(v) := \sin(jv), & \text{for } j = 1, 2, \ldots, 7. \end{cases}$$

To represent the action $a \equiv f$, we define the *base action feature map*

$$\boldsymbol{\phi}_f(f) := \big(1, f, f^2, f^3\big) \in \mathbb{R}^4.$$

The overall feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{3000}$ is constructed by taking the outer product of the three base feature maps $\boldsymbol{\phi}_p$, $\boldsymbol{\phi}_v$, and $\boldsymbol{\phi}_f$ as follows:

$$\boldsymbol{\phi}(s, a) := \mathrm{vec}\big\{\boldsymbol{\phi}_p(p) \otimes \boldsymbol{\phi}_v(v) \otimes \boldsymbol{\phi}_f(f)\big\} \in \mathbb{R}^{3000}\,. \tag{26}$$

Taking all possible triples of the three base features in the outer product leads to the overall dimension $d = 3000 = 50 \times 15 \times 4$. Given a weight vector $\boldsymbol{w} \in \mathbb{R}^{3000}$, we define the function $f_{\boldsymbol{w}}(s, a) := \langle \boldsymbol{w},\, \boldsymbol{\phi}(s, a)\rangle$, and we approximate the optimal Q-function using the function class $\mathscr{F} := \big\{f_{\boldsymbol{w}} \mid \boldsymbol{w} \in \mathbb{R}^{3000}\big\}$.

Fitted Q-iteration (FQI) We employed fitted Q-iteration with the linear feature $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{3000}$ to estimate an optimal policy $\widehat{\pi}$. The FQI process begins by initializing the weight vector as $\boldsymbol{w}_0 := \mathbf{0} \in \mathbb{R}^{3000}$. In each iteration, we first use the dataset $\mathcal{D} = \big\{(s_i, a_i, r_i, s_i')\big\}_{i=1}^n \subset \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$ to construct the pseudo-responses

$$y_i := r_i + \gamma \max_{a \in \mathcal{A}} \underbrace{\langle \boldsymbol{w}_t,\, \boldsymbol{\phi}(s_i', a)\rangle}_{f_{\boldsymbol{w}_t}(s_i', a)} \qquad \text{for } i = 1, \ldots, n, \tag{27}$$

corresponding to a stochastic estimate of the Bellman update applied to our current Q-function estimate $f_{\boldsymbol{w}_t}$. The polynomial form of the force feature $\boldsymbol{\phi}_f$ allows for a closed-form solution to the maximum operation required in equation (27). Given these pseudo-responses, we then update the weight vector $\boldsymbol{w}_t \to \boldsymbol{w}_{t+1}$ via the ridge regression

$$\boldsymbol{w}_{t+1} := \arg \min_{\boldsymbol{w} \in \mathbb{R}^{3000}} \Big\{ \frac{1}{n} \sum_{i=1}^n \big\{y_i - \langle \boldsymbol{w},\, \boldsymbol{\phi}(s_i, a_i)\rangle\big\}^2 + \lambda_n \|\boldsymbol{w}\|_2^2 \Big\}, \tag{28}$$

where $\lambda_n = \frac{0.01}{n}$ in all experiments reported here.

We terminate the procedure after at most 500 iterations, or when there have been 5 consecutive iterations with insignificant improvements in weights, where insignificant means that $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2 / \sqrt{3000} < 0.005$. Letting $\widehat{\boldsymbol{w}}$ represent the weight vector obtained from this procedure, the resulting policy $\widehat{\pi}$ is given by selecting the greedy action based on the Q-function estimate $\widehat{f}(s, a) := \langle \widehat{\boldsymbol{w}},\, \boldsymbol{\phi}(s, a)\rangle$.

D.3 Experimental configurations

Our experiments were based on an off-line dataset consisting of n i.i.d. tuples

$$\mathcal{D} = \big\{(s_i, a_i, r_i, s_i')\big\}_{i=1}^n \subset \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S},$$

where the state-action pairs $\big\{(s_i, a_i) = (p_i, v_i, f_i)\big\}_{i=1}^n$ were generated from a uniform distribution over the cube $[p_{\min}, p_{\max}] \times [v_{\min}, v_{\max}] \times [f_{\min}, f_{\max}]$. We performed independent experiments with the sample size n varying over the range

$$\begin{aligned} n &\in \big\{\lfloor e^k \rfloor \;\big|\; k = 10.5, 10.75, 11, \ldots, 13\big\} \\ &= \{36315, 46630, 59874, 76879, 98715, 126753, 162754, 208981, 268337, 344551, 442413\}\,. \end{aligned}$$

In each experiment, we generated a dataset $\mathcal{D}$, estimated an optimal policy $\widehat{\pi}$ based on the data, and evaluated the return $J(\widehat{\pi})$. For each sample size, we conducted 80 independent trials.

In order to evaluate the return $J(\widehat{\pi})$, for each initial position $p_0 = -0.5 + 0.2\,j/1000$ with $j = -500, -499, -498, \ldots, 499$, we simulated 30 independent 1000-step trajectories by executing the estimated policy $\widehat{\pi}$. The average return over the 30×1000 trajectories is used as the estimate of $J(\widehat{\pi})$.

In order to approximate the policy[7] $\pi^\dagger$ that represents "ground truth", we conducted a single experiment with sample size $n = 6.4 \times 10^6$ to obtain $\pi^\dagger$. We simulated 1000 trajectories for each initial position p_0 and calculated the average return, which serves as the reference value $J(\pi^\dagger)$. The value sub-optimality is then computed as the difference $J(\pi^\dagger) - J(\widehat{\pi})$.

E Verification of auxiliary claims

In this appendix, we collect the verification of various auxiliary claims made in the main text.

E.1 Properties of occupation measures

In this appendix, we prove a useful inequality

$$\big\|\mathcal{P}^\star_{h,h'} f\big\|_h \le \|f\|_{h'}\,, \tag{29}$$

which holds for the state-action occupation measures (3). This bound is used in the proof of bound (20) in Appendix B.2.

By definition, we have

$$\|\mathcal{P}^\star_h f\|_h^2 = \mathbb{E}_{\boldsymbol{\pi}^\star}\big[(\mathcal{P}^\star_h f)^2(S_h, A_h)\big] = \mathbb{E}_{\boldsymbol{\pi}^\star}\Big[\mathbb{E}_h\big[f(S_{h+1}, \pi^\star_{h+1}(S_{h+1})) \;\big|\; S_h, A_h\big]^2\Big]\,.$$

According to the property of variance, we can deduce

$$\mathbb{E}_{\boldsymbol{\pi}^\star}\Big[\mathbb{E}_h\big[f\big(S_{h+1}, \pi^\star_{h+1}(S_{h+1})\big) \;\big|\; S_h, A_h\big]^2\Big] \le \mathbb{E}_{\boldsymbol{\pi}^\star}\Big[f^2\big(S_{h+1}, \pi^\star_{h+1}(S_{h+1})\big)\Big] = \|f\|^2_{h+1}\,.$$

As a consequence, we find that $\|\mathcal{P}^\star_h f\|_h \le \|f\|_{h+1}$. Applying this inequality recursively leads to the conclusion that for any indices $1 \le h \le h' \le H$, we have

$$\big\|\mathcal{P}^\star_{h,h'} f\big\|_h = \big\|\mathcal{P}^\star_h \mathcal{P}^\star_{h+1,h'} f\big\|_h \le \big\|\mathcal{P}^\star_{h+1,h'} f\big\|_{h+1} \le \big\|\mathcal{P}^\star_{h+2,h'} f\big\|_{h+2} \le \cdots \le \|f\|_{h'}\,.$$

This establishes the bound (29).

E.2 Proof of the telescope inequality (6)

For completeness of this paper,[8] let us prove the telescope relation (6) stated in Section 2.3. For any policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_H)$ and sequence of functions $\boldsymbol{f} = (f_1, \ldots, f_H)$ with $f_H = r_H$, we have the "telescope" relation

$$V_1^{\boldsymbol{\pi}}(s) = f_1(s, \pi_1(s)) + \sum_{h=1}^{H-1} \mathbb{E}_{\boldsymbol{\pi}}\big[\big(\mathcal{T}_h^{\boldsymbol{\pi}} f_{h+1} - f_h\big)(S_h, A_h) \;\big|\; S_1 = s\big] \quad \text{for any state } s \in \mathcal{S}. \tag{30}$$

Here the value function $V_1^{\boldsymbol{\pi}}$ is given by $V_1^{\boldsymbol{\pi}}(s) := Q_1^{\boldsymbol{\pi}}(s, \pi_1(s))$. Taking $\boldsymbol{f} = \widehat{\boldsymbol{f}}$ in equation (30) yields

$$V_1^{\boldsymbol{\pi}}(s) = \widehat{f}_1(s, \pi_1(s)) + \sum_{h=1}^{H-1} \mathbb{E}_{\boldsymbol{\pi}}\big[\big(\mathcal{T}_h^{\boldsymbol{\pi}} \widehat{f}_{h+1} - \widehat{f}_h\big)(S_h, A_h) \;\big|\; S_1 = s\big]\,. \tag{31a}$$

Letting $\boldsymbol{\pi} = \widehat{\boldsymbol{\pi}}$ in equation (31a) yields

$$V_1^{\widehat{\boldsymbol{\pi}}}(s) = \widehat{f}_1(s, \widehat{\pi}_1(s)) + \sum_{h=1}^{H-1} \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\big[\big(\mathcal{T}_h^{\widehat{\boldsymbol{\pi}}} \widehat{f}_{h+1} - \widehat{f}_h\big)(S_h, A_h) \;\big|\; S_1 = s\big]\,. \tag{31b}$$

[7] In general, it is not guaranteed that $\pi^\dagger$ is equal to the optimal policy $\pi^\star$, due to approximation error that might arise from using the linear function class defined here.

[8] We are not claiming novelty here; see Theorem 2 of the paper [34]; or Lemma 3.2 in the paper [7] for analogous results.

https://doi.org/10.52202/079017-2279

Since $\widehat{\boldsymbol{\pi}}$ is a greedy policy with respect to function $\widehat{\boldsymbol{f}}$, we have

$$\widehat{f}_1(s, \widehat{\pi}_1(s)) \geq \widehat{f}_1(s, \pi_1(s)), \quad \text{and} \quad \mathcal{T}_h^{\widehat{\boldsymbol{\pi}}} \widehat{f}_{h+1} = \mathcal{T}_h^{\star} \widehat{f}_{h+1} \geq \mathcal{T}_h^{\boldsymbol{\pi}} \widehat{f}_{h+1} \quad \text{for any policy } \boldsymbol{\pi}.$$

Using this fact and subtracting equations (31a) and (31b), we obtain

$$V_1^{\boldsymbol{\pi}}(s) - V_1^{\widehat{\boldsymbol{\pi}}}(s) \leq \sum_{h=1}^{H-1} \left(\mathbb{E}_{\boldsymbol{\pi}} - \mathbb{E}_{\widehat{\boldsymbol{\pi}}}\right) \left[\left(\mathcal{T}_h^{\star} \widehat{f}_{h+1} - \widehat{f}_h\right)(S_h, A_h) \;\middle|\; S_1 = s \right].$$

Finally, taking the expectation over the initial distribution ξ_1 yields the claimed inequality (6).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper thoroughly discusses each point made in the abstract, including fast-rate convergence of RL in continuous state-action spaces, key stability properties, and the pessimism and optimism principles.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses the work's limitations and future directions, such as adapting the framework to linear quadratic control, extending beyond i.i.d. data and finite-dimensional linear function spaces, proving a lower bound to demonstrate sharpness, and examining model mis-specification.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

https://doi.org/10.52202/079017-2279

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results in this paper are mathematically rigorous, with intuitions in the main body and detailed justifications and proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix D explains how to reproduce the synthetic Mountain Car experiment from Section 1.1, including MDP structure, linear function space construction, FQI implementation, and experimental setups. No data is required.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The submission provides a supplementary code document to reproduce the Mountain Car experiment.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D explains how to reproduce the synthetic Mountain Car experiment from Section 1.1, including MDP structure, linear function space construction, FQI implementation, and experimental setups.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 1(b) in the paper includes error bars for the standard errors, and the estimated slope in Figure 1(b) uses a bootstrap confidence interval to confirm observation validity.

   Guidelines:

https://doi.org/10.52202/079017-2279

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 1.1 details the computer setup: “The experiment ran for 3 days on two laptops, each equipped with an Apple M2 Pro CPU and 16 GB RAM.”

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn’t make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research in the paper fully conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: The paper focuses on purely theoretical aspects of reinforcement learning and does not discuss direct societal impacts, given its theoretical nature.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

https://doi.org/10.52202/079017-2279

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2279