Instance-Dependent Confidence and Early Stopping for Reinforcement Learning

Eric Xia   ERICZXIA@MIT.EDU
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Koulik Khamaru   KK1241@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854 USA

Martin J. Wainwright   MJWAIN@MIT.EDU
Department of EECS and Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Michael I. Jordan   JORDAN@CS.BERKELEY.EDU
Department of EECS and Statistics
University of California, Berkeley
Berkeley, CA 94720 USA

Abstract

Reinforcement learning algorithms are known to exhibit a variety of convergence rates depending on the problem structure. Recent years have witnessed considerable progress in developing theory that is *instance-dependent*, along with algorithms that achieve such instance-optimal guarantees. However, important questions remain in how to utilize such notions for inferential purposes, or for early stopping, so that data and computational resources can be saved for "easy" problems. This paper develops data-dependent procedures that output instance-dependent confidence regions for evaluating and optimizing policies in a Markov decision process. Notably, our procedures require only black-box access to an instance-optimal algorithm, and re-use the samples used in the estimation algorithm itself. The resulting data-dependent stopping rule adapts instance-specific difficulty of the problem and allows for early termination for problems with favorable structure. We highlight benefit of such early stopping rules via some numerical studies.

Keywords: Reinforcement learning, policy evaluation, confidence intervals, instance dependence, instance optimality

1. Introduction

Reinforcement learning (RL) refers to a broad class of methods that are focused on learning how to make (near)-optimal decisions in dynamic environments. Although RL-based methods are now being deployed in various application domains (e.g., Tobin et al. (2017); Levine et al. (2016); Silver et al. (2016); Łukasz Kaiser et al. (2020); Schrittwieser et al. (2020)),

such deployments often lack a secure theoretical foundation. Given that RL involves making real-world decisions, often autonomously, the impact on humans can be significant, and there is an urgent need to shore up the foundations, providing practical and actionable guidelines for RL. A major part of the challenge is that popular RL algorithms exhibit a variety of behavior across domains and problem instances and existing methods and associated guarantees, generally tailored to the worst-case setting, fail to capture this variety. One way to move beyond worst-case bounds is to develop guarantees that adapt to the problem difficulty, helping to reveal what aspects of an RL problem make it an "easy" or "hard" problem. Indeed, in recent years, such a research agenda has begun to emerge and we have gained a refined understanding of the instance-dependent nature of various reinforcement learning problems (e.g., Simchowitz and Jamieson, 2019; Zanette and Brunskill, 2019; Zanette et al., 2019; Maillard et al., 2014; Khamaru et al., 2021; Pananjady and Wainwright, 2021).

The broader challenge—and the focus of this paper—is to recognize that RL involves decision-making under uncertainty, and to develop an inferential theory for RL problems. Such theory must not only be *instance-dependent*, but also *data-dependent*, meaning that quantities such as confidence intervals should be computable using available data. The latter property is not shared by most past instance-dependent guarantees in RL; with limited exceptions—e.g., Theorem 1(a) in Pananjady and Wainwright (2021) or the results on adaptive sampling leveraging instance-dependent structure in Zanette et al. (2019)—most results from past work depend on population-level objects—such as probability transition matrices, Bellman variances, or reward function bounds—that are not known to the user. Thus instance-dependent guarantees remain a *post hoc* justification of why certain problem are "easier" to solve than others, and somewhat inconsequential until after training.

Our work aims to close this gap between theory and practice by providing theoretical guarantees that are both instance-dependent and data-dependent. The resulting bounds are both sharp up to logarithmic factors and computable based on data. To illustrate the significance of our confidence intervals, we design an early-stopping procedure based on these intervals which lead to substantial reductions in the amount of data required for a target accuracy, in an effort to demonstrate how notions of instance-dependence can be used to aid the training process. In more detail, we make these contributions in the context of Markov decision processes (MDPs) with a finite number of states and actions, and problem-dependent confidence regions both for policy evaluation and optimal value function estimation. Contrary to prior work on instance-dependent analysis, our work allows a user to adapt their data requirements for the specific MDP at hand by exploiting the local difficulty of the MDP. As we show, doing so can lead to significant reductions in the sample sizes required for effective learning.

1.1 Related work

The problem of estimating the value function for a given policy in a Markov decision process (MDP) is a key subroutine in many modern-day RL algorithms. Examples include policy iteration (Howard, 1960), policy gradient, and actor-critic methods (Williams, 1992; Konda and Tsitsiklis, 2001; Silver et al., 2014; Mnih et al., 2016). Such use cases have provided the impetus for the recent interest in analyzing policy evaluation. Much of the focus in the past

has been on understanding TD-type algorithms with instance-dependent analyses: function approximation in the ℓ_2-norm (Bhandari et al., 2018; Dalal et al., 2018; Xu et al., 2020), tabular setting in the ℓ_∞-norm (Xia et al., 2021; Pananjady and Wainwright, 2021), or under kernel function approximation (Duan et al., 2021). Many of these results established instance-specific guarantees that improve upon global worst-case bounds (Azar et al., 2013). In particular, Khamaru et al. (2021) establish a local minimax lower-bound in the tabular setting, and develop a form of variance-reduced stochastic approximation that achieves it.

Policy optimization involves solving for an optimal policy within a given MDP. There exists a variety of different techniques for solving policy optimization; the one of interest here is Q-learning, introduced in Watkins and Dayan (1992). There has been much prior work on the theory of Q-learning, such as convergence of the standard updates (Tsitsiklis, 1994; Jaakkola et al., 1994; Even-Dar and Mansour, 2003; Wainwright, 2019a; Li et al., 2021), global minimax lower bounds for estimation of optimal Q-functions (Azar et al., 2013), variance-reduced versions of Q-learning and their worst-case optimality (Sidford et al., 2018b,a; Wainwright, 2019b), along with extensions to the asynchronous setting (Li et al., 2020b). Xia et al. (2021) establish a local non-asymptotic minimax lower bound for estimating the Q-function, and prove that it can be achieved by a variance-reduced form of Q-learning.

1.2 Contributions

The main contributions of this paper are to provide guarantees for policy evaluation and optimization that are both instance-optimal and data-dependent. Our first main result, stated as Theorem 1, applies to a meta-procedure that we propose. This meta-procedure takes as input any base procedure for policy evaluation that is instance-optimal up to constant factors; invokes this base procedure using the bulk of the dataset, and then uses the resulting output along with the full dataset to compute bounds on the estimated value function. We prove that these bounds—which are data-dependent by construction—are also instance-optimal up to constant factors. Thus, they can be inverted so as to produce a confidence region for the value function, and up to constant factors, the width of this confidence region is as small as possible for a pre-specified coverage. Next, based on the guarantees from Theorem 1, we introduce an early stopping protocol for policy evaluation, known as the EmpIRE procedure, that is guaranteed to output instance-optimal confidence regions up to constant factors upon stopping. As we show both theoretically and in simulation, use of this early stopping procedure can lead to significant reductions in sample sizes compared to the traditional worst-case criteria.

Figure 1 gives a preview of the results to come, including the behavior of this early stopping procedure (panel (a)), along with the attendant benefits of substantially reduced sample sizes (panel (b)). The EmpIRE method is an epoch-based protocol: within each epoch, it uses all currently available samples to estimate the value function along with its associated ℓ_∞-error, and it terminates when the error estimate drops below a pre-specified target. Our guarantees ensure that with high probability over all epochs, the error estimate is an upper bound on the true error, so that the final output of the procedure has guaranteed accuracy with the same probability. Figure 1(a) illustrates the behavior of the EmpIRE procedure over a simulated run. The predicted error incorporates the instance-dependent
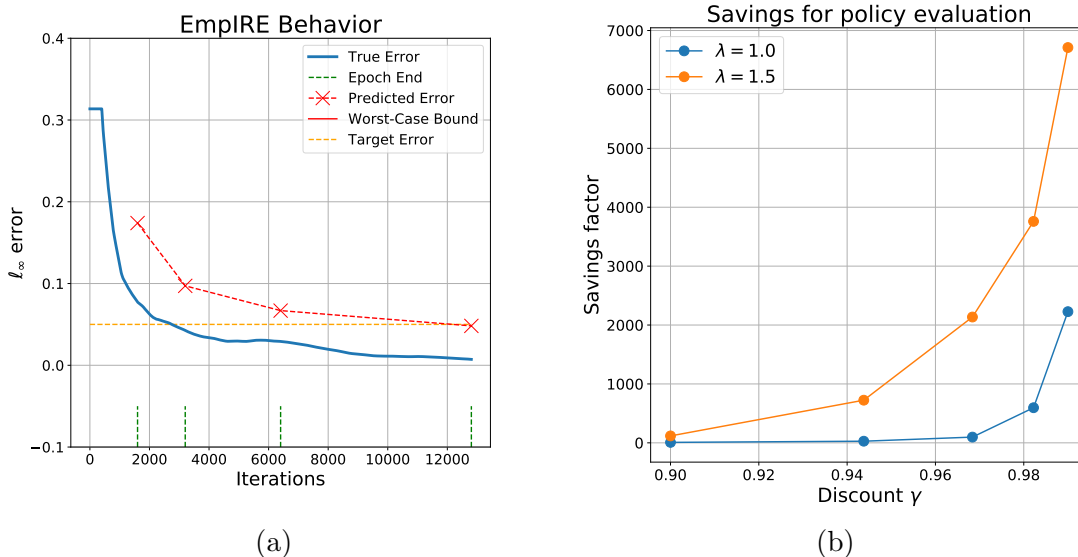
Figure 1. (a) Illustration of the behavior of the EmpIRE early stopping protocol when combined with an instance-optimal procedure for estimating value functions. The true error (blue line) of the value function estimate is plotted as a function of the number of samples used (or equivalently, iterations of the algorithm). The EmpIRE protocol checks the error at the end of a dyadically increasing sequence of epochs, as marked with vertical green lines; the associated error estimates are marked with red X's. For this particular run, the EmpIRE protocol terminates at the end of the 4th epoch when the predicted error is smaller than the target error (orange line). (b) Illustration of the savings in sample size afforded by the EmpIRE protocol for different choices of γ and λ. Plots show the ratio of worst-case sample size over the actual sample size (vertical axis) versus the log discount complexity (horizontal axis).

structure of the MDP at hand and allows us to terminate the procedure earlier for "easy" problems. Figure 1(b) highlights how difficulty can vary dramatically across different instances. We do so by constructing a class of MDPs for which the difficulty can be controlled by a parameter λ, with larger λ indicating an easier problem. For a given accuracy ϵ, let $n(\epsilon)$ be the number of samples required to achieve an estimate with this accuracy. We can bound this sample complexity using either global minimax theory (based on worst-case assumptions), and compare it to the instance-dependent results of the EmpIRE procedure. Figure 1(b) plots the ratio of the worst-case prediction (from global minimax) to the number of samples used by EmpIRE as a function of the discount parameter γ, for two different choices of the hardness parameter λ. We see that EmpIRE can yield dramatic reductions in sample complexity compared to a worst-case guarantee—on the order of 10^3 for larger discounts.

We also derive similar guarantees in the more challenging setting of policy optimization, where we again analyze a meta-procedure that takes as input any algorithm that returns an instance-optimal estimate of the optimal Q-value function. Theorem 3 gives the resulting data-dependent and instance-dependent guarantees enjoyed by this procedure, and inverting these bounds again leads confidence regions for the optimal value function. As before, these

guarantees can be combined with the EmpIRE protocol so as to perform early stopping while still retaining theoretical guarantees for policy optimization.

The remainder of this paper is organized as follows. Section 3 is devoted to results on the policy evaluation problem, and Section 4 discusses results related to policy optimization via optimal value function estimation. In Section 5, we provide the proofs of our main results, with some more technical results deferred to the appendices. We conclude with a discussion in Section 6.

1.3 Notation

For a positive integer n, let $[n] \coloneqq \{1, 2, \ldots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. We let $\mathbf{1}$ denote the all-ones vector in $\mathbb{R}^d$. Let e_j denote the jth standard basis vector in $\mathbb{R}^d$. For a vector $u \in \mathbb{R}^d$, we use $|u|$ to denote the entry-wise absolute value of a vector $u \in \mathbb{R}^d$; square-roots of vectors are, analogously, taken entry-wise. Given two vectors u, v of matching dimensions, we use $u \geq v$ to indicate that the difference vector $u - v$ is entry-wise non-negative; we define $u \leq v$ analogously. For a given matrix $A \in \mathbb{R}^{d \times d}$ we define the diagonal "norm" $\|A\|_{\mathrm{diag}} = \max_{i=1,\ldots,d} \left| e_i^T A e_i \right|$, i.e., the maximum diagonal entry in absolute terms. We use $\lesssim$ and $\mathcal{O}(\cdot)$ to denote relations that hold up to constant and logarithmic factors.

2. Background

In this section, we provide some background on tabular Markov decision processes (MDPs), policy evaluation, and optimal value estimation problems.

2.1 Markov decision processes

We start with a brief introduction to Markov decision processes (MDPs) with finite state $\mathcal{X}$ and action $\mathcal{U}$ spaces; see Puterman (2014); Bertsekas (2009); Sutton and Barto (2018) for an in-depth discussion. In a Markov decision process, the state x evolves dynamically in time under the influence of the actions. Concretely, there is a collection of probability transition kernels, $\{\boldsymbol{\mathcal{P}}_u(\cdot \mid x) \mid (x, u) \in \mathcal{X} \times \mathcal{U}\}$, where $\boldsymbol{\mathcal{P}}_u(x' \mid x)$ denotes the probability of a transition to the state x' when the action u is taken at the current state x. In addition, an MDP is equipped with a reward function $\mathbf{r}$ that maps every state-action pair (x, u) to a real number $\mathbf{r}(x, u)$. The reward $\mathbf{r}(x, u)$ is the reward received upon performing the action u in the state x. Overall, a given MDP is characterized by the pair $(\boldsymbol{\mathcal{P}}, \mathbf{r})$, along with a discount factor $\gamma \in (0, 1)$.

A deterministic policy π is a mapping $\mathcal{X} \to \mathcal{U}$: the quantity $\pi(x) \in \mathcal{U}$ indicates the action to be taken in the state x. The value of a policy is defined by the expected sum of discounted rewards in an infinite sample path. More precisely, for a given policy π and discount factor $\gamma \in (0, 1)$, the value function for policy π is given by

$$V^\pi(x) \coloneqq \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k \cdot \mathbf{r}(x_k, u_k) \mid x_0 = x \Big], \qquad \text{where } u_k = \pi(x_k) \text{ for all } k \geq 0. \tag{1}$$

A closely related object is the action-value or Q-function associated with the policy, which is given by

$$Q^\pi(x, u) \coloneqq \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k \cdot \mathbf{r}(x_k, u_k) \mid (x_0, u_0) = (x, u)\Big], \qquad \text{where } u_k = \pi(x_k) \text{ for all } k \geqslant 1. \tag{2}$$

Two core problems in reinforcement learning—and those that we analyze in this paper—are policy evaluation and policy optimization. In the problem of *policy evaluation*, we are given a fixed policy π, and our goal is to estimate its value function on the basis of samples. In the *policy optimization*, our goal is to estimate the optimal policy, along with the associated optimal Q-value function

$$Q^\star(x, u) \coloneqq \max_{\pi \in \mathbf{\Pi}} Q^\pi(x, u) \quad \text{for all } (x, u) \in \mathcal{X} \times \mathcal{U}, \tag{3}$$

again on the basis of samples. As we now describe, both of these problems have equivalent formulations as computing the fixed points of certain types of Bellman operators.

We conclude with some high-level remarks about our setting. The generative tabular MRP model is perhaps the most restrictive, and the "easiest" RL setting to study; however it is also one of the settings in which we have the most refined understanding of instance-dependence, with respect to sharp upper and lower bounds (Khamaru et al., 2021; Xia et al., 2021). It would be of considerable interest to study how one can utilize instance-dependence in other, more complicated, settings such as with function approximation or different sampling schemes.

2.2 Policy evaluation and Markov reward processes

We begin by formalizing the problem of policy evaluation. For a given MDP, if we fix some deterministic policy π, then the MDP reduces to a Markov reward process (MRP) over the state space $\mathcal{X}$. More precisely, the state evolution over time is determined by the set of transition functions $\{\boldsymbol{\mathcal{P}}_{\pi(x)}(\cdot \mid x), x \in \mathcal{X}\}$, whereas the reward received when at state x is given by $\mathbf{r}(x, \pi(x))$.

When the number of states is finite with $|\mathcal{X}|$, the transition functions $\{\boldsymbol{\mathcal{P}}_{\pi(x)}(\cdot \mid x), x \in \mathcal{X}\}$ and the rewards $\{\mathbf{r}(x, \pi(x)) \mid x \in \mathcal{X}\}$ can be conveniently represented as a $|\mathcal{X}| \times |\mathcal{X}|$ matrix and a $|\mathcal{X}|$ dimensional vector, respectively. For ease of notation, we use $\mathbf{P}$ to denote this $|\mathcal{X}| \times |\mathcal{X}|$ matrix, and r to denote this $|\mathcal{X}|$ dimensional vector. Concretely, for any state $x \in \mathcal{X}$, we define

$$r(x) \coloneqq \mathbf{r}(x, \pi(x)) \qquad \text{and} \qquad \mathbf{P}(x', x) \coloneqq \boldsymbol{\mathcal{P}}_{\pi(x)}(x' \mid x), \tag{4}$$

where $\mathbf{P}(x', x)$ denotes the row corresponding to x and the column corresponding to x'. We will often use $\mathbf{P}(x' \mid x)$ to denote $\mathbf{P}(x', x)$. With this formulation at hand, it is clear that evaluating the value of the policy π for the MDP $\mathcal{P}$ is the same as finding the value of the MRP $\mathcal{M} = (r, \mathbf{P}, \gamma)$ with reward r and transition $\mathbf{P}$ defined in equation (4).

Bellman evaluation operator: Given an MRP $\mathcal{M} = (r, \mathbf{P}, \gamma)$, its value $V^\star$ can be obtained as the unique fixed-point of the Bellman evaluation operator $\boldsymbol{\mathcal{T}}$. It acts on the set of value functions in the following way:

$$\boldsymbol{\mathcal{T}}(V)(x) = r(x) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}(x' \mid x) V(x') \quad \text{for all } x \in \mathcal{X}.$$

For finite-dimensional MDPs, the value and reward functions can be viewed as $|\mathcal{X}|$-dimensional vectors, and the transition function as a $|\mathcal{X}| \times |\mathcal{X}|$ dimensional matrix. As a result, the action of the Bellman operator $\boldsymbol{\mathcal{T}}$ can be written in the following compact form

$$\boldsymbol{\mathcal{T}}(V) = r + \gamma \mathbf{P} V, \tag{5}$$

and the associated fixed point relation is given by $V^\star = \boldsymbol{\mathcal{T}}(V^\star)$. See Puterman (2014); Sutton and Barto (2018); Bertsekas (2009) for more details.

Generative observation model for MRPs: In the learning setting, the pair $(\mathbf{P}, r)$ is unknown and we assume that we have access to *i.i.d. samples* $\{(R_k, \mathbf{Z}_k)\}_{k=1}^n$ from the reward vector r and from the transition matrix $\mathbf{P}$. Concretely, given a sample index $k \in \{1, 2, \ldots, n\}$, we have for all state $x \in \mathcal{X}$

$$\mathbf{Z}_k(\cdot, x) \sim \mathbf{P}(\cdot \mid x) \quad \text{and} \quad \mathbb{E}[R_k(x)] = r(x). \tag{6}$$

We also assume that the deviation of the reward sample $R_k(x)$ from the true reward $r(x)$ is bounded:

$$|R(x) - r(x)| \leqslant R_{\max} \quad \text{for all } x \in \mathcal{X}. \tag{7}$$

In other words, for every sample k, we observe for every state x a reward $R_k(x)$ and the next state $x' \sim \mathbf{P}(\cdot \mid x)$. Then we have $\mathbf{Z}_k(x', x) = 1$, and the remaining entries in the row corresponding to x are 0. Sometimes we will denote $\mathbf{Z}_k(x', x)$ as $\mathbf{Z}_k(x' \mid x)$.

2.3 Policy optimization via optimal Q-function estimation

Recall that the goal of policy optimization is to find an optimal policy along with the optimal Q-value function (cf. equation (3)). For MDPs with finite state space $\mathcal{X}$ and action space $\mathcal{U}$, any Q-value function can be represented as an element of $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|}$.

Moreover, the optimal Q-function $Q^\star$ is the *unique fixed point* of the Bellman (optimality) operator $\boldsymbol{\mathcal{J}}$, an operator on $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|}$ given via

$$\boldsymbol{\mathcal{J}}(Q)(x, u) := r(x, u) + \gamma \sum_{x' \in x} \mathbf{P}_u(x' \mid u) \max_{u' \in \mathcal{U}} Q(x', u'). \tag{8}$$

Given $Q^\star$, an optimal policy $\pi^\star$ is given by $\pi^\star(x) \in \arg\max_{u \in \mathcal{U}} Q^\star(x, u)$.

Generative observation model for MRPs: We operate in the *generative* observation model: we are given n i.i.d. samples of the form $\{(\mathcal{Z}_k, \mathcal{R}_k)\}_{k=1}^n$, where $\mathcal{R}_k$ is a matrix in $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|}$ and $\mathcal{Z}_k$ is a collection of $|\mathcal{U}|$ matrices in $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ indexed by $\mathcal{U}$. We denote by $\mathcal{Z}_k(x, u)$ the row-vector corresponding to state x and action u. The row-vector is computed via sampling from the transition kernel $\boldsymbol{\mathcal{P}}_u(\cdot \mid x)$, independently of all other (x, u), and

making the entry corresponding to the next state equal to one, and the remaining entries equal to zero. Concretely, for every state action pair $(x, u) \in \mathcal{X} \times \mathcal{U}$, we write

$$x' \sim \boldsymbol{\mathcal{P}}_u(\cdot \mid x) \qquad \text{and} \qquad \mathcal{Z}_u(\cdot \mid x) = \mathbf{1}_{x=x'}. \qquad (9)$$

We write $\mathcal{Z}_1 + \mathcal{Z}_2$ to indicate the collection of $|\mathcal{U}|$ matrices that are the sum of the matrices from $\mathcal{Z}_1$ and $\mathcal{Z}_2$ that match in the action. We also assume that $\mathcal{R}_k(x, u)$ is a bounded random variable with mean $\mathbf{r}(x, u)$ and that $|\mathcal{R}_k(x, u) - \mathbf{r}(x, u)| \leqslant \mathcal{R}_{\max}$. Additionally, $\mathcal{R}_k(x, u)$ is taken to be independent of all other state-action pairs and of the observations $\mathcal{Z}_k$.

3. Confidence Intervals for Policy Evaluation

In this section, we propose instance-dependent confidence regions for the policy evaluation problem. We start with a discussion of the problem-dependent functional that determines the fundamental difficulty of policy evaluation. As already detailed in Section 2.2, any MDP policy pair $(\mathcal{P}, \pi)$ naturally gives rise to an MRP, and the policy evaluation problem is equivalent to finding the value of that MRP. Accordingly, in the rest of the section, we discuss the problem of finding the value of a general MRP $\mathcal{M}$.

3.1 Optimal instance dependence for policy evaluation

Given access to a generative sample $(R, \mathbf{Z})$ from an MRP $\mathcal{M} = (r, \mathbf{P}, \gamma)$, define the single-sample empirical Bellman operator as

$$\widehat{\boldsymbol{\mathcal{T}}}(V)(x) := R(x) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{Z}(x' \mid x) V(x') \qquad \text{for all } x \in \mathcal{X}. \qquad (10)$$

By construction, for any fixed value function V, the quantity $\widehat{\boldsymbol{\mathcal{T}}}(V)$ is an unbiased estimate of $\boldsymbol{\mathcal{T}}(V)$. In the finite-dimensional setting, $\widehat{\boldsymbol{\mathcal{T}}}(V)$ can be considered a random vector with entries given by $\widehat{\boldsymbol{\mathcal{T}}}(V)(x)$, and we can talk about its covariance matrix. In a recent paper (Khamaru et al., 2021), the authors show that the quantity that determines the difficulty of estimating the value $V^\star$, given access to i.i.d. generative samples following the sampling mechanism (6), is the following $|\mathcal{X}| \times |\mathcal{X}|$ covariance matrix

$$\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star) := (\mathbf{I} - \gamma \mathbf{P})^{-1} \boldsymbol{\Sigma}_{\mathtt{val}}(\mathcal{M}, V^\star)(\mathbf{I} - \gamma \mathbf{P})^{-\top}. \qquad (11)$$

Intuitively, the covariance matrix $\boldsymbol{\Sigma}_{\mathtt{val}}(\mathcal{M}, V^\star) = \mathrm{Cov}(\widehat{\boldsymbol{\mathcal{T}}}(V^\star))$ captures the noise of the empirical Bellman operator, evaluated at the true value $V^\star$. This term is then compounded by powers of the discounted transition matrix $\gamma \mathbf{P}$, which captures how the perturbation in the Bellman operator propagates over time, and thus gives rise to the matrix $(\mathbf{I} - \gamma \mathbf{P})^{-1}$ via the Neumann series.

Xia et al. (2021), show that the functional $\|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}^{\frac{1}{2}}$ arises in lower bounds—in both asymptotic and non-asymptotic settings—on the error of any procedure for estimating the value function. In particular, any estimate $\tilde{V}_n$ of the value function must necessarily satisfy a lower bound of the form $\|\tilde{V}_n - V^\star\|_\infty \gtrsim \frac{1}{\sqrt{n}} \|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}^{\frac{1}{2}}$. Moreover Khamaru et al. (2021) provide a practical scheme that achieves this lower bound, modulo logarithmic

factors. In particular, in Theorem 1 and Proposition 1 of Khamaru et al. (2021) show that a variance-reduced version of policy evaluation, using a total of n samples, returns an estimate $\widehat{V}_n$ such that

$$\|\widehat{V}_n - V^\star\|_\infty \lesssim \frac{\sqrt{\log(|\mathcal{X}|/\delta)}}{\sqrt{n}}\|\mathbf{\Sigma}^\star(r,\mathbf{P},V^\star)\|_{\mathrm{diag}}^{1/2} + \mathcal{O}(\frac{1}{n}) \tag{12}$$

Notably, this upper bound can be much smaller than a worst case upper bound (see Section 3.2 in Khamaru et al. (2021)). In particular, the variance-reduced PE scheme provides far more accurate estimates for problems that are "easier," as measured by the size of the functional $\|\mathbf{\Sigma}^\star(r,\mathbf{P},V^\star)\|_{\mathrm{diag}}^{1/2}$.

3.2 A procedure for computing data-dependent bounds

Guarantees of the type (12)—showing that a certain procedure for policy evaluation is instance-optimal—are theoretically interesting, but are not practically useful. In particular, the bound (12) depends on the unknown population-level quantity $\|\mathbf{\Sigma}^\star(r,\mathbf{P},V^\star)\|_{\mathrm{diag}}^{1/2}$, so that for any given instance, we cannot actually evaluate the error bound achieved for a given sample size. Conversely, when data is being collected in a sequential fashion, as is often the case in practice, we cannot use this theory to decide when to stop.

The main result of this section rectifies this gap between theory and practice. We introduce a data-dependent upper bound on the error in policy evaluation (PE). In addition to the variance-reduced PE scheme described in Xia et al. (2021), there are other procedures that could also be used to achieve an instance-optimal bound, such as the ROOT-OP procedure applied to the Bellman operator (Li et al., 2020a; Mou et al., 2022). Rather than focus on the details of a specific PE method, our theory applies to a generic class of PE procedures, as we now define.

3.2.1 (φ_f, φ_s)-instance-optimal algorithms

Suppose we have access to some procedure $\mathcal{A}_{\mathtt{Eval}}$ that estimates the value function $V^\star$ of an MRP $\mathcal{M}$. More precisely, let $\widehat{V}_n$ denote the estimate obtained from the algorithm $\mathcal{A}_{\mathtt{Eval}}$ using n generative samples $\{R_k, \mathbf{Z}_k\}_{k=1}^n$. Our analysis applies to procedures that are instance-optimal in the following sense.

Given a failure probability $\delta \in (0,1)$, let φ_f and φ_s be non-negative functions of δ; depending on the procedure under consideration, these functions may also involve other known parameters (e.g. number of states, the discount factor γ, initialization of the algorithm $\mathcal{A}_{\mathtt{Eval}}$ etc.), but we omit such dependence so as to keep our notation streamlined. For any such pair, we say that the procedure $\mathcal{A}_{\mathtt{Eval}}$ is (φ_f, φ_s)-*instance optimal* if the estimate $\widehat{V}_n$ satisfies the ℓ_∞-bound

$$\|\widehat{V}_n - V^\star\|_\infty \leqslant \frac{\varphi_f(\delta)}{\sqrt{n}} \cdot \|\mathbf{\Sigma}^\star(r,\mathbf{P},V^\star)\|_{\mathrm{diag}}^{1/2} + \frac{\varphi_s(\delta)}{n}. \tag{13}$$

with probability at least $1 - \delta$. Thus, an (φ_f, φ_s)-instance-optimal procedure returns an estimate that, modulo any inflation of the error due to $\varphi_f(\delta)$ being larger than one, achieves the optimal instance-dependent bound. Note that the function φ_s is associated with the

higher order term (decaying as n^{-1}), so becomes negligible for larger sample sizes. As a concrete example, as shown by the bound (12), there is a variance-reduced version of policy evaluation that is instance-optimal in this sense with $\varphi_f(\delta) \asymp \sqrt{\log(|\mathcal{X}|/\delta)}$. Our results are based on the natural assumption that the functions $\varphi_f(\cdot)$ and $\varphi_s(\cdot)$ are non-negative, non-decreasing in δ, and $\varphi_f(\cdot)$ is uniformly lower bounded by 1.

In practice, even if an estimator enjoys the attractive guarantee (13), it cannot be exploited in practice since the right-hand side depends on the *unknown* problem parameters $(r, \mathbf{P}, V^\star)$. The main result of the following section is to introduce a data-dependent quantity that (a) also provides a high probability upper bound on $\|\widehat{V}_n - V^\star\|_\infty$; and (b) matches the guarantee (13) apart from changes to the leading pre-factor and the higher-order terms.

3.2.2 Constructing a data-dependent bound

In order to obtain a data-dependent bound on $\|\widehat{V}_n - V^\star\|_\infty$, it suffices to estimate the quantity $\|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}^{1/2}$. Our main result guarantees that there is a data-dependent estimate $\widehat{\mathbf{\Sigma}}(\widehat{V}_n, \mathcal{D})$ that leads to an upper bound that holds with high probability, and is within constant factors of the best possible bound. In terms of the shorthand $b(V) \coloneqq R_{\max} + \gamma\|V\|_\infty$, our *empirical error estimate* takes the form

$$\mathcal{E}_n(\widehat{V}_n, \mathcal{D}, \delta) \coloneqq \frac{2\sqrt{6} \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \|\widehat{\mathbf{\Sigma}}(\widehat{V}_n, \mathcal{D})\|_{\mathrm{diag}}^{1/2} + \frac{2\varphi_s(\delta)}{n} + \frac{6b(\widehat{V}_n)}{1-\gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n-1}. \quad (14)$$

This error estimate is a function of a dataset $\mathcal{D} = (\mathcal{D}_n, \mathcal{D}_{n_h})$ where:

- The base estimator acts on $\mathcal{D}_n$ with cardinality n to compute the value function estimate $\widehat{V}_n \coloneqq \mathcal{A}_{\mathtt{Eval}}(\mathcal{D}_n)$.
- We then re-use the larger data set $\mathcal{D}_n$ to compute a portion of the covariance estimate $\widehat{\mathbf{\Sigma}}(\widehat{V}_n, \mathcal{D})$.
- In addition, we make use of a smaller set $\mathcal{D}_{2n_h}$ corresponding to samples that are held-out, and also enter the covariance estimate $\widehat{\mathbf{\Sigma}}(\widehat{V}_n, \mathcal{D})$ as detailed in Section 3.2.3 following our theorem statement. This data set consists of $2n_h$ samples with $n_h \coloneqq \lceil \frac{24 \cdot \log(16|\mathcal{X}|^2/\delta)}{(1-\gamma)^2} \rceil$.

With this set-up, we have the following guarantee:

Theorem 1 *Given any (φ_f, φ_s)-instance-optimal algorithm $\mathcal{A}_{\mathtt{Eval}}$ as in (13), a tolerance probability $\delta \in (0,1)$ and a dataset $\mathcal{D} = (\mathcal{D}_n, \mathcal{D}_{2n_h})$, where $n \geqslant \varphi_f^2(\delta)\frac{24 \cdot 8 \cdot \log(8|\mathcal{X}|^2/\delta)}{(1-\gamma)^2}$, the empirical error estimate $\mathcal{E}_n$ from equation (14) has guarantees with probability at least $1 - 3\delta$:*

(a) The ℓ_∞-error is upper bounded by the empirical error estimate:

$$\|\widehat{V}_n - V^\star\|_\infty \leqslant \mathcal{E}_n(\widehat{V}_n, \mathcal{D}, \delta). \quad (15a)$$

(b) Moreover, this guarantee is order-optimal in the sense that

$$\mathcal{E}_n(\widehat{V}_n, \mathcal{D}, \delta) \leqslant \frac{14 \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}^{1/2} + \frac{4\varphi_s(\delta)}{n} + \frac{40b(V^\star)\varphi_f(\delta)}{1-\gamma} \cdot \frac{\sqrt{\log(\frac{8|\mathcal{X}|}{\delta})}}{n-1}. \quad (15b)$$

10

We prove this theorem in Section 5.1.

A few comments regarding Theorem 1 are in order. Since the empirical error estimate $\mathcal{E}_n$ can be computed based on the data, we obtain a data-dependent confidence interval on $\|\widehat{V}_n - V^\star\|_\infty$. More precisely, we are guaranteed the inclusion

$$\left[\widehat{V}_n(x) - \mathcal{E}_n, \widehat{V}_n(x) + \mathcal{E}_n\right] \ni V^\star(x) \qquad \text{uniformly for all } x \in \mathcal{X} \qquad (16)$$

with probability at least $1 - 3\delta$.

Note that the dominant term in the error estimate $\mathcal{E}_n$ is proportional to $n^{-1/2}\|\widehat{\boldsymbol{\Sigma}}(\widehat{V}_n, \mathcal{D}_{n,n_h})\|_{\text{diag}}^{1/2}$, and it corresponds to an estimate of the dominant term $n^{-1/2}\|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}^{1/2}$ from the bound (13). The second bound (15b) in Theorem 1 ensures that $\mathcal{E}_n$ provides an optimal approximation, up to constant pre-factors, of the leading order term on the right-hand side of the bound (13).

3.2.3 CONSTRUCTING THE EMPIRICAL COVARIANCE ESTIMATE

We now provide details on the construction of the empirical covariance estimate $\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V}_n; \mathcal{D})$. Recalling the definition (11), our procedure is guided by the decomposition

$$\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star) = \underbrace{(\mathbf{I} - \gamma\mathbf{P})^{-1}}_{\substack{\text{estimated using first} \\ \text{small holdout dataset}}} \underbrace{\boldsymbol{\Sigma}_{\text{val}}(\mathcal{M}, V^\star)}_{\substack{\text{estimated by re-using} \\ \text{the base procedure's dataset}}} \underbrace{(\mathbf{I} - \gamma\mathbf{P})^{-\top}}_{\substack{\text{estimated using second} \\ \text{small holdout dataset}}}.$$

Recall that we are given data of the form $\mathcal{D} = (\mathcal{D}_n, \mathcal{D}_{2n_h})$, where the dataset $\mathcal{D}_n$ was used by the base procedure $\mathcal{A}_{\text{Eval}}$ to compute the value function estimate $\widehat{V}_n$. Our first step is to re-use the samples $\mathcal{D}_n = \{(R_j, \mathbf{Z}_j)\}_{j \in \mathcal{F}}$ to compute the V-statistic

$$\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V}_n; \mathcal{D}_n) := \frac{1}{n(n-1)} \sum_{\substack{j,k \in \mathcal{F} \\ j < k}} (R_j - R_k + \gamma(\mathbf{Z}_j - \mathbf{Z}_k)\widehat{V}_n)(R_j - R_k + \gamma(\mathbf{Z}_j - \mathbf{Z}_k)\widehat{V}_n)^\top.$$

$$(17)$$

Due to our re-use of the samples, the estimate $\widehat{V}_n$ *is dependent* on the samples $(R_j, \mathbf{Z}_j)$ that define this V-statistic. Consequently, there are some technical innovations required in order to relate this term to $\boldsymbol{\Sigma}_{\text{val}}(\mathcal{M}, V^\star)$. Notably, we do not make use of uniform convergence results (which would lead to overly conservative requirements).

As for the remainder of our instance-dependent risk estimate, we partition the hold-out set $\mathcal{D}_{2n_h}$ into two subsets $(\mathcal{D}_{n_h}^1, \mathcal{D}_{n_h}^2)$, as indexed by $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. We use this data to compute the empirical averages

$$\widehat{\mathbf{Z}}_{\mathcal{H}_1} = \frac{1}{n_h} \sum_{i \in \mathcal{H}_1} \mathbf{Z}_i \qquad \text{and} \qquad \widehat{\mathbf{Z}}_{\mathcal{H}_2} = \frac{1}{n_h} \sum_{i \in \mathcal{H}_2} \mathbf{Z}_i. \qquad (18)$$

A standard approach would be to use the data to compute $\widehat{V}_n$ to estimate $\mathbf{P}$ and plug it in; while this approach would be asymptotically valid in the fixed dimension regime, it

introduces higher order error terms that would depend on the dimension $|\mathcal{X}|$. We utilize these holdout sets to provide non-asymptotic confidence guarantees that are dimension-free.

With these three ingredients, our empirical covariance estimate is given by

$$\widehat{\boldsymbol{\Sigma}}(\widehat{V}_n, \mathcal{D}) = (\mathbf{I} - \gamma \widehat{\mathbf{Z}}_{\mathcal{H}_1})^{-1} \, \widehat{\boldsymbol{\Sigma}}_{\mathtt{val}}(\widehat{V}_n, \mathcal{D}_n) \, (\mathbf{I} - \gamma \widehat{\mathbf{Z}}_{\mathcal{H}_2})^{-\top}. \tag{19}$$

Recalling that $n_h \coloneqq 2\lceil \frac{192 \cdot \log(8|\mathcal{X}|^2/\delta)}{(1-\gamma)^2} \rceil$, we note that the hold-out sets will typically be very small relative to the sample size n used to compute the estimate $\widehat{V}_n$ itself.

3.3 Early stopping with optimality: EmpIRE procedure

In practice, it is often the case that policy evaluation is carried out repeatedly, as an inner loop within some broader algorithm (e.g., policy iteration schemes or actor-critic methods). Due to repeated calls to a policy evaluation routine, it is especially desirable in these settings to minimize the amount of data used. In practice, it is adequate to terminate a policy evaluation routine once some target accuracy ϵ is achieved; guarantees in the ℓ_∞-norm, such as those analyzed here, are particularly attractive in this context.

In this section, we leverage the data-dependent bound from Theorem 1 so as to propose *early stopping procedure* for policy evaluation, applicable for any (φ_f, φ_s)-procedure for policy evaluation. We refer to it as the EmpIRE procedure, as a shorthand for **Emp**irical **I**nstance-optimal Markov **R**eward **E**valuation. Taking as input a desired error level ϵ and a tolerance probability δ_0, it is a sequential procedure that operates over a sequence of epochs, where the amount of data available increases dyadically as the epochs increase. Upon termination, it returns an estimate $\widehat{V}$ such that $\|\widehat{V} - V^\star\|_\infty \leqslant \epsilon$ with probability at least $1 - \delta_0$. Moreover, our theory guarantees that for any particular instance, the number of samples used to do so is within a constant factor of the minimum number of samples required for that particular instance. In this sense, the procedure is *adaptive* to the relative difficulty of the instance at hand.

The following corollary, which is a simplified version of Corollary 11 presented in Appendix B, guarantees that the EmpIRE procedure terminates with high probability, and returns an ϵ-accurate estimate of the value function while using a total sample size that is well-controlled. For ease of presentation, in Corollary 2 we assume that

$$\frac{\varphi_s(\delta)}{\varphi_f^2(\delta)} \leqslant c_0 \quad \text{for all} \quad \delta > 0. \tag{21}$$

where, c_0 is a universal constant. This assumption, although satisfied by both the instance-optimal algorithms (Khamaru et al., 2021; Mou et al., 2022) discussed in the paper, is not necessary; we impose this condition only to provide a simpler upper bound on the number of epochs and number of samples used by the protocol EmpIRE. A version of Corollary 2 without the assumption (21) can be be found in Corollary 11.

Given a tolerance probability $\delta_0 \in (0, 1)$ we define

$$M_{\max} \quad \coloneqq \quad \log_2 \max \left\{ \frac{(1-\gamma)^2 \, \|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}}{\epsilon^2}, \; \frac{c_0(1-\gamma)^2}{4\epsilon} + \frac{(1-\gamma)b(V^\star)}{\epsilon} \cdot \sqrt{\log(\tfrac{8|\mathcal{X}|}{\delta_0})} \right\}. \tag{22}$$

In the following statement, we use (c_1, c_2) to denote universal positive constants.

Algorithm EmpIRE Empirical **I**nstance-optimal Markov **R**eward **E**valuation

---

1: **Inputs:** (i) instance-optimal procedure $\mathcal{A}_{\texttt{Eval}}$; (ii) target accuracy ϵ and (iii) tolerance probability δ_0

2: Initialize $N_0 = \frac{32}{(1-\gamma)^2}$, $\delta_0 = \frac{\delta_{\text{target}}}{3}$, and $\mathcal{D}_{\mathcal{H}_1}, \mathcal{D}_{\mathcal{F}}$, and $\mathcal{D}_{\mathcal{H}_2}$ as empty sets.

3: **for** $m = 1, \dots$ **do**

4:   Set tolerance parameter $\delta_m = \frac{\delta_0}{2^m}$, along with

  Holdout size $\quad h_m := N_0 \cdot \log(4|\mathcal{X}|^2/\delta_m)$, and batch size $\quad N_m = 2^m N_0 \cdot \varphi_f^2(\delta_m) \cdot \log(4|\mathcal{X}|/\delta_m)$.

5:   Augment data sets $\mathcal{D}_{\mathcal{H}_1}, \mathcal{D}_{\mathcal{F}}$, and $\mathcal{D}_{\mathcal{H}_2}$ with additional i.i.d. samples such that $|\mathcal{H}_1| = |\mathcal{H}_2| = h_m$, and $|\mathcal{F}| = N_m$. Set $\mathcal{D} = \mathcal{D}_{\mathcal{H}_1} \cup \mathcal{D}_{\mathcal{F}} \cup \mathcal{D}_{\mathcal{H}_2}$.

6:   Compute estimate
$$\widehat{V} \leftarrow \mathcal{A}_{\texttt{Eval}}(\mathcal{D}_{\mathcal{F}}).$$

7:   Evaluate empirical error estimates

$$\hat{\varepsilon}_f := \frac{2\sqrt{6} \cdot \varphi_f(\delta_m)}{\sqrt{N_m}} \cdot \widehat{\boldsymbol{\Sigma}}(\widehat{V}, \mathcal{D}), \quad \text{and} \quad \hat{\varepsilon}_s := \frac{2\varphi_s(\delta_m)}{N_m} + \frac{6b(\widehat{V})}{1-\gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta_m)}}{N_m - 1}.$$
$$(20)$$

8:   **if** $\hat{\varepsilon}_f + \hat{\varepsilon}_s < \epsilon$ **then**

9:     Terminate

10:   **end if**

11: **end for**

Corollary 2 *Given any algorithm $\mathcal{A}_{\texttt{Eval}}$ satisfying condition (13), suppose that we run the EmpIRE protocol with target accuracy ϵ and tolerance probability δ_0, and assume that condition (21) is in force. Then with probability $1 - \delta_0$:*

(a) It terminates after a number of epochs M bounded as $M \leqslant c_1 + M_{\max}$.

(b) Upon termination, it returns an estimate $\widehat{V}$ such $\|\widehat{V} - V^\star\|_\infty \leqslant \epsilon$.

(c) The total number of samples used satisfies the bound

$$n \leqslant c_2 \max\left\{ \frac{\varphi_f^2(\delta_M)}{\epsilon^2} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}, \frac{1}{\epsilon}\left[\varphi_s(2\delta_M) + \frac{b(V^\star)}{1 - \gamma}\sqrt{\log(4|\mathcal{X}|/\delta_M)} \right] \right\}$$
(23)

where $\delta_M = \frac{\delta_0}{2^M}$.

See Appendix B for the the proof.

Let us make a few comments about this claim; in this informal discussion, we omit constants and terms that are logarithmic in the pair $(|\mathcal{X}|, 1/\delta)$. Corollary 2 guarantees that, with a controlled probability, the algorithm terminates and returns an accurate answer. Moreover, over all epochs, it uses of the order $\mathcal{O}\left(\frac{\|\mathbf{\Sigma}^\star(r,\mathbf{P},V^\star)\|_{\mathrm{diag}}}{\epsilon^2}\right)$ samples to achieve ϵ-accuracy in the ℓ_∞-norm. This guarantee should be compared to local minimax lower bounds from our past work (Khamaru et al., 2021): any procedure with an oracle that has access to the true error $\|\widehat{V} - V^\star\|_\infty$ requires at least $\frac{\|\mathbf{\Sigma}^\star(r,\mathbf{P},V^\star)\|_{\mathrm{diag}}}{\epsilon^2}$ samples. Thus, our procedure utilizes the same number of samples as any instance-optimal procedure equipped with such an oracle, up to constant and logarithmic factors.

Next, observe that the bound in the right hand side of the expression (22) depends on $\log(1/\epsilon)$, and is of higher order. Ignoring it for the moment, the number of epochs required to achieve a given target accuracy $\epsilon > 0$ is of the order $\log_2\left(\frac{(1-\gamma)^2}{\epsilon^2} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}\right)$ epochs, to an additive constant. Finally, we refer the reader to Corollary 11 in Appendix B for an analogue of Corollary 2 without assumption (21).

3.4 Numerical simulations

In this section, we use a simple two-state Markov reward process (MRP) $\mathcal{M} = (\mathbf{P}, r)$ to illustrate the behavior of our methods; this family of instances was introduced in past work by a subset of the current authors (Pananjady and Wainwright, 2021; Khamaru et al., 2021). For a parameter $p \in [0, 1]$, consider a transition matrix $\mathbf{P} \in \mathbb{R}^{2 \times 2}$ and reward vector $r \in \mathbb{R}^2$ given by

$$\mathbf{P} = \begin{bmatrix} p & 1 - p \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} 1 \\ \tau \end{bmatrix}.$$

where $\tau \in \mathbb{R}$ along with the discount $\gamma \in [0, 1)$ are additional parameters of the construction. Fix a scalar $\lambda \geqslant 0$, and then set

$$p = \frac{4\gamma - 1}{3\gamma}, \quad \text{and} \quad \tau = 1 - (1 - \gamma)^\lambda.$$

The probability p belongs to $[0, 1]$ as long as $\gamma \in [\frac{1}{4}, 1]$, so that we restrict our attention to this range. This is a Markov reward process induced via the more general construction of a Markov decision process given in Example 1 to follow in the sequel.

By calculations similar to those in Appendix A.2, we have

$$\|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}^{1/2} \leqslant c \cdot \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}.$$

Thus for larger choices of γ we see a "harder" problem, and for larger λ we see an "easier" problem, indicated by the complexity functional. If we knew that

$$\|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}^{1/2} = c \cdot \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda},$$

then we could easily see that an instance-optimal algorithm requires $n \gtrsim \frac{1}{\epsilon^2} \left(\frac{1}{1-\gamma}\right)^{3-2\lambda}$ samples to achieve ϵ accuracy in the ℓ_∞-norm. Typically we do not have knowledge of $\|\boldsymbol{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}^{1/2}$ prior to running the algorithm so we would need to use at least $n \gtrsim \frac{1}{\epsilon^2(1-\gamma)^3}$ samples to guarantee ϵ-accuracy. For $\lambda \geqslant 0.5$, it becomes apparent that an instance-optimal algorithm requires fewer samples than the worst-case guarantees indicate, and the differences become drastic for larger λ and γ. The EmpIRE algorithm allows the user to exploit the instance-specific difficulty of the problem at hand and avoid the worst-case scenario of requiring $n \gtrsim \frac{1}{\epsilon^2(1-\gamma)^3}$ samples.

We describe the details of the numerical simulations ran for the policy evaluation setting here. For every valid combination of (γ, λ), we ran Algorithm EmpIRE with the ROOT-SA algorithm (Mou et al., 2022) as our instance-optimal sub-procedure on the MRP. In each case, we performed $T = 1000$ trials, measuring the factor savings, i.e. the ratio of the number of samples required in the worst case to the number of samples actually used, as well as for our estimate $\widehat{V}$ the final predicted error for our estimated and the true error. The γ's were chosen to be uniformly spaced between 0.9 and 0.99 in the log-scale, and λ was chosen to be in the set $\{1.0, 1.5\}$. The desired tolerance was chosen to be $\epsilon = 0.1$. Our results are presented in Figure 1, as previously described. The initial point V_0 was chosen by setting aside $\frac{2}{(1-\gamma)^2}$ samples to construct a plug-in estimate of $V^\star$. As expected, increasing both γ and λ increases the savings in the number of samples used, as compared to the worst-case guarantees. Figure 1 highlights the improvement in sample size requirements of Algorithm EmpIRE, as compared to the using the worst-case guarantees and illustrates the benefits of exploiting the local structure of the problem at hand.

Figure 2 serves to verify that our theoretical guarantees describe the behavior observed in practice. For every combination of (γ, λ) we run 1000 trials of Algorithm EmpIRE and keeping track of the predicted error given by equation (15a) as well as the true error $\|\widehat{V} - V^\star\|_\infty$. Algorithm EmpIRE was run with chosen tolerance $\epsilon = 0.1$. Our theory ensures that the true error should be consistently below the predicted error for all combinations of γ and λ, which is the behavior illustrated in Figure 2. The plots also illustrate that the true error is consistently far below the predicted error (which itself is consistently below the specified tolerance ϵ), demonstrating that our predictions are relatively conservative, and that higher-order terms can potentially be dropped in error estimates while still remaining

a viable algorithm. Overall, Figure 2 illustrates that the bounds on Algorithm EmpIRE are correct and highlights its practical utility in an idealized setting.
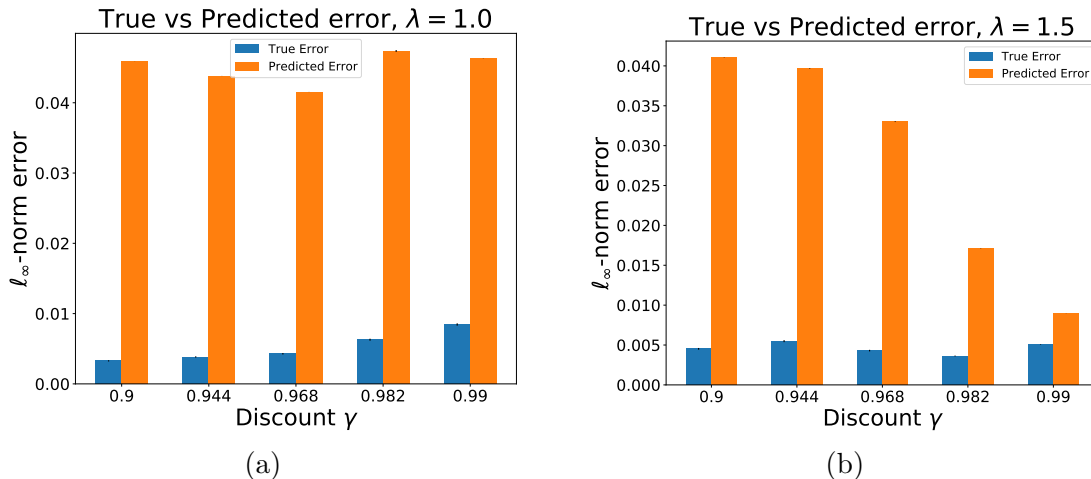
Figure 2. Illustration of the termination behavior of Algorithm EmpIRE applied on the MRP. Plots the average of the true error (blue) and predicted error (orange) along with error bars denoting the standard deviation for different choices of γ and for (a) $\lambda = 1.0$ and (b) $\lambda = 1.5$.

4. Confidence Regions for Optimal Q-function Estimation

In this section, we derive confidence regions for optimal Q-function estimation problem. Let us first describe the functional that characterizes the difficulty of the optimal Q-function estimation problem.

4.1 Instance-dependence for optimal Q-functions

Given a sample $(\mathcal{Z}, \mathcal{R})$ from our observation model (see equation (9)), we can define the single-sample empirical Bellman operator as

$$\widehat{\mathcal{J}}(Q) \coloneqq \mathcal{R}(x, u) + \gamma \sum_{x' \in \mathcal{X}} \mathcal{Z}_u(x' \mid x) \max_{u' \in \mathcal{U}} Q(x', u'), \tag{24}$$

where we have introduced $\mathcal{Z}_u(x' \mid x) \coloneqq \mathbb{1}_{\mathcal{Z}(x,u)=x'}$. Throughout this section, we use the shorthand $D = |\mathcal{X} \times \mathcal{U}|$.

In a recent paper (Xia et al., 2021), the authors show the quantity that determines the difficulty of estimating the optimal value function $Q^\star$ is[1]

$$\boldsymbol{\Sigma}_{\texttt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma) \coloneqq \max_{\pi \in \Pi^\star} \|(\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^\pi)^{-1} \boldsymbol{\Sigma}_{\texttt{opt}}^\star(Q^\star)(\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^\pi)^{-\top}\|_{\text{diag}}^{\frac{1}{2}}, \tag{25}$$

---

1. The quantity $\boldsymbol{\Sigma}_{\texttt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)$ is a the same of the term $\max_{\pi^\star \in \Pi^\star} \|\nu(\pi^\star; \boldsymbol{\mathcal{P}}, \mathbf{r}, \gamma)\|_\infty$ from Theorem 1 in Xia et al. (2021).

where the quality of an estimate is being measured by its ℓ_∞-distance from $Q^\star$. Here the set $\Pi^\star$ denotes the set of all optimal policies, the matrix $\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star) \coloneqq \mathrm{Cov}(\widehat{\boldsymbol{\mathcal{J}}}(Q^\star))$, and $\boldsymbol{\mathcal{P}}^\pi$ is a right-linear mapping of $\mathbb{R}^D$ to itself, whose action on a Q-function is given by

$$\boldsymbol{\mathcal{P}}^\pi(Q)(x,u) = \sum_{x' \in \mathcal{X}} \boldsymbol{\mathcal{P}}_u(x' \mid x) \cdot Q(x', \pi(x')) \qquad \text{for all} \quad (x,u) \in \mathcal{X} \times \mathcal{U}.$$

Note that by construction, the quantity $\widehat{\boldsymbol{\mathcal{J}}}(Q^\star)$ is an unbiased estimate of $\boldsymbol{\mathcal{J}}(Q^\star)$, and the covariance matrix $\mathrm{Cov}(\widehat{\boldsymbol{\mathcal{J}}}(Q^\star))$ in equation (25) captures the noise present in the empirical Bellman operator (24) as an estimate of the population Bellman operator (8), when evaluated at the optimal value function $Q^\star$. As for the pre-factor $(\mathbf{I} - \gamma\boldsymbol{\mathcal{P}}^\pi)^{-1}$, by a Neumann series expansion, we can write

$$(\mathbf{I} - \gamma\boldsymbol{\mathcal{P}}^\pi)^{-1} = \sum_{k=0}^{\infty} (\gamma\boldsymbol{\mathcal{P}}^\pi)^k.$$

The sum of the powers of $\gamma\boldsymbol{\mathcal{P}}^\pi$ accounts for the compounded effect of an initial perturbation when following the Markov chain specified by an optimal policy π. Put simply, the quantity (25) captures the noise accumulated when an optimal policy π is followed.

Xia et al. (2021) also show that under appropriate assumptions, the estimate $\widehat{Q}_n$, obtained from a variance reduced Q-learning algorithm using n i.i.d. samples, satisfies the bound

$$\|\widehat{Q}_n - Q^\star\|_\infty \leqslant \varphi_f(\delta) \cdot \frac{\boldsymbol{\Sigma}_{\mathtt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)}{\sqrt{n}} + \frac{c_2}{n} \tag{26}$$

with probability at least $1 - \delta$. Here the functions φ_f and φ_s depend on the tolerance parameter $\delta \in (0,1)$, along with logarithmic factors in the dimension. Furthermore, this algorithm achieves the non-asymptotic local minimax lower bound for the optimal Q-function estimation problem (see Theorem 1 in Xia et al. (2021)).

4.2 A conservative yet useful upper bound

Motivated by the success in Section 3, it is interesting to ask if we can prove a data-dependent estimate for the term $\boldsymbol{\Sigma}_{\mathtt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)$. Observe that the term $\boldsymbol{\Sigma}_{\mathtt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)$ from equation (25) depends on the set of *all* optimal policies $\Pi^\star$, and the authors are not aware of a data-dependent efficient estimate that is valid without imposing restrictive assumptions on the MDP $\mathcal{P} = (\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)$.

Instead, we study the following upper bound

$$\boldsymbol{\Sigma}_{\mathtt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma) \leqslant \frac{\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|_{\mathrm{diag}}^{\frac{1}{2}}}{1 - \gamma}. \tag{27}$$

Observe that all entries of the PSD matrix $\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)$ are upper bounded by the scalar $\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|_{\mathrm{diag}}$, the entries of $(\mathbf{I} - \gamma\boldsymbol{\mathcal{P}}^\pi)^{-1}$ are non-negative, and $\|(\mathbf{I} - \gamma\boldsymbol{\mathcal{P}}^\pi)^{-1}\|_{1,\infty} \leqslant \frac{1}{1-\gamma}$. Combining these three observations yield the claim (27).

It is interesting to compare the upper bound (27) with a worst case upper bound on $\mathbf{\Sigma}_{\texttt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)$. In light of Lemma 7 from Azar et al. (2013), assuming $\mathcal{R}_{\max} \leqslant 1$, and $\|\mathbf{r}\|_\infty \leqslant 1$ for simplicity, we have

$$\mathbf{\Sigma}_{\texttt{PolOpt}}(\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma) \leqslant \frac{1}{(1-\gamma)^{1.5}}. \tag{28}$$

By comparing the bounds (27) and (28), we see that the upper bound (27) is particularly useful when

$$\|\mathbf{\Sigma}_{\text{opt}}^\star(Q^\star)\|_{\text{diag}}^{\frac{1}{2}} \ll \frac{1}{\sqrt{1-\gamma}}.$$

As an illustration, we now describe an interesting sequence of MDPs for which the last condition is satisfied.

Example 1 (A continuum of illustrative examples) *Consider an MDP with two states* $\{x_1, x_2\}$, *two actions* $\{u_1, u_2\}$, *and with transition functions and reward functions given by*

$$\boldsymbol{\mathcal{P}}_{u_1} = \begin{bmatrix} p & 1-p \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{\mathcal{P}}_{u_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad and \quad \mathbf{r} = \begin{bmatrix} 1 & 0 \\ \tau & 0 \end{bmatrix}. \tag{29}$$

We assume that there is no randomness in the reward samples. Here the pair (p, τ) *along with the discount factor* γ *are parameters of the construction, and we consider a sub-family of these parameters indexed by a scalar* $\lambda \geqslant 0$. *For any such* λ *and discount factor* $\gamma \in (\frac{1}{4}, 1)$, *consider the setting*

$$p = \frac{4\gamma - 1}{3\gamma}, \quad and \quad \tau = 1 - (1-\gamma)^\lambda.$$

With these choices of parameters, the optimal Q-function $Q^\star$ *takes the form*

$$Q^\star = \begin{bmatrix} \frac{(1-\gamma)+\gamma\tau(1-p)}{(1-\gamma)(1-\gamma p)} & \gamma \cdot \frac{(1-\gamma)+\gamma\tau(1-p)}{(1-\gamma)(1-\gamma p)} \\ \frac{\tau}{1-\gamma} & \frac{\gamma\tau}{1-\gamma} \end{bmatrix},$$

with an unique optimal policy $\pi^\star(x_1) = \pi^\star(x_2) = u_1$. *We can then compute that*

$$\|\mathbf{\Sigma}_{\text{opt}}^\star(Q^\star)\|_{\text{diag}}^{\frac{1}{2}} = \frac{1-\tau}{1-\gamma p} \cdot \sqrt{p(1-p)} = c \cdot \left(\frac{1}{1-\gamma}\right)^{\frac{1}{2}-\lambda} < \frac{1}{\sqrt{1-\gamma}}. \tag{30}$$

See Appendix A.2 for the details of this calculation.

Example 1 shows that for any $\lambda > 0$, the bound $\frac{\|\mathbf{\Sigma}_{\text{opt}}^\star(Q^\star)\|_{\text{diag}}^{\frac{1}{2}}}{1-\gamma}$ is smaller than the worst case bound of $\frac{1}{(1-\gamma)^{1.5}}$ by a factor of $\frac{1}{(1-\gamma)^\lambda}$. We point out that this gap is significant when the discount factor γ is close to 1. For instance, for $\gamma = 0.99$ and $\lambda = 1$, the upper bound (27) is 10^2 times better than the worst case bound. Alternatively, the bound (26) yields an improvement of a factor of 10^4, when compared to a worst case value.

4.3 Data-dependent bounds for optimal Q-functions

In this section, we provide a data-dependent upper bound on $\|\widehat{Q} - Q^\star\|_\infty$, where the estimate $\widehat{Q}$ is obtained from any algorithm $\mathcal{A}_{\texttt{PolOpt}}$ satisfying certain convergence criterion.

4.3.1 Instance-valid algorithms

We assume that given access to n generative samples $\mathcal{D}_n \coloneqq \{\mathcal{R}_i, \mathcal{Z}_i\}_{i=1}^n$ from an MDP $\mathcal{P} = (\mathbf{r}, \boldsymbol{\mathcal{P}}, \gamma)$, the output $\widehat{Q}_n \coloneqq \mathcal{A}_{\texttt{PolOpt}}(\mathcal{D}_n)$, obtained using the data set $\mathcal{D}_n$, satisfies

$$\mathcal{A}_{\texttt{PolOpt}} \text{ condition:} \qquad \|\widehat{Q}_n - Q^\star\|_\infty \leqslant \frac{\varphi_f(\delta)}{\sqrt{n}} \cdot \frac{\|\boldsymbol{\Sigma}^\star_{\text{opt}}(Q^\star)\|_{\text{diag}}^{\frac{1}{2}}}{(1-\gamma)} + \frac{\varphi_s(\delta)}{n}. \tag{31}$$

with probability $1 - \delta$. Here φ_f and φ_s are functions of the tolerance level δ; in practical settings, they may in addition depend on the problem dimension in some cases, but we suppress this dependence for simplicity.

   We point out that any instance-optimal algorithm, i.e. algorithms satisfying the condition (26), automatically satisfies the condition (31). One added benefit of the condition (31) is that this condition is much easier to verify. To illustrate this point, in Proposition 12 in Appendix C, we show that the variance-reduced Q-learning algorithm from Xia et al. (2021) satisfies the condition (31) under a *milder assumption* (cf. Proposition 12 in Appendix C and Theorem 2 from Xia et al. (2021)). In the rest of the section, we focus on providing a data-dependent estimate for $\|\boldsymbol{\Sigma}^\star_{\text{opt}}(Q^\star)\|_{\text{diag}}^{\frac{1}{2}}$.

4.3.2 Constructing a data-dependent bound

In order to obtain a data-dependent bound on $\|\widehat{Q}_n - Q^\star\|_\infty$, it suffices to estimate the complexity term $\|\boldsymbol{\Sigma}^\star_{\text{opt}}(Q^\star)\|_{\text{diag}}^{1/2}$. Our next result guarantees that there is a data-dependent estimate $\widehat{\boldsymbol{\Sigma}}_{\text{opt}}(\widehat{Q}_n, \mathcal{D})$ which provides an upper bound that holds with high probability, and is within constant factors of $\|\boldsymbol{\Sigma}^\star_{\text{opt}}(Q^\star)\|_{\text{diag}}^{1/2}$. With a recycling of notation, we define $b(Q) \coloneqq R_{\max} + \gamma\|Q\|_\infty$, and construct the *empirical error estimate*

$$\mathcal{E}_n(\widehat{Q}_n, \mathcal{D}_n, \delta) \coloneqq \frac{2\sqrt{2} \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \frac{\|\widehat{\boldsymbol{\Sigma}}_{\text{opt}}(\widehat{Q}_n, \mathcal{D}_n)\|_{\text{diag}}^{1/2}}{(1-\gamma)} + \frac{2\varphi_s(\delta)}{n} + \frac{8b(\widehat{Q}_n)}{1-\gamma} \cdot \frac{\varphi_f(\delta)\sqrt{2\log(D/\delta)}}{n-1}. \tag{32}$$

In this error estimate, the base estimator uses the n-sample dataset $\mathcal{D}_n$ to compute the Q-value function estimate $\widehat{Q}_n \coloneqq \mathcal{A}_{\texttt{PolOpt}}(\mathcal{D}_n)$. It also uses the same data to estimate the covariance $\widehat{\boldsymbol{\Sigma}}_{\text{opt}}(\widehat{Q}_n, \mathcal{D}_n)$, according to the procedure given in Section 4.3.3.

The following result summarizes the guarantees associated with the empirical error estimate (32):

Theorem 3 *For any (φ_f, φ_s) instance-dependent algorithm $\mathcal{A}_{PolOpt}$ as in (31), a tolerance $\delta \in (0, 1)$ and a dataset $\mathcal{D}_n$ with $n \geqslant \varphi_f^2(\delta) \cdot \frac{32 \cdot \log(4D/\delta)}{(1-\gamma)^2}$, the following guarantees hold with probability at least $1 - 2\delta$:*

(a) *The ℓ_∞-error is upper bounded as*

$$\|\widehat{Q}_n - Q^\star\|_\infty \leqslant \mathcal{E}_n(\widehat{Q}_n, \mathcal{D}_n, \delta). \tag{33a}$$

(b) *Moreover, this guarantee is order-optimal in the sense that*

$$\mathcal{E}_n(\widehat{Q}_n, \mathcal{D}_n, \delta) \leqslant \frac{7 \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \frac{\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{1/2}_{\mathrm{diag}}}{1-\gamma} + \frac{6\varphi_s(\delta)}{n} + \frac{5b(Q^\star)}{1-\gamma} \cdot \frac{\sqrt{\log(D/\delta)}}{n-1}. \tag{33b}$$

See Section 5.2 for the proof.

A few comments regarding Theorem 3 are in order. Like in policy evaluation, the empirical error estimate $\mathcal{E}_n$ can be computed based on the data and we obtain a data-dependent confidence interval for $Q^\star$. In particular, equation (33a) guarantees that

$$\left[\widehat{Q}_n(x,u) - \mathcal{E}_n, \widehat{Q}_n(x,u) + \mathcal{E}_n\right] \ni Q^\star(x,u) \qquad \text{uniformly for all } x \in \mathcal{X}, u \in \mathcal{U}$$

with probability at least $1 - 2\delta$.

We point out that the dominating term in the error estimate $\mathcal{E}_n$ is proportional to $n^{-1/2}\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q}_n, \mathcal{D}_n)\|^{1/2}_{\mathrm{diag}}$, and this is an estimate of the dominating term $n^{-1/2}\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{1/2}_{\mathrm{diag}}$ from the bound (31). Additionally, the bound (33b) ensures that the proposed data-dependent bound in equation (33a) is a sharp approximation of $n^{-1/2}\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{1/2}_{\mathrm{diag}}$ from the bound (31).

4.3.3 ESTIMATION OF $\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{\frac{1}{2}}_{\mathrm{diag}}$

In this section, we provide details on the construction of the estimator $\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q}_n, \mathcal{D}_n)\|^{1/2}_{\mathrm{diag}}$ of $\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{\frac{1}{2}}_{\mathrm{diag}}$. Given a generative data set $\mathcal{D}_n := \{\mathcal{R}_i, \mathcal{Z}_i\}^n_{i=1}$, we first use $\widehat{Q}_n = \mathcal{A}_{\mathtt{PolOpt}}(\mathcal{D}_n)$ to estimate $Q^\star$, and use the same data set $\mathcal{D}_n$ to provide an empirical estimate of the quantity $\|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{\frac{1}{2}}_{\mathrm{diag}}$. Observe that it suffices to estimate the diagonal entries of the matrix $\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)$. Accordingly, our empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q}_n, \mathcal{D}_n)$, based on a data set $\mathcal{D}_n$, is a $D \times D$ *diagonal matrix*. Concretely, let $\widehat{\boldsymbol{\mathfrak{J}}}_i$ denote the empirical estimate of the Bellman optimality operator based on the ith generative sample, i.e.,

$$\widehat{\boldsymbol{\mathfrak{J}}}_i(Q)(x,u) = \mathcal{R}_i(x,u) + \gamma \sum_{x' \in \mathcal{X}} \mathcal{Z}^u_i(x' \mid x) \cdot \max_{u \in \mathcal{U}} Q(x', u), \quad \text{for all } (x,u) \in \mathcal{X} \times \mathcal{U}.$$

Recall that $\mathcal{Z}^u_i(x' \mid x)$ is 1 if and only if the sample in $\mathcal{Z}_i$ from (x,u) gives x' as the next state, and zero otherwise. We define the $(x,u)^{th}$ diagonal entry of the diagonal matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q}_n, \mathcal{D}_n)$ as

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q}_n, \mathcal{D}_n)((x,u),(x,u)) = \frac{1}{n(n-1)} \sum_{1 \leqslant i < j \leqslant k} \left\{\widehat{\boldsymbol{\mathfrak{J}}}_i(\widehat{Q}_n)(x,u) - \widehat{\boldsymbol{\mathfrak{J}}}_j(Q)(x,u)\right\}^2. \tag{34}$$

20

4.3.4 Early stopping for optimal Q-functions

Inspired by Algorithm EmpIRE, we can take an instance-dependent procedure $\mathcal{A}_{\texttt{PolOpt}}$ satisfying the condition (31) and construct a similar stopping procedure that terminates when the data-dependent upper bound of $\|\widehat{Q} - Q^\star\|_\infty$ estimate is below some user-specified threshold ϵ. Following Algorithm EmpIRE, one would iteratively recompute the estimate $\widehat{Q}$ using the algorithm $\mathcal{A}_{\texttt{PolOpt}}$, and check the condition

$$\frac{2\sqrt{2}\varphi_f(\delta_n)}{\sqrt{n}} \cdot \frac{\|\widehat{\boldsymbol{\Sigma}}_{\text{opt}}(\widehat{Q}_n, \mathcal{D}_n)\|_{\text{diag}}^{1/2}}{(1-\gamma)} + \frac{2\sqrt{2}\varphi_f(\delta_n) \cdot \sqrt{16\log(D/\delta)}}{(1-\gamma)} \cdot \frac{b(\widehat{Q}_n)}{n-1} + \frac{2\varphi_s\delta_n}{n} \leqslant \epsilon.$$

We terminate the algorithm $\mathcal{A}_{\texttt{PolOpt}}$ when this criterion is satisfied. Again, the correctness of such stopping procedure follows from Theorem 3 and a union bound.

4.4 Some numerical simulations

In the setting of Q-learning, the worst-case sample complexity for any MDP to achieve ϵ-accuracy is $\frac{1}{\epsilon^2(1-\gamma)^3}$. However Example 1 illustrates that even under our looser instance-dependent guarantee from equation (27), we only require $\frac{1}{\epsilon^2(1-\gamma)^{3-2\lambda}}$ samples. Like the example in policy evaluation, there can still be a substantial difference between the number of samples required for the worst-case versus the instance-dependent case. We illustrate the gains provided by Algorithm EmpIRE in the following numerical simulations.

For every combination of (γ, λ), we ran Algorithm EmpIRE with the ROOT-SA algorithm (Mou et al., 2022) as our base procedure on the MDP described in Example 1 for 1000 trials. We used values of γ that were uniformly spaced between 0.9 and 0.99 on the log-scale, and the two choices $\lambda \in \{1.0, 1.5\}$. The desired tolerance was set at $\epsilon = 0.05$. The initialization point Q_0 was chosen via setting aside $\frac{2}{(1-\gamma)^2}$ samples and estimating r and $\boldsymbol{\mathscr{P}}$ via averaging, and then solving for the optimal Q-function for this MDP. We measured the factor savings by computing the ratio of the worst-case sample complexity $\frac{1}{\epsilon^2(1-\gamma)^3}$ with the number of samples used by Algorithm EmpIRE. The results are presented in Figure 3. In order to verify the correctness of our guarantee, for each trial we also keep track of the predicted error given by equation (33a) and the true error $\|\widehat{Q} - Q^\star\|_\infty$. We see in Figure 4 that the true error is consistently below the predicted error and also consistently below the specified error threshold of 0.5. These plots illustrate the practical benefits that Algorithm EmpIRE brings when taking advantage of instance-dependent theory.
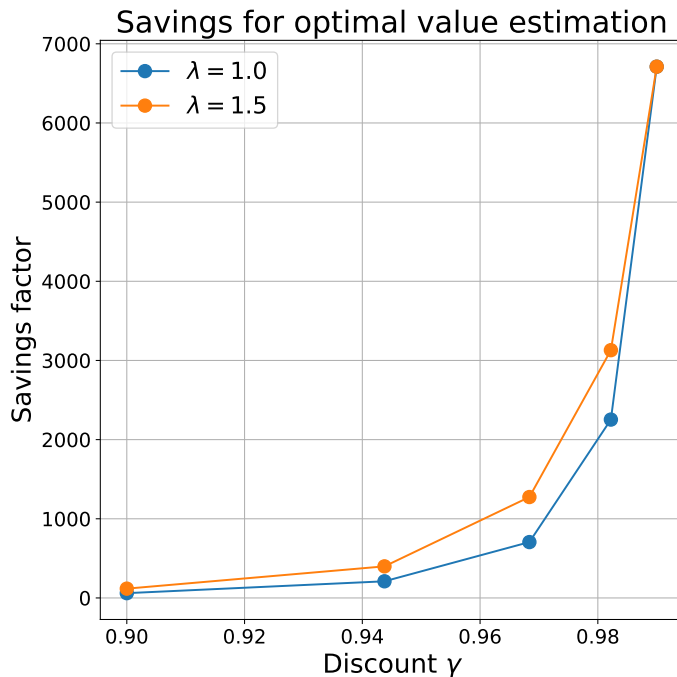
5. Proofs

In this section, we provide proofs of our main results, with Sections 5.1 and 5.2 devoted the proofs of Theorems 1 and 3, respectively.

5.1 Proof of Theorem 1

Throughout the proof, we suppress the dependence of $\widehat{V}_n$ on the sample size n. Additionally, we adopt the shorthand notations

$$\boldsymbol{\Sigma}_{\texttt{val}}^\star(V^\star) = \boldsymbol{\Sigma}_{\texttt{val}}(\mathcal{M}, V^\star), \qquad \widehat{\boldsymbol{\Sigma}}_{\texttt{val}}(V^\star) = \widehat{\boldsymbol{\Sigma}}_{\texttt{val}}(V^\star; \mathcal{D}_\mathcal{F}), \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}(\widehat{V}) = \widehat{\boldsymbol{\Sigma}}(\widehat{V}, \mathcal{D}_{n, n_h}).$$

Figure 3. Illustration of the savings in sample size requirements of Algorithm EmpIRE on the MDP in Example 1 for different choices of γ and λ. The figure plots the factor of savings, i.e. the ratio of the number of samples required in the worst case to the number of samples actually used against the log discount complexity factor for both $\lambda = 1.0$ (blue) and $\lambda = 1.5$ (orange).

The proof involves three auxiliary results that play key roles, which we state here. Our first result allows us to relate an idealized covariance estimate that involves the optimal value function $V^\star$ to one involving our estimate $\widehat{V}$.

Lemma 4 *We have*

$$\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{val}(V^\star)(\mathbf{I} - \gamma\mathbf{P})^{-1}\|_{\text{diag}}^{\frac{1}{2}} \leqslant \sqrt{2}\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{val}(\widehat{V})(\mathbf{I} - \gamma\mathbf{P})^{-1}\|_{\text{diag}}^{\frac{1}{2}}$$
$$+ \frac{\sqrt{8}}{1-\gamma}\|\widehat{V} - V^\star\|_\infty \quad (35)$$

It is worthwhile making some comments on this result. Recall that the estimate $\widehat{V}$ and the statistic defining the covariance $\widehat{\boldsymbol{\Sigma}}_{\text{val}}$ both depend on the same data. These kinds of dependencies—between data and estimators—arise when analyzing empirical risk minimization, in which context it is standard to make use of uniform laws of large numbers. In the current setting, however, we need only show that the two quantities are close when $\widehat{V}$ is close to $V^\star$. For this reason, it turns out that a relatively simple deterministic argument can be used to establish the claim (35). See Section 5.1.3 for the proof.

Our second auxiliary result controls the effect of using the hold-out sets as plug-in estimates of the probability transition matrix:

22
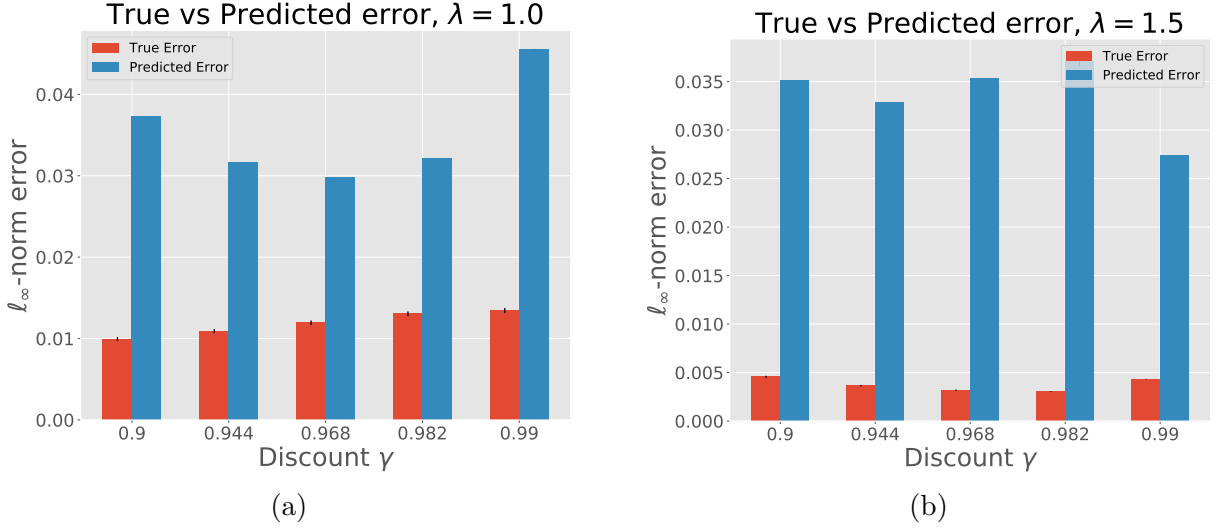
Figure 4. Illustration of the termination behavior of Algorithm EmpIRE applied on the MDP in Example 1. Plots the average of the true error (blue) and predicted error (orange) along with error bars denoting the standard deviation for different choices of γ and for (a) $\lambda = 1.0$ and (b) $\lambda = 1.5$.

Lemma 5 *For a given tolerance $\delta \in (0,1)$, consider a pair of hold-out estimates $\widehat{\boldsymbol{Z}}_{\mathcal{H}_1}$ and $\widehat{\boldsymbol{Z}}_{\mathcal{H}_2}$, each based on hold-out size $n_h \geqslant 24 \cdot \frac{\log(8|\mathcal{X}|^2/\delta)}{(1-\gamma)^2}$. For any positive semi-definite matrix $\mathbf{M}$, we have*

$$\|(\mathbf{I} - \gamma \mathbf{P})^{-1}\mathbf{M}(\mathbf{I} - \gamma \mathbf{P})^{-\top}\|_{\mathrm{diag}} \leqslant 3\|(\mathbf{I} - \gamma \widehat{\boldsymbol{Z}}_{\mathcal{H}_1})^{-1}\mathbf{M}(\mathbf{I} - \gamma \widehat{\boldsymbol{Z}}_{\mathcal{H}_2})^{-\top}\|_{\mathrm{diag}} \qquad (36)$$

with probability at least $1 - \delta$.

See Section 5.1.4 for a proof of this lemma.

Our third auxiliary lemma controls the error between the empirical covariance estimate and the true covariance.

Lemma 6 *For a given tolerance $\delta \in (0,1)$, we have*

$$\|(\mathbf{I} - \gamma \mathbf{P})^{-1}\boldsymbol{\Sigma}_{val}^{\star}(V^{\star})(\mathbf{I} - \gamma \mathbf{P})^{-\top}\|_{\mathrm{diag}}^{\frac{1}{2}} \leqslant \|(\mathbf{I} - \gamma \mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{val}(V^{\star})(\mathbf{I} - \gamma \mathbf{P})^{-\top}\|_{\mathrm{diag}}^{\frac{1}{2}}$$
$$+ \frac{b(V^{\star})}{1-\gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n-1}}$$

and

$$\|(\mathbf{I} - \gamma \mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{val}(V^{\star})(\mathbf{I} - \gamma \mathbf{P})^{-\top}\|_{\mathrm{diag}}^{\frac{1}{2}} \leqslant \|(\mathbf{I} - \gamma \mathbf{P})^{-1}\boldsymbol{\Sigma}_{val}^{\star}(V^{\star})(\mathbf{I} - \gamma \mathbf{P})^{-\top}\|_{\mathrm{diag}}^{\frac{1}{2}}$$
$$+ \frac{b(V^{\star})}{1-\gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n-1}},$$

with probability exceeding $1 - \delta$.

See Section 5.1.5 for a proof of this lemma.

Equipped with these auxiliary results, we now proceed with the proof of the theorem itself.

5.1.1 Proof of Theorem 1 part (a)

We begin by proving the bound (15a). Using Lemma 6, we have

$$\|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\text{diag}}^{\frac{1}{2}} = \|(\mathbf{I} - \gamma\mathbf{P})^{-1}\boldsymbol{\Sigma}_{\text{val}}^{\star}(V^{\star})(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}}^{\frac{1}{2}}$$

$$\leqslant \|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(V^{\star})(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}}^{\frac{1}{2}} + \frac{b(V^{\star})}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \quad (37)$$

with probability at at least $1 - \delta$. We use the shorthand

$$\boldsymbol{\Gamma} \coloneqq (\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(V^{\star})(\mathbf{I} - \gamma\mathbf{P})^{-\top}$$

to simplify notation. First, by applying Lemma 4, we have

$$\|\boldsymbol{\Gamma}\|_{\text{diag}}^{\frac{1}{2}} \leqslant \sqrt{2}\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\mathbf{P})^{-1}\|_{\text{diag}}^{\frac{1}{2}} + \tfrac{\sqrt{8}}{1-\gamma}\|\widehat{V} - V^{\star}\|_{\infty} \quad (38a)$$

Second, applying Lemma 5 with the choice $\mathbf{M} = \widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V})$, we have

$$\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}} \leqslant 3\|(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_1})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_2})^{-\top}\|_{\text{diag}} \quad (38b)$$

with probability at least $1 - \delta$.

Combining the preceding bounds yields

$$\|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\text{diag}}^{\frac{1}{2}} \leqslant \sqrt{6}\cdot\|(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_1})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_2})^{-\top}\|_{\text{diag}}^{\frac{1}{2}} + \frac{b(V^{\star})}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}}$$

$$+ \frac{\sqrt{8}}{1 - \gamma}\|\widehat{V} - V^{\star}\|_{\infty}$$

with probability at least $1 - 2\delta$.

The function $V \mapsto b(V)$ is 1-Lipschitz in the ℓ_{∞}-norm, so that we can write

$$\frac{b(V^{\star})}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \leqslant \frac{1}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \left\{ b(\widehat{V}) + \|\widehat{V} - V^{\star}\|_{\infty} \right\}.$$

Combining the pieces yields

$$\|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\text{diag}}^{\frac{1}{2}} \leqslant \sqrt{6}\cdot\|(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_1})^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_2})^{-\top}\|_{\text{diag}}^{\frac{1}{2}} + \frac{b(\widehat{V})}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}}$$

$$+ \|\widehat{V} - V^{\star}\|_{\infty} \cdot \left(\frac{\sqrt{8}}{1 - \gamma} + \frac{1}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \right). \quad (39)$$

We now recall the assumed bound (13) on the accuracy of $\widehat{V}$—viz.

$$\|\widehat{V} - V^\star\|_\infty \leqslant \frac{\varphi_f(\delta)}{\sqrt{n}} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}^{1/2} + \frac{\varphi_s(\delta)}{n}.$$

Substituting the bound (39) into the above, we have

$$\|\widehat{V} - V^\star\|_\infty \leqslant \frac{\varphi_f(\delta)}{\sqrt{n}} \left(\sqrt{6} \cdot \|\widehat{\mathbf{\Sigma}}(\widehat{V})\|_{\text{diag}}^{\frac{1}{2}} + \frac{b(\widehat{V})}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \right) + \frac{\varphi_s(\delta)}{n}$$

$$+ \frac{\varphi_f(\delta)}{\sqrt{n}} \cdot \left(\frac{\sqrt{8}}{1 - \gamma} + \frac{1}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \right) \|\widehat{V} - V^\star\|_\infty \quad (40)$$

Our choice of $n \geqslant \varphi_f^2(\delta) \cdot \frac{24 \cdot 8 \cdot \log(|\mathcal{X}|/\delta)}{(1-\gamma)^2}$ ensures that

$$\frac{\varphi_f(\delta)}{\sqrt{n}} \left(\frac{\sqrt{8}}{1 - \gamma} + \frac{1}{1 - \gamma} \cdot \sqrt{\frac{8\log(|\mathcal{X}|/\delta)}{n - 1}} \right) \leqslant \frac{1}{2},$$

and rearranging yields the claim (15a).

5.1.2 Proof of Theorem 1 part (b)

Here we prove the bound (15b). Like previously, we require the following lemma:

Lemma 7 *For a given tolerance $\delta \in (0, 1)$, consider a pair of hold-out estimates $\widehat{\mathbf{Z}}_{\mathcal{H}_1}$ and $\widehat{\mathbf{Z}}_{\mathcal{H}_2}$, each based on hold-out size $n_h \geqslant 4 \cdot \frac{\log(8|\mathcal{X}|^2/\delta)}{(1-\gamma)^2}$. For any positive semi-definite matrix $\mathbf{M}$, we have*

$$\|(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_1})^{-1}\mathbf{M}(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_2})^{-\top}\|_{\text{diag}} \leqslant 3\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\mathbf{M}(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}}, \qquad (41)$$

with probability at least $1 - \delta$.

This lemma is a "reversed" version of Lemma 5 needed for this proof; see Section 5.1.4 for its proof.

Overall, the argument is similar to that of bound (15a). We have

$$\|\widehat{\mathbf{\Sigma}}(\widehat{V})\|_{\text{diag}}^{\frac{1}{2}} = \|(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_1})^{-1}\widehat{\mathbf{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_2})^{-\top}\|_{\text{diag}}^{\frac{1}{2}}$$

$$\stackrel{(i)}{\leqslant} \sqrt{3} \cdot \|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\mathbf{\Sigma}}_{\text{val}}(\widehat{V})(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}}^{\frac{1}{2}}$$

$$\stackrel{(ii)}{\leqslant} \sqrt{6} \cdot \|(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\mathbf{\Sigma}}_{\text{val}}(V^\star)(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}}^{\frac{1}{2}} + \frac{\sqrt{24}}{1 - \gamma} \cdot \|\widehat{V} - V^\star\|_\infty$$

$$\stackrel{(iii)}{\leqslant} \sqrt{6} \cdot \|(\mathbf{I} - \gamma\mathbf{P})^{-1}\mathbf{\Sigma}_{\text{val}}^\star(V^\star)(\mathbf{I} - \gamma\mathbf{P})^{-\top}\|_{\text{diag}}^{\frac{1}{2}}$$

$$+ \frac{b(V^\star)}{1 - \gamma} \cdot \sqrt{\frac{48\log(|\mathcal{X}|/\delta)}{n - 1}} + \frac{\sqrt{24}}{1 - \gamma} \cdot \|\widehat{V} - V^\star\|_\infty$$

with probability at least $1 - 2\delta$. Inequality (i) follows from Lemma 5, inequality (ii) follows from the proof of bound (35), and inequality (iii) follows from Lemma 6.

Recall the definition

$$\mathcal{E}_n(\widehat{V}, \mathcal{D}, \delta) = \frac{2\sqrt{6} \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \|\widehat{\boldsymbol{\Sigma}}(\widehat{V})\|_{\mathrm{diag}}^{1/2} + \frac{2\varphi_s(\delta)}{n} + \frac{6b(\widehat{V}_n)}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1}.$$

Substituting the above in gives

$$\mathcal{E}_n(\widehat{V}, \mathcal{D}, \delta) \leqslant \frac{12 \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\mathrm{diag}}^{1/2} + \frac{2\sqrt{6} \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \frac{b(V^{\star})}{1 - \gamma} \cdot \sqrt{\frac{48 \log(|\mathcal{X}|/\delta)}{n - 1}}$$
$$+ \frac{24 \cdot \varphi_f(\delta)}{\sqrt{n}(1 - \gamma)} \cdot \|\widehat{V} - V^{\star}\|_\infty + \frac{2\varphi_s(\delta)}{n} + \frac{6b(\widehat{V}_n)}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1}.$$

Define the shorthand

$$T_n \coloneqq \frac{24 \cdot \varphi_f(\delta)}{\sqrt{n}(1 - \gamma)} \cdot \|\widehat{V} - V^{\star}\|_\infty + \frac{6b(\widehat{V}_n)}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1}.$$

We take as given for now that

$$T_n \leqslant \frac{2\varphi_f(\delta)}{\sqrt{n}} \cdot \|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\mathrm{diag}}^{1/2} + \frac{2\varphi_s(\delta)}{n} + \frac{6b(V^{\star})}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1}. \tag{42}$$

With this, we conclude

$$\mathcal{E}_n(\widehat{V}, \mathcal{D}, \delta) \leqslant \frac{14 \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\mathrm{diag}}^{1/2} + \varphi_f(\delta) \cdot \frac{40b(V^{\star})}{1 - \gamma} \cdot \frac{\sqrt{\log(|\mathcal{X}|/\delta)}}{n - 1} + \frac{4\varphi_s(\delta)}{n},$$

as desired.

Proof of Equation (42): Using the fact that $b(\cdot)$ is 1-Lipschitz in the ℓ_∞-norm, we have

$$\frac{6b(\widehat{V}_n)}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1} \leqslant \frac{6b(V^{\star})}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1}$$
$$+ \|\widehat{V}_n - V^{\star}\|_\infty \left(\frac{6}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1} \right).$$

Since $n \geqslant \varphi_f^2(\delta) \cdot \frac{24 \cdot 8 \cdot \log(8|\mathcal{X}|/\delta)}{(1 - \gamma)^2}$, we have

$$\frac{24 \cdot \varphi_f(\delta)}{\sqrt{n}(1 - \gamma)} + \left(\frac{6}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1} \right) \leqslant 2,$$

which implies

$$T_n \leqslant 2 \cdot \|\widehat{V} - V^{\star}\|_\infty + \frac{6b(V^{\star})}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1}$$
$$\leqslant \frac{2\varphi_f(\delta)}{\sqrt{n}} \cdot \|\boldsymbol{\Sigma}^{\star}(r, \mathbf{P}, V^{\star})\|_{\mathrm{diag}}^{1/2} + \frac{2\varphi_s(\delta)}{n} + \frac{6b(V^{\star})}{1 - \gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta)}}{n - 1},$$

using the bound (13).

5.1.3 Proof of Lemma 4

For any index $i \in [|\mathcal{X}|]$, we define the random variable of interest

$$U_i := e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathtt{val}}(V^\star)(\mathbf{I} - \gamma\mathbf{P})^{-1}e_i.$$

With this definition, substituting the expression for the matrix $\widehat{\boldsymbol{\Sigma}}_{\mathtt{val}}(V^\star)$ yields

$$U_i = \frac{1}{n(n-1)} \sum_{1 \leqslant j < k \leqslant n} \left(e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}(R_j - R_k + \gamma(\mathbf{Z}_j - \mathbf{Z}_k)V^\star)\right)^2$$

$$\overset{(i)}{\leqslant} \frac{2}{n(n-1)} \sum_{1 \leqslant j < k \leqslant n} \left(e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}(R_j - R_k + \gamma(\mathbf{Z}_j - \mathbf{Z}_k)\widehat{V})\right)^2$$

$$+ \frac{2\gamma^2}{n(n-1)} \sum_{1 \leqslant j < k \leqslant n} \left(e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}(\mathbf{Z}_j - \mathbf{Z}_k)(V^\star - \widehat{V})\right)^2$$

$$\overset{(ii)}{\leqslant} 2e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathtt{val}}(\widehat{V})(\mathbf{I} - \gamma\mathbf{P})^{-1}e_i + \frac{8}{(1-\gamma)^2} \cdot \|\widehat{V} - V^\star\|_\infty^2.$$

Here step (i) uses the fact that $(a + b)^2 \leqslant 2a^2 + 2b^2$ for scalars a, b, whereas step (ii) follows from the operator norm bound $\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\|_{1,\infty} \leqslant \frac{1}{1-\gamma}$ and Hölder's inequality. This completes the proof of the bound (35).

5.1.4 Proof of Lemmas 5 and 7

We begin by observing that the $\|\cdot\|_{\mathrm{diag}}$ operator, despite not being a norm, satisfies the triangle inequality. Indeed, for square matrices $\mathbf{A}$ and $\mathbf{B}$ with matching dimensions, we have

$$\begin{aligned}
\|\mathbf{A} + \mathbf{B}\|_{\mathrm{diag}} &= \max_i |e_i^T(\mathbf{A} + \mathbf{B})e_i| \\
&\leqslant \max_i \left(|e_i^T\mathbf{A}e_i| + |e_i^T\mathbf{B}e_i|\right) \\
&\leqslant \max_i |e_i^T\mathbf{A}e_i| + \max_i |e_i^T\mathbf{B}e_i| \\
&= \|\mathbf{A}\|_{\mathrm{diag}} + \|\mathbf{B}\|_{\mathrm{diag}}.
\end{aligned} \tag{43}$$

We also require the following simple lemma:

Lemma 8 *For holdout sample size satisfying $n_h \geqslant \frac{C\log(8|\mathcal{X}|/\delta)}{(1-\gamma)^2}$ and any vector $V \in \mathbb{R}^{|\mathcal{X}|}$, we have*

$$\|(\widehat{\mathbf{Z}}_{\mathcal{H}_1} - \mathbf{P})V\|_\infty \leqslant \frac{\sqrt{2}(1-\gamma)}{\sqrt{C}}\|V\|_\infty,$$

with probability at least $1 - \frac{\delta}{4}$.

Proof From Hoeffding's inequality, we have

$$[(\widehat{\mathbf{Z}}_{\mathcal{H}_1} - \mathbf{P}) \cdot V]_i = \frac{1}{n_h} \sum_{j=1}^{n_h} \langle \mathbf{Z}_j(i) - p_i, V \rangle \leqslant \sqrt{\frac{2\log(8|\mathcal{X}|/\delta)}{n_h}} \cdot \|V\|_\infty, \tag{44}$$

with probability at least $1 - \frac{\delta}{4|\mathcal{X}|}$. Here, $\mathbf{Z}_j(i)$ and p_i, respectively denote the ith row vector of the matrices $\mathbf{Z}_j$ and $\mathbf{P}$. Now, applying a union bound over $|\mathcal{X}|$ coordinates and using the lower bound on the holdout sample size n_h yields the claim of Lemma 8. ∎

We are now ready to prove the two lemmas. For ease of notation, we use the shorthands

$$\mathbf{A} = \mathbf{I} - \gamma\mathbf{P}, \quad \widetilde{\mathbf{A}} = \mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_1}, \quad \text{and} \quad \widehat{\mathbf{A}} := \mathbf{I} - \gamma\widehat{\mathbf{Z}}_{\mathcal{H}_2}.$$

In light of the triangle inequality (43), for Lemma 5 it suffices to prove that

$$\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\mathrm{diag}} \leqslant \frac{2}{3}\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}}$$

with probability at least $1 - \delta$. This would then imply

$$\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}} \leqslant 3\|\widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\mathrm{diag}},$$

as desired. Similarly, for Lemma 7 it suffices to show

$$\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\mathrm{diag}} \leqslant 2\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}}$$

with probability exceeding $1 - \delta$.

Simple algebra and another application of the triangle inequality (43) yields

$$
\begin{aligned}
\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\mathrm{diag}} &\leqslant \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}} + \|\widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})\|_{\mathrm{diag}} \\
&\leqslant \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}} + \|(\mathbf{A}^{-1} - \widehat{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}} \\
&\quad + \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})\|_{\mathrm{diag}} \\
&= \sum_{j=1}^{3} T_j,
\end{aligned}
\tag{45}
$$

where we define

$$T_1 := \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}}, \qquad T_2 := \|(\mathbf{A}^{-1} - \widehat{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}}, \qquad \text{and}$$
$$T_3 := \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})\|_{\mathrm{diag}}.$$

We bound these three terms individually.

Bounding T_1: We have

$$
\begin{aligned}
T_1 = \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}} &= \max_i |e_i^T \widetilde{\mathbf{A}}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}e_i| \\
&\leqslant \max_i \|\widetilde{\mathbf{A}}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}e_i\|_{\infty} \\
&\leqslant \sqrt{\frac{2}{C}} \cdot \max_i \|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}e_i\|_{\infty} \\
&= \frac{1}{\sqrt{12}} \cdot \max_{i,j} |e_j^T \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}e_i| \\
&\leqslant \frac{1}{\sqrt{12}}\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}},
\end{aligned}
$$

with probability at least $1 - \frac{\delta}{4}$. The third line follows from Lemma 8, a union bound over D coordinates and the operator norm bound $\|\mathbf{A}^{-1}\|_{1,\infty} = \|(\mathbf{I} - \gamma\mathbf{P})^{-1}\|_{1,\infty} \leqslant \frac{1}{1-\gamma}$. The final inequality follows from the fact that $\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}$ is a covariance matrix and the maximum entry of a covariance matrix is same as the maximum diagonal entry. Indeed, for a zero-mean random vector X with covariance matrix $\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}$, we have for all indices i, j,

$$|e_j^T \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} e_i| = |\mathrm{Cov}(X_j, X_i)| \leqslant \sqrt{\mathbb{E}[X_j^2] \cdot \mathbb{E}[X_i^2]} \leqslant \max_k \mathrm{Var}(X_k) = \|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}},$$

as claimed.

Bounding T_2: An identical calculations for the term T_2 in the bound (45) yields

$$\|(\mathbf{A}^{-1} - \widehat{\mathbf{A}}^{-1})\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}} \leqslant \sqrt{\frac{2}{C}}\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}}$$

with probability at least $1 - \frac{\delta}{4}$.

Bounding T_3: We have

$$\begin{aligned}
\|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})\|_{\mathrm{diag}} &= \max_i |e_i^T(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})e_i| \\
&\leqslant \max_i \|(\mathbf{A}^{-1} - \widetilde{\mathbf{A}}^{-1})\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})e_i\|_\infty \\
&= \max_i \|\widetilde{\mathbf{A}}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\mathbf{A}^{-1}\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})e_i\|_\infty \\
&\overset{(a)}{\leqslant} \sqrt{\frac{2}{C}} \cdot \max_i \|\mathbf{A}^{-1}\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-\top})e_i\|_\infty \\
&= \sqrt{\frac{2}{C}} \cdot \max_{i,j} |e_j^T \mathbf{A}^{-1}\mathbf{\Sigma}(\mathbf{A}^{-\top} - \widehat{\mathbf{A}}^{-\top})e_i| \\
&= \sqrt{\frac{2}{C}} \cdot \max_j \|(\mathbf{A}^{-1} - \widehat{\mathbf{A}}^{-\top})\mathbf{\Sigma}\mathbf{A}^{-\top}e_j\|_\infty \\
&\overset{(b)}{\leqslant} \frac{2}{C} \cdot \max_j \|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}e_j\|_\infty \\
&\leqslant \frac{2}{C}\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}},
\end{aligned}$$

Inequality (a) follows from Lemma 8 conditioned on the randomness of the matrix $\widehat{\mathbf{A}}$; recall that matrices $\widehat{\mathbf{A}}$ and $\widetilde{\mathbf{A}}$ are independent by construction. Inequality (b) follows again by applying the Lemma 8. Invoking a union bound, we have that the last bound holds with probability at least $1 - \frac{\delta}{2}$. Thus, we conclude with probability exceeding $1 - \delta$,

$$\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\mathrm{diag}} \leqslant \left(2\sqrt{\frac{2}{C}} + \frac{2}{C}\right)\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\mathrm{diag}},$$

for $n_h \geqslant \frac{C\log(8|\mathcal{X}|/\delta)}{(1-\gamma)^2}$.

Putting together the pieces, we have for $C = 24$ in Lemma 5

$$\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\text{diag}} \leqslant \left(\frac{2}{\sqrt{12}} + \frac{1}{12}\right)\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\text{diag}} \leqslant \frac{2}{3}\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\text{diag}},$$

with probability exceeding $1 - \delta$. Additionally for $C = 4$ in Lemma 7 we have

$$\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top} - \widetilde{\mathbf{A}}^{-1}\mathbf{\Sigma}\widehat{\mathbf{A}}^{-\top}\|_{\text{diag}} \leqslant \left(\sqrt{2} + \frac{1}{2}\right)\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\text{diag}} \leqslant 2\|\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-\top}\|_{\text{diag}},$$

with probability exceeding $1 - \delta$.

5.1.5 PROOF OF LEMMA 6

Like previously, we use the shorthand

$$U_i := e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}\widehat{\mathbf{\Sigma}}_{\text{val}}(V^\star)(\mathbf{I} - \gamma\mathbf{P})^{-T}e_i$$
$$= \frac{1}{n(n-1)}\sum_{1 \leqslant j < k \leqslant n}\left(e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}((R_j - R_k) + \gamma(\mathbf{Z}_j - \mathbf{Z}_k)V^\star)\right)^2.$$

A straightforward calculation yields

$$\mathbb{E}[U_i] = e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}\mathbf{\Sigma}_{\text{val}}^\star(V^\star)(\mathbf{I} - \gamma\mathbf{P})^{-T}e_i.$$

Define the random variable

$$V_j := e_i^T(\mathbf{I} - \gamma\mathbf{P})^{-1}(R_j + \gamma\mathbf{Z}_jV^\star),$$

we then have $U_i = \frac{1}{n(n-1)}\sum_{1 \leqslant j < k \leqslant n}(V_j - V_k)^2$. Thus, applying the empirical Bernstein lemma 10 along with a union bound we get

$$\left|\sqrt{U_i} - \sqrt{\mathbb{E}[U_i]}\right| \leqslant \frac{2b(V^\star)}{1 - \gamma} \cdot \sqrt{\frac{2\log(|\mathcal{X}|/\delta)}{n - 1}}$$

with probability exceeding $1 - \delta$ for all i. The claim then follows from the fact that

$$\left|\max_i \sqrt{U_i} - \max_i \sqrt{\mathbb{E}[U_i]}\right| \leqslant \max_i \left|\sqrt{U_i} - \sqrt{\mathbb{E}[U_i]}\right|.$$

5.2 Proof of Theorem 3

The proof of this theorem is similar to the proof of Theorem 1, and it is based on a Lipschitz property of the empirical covariance matrix operator $\|\widehat{\mathbf{\Sigma}}_{\text{opt}}(\cdot)\|_{\text{diag}}^{\frac{1}{2}}$ (see Lemma 9) and an empirical Bernstein lemma (see Lemma 10). For notational simplicity, we use $\widehat{Q}$ as a shorthand for $\widehat{Q}_n$, and similarly, $\widehat{\mathbf{\Sigma}}_{\text{opt}}(\widehat{Q})$ in place of $\widehat{\mathbf{\Sigma}}_{\text{opt}}(\widehat{Q}_n, \mathcal{D}_n)$.

It suffices to prove that the following two bounds hold, each with probability at least $1 - \delta$:

$$\|\mathbf{\Sigma}_{\text{opt}}^\star(Q^\star)\|_{\text{diag}}^{\frac{1}{2}} \leqslant \|\widehat{\mathbf{\Sigma}}_{\text{opt}}(\widehat{Q})\|_\infty^{\frac{1}{2}} + b(\widehat{Q}) \cdot \sqrt{\frac{8\log(D/\delta)}{n - 1}} + \|\widehat{Q} - Q^\star\|_\infty \cdot \left(\sqrt{\frac{8\log(D/\delta)}{n-1}} + \sqrt{8}\right) \tag{46a}$$

$$\|\widehat{\mathbf{\Sigma}}_{\text{opt}}(\widehat{Q})\|_\infty^{\frac{1}{2}} \leqslant \sqrt{2} \cdot \|\mathbf{\Sigma}_{\text{opt}}^\star(Q^\star)\|_{\text{diag}}^{\frac{1}{2}} + b(Q^\star) \cdot \sqrt{\frac{16\log(D/\delta)}{n-1}} + \sqrt{8}\|\widehat{Q} - Q^\star\|_\infty, \tag{46b}$$

30

Indeed, combining the bound (46a) with the condition (31) of the algorithm $\mathcal{A}_{\texttt{PolOpt}}$ and the sample size lower bound $n \geqslant c_1(\mathcal{A}_{\texttt{Eval}}, \delta)^2 \cdot \frac{32 \cdot \log(D/\delta)}{(1-\gamma)^2}$, we have

$$\|\widehat{Q} - Q^\star\|_\infty \leqslant \frac{2\sqrt{2} \cdot \varphi_f(\delta)}{\sqrt{n}} \cdot \frac{\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q})\|_{\mathrm{diag}}^{\frac{1}{2}}}{1-\gamma} + \frac{2c_1(\mathcal{A}_{\texttt{Eval}})b(\widehat{Q})}{1-\gamma} \cdot \frac{\sqrt{16 \log(D/\delta)}}{n-1} + \frac{2\varphi_s(\delta)}{n}.$$

with probability at least $1 - 2\delta$. Furthermore, since function b is γ-Lipschitz in the ℓ_∞-norm, we have

$$|b(Q) - b(Q^\star)| \leqslant \gamma \cdot \|Q - Q^\star\|_\infty.$$

Combining this inequality with the bound (46b) and the condition (31) yields the claimed bound (33b).

It remains to prove the bounds (46a) and (46b). In doing so, we make use of the following auxiliary lemma:

Lemma 9 *For any pair Q_1, Q_2, we have*

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q_1)\|_{\mathrm{diag}}^{\frac{1}{2}} \leqslant \sqrt{2}\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q_2)\|_{\mathrm{diag}}^{\frac{1}{2}} + \sqrt{8}\|Q_1 - Q_2\|_\infty.$$

See Section 5.2.3 for the proof of this claim.

5.2.1 PROOF OF THE BOUND (46a)

We have

$$\|\boldsymbol{\Sigma}_{\mathrm{opt}}^\star(Q^\star)\|_{\mathrm{diag}}^{\frac{1}{2}} \overset{(i)}{\leqslant} \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q^\star)\|_\infty^{\frac{1}{2}} + \frac{b(Q^\star)}{1-\gamma} \cdot \sqrt{\frac{8 \log(D/\delta)}{n-1}}$$

$$\overset{(ii)}{\leqslant} \sqrt{2}\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q})\|_{\mathrm{diag}}^{\frac{1}{2}} + \sqrt{8} \cdot \|\widehat{Q} - Q^\star\|_\infty + b(\widehat{Q}) \cdot \sqrt{\frac{8 \log(D/\delta)}{n-1}}$$

$$+ \|\widehat{Q} - Q^\star\|_\infty \cdot \sqrt{\frac{8 \log(D/\delta)}{n-1}},$$

with probability at least $1 - \delta$. Inequality (i) follows by applying an empirical Bernstein bound (cf. Lemma 10 in Appendix A.1) to each diagonal entry of the matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(V^\star)$, combined with a union bound on all D diagonal entries. (See the proof of Theorem 1 for an analogous calculation.) Inequality (ii) follows Lemma 9 on Lipschitz properties of the empirical variance estimate.

5.2.2 PROOF OF THE BOUND (46b)

The proof of this claim is similar to that of the bound (46a). We have

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(\widehat{Q})\|_{\mathrm{diag}}^{\frac{1}{2}} \overset{(i)}{\leqslant} \sqrt{2} \cdot \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q^\star)\|_{\mathrm{diag}}^{\frac{1}{2}} + \sqrt{8}\|\widehat{Q} - Q^\star\|_\infty$$

$$\overset{(ii)}{\leqslant} \sqrt{2} \cdot \|\boldsymbol{\Sigma}_{\mathrm{opt}}^\star(Q^\star)\|_{\mathrm{diag}}^{\frac{1}{2}} + b(Q^\star) \cdot \sqrt{\frac{16 \log(D/\delta)}{n-1}} + \sqrt{8} \cdot \|\widehat{Q} - Q^\star\|_\infty$$

with probability at least $1 - \delta$. Here the inequality (i) follows from the Lemma 9, and inequality (ii) follows from an empirical Bernstein bound (see Lemma 10 in Appendix A.1) combined with the union bound.

5.2.3 Proof of Lemma 9

Given a square matrix $\boldsymbol{\Sigma}$, the quantity $\|\boldsymbol{\Sigma}\|_{\mathrm{diag}}$ only depends on the diagonal elements of the matrix $\boldsymbol{\Sigma}$. For any state-action pair $z = (x, u)$, let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q)(z)$ denote the diagonal entry of the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q)$ associated with the state action pair z. Substituting the definition of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q)$ from equation (34) yields

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q_1)(z) &= \tfrac{1}{n(n-1)} \sum_{1 \leqslant i < \ell \leqslant n} \left(\widehat{\boldsymbol{\mathcal{J}}}_i(Q_1)(z) - \widehat{\boldsymbol{\mathcal{J}}}_\ell(Q_1)(z) \right)^2 \\
&= \tfrac{1}{n(n-1)} \sum_{1 \leqslant i < \ell \leqslant n} \left(\widehat{\boldsymbol{\mathcal{J}}}_i(Q_2)(z) - \widehat{\boldsymbol{\mathcal{J}}}_\ell(Q_2)(z) + \widehat{\boldsymbol{\mathcal{J}}}_i(Q_1)(z) - \widehat{\boldsymbol{\mathcal{J}}}_i(Q_2)(z) + \widehat{\boldsymbol{\mathcal{J}}}_\ell(Q_2)(z) - \widehat{\boldsymbol{\mathcal{J}}}_\ell(Q_1)(z) \right)^2 \\
&\overset{(i)}{\leqslant} 2\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q_2)(z) + \tfrac{4}{n(n-1)} \sum_{1 \leqslant i < \ell \leqslant n} \left\{ ((\widehat{\boldsymbol{\mathcal{J}}}_i(Q_2) - \widehat{\boldsymbol{\mathcal{J}}}_i(Q_1))(z))^2 + ((\widehat{\boldsymbol{\mathcal{J}}}_\ell(Q_2) - \widehat{\boldsymbol{\mathcal{J}}}_\ell(Q_1)(z))^2 \right\} \\
&\overset{(ii)}{\leqslant} 2\widehat{\boldsymbol{\Sigma}}_{\mathrm{opt}}(Q_2)(z) + 8\gamma^2 \|Q_1 - Q_2\|_\infty^2,
\end{aligned}
$$

where step (i) follows from the elementary inequality $(a + b)^2 \leqslant 2a^2 + 2b^2$; and step (ii) uses the fact that the noisy Bellman operator $\widehat{\boldsymbol{\mathcal{J}}}_\ell$ is γ-Lipschitz in the ℓ_∞-norm. This completes the proof.

6. Discussion

Our work addresses the problem of obtaining instance-dependent confidence regions for the policy evaluation problem and the optimal value estimation problem of an MDP, given access to an instance optimal algorithm. The confidence regions are constructed by estimating the instance-dependent functionals that control problem difficulty in a local neighborhood of the given problem instance. For both problems, the instance-dependent confidence regions are shown to be significantly shorter for problems with favorable structure.

Our results also leave a few interesting questions. For instance, one could be improving the bound from Theorem 3 for the optimal value estimation problem. Additionally in the setting of policy evaluation, we believe that the need for two independent holdout sets is unnecessary and more likely a proof defect; we conjecture it suffices to use one holdout set to estimate $(\mathbf{I} - \gamma \mathbf{P})^{-1}$ on both the left and right side. However what is most interesting is extending these results to more complicated settings, such as in scenarios involving function approximation or different sampling schemes such as offline RL.

Acknowledgments

Appendix A. Auxiliary Lemmas

In this section, we state the auxiliary lemmas that are used in the main part of the paper.

A.1 Empirical Bernstein

The following empirical Bernstein bound is a re-statement of Theorem 10 from Maurer and Pontil (2009):

Lemma 10 *Let $\{Z_i\}_{i=1}^n$ be an i.i.d. sequence of real valued random variables taking values in the unit interval $[0,1]$, and define the variance estimate $\widehat{\boldsymbol{\Sigma}}(Z) = \frac{1}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} (Z_i - Z_j)^2$. Then for any $\delta \in (0,1)$, we have*

$$\left| \sqrt{\mathbb{E}[\widehat{\boldsymbol{\Sigma}}(Z)]} - \sqrt{\widehat{\boldsymbol{\Sigma}}(Z)} \right| \leqslant \sqrt{\frac{2 \log(1/\delta)}{n-1}} \tag{47}$$

with probability at least $1 - 2\delta$.

A.2 Calculations for Example 1

Here we derive the bound (30). Letting $Q^\star$ denote the value function of the optimal policy $\pi^\star$, we have

$$(\mathcal{Z}^{\pi^\star} - \boldsymbol{\mathcal{P}}^{\pi^\star})Q^\star = \left[\begin{array}{cc} \vert & \vert \\ (\mathcal{Z}_{u_1} - \boldsymbol{\mathcal{P}}_{u_1})V^\star & 0 \\ \vert & \vert \end{array} \right]. \tag{48}$$

Letting $\mathbf{W} = (\mathbf{I} - \gamma\boldsymbol{\mathcal{P}}_{u_1})^{-1}(\mathcal{Z}_{u_1} - \boldsymbol{\mathcal{P}}_{u_1})Q_{\pi^\star}$ and solving for $(\mathbf{I} - \gamma\boldsymbol{\mathcal{P}}^{\pi^\star})\mathbf{Y} = \gamma(\mathcal{Z}^{\pi^\star} - \boldsymbol{\mathcal{P}}^{\pi^\star})Q^\star$ gives

$$\mathbf{Y} = \gamma \cdot \left[\begin{array}{cc} \vert & \vert \\ \mathbf{W} & \gamma\mathbf{W} \\ \vert & \vert \end{array} \right]. \tag{49}$$

Finally, a simple calculation yields

$$\mathrm{Var}(\widehat{\boldsymbol{\mathcal{g}}}(Q^\star)(x_1, u_1)) = p(1-p) \cdot \frac{(1-\tau)^2}{(1-\gamma p)^2}, \qquad \mathrm{Var}(\widehat{\boldsymbol{\mathcal{g}}}(Q^\star)(x_2, u_1)) = 0,$$

$$\mathrm{Var}(\widehat{\boldsymbol{\mathcal{g}}}(Q^\star)(x_1, u_2)) = 0, \quad \text{and} \quad \mathrm{Var}(\widehat{\boldsymbol{\mathcal{g}}}(Q^\star)(x_2, u_2)) = 0.$$

Substituting $\tau = 1 - (1-\gamma)^\lambda$ yields the claimed bound.

Appendix B. Proof of Corollary 2

The claim of Corollary 2 follows from Corollary 11, the condition (21) and the fact that the functions φ_f and φ_s are lower bounded by 1.

Corollary 11 *Given any algorithm $\mathcal{A}_{\mathtt{Eval}}$ satisfying condition (13), target accuracy ϵ, and tolerance probability δ_0. Let $\widehat{V}$ denote the output of algorithm EmpIRE with input pair (ϵ, δ_0). Then the following statements hold:*

(a) *The estimate $\widehat{V}$ is ϵ-accurate in the ℓ_∞-norm:*

$$\|\widehat{V} - V^\star\|_\infty \leqslant \epsilon, \tag{50a}$$

with probability exceeding $1 - \delta_0$.

(b) *Define the non-negative integer sets*

$$A = \left\{ m : m + \log_2(\log(4|\mathcal{X}|/\delta_m)) \geqslant \log_2\left(\frac{(1-\gamma)^2}{\epsilon^2} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}} \right) + 4 \right\}, \qquad \text{and}$$

$$B = \{ m : m + \log_2(\log(4|\mathcal{X}|/\delta_m)) \geqslant \tag{50b}$$

$$\log_2\left(\frac{(1-\gamma)^2}{\epsilon \varphi_f^2(\delta_m)} \cdot \left[\frac{\varphi_s(\delta_m)}{4} + \frac{64 b(V^\star)}{1-\gamma} \cdot \sqrt{\log(8|\mathcal{X}|/\delta_m)} \right] \right) \right\}. \tag{50c}$$

The algorithm EmpIRE terminates in at most

$$M = \inf \{A \cap B\} \tag{50d}$$

epochs with probability exceeding $1 - \delta_0$.

(c) *For universal constants (c_1, c_2, c_3), the algorithm EmpIRE requires at most*

$$n \leqslant \max\left\{ \frac{c_1 \varphi_f^2(\delta_M)}{\epsilon^2} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}, \frac{1}{\epsilon} \left[c_2 \varphi_s(2\delta_M) + c_3 \frac{b(V^\star)}{1-\gamma} \sqrt{\log(4|\mathcal{X}|/\delta_M)} \right] \right\} \tag{50e}$$

samples with probability exceeding $1 - \delta_0$.

Remarks: Note that in instance-optimal algorithms $\varphi_f(\delta)$ and $\varphi_f(\delta)$ are typically logarithmic functions of $1/\delta$ such as variance-reduced policy evaluation (Khamaru et al., 2021) or ROOT-SA (Mou et al., 2022), ensuring that M exists and our algorithm terminates. In both equations (50a) and (50b), typically the first term involving $\frac{\|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}}{\epsilon^2}$ is the dominant term and establishes a sample complexity of $\mathcal{O}\left(\frac{\|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}}{\epsilon^2} \right)$.

Proof of Corollary 11: Taking for granted now that the algorithm terminates, observe that equation (50a) follows from algorithm's termination in M epochs, applying a union bound over equation (15a) from Theorem 1 for epochs $m = 1, \dots, M$ and equation (15b) for the final epoch M yields the claim with probability exceeding $1 - \delta_0$.

We now turn to the claim (50b). Note that the algorithm terminates at epoch m only if

$$\hat{\varepsilon}_f + \hat{\varepsilon}_s = \frac{2\sqrt{6} \cdot \varphi_f(\delta_m)}{\sqrt{N_m}} \cdot \widehat{\mathbf{\Sigma}}(\widehat{V}, \mathcal{D}) + \frac{2\varphi_s(\delta_m)}{N_m} + \frac{6b(\widehat{V})}{1-\gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta_m)}}{N_m - 1} \leqslant \epsilon.$$

By equation (15b), we have

$$\hat{\varepsilon}_f + \hat{\varepsilon}_s \leqslant \frac{13\varphi_f(\delta_m)}{\sqrt{N_m}} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\text{diag}}^{\frac{1}{2}} + \frac{2\varphi_s(\delta_m)}{N_m} + \frac{33b(V^\star)}{1-\gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta_m)}}{N_m}.$$

Consequently, the stopping condition $\hat{\varepsilon}_f + \hat{\varepsilon}_s \leqslant \epsilon$ holds as long as

$$\frac{13\varphi_f(\delta_m)}{\sqrt{N_m}} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}^{\frac{1}{2}} \leqslant \frac{\epsilon}{2}, \qquad \text{and} \tag{51a}$$

$$\frac{2\varphi_s(\delta_m)}{N_m} + \frac{33b(V^\star)}{1-\gamma} \cdot \frac{\sqrt{\log(8|\mathcal{X}|/\delta_m)}}{N_m} \leqslant \frac{\epsilon}{2}. \tag{51b}$$

Since $N_m = 2^m \varphi_f^2(\delta_m) \cdot \frac{32 \log(4|\mathcal{X}|/\delta_m)}{(1-\gamma)^2}$ by definition, equation (50b) follows from plugging in N_m and solving for when the above expression is less than ϵ. The termination condition comes from M being the first m such that both conditions above hold.

When the algorithm terminates in M epochs, then the total number of samples used can be bounded as

$$\sum_{m=1}^M (N_m + 2h_m) \leqslant c_1 \cdot \frac{2^M}{(1-\gamma)^2} \cdot \varphi_f^2(\delta_M) \cdot \log(4|\mathcal{X}|/\delta_M) \;=\; c_1 \cdot \frac{2^{M+\log_2(\log(4|\mathcal{X}|/\delta_M))}}{(1-\gamma)^2} \cdot \varphi_f^2(\delta_M). \tag{52}$$

for some universal constant c_1. Here, we have used the assumption that the maps $\delta \mapsto \varphi_f(\delta)$ and $\delta \mapsto \varphi_s(\delta_m)$ are increasing functions of δ; consequently for all $m = 1, 2, \ldots, M$,

$$\varphi_f(\delta_m) \leqslant \varphi_f(\delta_M) \qquad \text{and} \qquad \varphi_s(\delta_m) \leqslant \varphi_s(\delta_M).$$

Recall that of M is the infimum of the two sets A and B, defined in (50b), thus any integer smaller than M, in particular $M-1$ satisfies the following upper bound

$$M - 1 + \log_2\big(\log(4|\mathcal{X}|/\delta_{M-1})\big)$$
$$\leqslant c_2 + \max\left\{\log_2\left(\frac{(1-\gamma)^2}{\epsilon^2} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}\right),\right.$$
$$\left.\log_2\left(\frac{(1-\gamma)^2}{\epsilon\varphi_f^2(\delta_{M-1})} \cdot \left[c_3\varphi_s(\delta_{M-1}) + c_4\frac{b(V^\star)}{1-\gamma} \cdot \sqrt{\log(8|\mathcal{X}|/\delta_{M-1})}\right]\right)\right\}$$

for some universal constants (c_2, c_3, c_4). Rewriting the last bound in terms of M and using the relation $\delta_{M-1} = 2\delta_M$ we obtain

$$M + \log_2\big(\log(2|\mathcal{X}|/\delta_M)\big)$$
$$\leqslant c_2 + \max\left\{\log_2\left(\frac{(1-\gamma)^2}{\epsilon^2} \cdot \|\mathbf{\Sigma}^\star(r, \mathbf{P}, V^\star)\|_{\mathrm{diag}}\right),\right.$$
$$\left.\log_2\left(\frac{(1-\gamma)^2}{\epsilon\varphi_f^2(2\delta_M)} \cdot \left[c_3\varphi_s(2\delta_M) + c_4\frac{b(V^\star)}{1-\gamma} \cdot \sqrt{\log(4|\mathcal{X}|/\delta_M)}\right]\right)\right\}.$$

Substituting the last bound on M into equation (52) and using the fact that $\delta \mapsto \varphi_f(\delta)$ is a decreasing function of δ yields equation (50e). This completes the proof of Corollary 11.

Appendix C. Policy Optimization: Further Details

In this section, we state and prove a bound for the variance reduced Q-learning algorithm studied in papers (Wainwright, 2019b; Xia et al., 2021). The goal is to to show that the variance reduced Q-learning algorithm satisfies the condition in equation (31). Much of the content of this section is directly borrowed from the paper (Xia et al., 2021). Throughout this section, we use the shorthand $D := |\mathcal{X}| \cdot |\mathcal{U}|$.

C.1 Variance-reduced Q-learning

We start by restating the variance reduced Q-learning algorithm from Xia et al. (2021). See Xia et al. (2021); Wainwright (2019b) for a motivation of the algorithm.

A single epoch: The epochs are indexed with integers $m = 1, 2, \ldots, M$, where M corresponds to the total number of epochs to be run. Each epoch m requires the following four inputs:

- an element $\bar{Q}$, which is chosen to be the output of the previous epoch $m - 1$;

- a positive integer K denoting the number of steps within the given epoch;

- a positive integer B_m denoting the batch size used to calculate the Monte Carlo update:

$$\bar{\boldsymbol{\mathcal{J}}}_{n_m}(\bar{Q}_m) := \frac{1}{B_m} \sum_{i \in \mathcal{D}_m} \widehat{\boldsymbol{\mathcal{J}}}_i(\bar{Q}_m). \tag{53}$$

- a set of fresh operators $\{\widehat{\boldsymbol{\mathcal{J}}}_i\}_{i \in \mathcal{C}_m}$, with $|\mathcal{C}_m| = B_m + K$. The set $\mathcal{C}_m$ is partitioned into two subsets having sizes B_m and K, respectively. The first subset, of size B_m, which we call $\mathcal{D}_m$, is used to construct the Monte Carlo approximation (53). The second subset, of size K is used to run the K steps within the epoch.

We summarize a single epoch in pseudo code form in Algorithm SingleEpoch.
Overall algorithm: The overall algorithm, denoted by VR-QL for short, has five inputs: (a) an initialization $\bar{Q}_1$, (b) an integer M, denoting the number of epochs to be run, (c) an integer K, denoting the length of each epoch, (d) a sequence of batch sizes $\{B_m\}_{m=1}^K$, denoting the number of operators used for re-centering in the M epochs, and (e) sample batches $\{\{\widehat{\boldsymbol{\mathcal{J}}}_i\}_{i \in \mathcal{C}_m}\}_{m=1}^M$ to be used in the M epochs. Given these five inputs, the overall procedure can be summarized as in Algorithm VR-QL.

Input parameters Given a tolerance probability $\delta \in (0, 1)$ and the number of available i.i.d. samples n, we run Algorithm VR-QL with a total of $M := \log\left(\frac{n(1-\gamma)^2}{8 \log((16D/\delta) \cdot \log n)}\right)$ epochs, along with the following parameter choices:
`Re-centering sizes:`

$$n_m = c_1 \frac{4^m}{(1-\gamma)^2} \cdot \log_4(16MD/\delta) \tag{54a}$$

Algorithm SingleEpoch   RunEpoch $(\bar{Q}; K, B_m, \{\widehat{\boldsymbol{\jmath}}_i\}_{i \in \mathcal{C}_m})$

1: Given (a) Epoch length K, (b) Re-centering vector $\bar{Q}$, (c) Re-centering batch size B_m, (d) Operators $\{\widehat{\boldsymbol{\jmath}}_i\}_{i \in \mathcal{C}_m}$
2: Compute the re-centering quantity

$$\overline{\boldsymbol{\jmath}}_{n_m}(\bar{Q}) \coloneqq \frac{1}{B_m} \sum_{i \in \mathcal{D}_m} \widehat{\boldsymbol{\jmath}}_i(\bar{Q})$$

3: Initialize $Q_1 = \bar{Q}$
4: **for** $k = 1, 2, \ldots, K$ **do**
5:     Compute the variance-reduced update:

$$Q_{k+1} = (1 - \alpha_k)Q_k + \alpha_k \left\{ \widehat{\boldsymbol{\jmath}}_k(Q_k) - \widehat{\boldsymbol{\jmath}}_k(\bar{Q}_m) + \overline{\boldsymbol{\jmath}}_{n_m}(\bar{Q}_m) \right\}, \quad \text{with stepsize } \alpha_k = \frac{1}{1 + (1 - \gamma)k}.$$

6: **end for**
7: **return** θ_{K+1}

Algorithm VR-QL

1: Given (a) Initialization $\bar{Q}_1$, (b) Number of epochs, M, (c) Epoch length K, (d) Re-centering sample sizes $\{B_m\}_{m=1}^M$, (e) Sample batches $\{\widehat{\boldsymbol{\jmath}}_i\}_{i \in \mathcal{C}_m}$ for $m = 1, \ldots, M$
2: Initialize at $\bar{Q}_1$
3: **for** $m = 1, 2, \ldots, M$ **do**
4:     $\bar{Q}_{m+1} = \text{RunEpoch}(\bar{Q}_m; K, B_m, \{\widehat{\boldsymbol{\jmath}}_i\}_{i \in \mathcal{C}_m})$
5: **end for**
6: **return** $\bar{Q}_{M+1}$ as final estimate

`Sample batches:`

$$\text{Partition the } n \text{ samples to obtain } \{\widehat{\boldsymbol{\mathcal{J}}}_i\}_{i \in \mathcal{C}_m} \text{ for } m = 1, \ldots, M \tag{54b}$$

`Epoch length:`

$$K = \frac{n}{2M}. \tag{54c}$$

Proposition 12 *Suppose the inputs parameters of Algorithm* VR-QL *are chosen according to parameter choices* (54), *and the sample size satisfies the lower bound* $\frac{n}{M} \geqslant c_1 \frac{\log(8DM/\delta)}{(1-\gamma)^3}$. *Then, for any initialization* $\bar{Q}_1$, *the output* $\widehat{Q}_n \equiv \bar{Q}_{M+1}$ *satisfies the*

$$\|\widehat{Q}_n - Q^\star\|_\infty \leqslant c_1 \cdot \frac{\|\boldsymbol{\Sigma}^\star_{\text{opt}}(Q^\star)\|_{\text{diag}}^{\frac{1}{2}}}{1 - \gamma} \cdot \sqrt{\frac{\log(8DM/\delta)}{n}} + c_2 \cdot \frac{b(Q^\star)}{1 - \gamma} \cdot \frac{\log(8DM/\delta)}{n}$$
$$+ c_3 \cdot \|Q_1 - Q^\star\|_\infty \cdot \frac{\log^2((16DM/\delta) \cdot \log n)}{n^2(1-\gamma)^4},$$

with probability at least $1 - \delta$. *Here* c_1, c_2, c_3 *are universal constants.*

C.2 Proof of Proposition 12

The proof is this similar to that of the Theorem 2 in the paper (Xia et al., 2021); in particular, we use a modified version of Lemma 7 from the paper (Xia et al., 2021).

C.2.1 Proof set-up

We start by introducing some notation used in the paper (Xia et al., 2021). Recall that the update within an epoch takes the form (cf. SingleEpoch)

$$Q_{k+1} = (1 - \alpha_k)Q_k + \alpha_k \left\{ \widehat{\boldsymbol{\mathcal{J}}}_k(\theta) - \widehat{\boldsymbol{\mathcal{J}}}_k(\bar{Q}_m) + \overline{\boldsymbol{\mathcal{J}}}_{n_m}(\bar{Q}_m) \right\},$$

where $\bar{Q}_m$ represents the input into epoch m. We define the shifted operators and noisy shifted operators for epoch m by

$$\mathbf{J}(Q) = \boldsymbol{\mathcal{J}}(Q) - \boldsymbol{\mathcal{J}}(\bar{Q}_m) + \overline{\boldsymbol{\mathcal{J}}}_{n_m}(\bar{Q}_m) \quad \text{and} \quad \widehat{\mathbf{J}}_k(Q) = \widehat{\boldsymbol{\mathcal{T}}}_k(Q) - \widehat{\boldsymbol{\mathcal{J}}}_k(\bar{Q}_m) + \overline{\boldsymbol{\mathcal{J}}}_{n_m}(\bar{Q}_m). \tag{55}$$

Since both of the operators $\boldsymbol{\mathcal{J}}$ and $\widehat{\boldsymbol{\mathcal{J}}}_k$ are γ-contractive in the ℓ_∞-norm, the operators $\mathbf{J}$ and $\widehat{\mathbf{J}}_k$ are also γ-contractive operators in the same norm. Let $\widehat{Q}_m$ denote the unique fixed point of the operator $\mathbf{J}$.

With this set-up, it suffices to prove the following modification of Lemma 7 from Xia et al. (2021).

Lemma 13 *Assume that* n_m *satisfies the bound* $n_m \geqslant \frac{c\log(8DM/\delta)}{(1-\gamma)^2}$. *Then we have*

$$\|\widehat{Q}_m - Q^\star\|_\infty \leqslant \frac{\|\bar{Q}_m - Q^\star\|_\infty}{33} + c_4 \left\{ \frac{\|\boldsymbol{\Sigma}^\star_{\text{opt}}(Q^\star)\|_{\text{diag}}^{\frac{1}{2}}}{1 - \gamma} \cdot \sqrt{\frac{\log(8DM/\delta)}{n_m}} + \frac{b(Q^\star)}{1 - \gamma} \cdot \frac{\log(8DM/\delta)}{n_m} \right\},$$

with probability at least $1 - \frac{\delta}{2M}$.

Indeed, the proof of Proposition 12 follows directly from the proof of Theorem 2 in Xia et al. (2021) by replacing Lemma 7 by Lemma 13.

C.2.2 Proof of Lemma 13

In order to simplify notation, we drop the epoch number m from $\widehat{Q}_m$ and $\bar{Q}_m$ throughout the remainder of the proof. Let $\widehat{\pi}$ and $\pi^\star$ denote the greedy policies with respect to the Q functions $\widehat{Q}$ and $Q^\star$, respectively. Concretely,

$$\pi^\star(x) = \arg\max_{u\in\mathcal{U}} \; Q^\star(x, u) \qquad \widehat{\pi}(x) = \arg\max_{u\in\mathcal{U}} \; \widehat{Q}(x, u). \tag{56}$$

Ties in the $\arg\max$ are broken by choosing the action u with smallest index.

By the optimality of the policies $\widehat{\pi}$ and $\pi^\star$ for the Q-functions $\widehat{Q}$ and $Q^\star$, respectively, we have

$$Q^\star = \mathbf{r} + \gamma\boldsymbol{\mathcal{P}}^{\pi^\star}Q^\star \quad\text{and}\quad \widehat{Q} = \widetilde{\mathbf{r}} + \gamma\boldsymbol{\mathcal{P}}^{\widehat{\pi}}\widehat{Q}, \quad\text{where}\quad \widetilde{\mathbf{r}} := \mathbf{r} + \bar{\boldsymbol{\mathcal{J}}}_{n_m}(\bar{Q}) - \boldsymbol{\mathcal{J}}(\bar{Q}). \tag{57}$$

Our proof is based on the following intermediate inequality which we prove at the end of this section.

$$\|\widehat{Q} - Q^\star\|_\infty \leqslant \frac{1}{1-\gamma}\|\widetilde{\mathbf{r}} - \mathbf{r}\|_\infty. \tag{58}$$

With the last inequality at hand it suffices to prove an upper bound on the term $\|\mathbf{r} - \mathbf{r}\|_\infty$.

Recall the definition $\widetilde{\mathbf{r}} := \widehat{\mathcal{R}} + \gamma(\widehat{\mathcal{Z}}^{\overline{\pi}} - \boldsymbol{\mathcal{P}}^{\overline{\pi}})\bar{Q}$, where $\overline{\pi}$ a policy greedy with respect to $\bar{Q}$; that is, $\overline{\pi}(x) = \arg\max_{u\in\mathcal{U}} \bar{Q}(x, u)$, where we break ties in the $\arg\max$ by choosing the action u with smallest index. We have

$$\|\widetilde{\mathbf{r}} - \mathbf{r}\|_\infty \leqslant \|(\widehat{\mathcal{R}} - \mathbf{r}) + \gamma(\widehat{\mathcal{Z}}^{\pi^\star} - \boldsymbol{\mathcal{P}}^{\pi^\star})Q^\star\|_\infty + \gamma\|(\widehat{\mathcal{Z}}^{\overline{\pi}}\bar{Q} - \widehat{\mathcal{Z}}^{\pi^\star}Q^\star) - (\boldsymbol{\mathcal{P}}^{\overline{\pi}}\bar{Q} - \boldsymbol{\mathcal{P}}^{\pi^\star}Q^\star)\|_\infty.$$

Observe that the random variable $\widehat{\mathcal{R}}$ and $\widehat{\mathcal{Z}}$ are averages of n_m i.i.d. random variables $\{R_i\}$ and $\{\mathcal{Z}_i\}$, respectively. Consequently, applying Bernstein bound along with a union bound we have the following bound with probability least $1 - \frac{\delta}{4M}$:

$$\|(\widehat{\mathcal{R}} - \mathbf{r}) + \gamma(\widehat{\mathbf{Z}}^{\pi^\star} - \boldsymbol{\mathcal{P}}^{\pi^\star})Q^\star\|_\infty \leqslant \frac{4}{\sqrt{n_m}} \cdot \|\boldsymbol{\Sigma}^\star_{\mathrm{opt}}(Q^\star)\|^{\frac{1}{2}}_{\mathrm{diag}} \cdot \sqrt{\log(8DM/\delta)}$$

$$+ \frac{4b(Q^\star)}{(1-\gamma)n_m} \cdot \log(8DM/\delta).$$

Finally, for each state-action pair (x, u) the random variable $(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{Q} - \widehat{\mathbf{Z}}^{\pi^\star}Q^\star)(x, u)$ has expectation $(\boldsymbol{\mathcal{P}}^{\overline{\pi}}\bar{Q} - \boldsymbol{\mathcal{P}}^{\pi^\star})(x, u)$ with entries uniformly bounded by $2\|\bar{Q} - Q^\star\|_\infty$. Consequently, by a standard application of Hoeffding's inequality combined with the lower bound $n_m \geqslant c_3\frac{4^m}{(1-\gamma)^2}\log(8DM/\delta)$, we have

$$\frac{\gamma}{1-\gamma} \cdot \|(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{Q} - \widehat{\mathbf{Z}}^{\pi^\star}Q^\star) - (\boldsymbol{\mathcal{P}}^{\overline{\pi}}\bar{Q} - \boldsymbol{\mathcal{P}}^{\pi^\star}Q^\star)\|_\infty \leqslant \frac{\|\bar{Q} - Q^\star\|_\infty}{33},$$

with probability at least $1 - \frac{\delta}{4M}$. The statement of Lemma 13 then follows from combining these two high-probability statements with a union bound. It remains to prove the claim (58).

Proof of equation (58)**:** By optimality of the policies $\widehat{\pi}$ and $\pi^{\star}$ for the Q-functions $\widehat{Q}$ and $Q^{\star}$, respectively, we have

$$Q^{\star} = \mathbf{r} + \gamma \boldsymbol{\mathcal{P}}^{\pi^{\star}} Q^{\star} \geq \mathbf{r} + \gamma \boldsymbol{\mathcal{P}}^{\pi^{\star}} \widehat{Q} \quad \text{and} \quad \widehat{Q} = \widetilde{\mathbf{r}} + \gamma \boldsymbol{\mathcal{P}}^{\widehat{\pi}} \widehat{Q} \geq \widetilde{\mathbf{r}} + \gamma \boldsymbol{\mathcal{P}}^{\widehat{\pi}} Q^{\star}. \tag{59}$$

Thus, we have

$$Q^{\star} - \widehat{Q} = \mathbf{r} - \widetilde{\mathbf{r}} + \gamma(\boldsymbol{\mathcal{P}}^{\pi^{\star}} Q^{\star} - \boldsymbol{\mathcal{P}}^{\widehat{\pi}} \widehat{Q}) \leq \mathbf{r} - \widetilde{\mathbf{r}} + \gamma \boldsymbol{\mathcal{P}}^{\pi^{\star}}(Q^{\star} - \widehat{Q}). \tag{60}$$

Rearranging the last inequality, and using the non-negativity of the entries of $(\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^{\pi^{\star}})^{-1}$ we conclude

$$(Q^{\star} - \widehat{Q}) \leq (\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^{\pi^{\star}})^{-1}(\mathbf{r} - \widetilde{\mathbf{r}}).$$

This completes the proof of the bound; a similar argument gives

$$(\widehat{Q} - Q^{\star}) \leq (\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^{\widehat{\pi}})^{-1}(\mathbf{r} - \widetilde{\mathbf{r}}).$$

Collecting the two bounds, we have

$$|\widehat{Q} - Q^{\star}| \leq \max\left\{(\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^{\pi^{\star}})^{-1}(\mathbf{r} - \widetilde{\mathbf{r}}), (\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^{\widehat{\pi}})^{-1}(\mathbf{r} - \widetilde{\mathbf{r}})\right\},$$

where max denotes the entry-wise maximum. The desired then follows from the fact the bound $\|(\mathbf{I} - \gamma \boldsymbol{\mathcal{P}}^{\pi})^{-1}\|_{1,\infty} \leq \frac{1}{1-\gamma}$ for any policy π. This completes the proof.

References

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Dimitri P. Bertsekas. *Neuro-Dynamic Programming*. Springer US, Boston, MA, 2009. ISBN 978-0-387-74759-0.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692. PMLR, 2018.

Gal Dalal, Szörényi Balázs, Thoppe Gugan, and Shie. Mannor. Finite sample analyses for TD (0) with function approximation. In *AAAI Conference on Artificial Intelligence*, volume 32, pages 6144–6153, 2018.

Yaqi Duan, Mengdi Wang, and Martin J Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5:1–25, 2003.

R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6), November 1994.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? An instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040, 2021.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Society for Industrial and Applied Mathematics*, 42, 04 2001.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

C. J. Li, W. Mou, M. J. Wainwright, and M. I. Jordan. ROOT-SGD: Sharp Nonasymptotics and Asymptotic Efficiency in a Single Algorithm. Technical report, UC Berkeley, August 2020a. Arxiv technical report 2008.12690.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33, 2020b.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Tightening the dependence on horizon in the sample complexity of Q-learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6296–6306, 2021.

Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. How hard is my MDP? The distribution-norm to the rescue. In *Advances in Neural Information Processing Systems*, volume 27, pages 1835–1843, 2014.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

Wenlong Mou, Koulik Khamaru, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. Optimal variance-reduced stochastic approximation in Banach spaces. *arXiv preprint*, 2022.

Ashwin Pananjady and Martin J. Wainwright. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2021. doi: 10.1109/TIT.2020.3027316.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature.*, 588(7839):604–609, 2020. ISSN 0028-0836.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, pages 5192–5202, 2018a.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *ACM-SIAM Symposium on Discrete Algorithms*, volume 29, pages 770–787. SIAM, 2018b.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. *31st International Conference on Machine Learning, ICML 2014*, 1, 06 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 1153–1162, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30. IEEE, 2017.

J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

Martin J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. Technical report, Arxiv, 2019a. arXiv:1905.06265.

Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. Technical report, Arxiv, 2019b. arXiv:1906.04697.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229 – 256, 1992.

Eric Xia, Koulik Khamaru, Martin J Wainwright, and Michael I Jordan. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning. *arXiv preprint arXiv:2106.14352*, 2021.

Tengyu Xu, Zhe Wang, Yi Zhou, and Yingbin Liang. Reanalysis of variance reduced temporal difference learning, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, volume 32, pages 5625–5634, 2019.

Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for Atari. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1xCPJHtDB`.