Koulik Khamaru[*], Ashwin Pananjady[†], Feng Ruan[‡], Martin J. Wainwright[§], and Michael I. Jordan[¶]

Abstract. We address the problem of policy evaluation in discounted, tabular Markov decision processes, and provide instance-dependent guarantees on the ℓ_∞-error under a generative model. We establish both asymptotic and non-asymptotic versions of local minimax lower bounds for policy evaluation, thereby providing an instance-dependent baseline by which to compare algorithms. Theory-inspired simulations show that the widely-used temporal difference (TD) algorithm is strictly suboptimal when evaluated in a non-asymptotic setting, even when combined with Polyak-Ruppert iterate averaging. We remedy this issue by introducing and analyzing variance-reduced forms of stochastic approximation, showing that they achieve non-asymptotic, instance-dependent optimality up to logarithmic factors.

Key words. Temporal difference learning, Polyak-Ruppert averaging, variance reduction.

AMS subject classifications. 68Q25, 68R10, 68U05

1. Introduction.

Reinforcement learning (RL) refers to a class of methods for the optimal control of dynamical systems [7, 6, 46, 8] that has begun to make inroads in a wide range of applied problem domains. However, this empirical research has revealed the limitations of our theoretical understanding of this class of methods: more precisely, popular RL algorithms exhibit a variety of behavior across domains and problem instances, and existing theoretical bounds, which are generally based on worst-case assumptions, fail to capture this variety. An important theoretical goal is to develop *instance-specific* analyses that help to reveal what aspects of a given problem make it "easy" or "hard," and allow distinctions to be drawn between ostensibly similar algorithms in terms of their performance profiles. The focus of this paper is on developing such a theoretical understanding for a class of popular stochastic approximation algorithms used for policy evaluation.

RL methods are generally formulated in terms of a Markov decision process (MDP). An agent operates in an environment whose dynamics are described by an MDP but are unknown: at each step, it observes the current state of the environment, and takes an action that changes the state according to some stochastic transition function. The eventual goal of the agent is to learn a *policy*—a mapping from states to actions—that optimizes the reward accrued over time. In the typical setting, rewards are assumed to be additive over time, and are also discounted over time. Within this broad context, a key sub-problem is that of *policy evaluation*, where the goal is estimate the long-term expected reward of a fixed policy based on observed state-to-state transitions and one-step rewards. It is often preferable to have ℓ_∞-norm guarantees for such an estimate, since these are particularly compatible with policy iteration methods. In particular, policy iteration can be shown to converge at a geometric rate when combined with policy evaluation methods that are accurate in ℓ_∞-norm (see, e.g.,

[*]Department of Statistics, UC Berkeley (koulik@berkeley.edu).
[†]Industrial & Systems Engineering and Electrical & Computer Engineering, Georgia Tech (ashwinpm@gatech.edu).
[‡]Department of Statistics, UC Berkeley (fengruan@berkeley.edu).
[§]Department of Statistics and EECS, UC Berkeley (wainwrig@berkeley.edu).
[¶]Department of Statistics and EECS, UC Berkeley (jordan@cs.berkeley.edu).

the sources [1, 8]).

In this paper, we study a class of stochastic approximation algorithms for this problem under a generative model for the underlying MDP, with a focus on developing instance-dependent bounds. Our results complement an earlier paper by a subset of the authors [38], which studied the least squares temporal difference (LSTD) method through such a lens.

1.1. Related work. We begin with a broad overview of related work, categorizing that work as involving asymptotic analysis, non-asymptotic analysis, or instance-dependent analysis.

Asymptotic theory. Markov reward processes have been the subject of considerable classical study [22, 21]. In the context of reinforcement learning and stochastic control, the policy evaluation problem for such processes has been tackled by various approaches based on stochastic approximation. Here we focus on past work that studies the *temporal difference* (TD) update and its relatives; see the paper [16] for a comprehensive survey. The TD update was originally proposed by Sutton [45], and is typically used in conjunction with an appropriate parameterization of value functions. Classical results on the algorithm are typically asymptotic, and include both convergence guarantees [24, 11, 12] and examples of divergence [5]; see the paper [48] for conditions that guarantee asymptotic convergence.

It is worth noting that the TD algorithm is a form of linear stochastic approximation, and can be fruitfully combined with the iterate-averaging procedure put forth independently by Polyak [39] and Ruppert [42]. The subsequent work of Polyak and Juditsky [40] deserves special mention, since it shows that under fairly mild conditions, the TD algorithm converges when combined with Polyak-Ruppert iterate averaging. To be clear, in the specific context of the policy evaluation problem, the results in the Polyak-Juditsky paper [40] allow noise only in the observations of rewards (i.e., the transition function is assumed to be known). However, the underlying techniques can be extended to derive results in the setting in which we only observe samples of transitions; for instance, see the work of Tadic [47] for results of this type.

Non-asymptotic theory. Recent years have witnessed significant interest in understanding TD-type algorithms from the non-asymptotic standpoint. Bhandari et al. [9] focus on proving ℓ_2-guarantees for the TD algorithm when combined with Polyak-Ruppert iterate averaging. They consider both the generative model as well as the Markovian noise model, and provide non-asymptotic guarantees on the expected error. Their results also extend to analyses of the popular TD(λ) variant of the algorithm, as well as to Q-learning in specific MDP instances. Also noteworthy is the analysis of Lakshminarayanan and Szepesvari [29], carried out in parallel with Bhandari et al. [9]; it provides similar guarantees on the TD(0) algorithm with constant stepsize and averaging. Note that both of these analyses focus on ℓ_2-guarantees (equipped with an associated inner product), and thus can directly leverage proof techniques for stochastic optimization [4, 37].

Other related results[1] include those of Dalal et al. [15], Doan et al. [17], Korda and La [28], and also more contemporary papers [55, 51]. The latter three of these papers introduce a variance-reduced form of temporal difference learning, a variant of which we analyze in this

[1]It should be noted that there were some errors in the results of Korda and La [28] that were pointed out by both Lakshminarayanan and Szepesvari [29] and Xu et al. [55].

paper.

Instance-dependent results. The focus on instance-dependent guarantees for TD algorithms is recent, and results are available both in the ℓ_2-norm setting [9, 29, 15, 55] and the ℓ_∞-norm settings [38]. In general, however, the guarantees provided by work to date are not sharp. For instance, the bounds in [15] scale exponentially in relevant parameters of the problem, whereas the papers [9, 29, 55] do not capture the correct "variance" of the problem instance at hand. A subset of the current authors [38] derived ℓ_∞ bounds on policy evaluation for the plug-in estimator. These results were shown to be locally minimax optimal in certain regions of the parameter space. There has also been some recent focus on obtaining instance-dependent guarantees in online reinforcement learning settings [34]. This has resulted in more practically useful algorithms that provide, for instance, horizon-independent regret bounds for certain episodic MDPs [56, 25], thereby improving upon worst-case bounds [3]. Recent work has also established some instance-dependent bounds, albeit not sharp over the whole parameter space, for the problem of state-action value function estimation in Markov decision processes, for both ordinary Q-learning [53] and a variance-reduced improvement [54].

1.2. Contributions.

In this paper, we study stochastic approximation algorithms for evaluating the value function of a tabular Markov reward process in the discounted setting. Our goal is to provide a sharp characterization of performance in the ℓ_∞-norm, for procedures that are given access to state transitions and reward samples under the generative model. In practice, temporal difference learning is typically applied with an additional layer of (linear) function approximation. In the current paper, so as to bring the instance dependence into sharp focus, we study the algorithms without this function approximation step. In this context, we tell a story with three parts, as detailed below:

Local minimax lower bounds. Global minimax analysis provides bounds that hold uniformly over large classes of models. In this paper, we seek to gain a more refined understanding of how the difficulty of policy evaluation varies as a function of the instance. In order to do so, we undertake an analysis of the local minimax risk associated with a problem. We first prove an asymptotic statement (Proposition 3.1) that characterizes the local minimax risk up to a logarithmic factor; it reveals the relevance of two functionals of the instance that we define. In proving this result, we make use of the classical asymptotic minimax theorem [23, 31, 32]. We then refine this analysis by deriving a *non-asymptotic* local minimax bound, as stated in Theorem 3.2, which is derived using the non-asymptotic local minimax framework of Cai and Low [14], an approach that builds upon the seminal concept of hardest local alternatives that can be traced back to Stein [44].

Non-asymptotic suboptimality of iterate averaging. Our local minimax lower bounds raise a natural question: Do standard procedures for policy evaluation achieve these instance-specific bounds? In Section 3.2, we address this question for the TD(0) algorithm with iterate averaging. Via a careful simulation study, we show that for many popular stepsize choices, the algorithm *fails* to achieve the correct instance-dependent rate in the non-asymptotic setting, even when the sample size is quite large. This is true for both the constant stepsize, as well as polynomial stepsizes of various orders. Notably, the algorithm with polynomial stepsizes of certain orders achieves the local risk in the asymptotic setting (see Proposition 3.1).

Non-asymptotic optimality of variance reduction. In order to remedy this issue with iterate averaging, we propose and analyze a variant of TD learning with variance reduction, showing both through theoretical (see Theorem 2) and numerical results (see Figure 3) that this algorithm achieves the correct instance-dependent rate provided the sample size is larger than an explicit threshold. Thus, this algorithm is provably better than TD(0) with iterate averaging.

1.3. Notation. For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. We use $c, C, c_1, c_2, \ldots$ to denote universal constants that may change from line to line. We let $\mathbf{1}$ denote the all-ones vector in $\mathbb{R}^D$. Let e_j denote the jth standard basis vector in $\mathbb{R}^D$. We let $v_{(i)}$ denote the i-th order statistic of a vector v, i.e., the i-th largest entry of v. For a pair of vectors (u, v) of compatible dimensions, we use the notation $u \preceq v$ to indicate that the difference vector $v - u$ is entrywise non-negative. The relation $u \succeq v$ is defined analogously. We let $|u|$ denote the entrywise absolute value of a vector $u \in \mathbb{R}^D$; squares and square-roots of vectors are, analogously, taken entrywise. Note that for a positive scalar λ, the statements $|u| \preceq \lambda \cdot \mathbf{1}$ and $\|u\|_\infty \leq \lambda$ are equivalent. Finally, we let $\|\mathbf{M}\|_{1,\infty}$ denote the maximum ℓ_1-norm of the rows of a matrix $\mathbf{M}$, and refer to it as the $(1, \infty)$-operator norm of a matrix.

2. Background and problem formulation.

We begin by introducing the basic mathematical formulation of Markov reward processes (MRPs) and generative observation models.

2.1. Markov reward processes and value functions. We study MRPs defined on a finite set of D states, which we index by the set $[D] \equiv \{1, 2, \ldots, D\}$. The state evolution over time is determined by a set of transition functions, $\{P(\cdot|i),\ i \in [D]\}$. Note that each such transition function can be naturally associated with a D-dimensional vector; denote the i-th such vector as p_i. We let $\mathbf{P} \in [0, 1]^{D \times D}$ denote a row-stochastic (Markov) transition matrix, where row i of this matrix contains the vector p_i. Also associated with an MRP is a population reward function, $r : [D] \mapsto \mathbb{R}$, possessing the semantics that a transition from state i results in the reward $r(i)$. For convenience, we engage in a minor abuse of notation by letting r also denote a vector of length D; here r_i corresponds to the reward obtained at state i.

We formulate the long-term value of a state in the MRP in terms of the infinite-horizon, discounted reward. This value function (denoted here by the vector $\theta^* \in \mathbb{R}^D$) can be computed as the unique solution of the Bellman fixed-point relation, $\theta^* = r + \gamma \mathbf{P}\theta^*$.

2.2. Observation model. In the learning setting, the pair $(\mathbf{P}, r)$ is unknown, and we accordingly assume access to a black box that generates samples from the transition and reward functions. In this paper, we operate under a setting known as the synchronous[2] or generative setting [27]; this setting is also often referred to as the "i.i.d. setting" in the policy evaluation literature. For a given sample index, $k \in \{1, 2, \ldots, N\}$ and for each state $j \in [D]$, we observe a random next state

$$X_{k,j} \sim P(\cdot|j) \qquad \text{for } j \in [D]. \tag{2.1a}$$

[2]With standard arguments, our results can be extended to the setting in which the noise is the problem evolves according to a martingale.

We collect these transitions in a matrix $\mathbf{Z}_k$, which by definition contains one 1 in each row: the 1 in the j-th row corresponds to the index of state $X_{k,j}$. We also observe a random reward vector $R_k \in \mathbb{R}^D$, where the rewards are generated independently across states with[3]

$$R_{k,j} \sim \mathcal{N}(r_j, \sigma_r^2). \tag{2.1b}$$

Given these samples, define the k-th (noisy) linear operator $\widehat{\mathcal{T}}_k : \mathbb{R}^D \mapsto \mathbb{R}^D$ whose evaluation at the point θ is given by

$$\widehat{\mathcal{T}}_k(\theta) = R_k + \gamma \mathbf{Z}_k \theta. \tag{2.2}$$

The construction of these operators is inspired by the fact that we are interested in computing the fixed point of the population operator,

$$\mathcal{T} : \theta \mapsto r + \gamma \mathbf{P}\theta, \tag{2.3}$$

and a classical and natural way to do so is via a form of stochastic approximation known as temporal difference learning, which we describe next.

2.3. Temporal difference learning and its variants. Classical temporal difference (TD) learning algorithms are parametrized by a sequence of stepsizes, $\{\alpha_k\}_{k \geq 1}$, with $\alpha_k \in (0, 1]$. Starting with an initial vector $\theta_1 \in \mathbb{R}^D$, the TD updates take the form

$$\theta_{k+1} = (1 - \alpha_k)\theta_k + \alpha_k \widehat{\mathcal{T}}_k(\theta_k) \quad \text{for } k = 1, 2, \ldots. \tag{2.4}$$

In the sequel, we discuss three popular stepsize choices:

$$\texttt{Constant stepsize:} \qquad \alpha_k = \alpha, \qquad \text{where} \qquad 0 < \alpha \leq \alpha_{\max}. \tag{2.5a}$$

$$\texttt{Polynomial stepsize:} \qquad \alpha_k = \frac{1}{k^\omega} \qquad \text{for some } \omega \in (0, 1). \tag{2.5b}$$

$$\texttt{Recentered-linear stepsize:} \qquad \alpha_k = \frac{1}{1 + (1 - \gamma)k}. \tag{2.5c}$$

In addition to the TD sequence (2.4), it is also natural to perform *Polyak-Ruppert averaging*, which produces a parallel sequence of averaged iterates

$$\widetilde{\theta}_k = \frac{1}{k} \sum_{j=1}^{k} \theta_j \quad \text{for } k = 1, 2, \ldots. \tag{2.6}$$

Such averaging schemes were introduced in the context of general stochastic approximation by Polyak [40] and Ruppert [42]. A large body of theoretical literature demonstrates that such an averaging scheme improves the rates of convergence of stochastic approximation when run with overly "aggressive" stepsizes [4, 40, 42].

[3] All of our upper bounds extend with minor modifications to the sub-Gaussian reward setting.

3. Main results. We turn to the statements of our main results and discussion of their consequences. All of our statements involve certain measures of the local complexity of a given problem, which we introduce first. We then turn to the statement of lower bounds on the ℓ_∞-norm error in policy evaluation. In Section 3.1, we prove two lower bounds. Our first result, stated as Proposition 3.1, is asymptotic in nature (holding as the sample size $N \to +\infty$). Our second lower bound, stated as Theorem 3.2, provides a result that holds for a range of finite sample sizes. Given these lower bounds, it is then natural to wonder about known algorithms that achieve them. Concretely, does the TD(0) algorithm combined with Polyak-Ruppert averaging achieve these instance-dependent bounds? In Section 3.2, we undertake a careful empirical study of this question, and show that in the non-asymptotic setting, this algorithm fails to match the instance-dependent bounds. This finding sets up the analysis in Section 3.3, where we introduce a variance-reduced version of TD(0), and prove that it does achieve the instance-dependent lower bounds from Theorem 3.2 up to a logarithmic factor in dimension.

Local complexity measures. Recall the generative observation model described in Section 2.2. For a transition matrix $\mathbf{P}$, we write $\mathbf{Z} \sim \mathbf{P}$ to mean a random matrix with $\{0,1\}$ entries, and a single one in each row (with the position of the one in row $\mathbf{Z}_j$ determined by sampling from the transition distribution specified by row $\mathbf{P}_j$). Also recall that the random reward vector $R \in \mathbb{R}^D$ such that $R_k \sim \mathcal{N}(r_j, \sigma_r^2)$. As we show shortly, the complexity of estimating the value function θ^* depends on the covariance matrix

$$\mathbf{\Sigma}^*(\mathbf{P}, r) = (\mathbf{I} - \gamma\mathbf{P})^{-1} \operatorname{cov}(R + \gamma\mathbf{Z}\theta^*)(\mathbf{I} - \gamma\mathbf{P})^{-\top}. \tag{3.1a}$$

The term $\operatorname{cov}(R + \gamma\mathbf{Z}\theta^*) = \operatorname{cov}(\widehat{\mathcal{T}}_k(\theta^*))$ denotes the variance of the empirical Bellman operator (2.2) applied to the true value function, and it captures the effect of noise. This error is compounded by powers of the discounted transition matrix,[4] which captures how perturbations propagate over time, and thus gives rise to the matrix $(\mathbf{I} - \gamma\mathbf{P})^{-1}$. In Section 3.1, we argue that local complexity of estimating the value function θ^* depends on $\|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}}$, i.e. the maximal diagonal entry of the matrix $\mathbf{\Sigma}^*(\mathbf{P}, r)$.

Since the transition and reward samples are assumed to be independent under the generative observation model 2.2, we can decompose the covariance matrix $\mathbf{\Sigma}^*(\mathbf{P}, r)$ into two parts:

$$\mathbf{\Sigma}^*(\mathbf{P}, r) = (\mathbf{I} - \gamma\mathbf{P})^{-1} \operatorname{cov}(\gamma\mathbf{Z}\theta^*)(\mathbf{I} - \gamma\mathbf{P})^{-\top} + (\mathbf{I} - \gamma\mathbf{P})^{-1} \operatorname{cov}(R)(\mathbf{I} - \gamma\mathbf{P})^{-\top}.$$

Throughout the paper, we use the shorthand notation

$$\nu(\mathbf{P}, \theta^*) := \|\operatorname{diag}\Big((\mathbf{I} - \gamma\mathbf{P})^{-1} \operatorname{cov}(\gamma\mathbf{Z}\theta^*)(\mathbf{I} - \gamma\mathbf{P})^{-\top}\Big)\|_\infty^{\frac{1}{2}} \tag{3.1b}$$

$$\rho(\mathbf{P}, r) := \|\operatorname{diag}\Big((\mathbf{I} - \gamma\mathbf{P})^{-1} \operatorname{cov}(R)(\mathbf{I} - \gamma\mathbf{P})^{-\top}\Big)\|_\infty^{\frac{1}{2}}. \tag{3.1c}$$

[4] Observe that we have the von Neumann expansion $\sum_{j=0}^{\infty} (\gamma\mathbf{P})^j = (\mathbf{I} - \gamma\mathbf{P})^{-1}$.

In terms of the above notation, we have the following convenient sandwich relation:

$$\frac{1}{2} \cdot \{\nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\} \;\leq\; \|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}} \;\leq\; 2 \cdot \{\nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\}\,. \tag{3.1d}$$

A portion of our results also involves the quantity

$$b(\theta) := \frac{\|\theta\|_{\text{span}}}{1-\gamma}, \tag{3.1e}$$

where $\|\theta\|_{\text{span}} = \max_{j \in [D]} \theta_j - \min_{j \in [D]} \theta_j$ is the span seminorm.

3.1. Local minimax lower bound. Throughout this section, we use the letter $\mathcal{P}$ to denote an individual problem instance, $\mathcal{P} = (\mathbf{P}, r)$, and use $\theta(\mathcal{P}) := \theta^* = (\mathbf{I} - \gamma \mathbf{P})^{-1} r$ to denote the *target* of interest. The aim of this section is to establish *instance-specific* lower bounds for estimating $\theta(\mathcal{P})$ under the observation model (2.1). In order to do so, we adopt a local minimax approach.

The remainder of this the section is organized as follows. In Section 3.1.1, we prove an asymptotic local minimax lower bound, valid as the sample size N tends to infinity. It gives an explicit Gaussian limit for the rescaled error that can be achieved by any procedure. The asymptotic covariance in this limit law depends on the problem instance, and is very closely related to the functional $\mathbf{\Sigma}^*(\mathbf{P}, r)$. Moreover, we show that this limit can be achieved—in the asymptotic sense—by the TD algorithm combined with Polyak-Ruppert averaging. While this provides a useful sanity check, in practice we implement estimators using a finite number of samples N, so it is important to obtain non-asymptotic lower bounds for a full understanding. With this motivation, Section 3.1.2 provides a new, *non-asymptotic* instance-specific lower bound for the policy evaluation problem. We show that the functional $\mathbf{\Sigma}^*(\mathbf{P}, r)$ also covers the instance-specific complexity in the finite-sample setting. In proving this non-asymptotic lower bound, we build upon techniques in the statistical literature based on constructing hardest one-dimensional alternatives [44, 10, 18, 19, 13]. As we shall see in later sections, while the TD algorithm with averaging is instance-specific optimal in the asymptotic setting, it *fails* to achieve our non-asymptotic lower bound.

3.1.1. Asymptotic local minimax lower bound. Our first approach towards an instance-specific lower bound is an asymptotic one, based on classical local asymptotic minimax theory. For regular and parametric families, the Hájek–Le Cam local asymptotic minimax theorem [23, 31, 32] shows that the Fisher information—an instance-specific functional—characterizes a fundamental asymptotic limit. Our model class is both parametric and regular (cf. equation (2.1)), and so this classical theory applies to yield an asymptotic local minimax bound. Some additional work is needed to relate this statement to the more transparent complexity measure $\mathbf{\Sigma}^*(\mathbf{P}, r)$ that we have defined.

In order to state our result, we require some additional notation. Fix an instance $\mathcal{P} = (\mathbf{P}, r)$. For any $\epsilon > 0$, we define an ϵ-neighborhood of problem instances by

$$\mathfrak{N}(\mathcal{P}; \epsilon) = \left\{\mathcal{P}' = (\mathbf{P}', r') : \left\|\mathbf{P} - \mathbf{P}'\right\|_F + \left\|r - r'\right\|_2 \leq \epsilon\right\}.$$

Adopting the ℓ_∞-norm as the loss function, the *local asymptotic minimax risk* is given by

$$\mathfrak{M}_\infty(\mathcal{P}) \equiv \mathfrak{M}_\infty(\mathcal{P}; \|\cdot\|_\infty) = \lim_{c\to\infty} \liminf_{N\to\infty} \inf_{\hat{\theta}_N} \sup_{\mathcal{Q}\in\mathfrak{N}(\mathcal{P};c/\sqrt{N})} \mathbb{E}_{\mathcal{Q}}\left[\sqrt{N}\left\|\hat{\theta}_N - \theta(\mathcal{Q})\right\|_\infty\right]. \tag{3.2}$$

Here the infimum is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N i.i.d. observations drawn according to the observation model (2.1).

Our first main result characterizes the local asymptotic risk $\mathfrak{M}_\infty(\mathcal{P})$ exactly, and shows that it is attained by stochastic approximation with Polyak-Ruppert averaging. Recall the Polyak-Ruppert (PR) sequence $\{\widetilde{\theta}_k\}_{k\geq 1}$ defined in equation (2.6), and let $\{\widetilde{\theta}_k^\omega\}_{k\geq 1}$ denote this sequence when the underlying SA algorithm is the TD update with the polynomial stepsize sequence (2.5b) with exponent ω.

Proposition 3.1. *Let $Z \in \mathbb{R}^D$ be a multivariate Gaussian with zero mean and covariance matrix $\mathbf{\Sigma}^*(\mathbf{P}, r)$, then the local asymptotic minimax risk at problem instance $\mathcal{P}$ is given by*

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty]. \tag{3.3a}$$

Furthermore, for each problem instance $\mathcal{P}$ and scalar $\omega \in (1/2, 1)$, this limit is achieved by the TD algorithm with an ω-polynomial stepsize and PR-averaging:

$$\lim_{N\to\infty} \sqrt{N} \cdot \mathbb{E}\left[\|\widetilde{\theta}_N^\omega - \theta(\mathcal{P})\|_\infty\right] = \mathbb{E}[\|Z\|_\infty]. \tag{3.3b}$$

With the convention that $\theta^* \equiv \theta(\mathcal{P})$, a short calculation bounding the maximum absolute value of sub-Gaussian random variables (see, e.g., Ex. 2.11 in Wainwright [52]) yields the sandwich relation

$$\|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}} \leq \mathbb{E}[\|Z\|_\infty] \leq \sqrt{2\log D} \cdot \|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}},$$

so that Proposition 3.1 shows that, up to a logarithmic factor in dimension D, the local asymptotic minimax risk is entirely characterized by the functional $\|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}}$.

It should be noted that lower bounds similar to equation (3.3a) have been shown for specific classes of stochastic approximation algorithms [49]. However, to the best of our knowledge, a local minimax lower bound—one applying to any procedure that is a measurable function of the observations—is not available in the existing literature.

Furthermore, equation (3.3b) shows that stochastic approximation with polynomial stepsizes and averaging attains the exact local asymptotic risk. Our proof of this result essentially mirrors that of Polyak and Juditsky [40], and amounts to verifying their assumptions under the policy evaluation setting. Given this result, it is natural to ask if averaging is optimal also in the non-asymptotic setting; answering this question is the focus of the next two sections of the paper.

3.1.2. Non-asymptotic local minimax lower bound. Proposition 3.1 provides an instance-specific lower bound on $\theta(\mathcal{P})$ that holds asymptotically. In order to obtain a non-asymptotic guarantee, we borrow ideas from the non-asymptotic framework introduced by Cai and Low [13] for nonparametric shape-constrained inference. Adapting their definition of local minimax risk

to our problem setting, given the loss function $L(\theta - \theta^*) = \|\theta - \theta^*\|_\infty$, the (normalized) *local non-asymptotic minimax risk* for $\theta(\cdot)$ at instance $\mathcal{P} = (\mathbf{P}, r)$ is given by

$$\mathfrak{M}_N(\mathcal{P}) = \sup_{\mathcal{P}'} \inf_{\hat{\theta}_N} \max_{\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}'\}} \sqrt{N} \cdot \mathbb{E}_{\mathcal{Q}} \left[\|\widehat{\theta}_N - \theta(\mathcal{Q})\|_\infty \right]. \tag{3.4}$$

Here the infimum is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N i.i.d. observations drawn according to the observation model (2.1), and the normalization by $\sqrt{N}$ is for convenience. The definition (3.4) is motivated by the notion of the hardest one-dimensional alternative [50, Ch. 25]. Indeed, given an instance $\mathcal{P}$, the local non-asymptotic risk $\mathfrak{M}_N(\mathcal{P})$ first looks for the hardest alternative $\mathcal{P}'$ against $\mathcal{P}$ (which should be local around $\mathcal{P}$), then measures the worst-case risk over $\mathcal{P}$ and its (local) hardest alternative $\mathcal{P}'$. As explained in detail in the paper [20], this instance-specific local minimax risk thus defined imposes a fundamental limit on all learning procedures: any algorithm achieving better behavior than the lower bound at one instance must have substantially worse behavior at some other instances.

With this definition in hand, we lower bound the local non-asymptotic minimax risk using the complexity measure $\|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}}$ defined in equation (3.1):

Theorem 3.2. *There exists a universal constant $c > 0$ such that for any instance $\mathcal{P} = (\mathbf{P}, r)$, the local non-asymptotic minimax risk is lower bounded as*

$$\mathfrak{M}_N(\mathcal{P}) \geq c \cdot \|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}}. \tag{3.5}$$

This bound is valid for all sample sizes N that satisfy

$$N \geq N_0 := \max \left\{ \frac{\gamma^2}{(1-\gamma)^2}, \frac{b^2(\theta^*)}{\nu^2(\mathbf{P}, \theta^*)} \right\}. \tag{3.6}$$

A few comments are in order. First, it is natural to wonder about the necessity of condition (3.6) on the sample size N in our lower bound. Our past work provides upper bounds on the ℓ_∞-error of the plug-in estimator [38], and these results also require a bound of this type. In fact, when the rewards are observed with noise (i.e., for any $\sigma_r > 0$), the condition $N \gtrsim \frac{\gamma^2}{(1-\gamma)^2}$ is natural, since it is necessary in order to obtain an estimate of the value function with $\mathcal{O}(1)$ error. On the other hand, in the special case of deterministic rewards ($\sigma_r = 0$), it is interesting to ask how the fundamental limits of the problem behave in the absence of this condition (see Section 5 for further discussion of this point).

Second, note that Theorem 3.2 may be viewed as a strengthening of local minimax lower bounds established in prior work by a subset of the current authors [38], which held over sub-classes of MRPs satisfying certain conditions. Theorem 3.2, on the other hand, is a lower bound that holds in the neighborhood of every instance. Having said that, the lower bounds in the paper [38] are able to capture logarithmic factors in the dimension, but Theorem 3.2, owing to the two-point nature of the construction, is not.

Finally, note that the non-asymptotic lower bound (3.5) is closely connected to the asymptotic local minimax bound from Proposition 3.1. In particular, for any sample size N satisfying the lower bound (3.6), our non-asymptotic lower bound (3.5) coincides with the

asymptotic lower bound (3.3a) up to a constant factor. Thus, it cannot be substantially sharpened. The finite-sample nature of the lower bound (3.5) is a powerful tool for assessing optimality of procedures: it provides a performance benchmark that holds over a large range of finite sample sizes N. Indeed, in the next section, we study the performance of the TD learning algorithm with Polyak-Ruppert averaging. While this procedure achieves the local minimax lower bound asymptotically, as guaranteed by equation (3.3b) in Proposition 3.1, it falls short of doing so in natural *finite-sample* scenarios.

3.2. Suboptimality of averaging. Polyak and Juditsky [40] provide a general set of conditions under which a given stochastic-approximation (SA) algorithm, when combined with Polyak-Ruppert averaging, is guaranteed to have asymptotically optimal behavior. For the current problem, the bound (3.3b) in Proposition 3.1, which is proved using the Polyak-Juditsky framework, shows that SA with polynomial stepsizes and averaging have this favorable asymptotic property.

However, asymptotic theory of this type gives no guarantees in the finite-sample setting. In particular, suppose that we are given a sample size N that scales as $(1-\gamma)^{-2}$, as specified in our lower bounds. Does the averaged TD(0) algorithm exhibit optimal behavior in this non-asymptotic setting? In this section, we answer this question in the negative. More precisely, we describe a parameterized family of Markov reward processes, and provide careful simulations that reveal the suboptimality of TD without averaging.

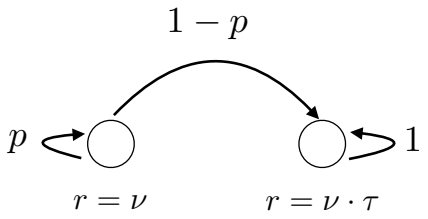


Figure 1. Illustration of the 2-state MRP used in the simulation. The triple of scalars (p, ν, τ), along with the discount factor γ, are parameters of the construction. The chain remains in state 1 with with probability p and transitions to state 2 with probability $1-p$; on the other hand, state 2 is absorbing. The rewards in states 1 and 2 are deterministic, specified by ν and $\nu\tau$, respectively.

3.2.1. A simple construction. The lower bound in Theorem 3.2 predicts a range of behaviors depending on the quantity $\|\operatorname{diag}(\mathbf{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}}$, and equivalently on the pair $\nu(\mathbf{P}, \theta^*)$ and $\rho(\mathbf{P}, r)$ (Cf. equation (3.1d)). In order to observe a large subset of these behaviors, it suffices to consider a very simple MRP, $\mathcal{P} = (\mathbf{P}, r)$ with $D = 2$ states, as illustrated in Figure 1. In this MRP, the transition matrix $\mathbf{P} \in \mathbb{R}^{2\times 2}$ and reward vector $r \in \mathbb{R}^2$ take the form

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

Here the triple (p, ν, τ), along with the discount factor $\gamma \in [0, 1)$, are parameters of the construction.

In order to parameterize this MRP in a scalarized manner, we vary the triple (p, ν, τ) in the following way. First, we fix a scalar $\lambda \geq 0$, and then we set

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

Note that this sub-family of MRPs is fully parametrized by the pair (γ, λ). Let us clarify why this particular scalarization is interesting. It can be shown via simple calculations that the underlying MRP satisfies

$$\nu(\mathbf{P}, \theta^*) \sim \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}, \quad \rho(\mathbf{P}, r) = 0 \quad \text{and} \quad b(\theta^*) \sim \left(\frac{1}{1-\gamma}\right)^{2-\lambda},$$

where $\sim$ denotes equality that holds up to a constant pre-factor. Consequently, by Theorem 3.2 the minimax risk, measured in terms of the ℓ_∞-norm, satisfies

$$\mathfrak{M}_N(\mathcal{P}) \geq c \cdot \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}. \tag{3.7}$$

Thus, it is natural to study whether the TD(0) algorithm with PR averaging achieves this error.

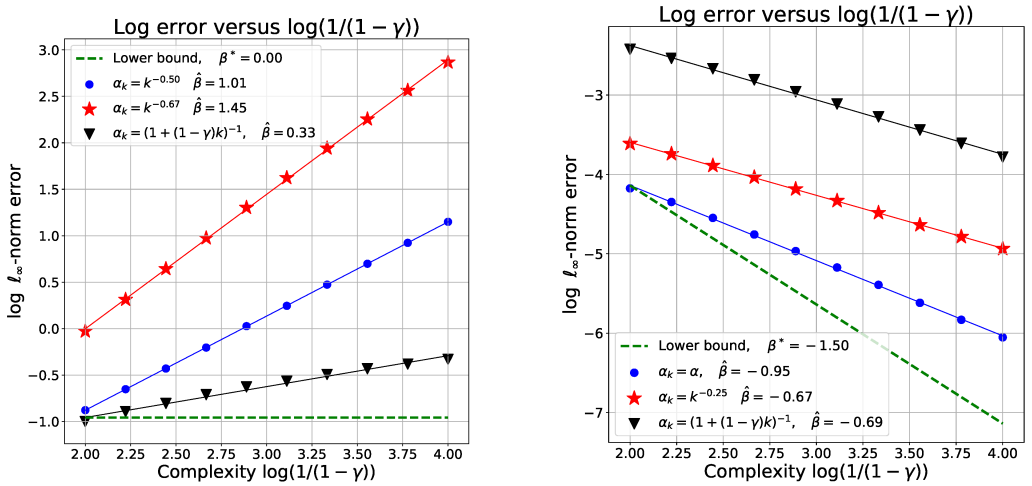
We note in passing that conceptually similar (special cases of such) instances with two-state Markov chains have been used to obtain other worst-case lower bounds in reinforcement learning [2, 30]. A previous paper by a subset of the authors [38] introduced the current family of instances to interpolate smoothly between the most trivial and most difficult problems as the discount factor is varied, but the motivation there was still to provide worst-case lower bounds holding over a sub-class of problems. The current paper takes this a step further, and uses this family to evaluate *local* notions of optimality.

3.2.2. A simulation study. In order to compare the behavior of averaged TD with the lower bound (3.7), we performed a series of experiments of the following type. For a fixed parameter λ in the range $[0, 1.5]$, we generated a range of MRPs with different values of the discount factor γ. For each value of the discount parameter γ, we consider the problem of estimating θ^* using a sample size N set to be one of two possible values: namely, $N \in \left\{ \lceil \frac{8}{(1-\gamma)^2} \rceil, \lceil \frac{8}{(1-\gamma)^3} \rceil \right\}$.

In Figure 2, we plot the ℓ_∞-error of the averaged SA, for constant stepsize (2.5a), polynomial-decay stepsize (2.5b) and recentered linear stepsize (2.5c), as a function of γ. The plots show the behavior for $\lambda \in \{0.5, 1.5\}$. Each point on each curve is obtained by averaging 1000 Monte Carlo trials of the experiment. Note that from our lower bound calculations above (3.7), the log ℓ_∞-error is related to the complexity $\log\left(\frac{1}{1-\gamma}\right)$ in a linear fashion; we use β^* to denote the slope of this idealized line. Simple algebra yields

$$\beta^* = \frac{1}{2} - \lambda \quad \text{for} \quad N = \frac{1}{(1-\gamma)^2}, \quad \text{and} \quad \beta^* = -\lambda \quad \text{for} \quad N = \frac{1}{(1-\gamma)^3}. \tag{3.8}$$

In other words, for an algorithm which achieves the lower bound predicted by our theory, we expect a linear relationship between the log ℓ_∞-error and log discount complexity $\log\left(\frac{1}{1-\gamma}\right)$, with the slope β^*.



(a) $\lambda = 0.5, \quad N = \lceil \frac{8}{(1-\gamma)^2} \rceil$. (b) $\lambda = 1.5, \quad N = \lceil \frac{8}{(1-\gamma)^3} \rceil$.

Figure 2. Log-log plots of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for various algorithms. Each point represents an average over 1000 trials, with each trial simulations are for the 2-state MRP depicted in Figure 1 with the parameter choices $p = \frac{4\gamma-1}{3\gamma}$, $\nu = 1$ and $\tau = 1 - (1-\gamma)^\lambda$. We have also plotted the least-squares fits through these points, and the slopes of these lines are provided in the legend. In particular, the legend contains the stepsize choice for averaged SA (denoted as α_k), the slope $\hat{\beta}$ of the least-squares line, and the ideal value β^* of the slope computed in equation 3.8. We also include the lower bound predicted by Theorem 3.2 for these examples as a dotted line for comparison purposes. Logarithms are to the natural base.

Accordingly, for the averaged SA estimators with the stepsize choices in (2.5a)-(2.5c), we performed a linear regression to estimate the slopes between the log ℓ_∞-error and the log discount-complexity $\log\left(\frac{1}{1-\gamma}\right)$. The plot legend reports the stepsize choices α_k and the slope $\hat{\beta}$ of the fitted regression line. We also include the lower bound in the plots, as a dotted line along with its slope, for a visual comparison. We see that the slopes corresponding to the averaged SA algorithm are higher compared to the ideal slopes of the dotted lines. Stated differently, this means that the averaged SA algorithm does not achieve the lower bound with either the constant step or the polynomial-decay step. Overall, the simulations provided in this section demonstrate that the averaged SA algorithm, although guaranteed to be asymptotically optimal by equation (3.3b) in Proposition 3.1, does not yield the ideal non-asymptotic behavior.

3.3. Variance-reduced policy evaluation. In this section, we propose and analyze a variance-reduced version of the TD learning algorithm. As in standard variance-reduction schemes, such as SVRG [26], our algorithm proceeds in epochs. In each epoch, we run a standard stochastic approximation scheme, but we recenter our updates in order to reduce their variance. The recentering uses an empirical approximation to the population Bellman operator $\mathcal{T}$.

We describe the behavior of the algorithm over epochs by a sequence of operators, $\{\mathcal{V}_m\}_{m \geq 1}$,

which we define as follows. At epoch m, the method uses a vector $\bar{\theta}_m$ in order to recenter the update, where the vector $\bar{\theta}_m$ should be understood as the best current approximation to the unknown vector θ^*. In the ideal scenario, such a recentering would involve the quantity $\mathcal{T}(\bar{\theta}_m)$, where $\mathcal{T}$ denotes the population operator previously defined in equation (2.3). Since we lack direct access to the population operator $\mathcal{T}$, however, we use the Monte Carlo approximation

$$\widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m) := \frac{1}{N_m} \sum_{i \in \mathfrak{D}_m} \widehat{\mathcal{T}}_i(\bar{\theta}_m), \tag{3.9}$$

where the empirical operator $\widehat{\mathcal{T}}_i$ is defined in equation (2.2). Here the set $\mathfrak{D}_m$ is a collection of N_m i.i.d. samples, independent of all other randomness.

Given the pair $(\bar{\theta}_m, \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m))$ and a stepsize $\alpha \in (0,1)$, we define the operator $\mathcal{V}_k$ on $\mathbb{R}^D$ as follows:

$$\theta \mapsto \mathcal{V}_k\left(\theta; \alpha, \bar{\theta}_m, \widetilde{\mathcal{T}}_{N_m}\right) := (1-\alpha)\theta + \alpha\left\{\widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\bar{\theta}_m) + \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m)\right\}. \tag{3.10}$$

As defined in equation (2.2), the quantity $\widehat{\mathcal{T}}_k$ is a stochastic operator, where the randomness is independent of the set of samples $\mathfrak{D}_m$ used to define $\widetilde{\mathcal{T}}_{N_m}$. Consequently, the stochastic operator $\widehat{\mathcal{T}}_k$ is independent of the recentering vector $\widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m)$. Moreover, by construction, for each $\theta \in \mathbb{R}^D$, we have

$$\mathbb{E}\left[\widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\bar{\theta}_m) + \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m)\right] = \mathcal{T}(\theta).$$

Thus, we see that $\mathcal{V}_k$ can be seen as an unbiased stochastic approximation of the population-level Bellman operator. As will be clarified in the analysis, the key effect of the recentering steps is to reduce its associated variance.

3.3.1. A single epoch. Based on the variance-reduced policy evaluation update defined in equation (3.10), we are now ready to define a single epoch of the overall algorithm. We index epochs using the integers $m = 1, 2, \ldots, M$, where M corresponds to the total number of epochs to be run. Epoch m requires as inputs the following quantities:

- a vector $\bar{\theta}$, which is chosen to be the output of the previous epoch,
- a positive integer K denoting the number of steps within the given epoch,
- a positive integer N_m denoting the number of samples used to calculate the Monte Carlo update (3.9),
- a sequence of stepsizes $\{\alpha_k\}_{k\geq 1}^K$ with $\alpha_k \in (0,1)$, and
- a set of fresh samples $\{\widehat{\mathcal{T}}_i\}_{i \in \mathfrak{E}_m}$, with $|\mathfrak{E}_m| = N_m + K$. The first N_m samples are used to define the dataset $\mathfrak{D}_m$ that underlies the Monte Carlo update (3.9), whereas the remaining K samples are used in the K steps within each epoch.

We summarize the operations within a single epoch in Algorithm 1.

The choice of the stepsize sequence $\{\alpha_k\}_{k\geq 1}$ is crucial, and it also determines the epoch length K. Roughly speaking, it is sufficient to choose a large enough epoch length to ensure that the error is reduced by a constant factor in each epoch. In Section 3.3.3 to follow, we study three popular stepsize choices—the constant stepsize (2.5a), the polynomial stepsize (2.5b) and the recentered linear stepsize (2.5c)—and provide lower bounds on the requisite epoch length in each case.

Algorithm 1 RunEpoch $(\bar{\theta}; K, N_m, \{\alpha_k\}_{k=1}^K, \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m})$

Given (a) Epoch length K , (b) Recentering vector $\bar{\theta}$, (c) Recentering sample size N_m, (d) Stepsize sequence $\{\alpha_k\}_{k\geq 1}^K$, (e) Samples $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$

Compute the recentering quantity $\widetilde{\mathcal{T}}_{N_m}(\bar{\theta}) := \frac{1}{N_m} \sum_{i\in\mathfrak{D}_m} \widehat{\mathcal{T}}_i(\bar{\theta})$

Initialize $\theta_1 = \bar{\theta}$

for $k = 1, 2, \ldots, K$ **do**

Compute the variance-reduced update:

$$\theta_{k+1} = \mathcal{V}_k\left(\theta_k; \alpha_k, \bar{\theta}, \widetilde{\mathcal{T}}_{N_m}\right)$$

end for

3.3.2. Overall algorithm. We are now ready to specify our variance-reduced policy-evaluation (VRPE) algorithm. The overall algorithm has five inputs: (a) an integer M, denoting the number of epochs to be run, (b) an integer K, denoting the length of each epoch, (c) a sequence of sample sizes $\{N_m\}_{m=1}^M$ denoting the number of samples used for recentering, (d) Sample batches $\{\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}\}_{m=1}^M$ to be used in m epochs, and (e) a sequence of stepsize $\{\alpha_k\}_{k\geq 1}$ to be used in each epoch. Given these five inputs, we summarize the overall procedure in Algorithm 2:

Algorithm 2 Variance-reduced policy evaluation (VRPE)

Given (a) Number of epochs M, (b) Epoch length K , (c) Recentering sample sizes $\{N_m\}_{m=1}^M$, (d) Sample batches $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$, for $m = 1, \ldots, M$, (e) Stepsize $\{\alpha_k\}_{k=1}^K$

Initialize at $\bar{\theta}_1$

for $m = 1, 2, \ldots, M$ **do**

$\bar{\theta}_{m+1} = \text{RunEpoch}\left(\bar{\theta}_m; K, N_m, \{\alpha\}_{k=1}^K, \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}\right)$

end for

Return $\bar{\theta}_{M+1}$ as the final estimate

In the next section, we provide a detailed description on how to choose these input parameters for three popular stepsize choices (2.5a)–(2.5c). Finally, we reiterate that at epoch m, the algorithm uses $N_m + K$ new samples, and the samples used in the epochs are independent of each other. Accordingly, the total number of samples used in M epochs is given by $KM + \sum_{m=1}^M N_m$.

3.3.3. Instance-dependent guarantees. Given a desired failure probability, $\delta \in (0, 1)$, and a total sample size N, we specify the following choices of parameters in Algorithm 2:

(3.11a) `Number of epochs` $M := \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$

(3.11b)

$$\texttt{Recentering sample sizes :} \quad N_m := 2^m \frac{4^2 \cdot 9^2 \cdot \log(8MD/\delta)}{(1-\gamma)^2} \qquad \text{for } m = 1, \ldots, M$$

(3.11c) `Sample batches:` Partition the N samples to obtain $\{\widehat{\mathcal{T}}_i\}_{i \in \mathfrak{E}_m}$ for $m = 1, \ldots M$

(3.11d)
$$\texttt{Epoch length:} \quad K = \frac{N}{2M}$$

In the following theorem statement, we use (c_1, c_2, c_3, c_4) to denote universal constants.

Theorem 3.3. *(a) Suppose that the input parameters of Algorithm 2 are chosen according to equation (3.11). Furthermore, suppose that the sample size N satisfies one of the following three stepsize-dependent lower bounds:*

(a) $\frac{N}{M} \geq c_1 \frac{\log(8ND/\delta)}{(1-\gamma)^3}$ *for recentered linear stepsize* $\alpha_k = \frac{1}{1+(1-\gamma)k}$,

(b) $\frac{N}{M} \geq c_2 \log(8ND/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$ *for polynomial stepsize* $\alpha_k = \frac{1}{k^\omega}$ *with* $0 < \omega < 1$,

(c) $\frac{N}{M} \geq \frac{c_3}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}$ *for constant stepsize* $\alpha_k = \alpha \leq \frac{1}{5^2 \cdot 32^2} \cdot \frac{(1-\gamma)^2}{\log(8ND/\delta)}$.

Then for any initilization $\bar{\theta}_1$, the output $\bar{\theta}_{M+1}$ satisfies

$$\|\bar{\theta}_{M+1} - \theta^*\|_\infty \leq c_4 \cdot \left\|\bar{\theta}_1 - \theta^*\right\|_\infty \cdot \frac{\log^2((8D/\delta) \cdot \log N)}{N^2(1-\gamma)^4}$$

$$+ c_4 \cdot \left\{ \sqrt{\frac{\log(8DM/\delta)}{N}} \cdot \| \operatorname{diag}(\boldsymbol{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}} + \frac{\log(8DM/\delta)}{N} \cdot b(\theta^*) \right\}, \tag{3.12}$$

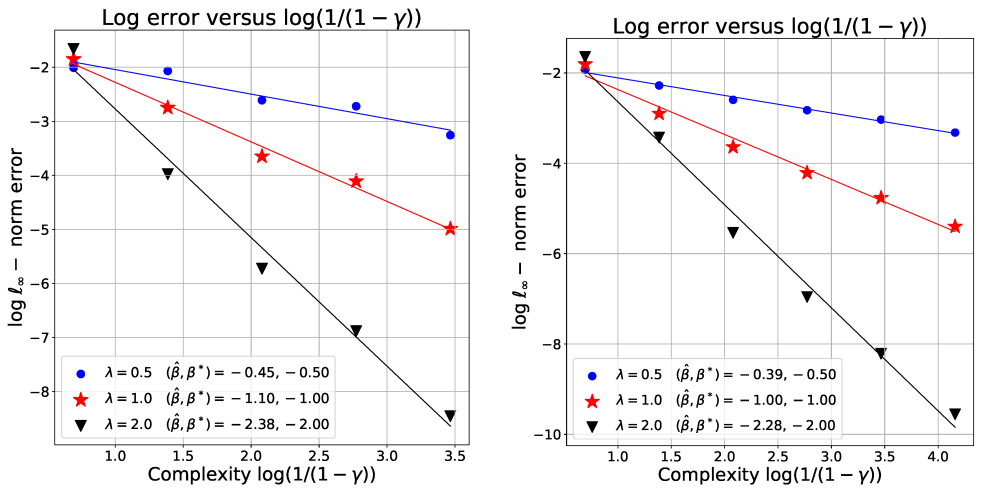
with probability exceeding $1 - \delta$.

See Section 4.3 for the proof of this theorem.

A few comments on the upper bound provided in Theorem 3.3 are in order. In order to facilitate a transparent discussion in this section, we use the notation $\gtrsim$ in order to denote a relation that holds up to logarithmic factors in the tuple $\left(N, D, (1-\gamma)^{-1}\right)$.

Initialization dependence. The first term on the right-hand side of the upper bound (3.12) depends on the initialization $\bar{\theta}_1$. It should be noted that when viewed as a function of the sample size N, this initialization-dependent term decays at a faster rate compared to the other two terms. This indicates that the performance of Algorithm 2 does not depend on the initialization $\bar{\theta}_1$ in a significant way. A careful look at the proof (cf. Section 4.3) reveals that the coefficient of $\|\bar{\theta}_1 - \theta^*\|_\infty$ in the bound (3.12) can be made significantly smaller. In particular, for any $p \geq 1$ the first term in the right-hand side of bound (3.12) can be replaced by

$$c_4 \cdot \frac{\|\bar{\theta}_1 - \theta^*\|_\infty}{N^p} \cdot \frac{\log^p((8D/\delta) \cdot \log N)}{(1-\gamma)^{2p}},$$



(a) Recentered-linear stepsize (b) Polynomial stepsize

Figure 3. Log-log plots of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for the VRPE algorithm. Each point is computed from an average over 1000 trials. Each trial entails drawing $N = \lceil \frac{8}{(1-\gamma)^3} \rceil$ samples from the 2-state MRP in Figure 1 with the parameter choices $p = \frac{4\gamma-1}{3\gamma}$, $\nu = 1$ and $\tau = 1-(1-\gamma)^\lambda$. Each line on each plot represents a different value of λ, as labeled in the legend. We have also plotted the least-squares fits through these points, and the slopes of these lines are also provided in the legend. We also report the pair $(\hat{\beta}, \beta^*)$, where the coefficient $\hat{\beta}$ denotes the slope of the least-squares fit and β^* denotes the slope predicted from the lower bound calculation (3.8). (a) Performance of VRPE for the recentered linear stepsize (2.5c). (b) Performance of VRPRE with polynomially decaying stepsizes (2.5b) with $\omega = 2/3$.

by increasing the recentering sample size (3.11b) by a constant factor and changing the values of the absolute constants (c_1, c_2, c_3, c_4), with these values depending only on the value of p. We have stated and proved a version for $p = 2$. Assuming the number of samples N satisfies $N \geq (1-\gamma)^{-(2+\Delta)}$ for some $\Delta > 0$, the first term on the right-hand side of bound (3.12) can always be made smaller than the other two terms. In the sequel we show that each of the lower bound conditions (a)-(c) in the statement of Theorem 3.3 requires a lower bound of the form $N \gtrsim (1-\gamma)^{-3}$.

Comparing the upper and lower bounds. The second and the third terms in (3.12) show the instance-dependent nature of the upper bound, and they are the dominating terms. Furthermore, assuming that the minimum sample size requirements from Theorems 3.2 and 3.3 are met, we find that the upper bound (3.12) matches the lower bound (3.5) up to logarithmic terms.

It is worthwhile to explicitly compute the minimum sample size requirements in Theorems 3.2 and 3.3. Ignoring the logarithmic terms and constant factors for the moment, unwrapping the lower bound conditions (a)-(c) in Theorem 3.3, we see that for both the constant stepsize and the recentered linear stepsize the sample size needs to satisfy $N \gtrsim (1-\gamma)^{-3}$. For the polynomial stepsize $\lambda_k = \frac{1}{k^\omega}$, the sample size has to be at least $(1-\gamma)^{-\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$. Minimizing the last bound for different values of $\omega \in (0,1)$, we see that the minimum value is

attained at $\omega = 2/3$, and in that case the bound (3.12) is valid when $N \gtrsim (1-\gamma)^{-3}$. Overall, for all the three stepsize choices discussed in Theorem 3.3 we require $N \gtrsim (1-\gamma)^{-3}$ in order to certify the upper bound. Returning to Theorem 3.2, from assumption (3.6) we see that in the best case scenario, Theorem 3.2 is valid as soon as $N \gtrsim (1-\gamma)^{-2}$. Putting together the pieces we find that the sample size requirement for Theorem 3.3 is more stringent than that of Theorem 3.2. Currently, we do not know whether the minimum sample size requirements in Theorems 3.2 and 3.3 are necessary; answering this question is an interesting direction for future research.

Simulation study. It is interesting to demonstrate the sharpness of our bounds via a simulation study, using the same scheme as our previous study of TD(0) with averaging. In Figure 3, we report the results of this study; see the figure caption for further details. At a high level, we see that the VRPE algorithm, with either the recentered linear stepsize (panel (a)) or the polynomial stepsize $t^{-2/3}$, produces errors that decay with the exponents predicted by our instance-dependent theory for $\lambda \in \{0.5, 1.0, 2.0\}$. See the figure caption for further details.

4. Proofs. We now turn to the proofs of our main results. Throughout, we use the shorthand

$$\Sigma_{\mathbf{P}}(\theta) = \operatorname{cov}_{\mathbf{Z}\sim\mathbf{P}}((\mathbf{Z}-\mathbf{P})\theta). \tag{4.1}$$

We also make frequent use of the sandwich relation (3.1d), restated below for convenience:

$$\frac{1}{2}\cdot\{\nu(\mathbf{P},\theta^*)+\rho(\mathbf{P},r)\} \;\le\; \|\operatorname{diag}(\boldsymbol{\Sigma}^*(\mathbf{P},r))\|_\infty^{\frac{1}{2}} \;\le\; 2\cdot\{\nu(\mathbf{P},\theta^*)+\rho(\mathbf{P},r)\} \tag{4.2}$$

4.1. Proof of Proposition 3.1. Recall the definition of the matrix $\Sigma_{\mathbf{P}}(\theta)$ from equation (4.1), and define the covariance matrix

$$\boldsymbol{\Sigma}^*(\mathbf{P},r) = (\mathbf{I}-\gamma\mathbf{P})^{-1}(\gamma^2\Sigma_{\mathbf{P}}(\theta*)+\sigma_r^2\mathbf{I})(\mathbf{I}-\gamma\mathbf{P})^{-T}. \tag{4.3}$$

Recall that we use Z to denote a multivariate Gaussian random vector $Z\sim\mathcal{N}(0,\boldsymbol{\Sigma}^*(\mathbf{P},r))$, and that the sequence $\{\widetilde{\theta}_k^\omega\}_{k\ge1}$ is generated by averaging the iterates of stochastic approximation with polynomial stepsizes (2.5b) with exponent ω. With this notation, the two claims of the theorem are:

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty], \quad \text{and} \tag{4.4a}$$

$$\lim_{N\to\infty}\mathbb{E}\left[\sqrt{N}\cdot\|\widetilde{\theta}_N^\omega-\theta^*\|_\infty\right] = \mathbb{E}[\|Z\|_\infty]. \tag{4.4b}$$

We now prove each of these claims separately.

4.1.1. Proof of equation (4.4a). For the reader's convenience, let us state a version of the Hájek–Le Cam local asymptotic minimax theorem [50, Ch.8, Ch.25]:

Theorem 4.1. *Let $\{P_{\vartheta'}\}_{\vartheta'\in\Theta}$ be a family of parametric models, quadratically mean differentiable with Fisher information matrices $J_{\vartheta'}$. Fix some parameter $\vartheta\in\operatorname{int}(\Theta)$, and consider a*

function $\psi : \Theta \to \mathbb{R}^D$ that is differentiable at ϑ. Then for any quasi-convex loss $L : \mathbb{R}^D \to \mathbb{R}$, we have:

$$\lim_{c\to\infty} \lim_{N\to\infty} \inf_{\hat{\vartheta}_N} \sup_{\substack{\vartheta' \\ \|\vartheta'-\vartheta\|_2 \le c/\sqrt{N}}} \mathbb{E}_{\vartheta'}\left[L\big(\sqrt{N}\cdot(\hat{\vartheta}_N - \psi(\vartheta'))\big)\right] = \mathbb{E}[L(Z)], \tag{4.5}$$

where the infimum is taken over all estimators $\hat{\vartheta}_N$ that are measurable functions of N i.i.d. data points drawn from P_ϑ, and the expectation is taken over a multivariate Gaussian $Z \sim \mathcal{N}(0, \nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta))$.

Returning to the problem at hand, let $\vartheta = (\mathbf{P}, r)$ denote the unknown parameters of the model and let $\psi(\vartheta) = \theta(\mathcal{P}) = (\mathbf{I} - \gamma\mathbf{P})^{-1} r$ denote the target vector.

In the first case where $\vartheta = (\mathbf{P}, r)$ lies in the interior of the parameter space[5], a direct application of Theorem 4.1 shows that

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty] \quad \text{where } Z = \mathcal{N}(0, \nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta)), \tag{4.6}$$

where J_ϑ is the Fisher information at ϑ. The following result provides a more explicit form of the covariance of Z:

Lemma 4.2. *We have the identity*

$$\nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta) = \mathbf{\Sigma}^*(\mathbf{P}, r) := (\mathbf{I} - \gamma\mathbf{P})^{-1}(\gamma^2 \Sigma_{\mathbf{P}}(\theta^*) + \sigma_r^2 \mathbf{I})(\mathbf{I} - \gamma\mathbf{P})^{-T}. \tag{4.7}$$

Although the proof of this claim is relatively straightforward, it involves some lengthy and somewhat tedious calculations; we refer the reader to Appendix SM1.1 for the proof.

Given the result from Lemma 4.2, the claim (4.4a) follows by substituting the relation (4.7) into (4.6). This proves the case when ϑ is in the interior of the parameter space.

In the second case where ϑ lies on the boundary of the parameter space, with some diligent work, one can use the same arguments to prove the claim (4.4a). In fact, we need to show additionally that Theorem 4.1 also holds when ϑ lies on the boundary. The classical delta-method allows us to reduce the problem to showing that the local asymptotic minimax result holds for estimating $\mathbf{P}$ when $\mathbf{P}$ lies on the boundary, i.e., $\mathbf{P}_{i,j} \in \{0, 1\}$ for some $\{i, j\}$. This requires a direct and tedious verification, which we leave the details to the reader. Here we provide only the basic intuition. The key observation is (i) $\mathbf{P}_{i,j}$ is the mean of a Bernoulli random variable, and (ii) one can verify easily the local asymptotic minimax lower and upper bound for estimating $\mathbf{P}_{i,j}$ are precisely equal to each other, and in fact, both are equal to zero when $\mathbf{P}_{i,j} \in \{0, 1\}$, since the Bernoulli variable becomes deterministic when $\mathbf{P}_{i,j} \in \{0, 1\}$.

4.1.2. Proof of equation (4.4b). The proof of this claim follows from the results of Polyak and Juditsky [40, Theorem 1], once their assumptions are verified for TD(0) with polynomial stepsizes. Recall that the TD iterates in equation (2.4) are given by the sequence $\{\theta_k\}_{k\ge 1}$, and that $\widetilde{\theta}_k^\omega$ denotes the k-th iterate generated by averaging.

[5] More precisely, this means that $\mathbf{P}$ lies in the relative interior of the convex set $\{\mathbf{P} : \mathbf{P}\mathbf{1} = \mathbf{1}, \mathbf{P} \ge 0\}$.

For each $k \geq 1$, note the following equivalence between the notation of our paper and that of Polyak and Juditsky [40], or PJ for short:

$$x_k \equiv \theta_k, \qquad \gamma_k \equiv \alpha_k, \qquad \mathbf{A} \equiv \mathbf{I} - \gamma \mathbf{P}, \quad \text{and} \quad \xi_k = (R_k - r) + (\mathbf{Z}_k - \mathbf{P})\theta_k.$$

Let us now verify the various assumptions in the PJ paper. Assumption 2.1 in the PJ paper holds by definition, since the matrix $\mathbf{I} - \gamma \mathbf{P}$ is Hurwitz. Assumption 2.2 in the PJ paper is also satisfied by the polynomial stepsize sequence for any exponent $\omega \in (0, 1)$.

It remains to verify the assumptions that must be satisfied by the noise sequence $\{\xi_k\}_{k \geq 1}$. In order to do so, write the k-th such iterate as

$$\xi_k = (R_k - r) + (\mathbf{Z}_k - \mathbf{P})\theta^* + (\mathbf{Z}_k - \mathbf{P})(\theta_k - \theta^*).$$

Since $\mathbf{Z}_k$ is independent of the sequence $\{\theta_i\}_{i=1}^k$, it follows that the condition

$$\lim_{N \to \infty} \mathbb{E}\left[\|\theta_N - \theta^*\|_2^2\right] = 0 \tag{4.8}$$

suffices to guarantee that Assumptions 2.3–2.5 in the PJ paper are satisfied. We now claim that for each $\omega \in (1/2, 1]$, condition (4.8) is satisfied by the TD iterates. Taking this claim as given for the moment, note that applying Theorem 1 of Polyak and Juditsky [40] establishes claim (4.4b), for any exponent $\omega \in (1/2, 1)$.

It remains to establish condition (4.8). For any $\omega \in (1/2, 1]$, the sequence of stepsizes $\{\alpha_k\}_{k \geq 1}$ satisfies the conditions

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Consequently, classical results due to Robbins and Monro [41, Theorem 2] guarantee ℓ^2-convergence of θ_N to θ^*.

4.2. Proof of Theorem 3.2. Throughout the proof, we use the notation $\mathcal{P} = (\mathbf{P}, r)$ and $\mathcal{P}' = (\mathbf{P}', r')$ to denote, respectively, the problem instance at hand and its alternative. Moreover, we use $\theta^* \equiv \theta(\mathcal{P})$ and $\theta(\mathcal{P}')$ to denote the associated target parameters for each of the two problems $\mathcal{P}$ and $\mathcal{P}'$. We use $\Delta_{\mathbf{P}} = \mathbf{P} - \mathbf{P}'$ and $\Delta_r = r - r'$ to denote the differences of the parameters. For probability distributions, we use P and P' to denote the marginal distribution of a single observation under $\mathcal{P}$ and $\mathcal{P}'$, and use P^N and $(P')^N$ to denote the distribution of N i.i.d observations drawn from P or P', respectively.

4.2.1. Proof structure. We introduce two special classes of alternatives of interest, denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively:

$$\mathcal{S}_1 = \left\{\mathcal{P}' = (\mathbf{P}', r') \mid r' = r\right\}, \quad \text{and} \quad \mathcal{S}_2 = \left\{\mathcal{P}' = (\mathbf{P}', r') \mid \mathbf{P}' = \mathbf{P}\right\}.$$

In words, the class $\mathcal{S}_1$ consists of alternatives $\mathcal{P}'$ that have the same reward vector r as $\mathcal{P}$, but a different transition matrix $\mathbf{P}'$. Similarly, the class $\mathcal{S}_2$ consists of alternatives $\mathcal{P}'$ with

the same transition matrix $\mathbf{P}$, but a different reward vector. By restricting the alternative $\mathcal{P}'$ within class $\mathcal{S}_1$ and $\mathcal{S}_2$, we can define *restricted versions* of the local minimax risk, namely

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_1) \equiv \sup_{\mathcal{P}'\in\mathcal{S}_1}\inf_{\hat{\theta}_N}\max_{\mathcal{P}\in\{\mathcal{P},\mathcal{P}'\}} \mathbb{E}_{\mathcal{P}}\left[\sqrt{N}\cdot\left\|\hat{\theta}_N-\theta(\mathcal{P})\right\|_\infty\right], \quad \text{and} \tag{4.9a}$$

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_2) \equiv \sup_{\mathcal{P}'\in\mathcal{S}_2}\inf_{\hat{\theta}_N}\max_{\mathcal{P}\in\{\mathcal{P},\mathcal{P}'\}} \mathbb{E}_{\mathcal{P}}\left[\sqrt{N}\cdot\left\|\hat{\theta}_N-\theta(\mathcal{P})\right\|_\infty\right]. \tag{4.9b}$$

The main part of the proof involves showing that there is a universal constant $c>0$ such that the lower bounds

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_1) \geq c\cdot\gamma\nu(\mathbf{P},\theta^*), \quad \text{and} \tag{4.10a}$$

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_2) \geq c\cdot\rho(\mathbf{P},r) \tag{4.10b}$$

both hold (assuming that the sample size N is sufficiently large to satisfy the condition (3.6)). Since we have $\mathfrak{M}_N(\mathcal{P}) \geq \max\{\mathfrak{M}_N(\mathcal{P};\mathcal{S}_1),\mathfrak{M}_N(\mathcal{P};\mathcal{S}_2)\}$, these lower bounds in conjunction with the sandwich relation (4.2) imply the claim Theorem 3.2. The next section shows how to prove these two bounds.

4.2.2. Proof of the lower bounds (4.10a) and (4.10b):. Our first step is to lower bound the local minimax risk for each problem class in terms of a modulus of continuity between the Hellinger distance and the ℓ_∞-norm.

Lemma 4.3. *For each $\mathcal{S}\in\{\mathcal{S}_1,\mathcal{S}_2\}$, we have the lower bound $\mathfrak{M}_N(\mathcal{P};\mathcal{S}) \geq \frac{1}{8}\cdot\underline{\mathfrak{M}}_N(\mathcal{P};\mathcal{S})$, where we define*

$$\underline{\mathfrak{M}}_N(\mathcal{P};\mathcal{S}) := \sup_{\mathcal{P}'\in\mathcal{S}}\left\{\sqrt{N}\cdot\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty \mid d_{\mathrm{hel}}(P,P')\leq\frac{1}{2\sqrt{N}}\right\}, \tag{4.11}$$

where $d_{\mathrm{hel}}(P,P')$ denotes the Hellinger distance between the two distributions P and P'. The proof of Lemma 4.3 follows a relatively standard argument, one which reduces estimation to testing; see Appendix SM2.1 for details.

This lemma allows us to focus our remaining attention on lower bounding the quantity $\underline{\mathfrak{M}}_N(\mathcal{P};\mathcal{S})$. In order to do so, we need both a lower bound on the ℓ_∞-norm $\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty$ and an upper bound on the Hellinger distance $d_{\mathrm{hel}}(P,P')$. These two types of bounds are provided in the following two lemmas. We begin with lower bounds on the ℓ_∞-norm:

Lemma 4.4. *(a) For any $\mathcal{P}$ and for all $\mathcal{P}'\in\mathcal{S}_1$, we have*

$$\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty \geq \left(1-\frac{\gamma}{1-\gamma}\left\|\Delta_{\mathbf{P}}\right\|_\infty\right)_+\cdot\left\|\gamma(\mathbf{I}-\gamma\mathbf{P})^{-1}\Delta_{\mathbf{P}}\theta^*\right\|_\infty. \tag{4.12a}$$

(b) For any $\mathcal{P}$ and for all $\mathcal{P}'\in\mathcal{S}_2$, we have

$$\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty \geq \left\|(\mathbf{I}-\gamma\mathbf{P})^{-1}\Delta_r\right\|_\infty. \tag{4.12b}$$

See Appendix SM2.2 for the proof of this claim.

Next, we require upper bounds on the Hellinger distance:

Lemma 4.5. *(a) For each $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_1$, we have*

$$d_{\text{hel}}(P, P')^2 \leq \frac{1}{2} \sum_{i,j} \frac{((\Delta_{\mathbf{P}})_{i,j})^2}{\mathbf{P}_{i,j}}. \tag{4.13a}$$

(b) For each $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_2$, we have

$$d_{\text{hel}}(P, P')^2 \leq \frac{1}{2\sigma_r^2} \|r_1 - r_2\|_2^2. \tag{4.13b}$$

See Appendix SM2.3 for the proof of this upper bound.

Using Lemmas 4.4 and 4.5, we can derive two different lower bounds. First, we have the lower bound $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}_1) \geq \underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$, where

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_1} \left\{ \sqrt{N} \cdot \left(1 - \frac{\gamma \|\Delta_{\mathbf{P}}\|_\infty}{1-\gamma}\right)_+ \cdot \left\|\gamma(\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} \theta^*\right\|_\infty \mid \sum_{i,j} \frac{\left((\Delta_{\mathbf{P}})_{i,j}\right)^2}{\mathbf{P}_{i,j}} \leq \frac{1}{2N} \right\}. \tag{4.14a}$$

Second, we have the lower bound $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}_2) \geq \underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2)$, where

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_2} \left\{ \sqrt{N} \cdot \| \left(\mathbf{I} - \gamma \mathbf{P}\right)^{-1} \Delta_r \|_\infty \mid \frac{1}{\sigma_r^2} \|r_1 - r_2\|_2 \leq \frac{1}{2N} \right\}. \tag{4.14b}$$

In order to complete the proofs of the two lower bounds (4.10a) and (4.10b), it suffices to show that

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) \geq \frac{1}{\sqrt{2}} \cdot \rho(\mathbf{P}, r), \quad \text{and} \tag{4.15a}$$

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) \geq \frac{1}{2\sqrt{2}} \cdot \gamma \nu(\mathbf{P}, \theta^*). \tag{4.15b}$$

Proof of the bound (4.15a). This lower bound is easy to show—it follows from the definition:

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) = \frac{\sigma_r}{\sqrt{2}} \left\|(\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_r\right\|_\infty = \frac{1}{\sqrt{2}} \rho(\mathbf{P}, r).$$

Proof of the bound (4.15b). The proof of this claim is much more delicate. Our strategy is to construct a special "hard" alternative, $\overline{\mathcal{P}} \in \mathcal{S}_1$, that leads to a good lower bound on $\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$. Lemma 4.6 below is the main technical result that we require:

Lemma 4.6. *There exists some probability transition matrix $\bar{\mathbf{P}}$ with the following properties:*

(a) It satisfies the constraint $\sum_{i,j} \frac{\left((\bar{\mathbf{P}} - \mathbf{P})_{i,j}\right)^2}{\mathbf{P}_{i,j}} \leq \frac{1}{2N}$.

(b) It satisfies the inequalities

$$\left\|\bar{\mathbf{P}} - \mathbf{P}\right\|_\infty \leq \frac{1}{\sqrt{2N}}, \quad \textit{and} \quad \left\|\gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}(\bar{\mathbf{P}} - \mathbf{P})\theta^*\right\|_\infty \geq \frac{\gamma}{\sqrt{2N}} \cdot \nu(\mathbf{P}, \theta^*).$$

See Appendix SM2.4 for the proof of this claim.

Given the matrix $\bar{\mathbf{P}}$ guaranteed by this lemma, we consider the "hard" problem $\bar{\mathcal{P}} := (\bar{\mathbf{P}}, r) \in \mathcal{S}_1$. From the definition of $\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$ in equation (4.14a), we have that

$$\begin{aligned}\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) &\geq \sqrt{N} \cdot \left(1 - \frac{\gamma}{1-\gamma}\left\|\mathbf{P} - \bar{\mathbf{P}}\right\|_\infty\right)_+ \cdot \left\|\gamma(\mathbf{I} - \gamma\bar{\mathbf{P}})^{-1}(\mathbf{P} - \bar{\mathbf{P}})\theta^*\right\|_\infty \\ &\geq \sqrt{N} \cdot \left(1 - \frac{\gamma}{1-\gamma} \cdot \frac{1}{\sqrt{2N}}\right)_+ \cdot \frac{\gamma}{\sqrt{2N}} \cdot \nu(\mathbf{P}, \theta^*) \geq \frac{1}{2\sqrt{2}} \cdot \gamma\nu(\mathbf{P}, \theta^*),\end{aligned}$$

where the last inequality follows by the assumed lower bound $N \geq \frac{4\gamma^2}{(1-\gamma)^2}$. This completes the proof of the lower bound (4.15b).

4.3. Proof of Theorem 3.3. This section is devoted to the proof of Theorem 3.3, which provides the achievability results for variance-reduced policy evaluation.

4.3.1. Proof of part (a). We begin with a lemma that characterizes the progress of Algorithm 2 over epochs:

Lemma 4.7. *Under the assumptions of Theorem 3.3 (a), there is an absolute constant c such that for each epoch $m = 1, \ldots, M$, we have:*

$$\begin{aligned}\left\|\bar{\theta}_{m+1} - \theta^*\right\|_\infty &\leq \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{4} \\ &\quad + c\left\{\sqrt{\frac{\log(8DM/\delta)}{N_m}}\Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big) + \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*)\right\}, \qquad (4.16)\end{aligned}$$

with probability exceeding $1 - \frac{\delta}{M}$.

Taking this lemma as given for the moment, let us complete the proof. We use the shorthand

$$\tau_m := \sqrt{\frac{\log(8DM/\delta)}{N_m}}\Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big) \quad \text{and} \quad \eta_m := \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*) \qquad (4.17)$$

to ease notation, and note that $\frac{\tau_m}{\sqrt{2}} \leq \tau_{m+1}$ and $\frac{\eta_m}{2} \leq \eta_{m+1}$, for each $m \geq 1$. Using this notation and unwrapping the recursion relation from Lemma 4.7, we have

$$\begin{aligned}\left\|\bar{\theta}_{M+1} - \theta^*\right\|_\infty &\leq \frac{\left\|\bar{\theta}_M - \theta^*\right\|_\infty}{4} + c(\tau_M + \eta_M) \\ &\overset{(i)}{\leq} \frac{\left\|\bar{\theta}_{M-1} - \theta^*\right\|_\infty}{4^2} + \frac{c}{2}(\tau_M + \eta_M) + c(\tau_M + \eta_M) \\ &\overset{(ii)}{\leq} \frac{\left\|\bar{\theta}_1 - \theta^*\right\|_\infty}{4^M} + 2c(\tau_M + \eta_M).\end{aligned}$$

Here, step (i) follows by applying the one-step application of the recursion (4.16), and by using the upper bounds $\frac{\tau_m}{\sqrt{2}} \leq \tau_{m+1}$ and $\frac{\eta_m}{2} \leq \eta_{m+1}$. Step (ii) follows by repeated application of the recursion (4.16). The last inequality holds with probability at least $1-\delta$ by a union bound over the M epochs.

It remains to express the quantities 4^M, τ_M and η_M—all of which are controlled by the recentering sample size N_M—in terms of the total number of available samples N. Towards this end, observe that the total number of samples used for recentering at M epochs is given by

$$\sum_{m=1}^{M} N_m \asymp 2^M \cdot \frac{\log(8MD/\delta)}{(1-\gamma)^2}.$$

Substituting the value of $M = \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$ we have

$$c_1 N \leq N_M \asymp \sum_{m=1}^{M} N_m \leq \frac{N}{2},$$

where c_1 is a universal constant. Consequently, the total number of samples used by Algorithm 2 is given by

$$MK + \sum_{m=1}^{M} N_m \leq \frac{N}{2} + \frac{N}{2} = N,$$

where in the last equation we have used the fact that $MK = \frac{N}{2}$. Finally, using $M = \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$ we have the following relation for some universal constant c:

$$4^M = c \cdot \frac{N^2(1-\gamma)^4}{\log^2((8D/\delta)\cdot\log N)}$$

Putting together the pieces and using the sandwich relation (4.2), we conclude that

$$\begin{aligned}\|\bar{\theta}_{M+1} - \theta^*\|_\infty \leq c_2 \left\|\bar{\theta}_1 - \theta^*\right\|_\infty \cdot \frac{\log^2((8D/\delta)\cdot\log N)}{N^2(1-\gamma)^4} \\ + c_2\left\{\sqrt{\frac{\log(8DM/\delta)}{N}} \cdot \|\operatorname{diag}(\boldsymbol{\Sigma}^*(\mathbf{P}, r))\|_\infty^{\frac{1}{2}} + \frac{\log(8DM/\delta)}{N} \cdot b(\theta^*)\right\},\end{aligned}$$

for a suitable universal constant c_2. The last bound is valid with probability exceeding $1-\delta$ via the union bound. In order to complete the proof, it remains to prove Lemma 4.7, which we do in the following subsection.

4.3.2. Proof of Lemma 4.7. We now turn to the proof of the key lemma within the argument. We begin with a high-level overview in order to provide intuition. In the m-th epoch that updates the estimate from $\bar{\theta}_m$ to $\bar{\theta}_{m+1}$, the vector $\bar{\theta} \equiv \bar{\theta}_m$ is used to recenter the updates. Our analysis of the m-th epoch is based on a sequence of recentered operators

$\{\mathcal{J}_k^m\}_{k\geq 1}$ and their population analogs $\mathcal{J}^m(\theta)$, analyzed conditionally on $\widetilde{\mathcal{T}}_N(\bar{\theta}_m)$. The action of these operators on a point θ is given by the relations

$$\text{(4.18a)} \quad \mathcal{J}_k^m(\theta) := \widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\bar{\theta}_m) + \widetilde{\mathcal{T}}_N(\bar{\theta}_m), \quad \text{and} \quad \mathcal{J}^m(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\bar{\theta}_m) + \widetilde{\mathcal{T}}_N(\bar{\theta}_m).$$

By definition, the updates within epoch m can be written as

$$\text{(4.18b)} \qquad \theta_{k+1} = (1-\alpha_k)\,\theta_k + \alpha_k \mathcal{J}_k^m(\theta_k).$$

Note that the operator $\mathcal{J}^m$ is γ-contractive in $\|\cdot\|_\infty$-norm, and as a result it has a unique fixed point, which we denote by $\widehat{\theta}_m$. Since $\mathcal{J}^m(\theta) = \mathbb{E}[\mathcal{J}_k^m(\theta)]$ by construction, when studying epoch m, it is natural to analyze the convergence of the sequence $\{\theta_k\}_{k\geq 1}$ to $\widehat{\theta}_m$.

Suppose that we have taken K steps within epoch m. Applying the triangle inequality yields the bound

$$\text{(4.18c)} \qquad \|\bar{\theta}_{m+1} - \theta^*\|_\infty = \|\theta_{K+1} - \theta^*\|_\infty \leq \left\|\theta_{K+1} - \widehat{\theta}_m\right\|_\infty + \left\|\widehat{\theta}_m - \theta^*\right\|_\infty.$$

With this decomposition, our proof of Lemma 4.7 is based on two auxiliary lemmas that provide high-probability upper bounds on the two terms on the right-hand side of inequality (4.18c).

Lemma 4.8. *Let (c_1, c_2, c_3) be positive numerical constants, and suppose that the epoch length K satisfies one the following three stepsize-dependent lower bounds:*

(a) $K \geq c_1 \frac{\log(8KMD/\delta)}{(1-\gamma)^3}$ *for recentered linear stepsize* $\alpha_k = \frac{1}{1+(1-\gamma)k}$,

(b) $K \geq c_2 \log(8KMD/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$ *for polynomial stepsize* $\alpha_k = \frac{1}{k^\omega}$ *with* $0 < \omega < 1$,

(c) $K \geq \frac{c_3}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}$ *for constant stepsize* $\alpha_k = \alpha \leq \frac{(1-\gamma)^2}{\log(8KMD/\delta)} \cdot \frac{1}{5^2 \cdot 32^2}$.

Then after K update steps with epoch m, the iterate θ_{K+1} satisfies the bound

$$\text{(4.19)} \quad \|\theta_{K+1} - \widehat{\theta}_m\|_\infty \leq \frac{1}{8}\|\bar{\theta}_m - \theta^*\|_\infty + \frac{1}{8}\|\widehat{\theta}_m - \theta^*\|_\infty \quad \textit{with probability at least } 1 - \tfrac{\delta}{2M}.$$

See Appendix SM3.1 for the proof of this claim.

Our next auxiliary result provides a high-probability bound on the difference $\|\widehat{\theta}_m - \theta^*\|_\infty$.

Lemma 4.9. *There is an absolute constant c_4 such that for any recentering sample size satisfying $N_m \geq 4^2 \cdot 9^2 \cdot \frac{\log(MD/\delta)}{(1-\gamma)^2}$, we have*

$$\|\widehat{\theta}_m - \theta^*\|_\infty \leq \tfrac{1}{9}\|\bar{\theta}_m - \theta^*\|_\infty + c_4 \left\{ \sqrt{\frac{\log(8DM/\delta)}{N_m}} \Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r) \Big) + \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*) \right\},$$

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

See Appendix SM3.2 for the proof of this claim.

With Lemmas 4.8 and 4.9 in hand, the remainder of the proof is straightforward. Recall from equation (4.17) the shorthand notation τ_m and η_m. Using our earlier bound (4.18c), we have that at the end of epoch m (which is also the starting point of epoch $m+1$),

$$\begin{aligned}
\left\|\bar{\theta}_{m+1} - \theta^*\right\|_\infty &\leq \|\theta_{K+1} - \widehat{\theta}_m\|_\infty + \|\widehat{\theta}_m - \theta^*\|_\infty \\
&\overset{(i)}{\leq} \left\{ \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{8} + \frac{1}{8}\left\|\widehat{\theta}_m - \theta^*\right\|_\infty \right\} + \left\|\widehat{\theta}_m - \theta^*\right\|_\infty \\
&= \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{8} + \frac{9}{8} \cdot \left\|\widehat{\theta}_m - \theta^*\right\|_\infty \\
&\overset{(ii)}{\leq} \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{8} + \frac{1}{8}\left\{\|\bar{\theta}_m - \theta^*\|_\infty + c_4(\tau_m + \eta_m)\right\} \\
&\leq \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{4} + c_4(\tau_m + \eta_m),
\end{aligned}$$

where inequality (i) follows from Lemma 4.8(a), and inequality (ii) from Lemma 4.9. Finally, the sequence of inequalities above holds with probability at least $1 - \frac{\delta}{M}$ via a union bound. This completes the proof of Lemma 4.7.

4.3.3. Proof of Theorem 3.3, parts (b) and (c). The proofs of Theorem 3.3 parts (b) and (c) require versions of Lemma 4.7 for the polynomial stepsize (2.5b) and constant stepsize (2.5a), respectively. These two versions of Lemma 4.7 can be obtained by simply replacing Lemma 4.8, part (a), by Lemma 4.8, parts (b) and (c), respectively, in the proof of Lemma 4.7.

5. Discussion. In this paper, we have undertaken an instance-specific analysis of the problem of policy evaluation in discounted Markov decision processes. Our contribution is three-fold. First, we provided a non-asymptotic instance-dependent local-minimax bound on the ℓ_∞-error for the policy evaluation problem under the generative model. Next, via careful simulations, we showed that the standard TD-learning algorithm—even when combined with Polyak-Ruppert iterate averaging—does not yield ideal non-asymptotic behavior as captured by our lower bound. In order to remedy this difficulty, we introduced and analyzed a variance-reduced (VR) version of the standard TD-learning algorithm which achieves our non-asymptotic instance-dependent lower bound up to logarithmic factors. We close with some discussions of interesting open directions.

Exploring the connection between variance and higher-order terms. Underlying our results is an exploration of the variance of various algorithms together with their higher-order error terms. Note that both Polyak–Ruppert averaging and the variance reduction (VR) device are methods by which the natural stochastic approximation iterates are stabilized; in that sense, the abstract phenomenon of "variance reduction" is common to both algorithms. On the other hand, the higher-order terms in the error of the averaging estimator (which vanish as $N \to \infty$) end up dominating the risk for small sample sizes, but the VR update is more effective at controlling these higher-order error terms. Another typical method that achieves variance reduction is adding minibatching to stochastic approximation. Now one would still expect that if minibatching were employed in conjunction with averaging, the issue above of large higher-order terms would persist unless the batch size was chosen to grow with the effective horizon $1/(1-\gamma)$; indeed, in our VR update, we use a large sample size (of the order $(1-\gamma)^{-3}$)

to recenter our updates in each epoch. A deeper exploration of the interaction between the variance of an algorithm with its higher-order terms is an interesting open direction in related problems.

Sharp characterization of sample size threshold. Both the upper and lower bounds discussed in this paper hold when the sample size is bigger than an explicit threshold; relaxing this minimum sample size requirement is an interesting future research direction. We note that this question is quite a delicate one; indeed, deriving sharp thresholds on such a worst-case sample size "barrier" in tabular reinforcement learning has been a topic of recent focus (see, e.g. the very recent paper [33]).

Accommodating Markov noise and function approximation. In this paper, our study of of policy evaluation was restricted to the tabular case, and focused on providing ℓ_∞-bounds under the generative model. Arguably, the more relevant setting in practice is where we do not have access to a simulator and instead observe a trajectory of observations from the Markov chain. In this setting, the most natural guarantees are usually obtained in a weighted ℓ_2-norm (see, e.g., Tsitsiklis and Van Roy [48]). It is an interesting open question as to whether variance-reduced policy evaluation still has good performance in this setting. Our current analysis—much of our which relies on specific contraction properties that hold for the empirical Bellman update—does not immediately apply. Understanding the instance-specific (sub-)optimality of Polyak—Ruppert averaging in the Markov setting is also an interesting problem in its own right. We note that preliminary progress in this direction has been made by a subset of the current authors in papers written shortly after ours [35, 36]; they provided upper bounds on averaged stochastic approximation with control on higher-order terms. However, this analysis does not offer any guidance about the optimality of these terms.

Extensions to the Hurwitz case and other error metrics. Finally, let us briefly comment on the case of solving the linear system $\mathbf{A}\theta^* = b$ from noisy observations of the pair $(\mathbf{A}, b)$. This problem has received significant attention in the case when the matrix $-\mathbf{A}$ is Hurwitz[6] [40, 4, 29, 35], and the policy evaluation setting considered in this paper is a special case. Indeed, Proposition 3.1 has a direct analog in the more general Hurwitz setting. On the other hand, while our non-asymptotic lower bound in Theorem 3.2 uses some properties that are specific to the MRP setting, we expect that a similar lower bound ought to apply to a large subclass of Hurwitz matrices. Our simulations in Section 3.2 were shown for a particular family of Hurwitz matrices, corresponding to a collection of MRPs with varying discount factor, but we expect that a similar phenomenon ought to hold for other carefully constructed families. Note that while the simulations in Section 3.2 were illustrated for the ℓ_∞-norm, Polyak–Ruppert averaging is also clearly suboptimal for this family in other norms (e.g. the ℓ_2 norm) since our instance made use of a two-dimensional example and the error would behave equivalently in all ℓ_p norms. Finally, as alluded to above, our proof of Theorem 3.3 leverages the MRP setting and contractivity properties with respect to the ℓ_∞ norm, and it is an interesting open question whether a similar result can be proved in the general Hurwitz case for arbitrary norms.

REFERENCES

[6] A Hurwitz matrix is one whose eigenvalues all have real part strictly negative.

[1] A. Agarwal, N. Jiang, and S. M. Kakade, *Reinforcement learning: Theory and algorithms*, Technical Report, Department of Computer Science, University of Washington, (2019).

[2] M. G. Azar, R. Munos, and H. J. Kappen, *Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model*, Machine Learning, 91 (2013), pp. 325–349.

[3] M. G. Azar, I. Osband, and R. Munos, *Minimax regret bounds for reinforcement learning*, in Proceedings of the International Conference on Machine Learning, 2017.

[4] F. Bach and E. Moulines, *Non-asymptotic analysis of stochastic optimization algorithms for machine learning*, in Advances in Neural Information Processing Systems, December 2011.

[5] L. Baird, *Residual algorithms: Reinforcement learning with function approximation*, in Machine Learning Proceedings 1995, Elsevier, 1995, pp. 30–37.

[6] D. Bertsekas, *Dynamic Programming and Stochastic Control*, vol. 2, Athena Scientific, Belmont, MA, 1995.

[7] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, vol. 1, Athena Scientific, Belmont, MA, 1995.

[8] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[9] J. Bhandari, D. Russo, and R. Singal, *A finite time analysis of temporal difference learning with linear function approximation*, arXiv preprint arXiv:1806.02450, (2018).

[10] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation*, Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet, 65 (1983), pp. 181–238.

[11] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Springer, 2009.

[12] V. S. Borkar and S. P. Meyn, *The ODE method for convergence of stochastic approximation and reinforcement learning*, SIAM Journal on Control and Optimization, 38 (2000), pp. 447–469.

[13] T. Cai and M. Low, *A framework for estimating convex functions*, Statistica Sinica, 25 (2015), pp. 423–456.

[14] T. T. Cai and M. G. Low, *An adaptation theory for nonparametric confidence intervals*, Annals of Statistics, 32 (2004), pp. 1805–1840.

[15] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, *Finite sample analyses for TD(0) with function approximation*, in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[16] C. Dann, G. Neumann, and J. Peters, *Policy evaluation with temporal differences: A survey and comparison*, The Journal of Machine Learning Research, 15 (2014), pp. 809–883.

[17] T. T. Doan, S. T. Maguluri, and J. Romberg, *Finite-time performance of distributed temporal difference learning with linear function approximation*, arXiv preprint arXiv:1907.12530, (2019).

[18] D. L. Donoho and R. C. Liu, *Geometrizing rates of convergence I*, Tech. Report 137, University of California, Berkeley, Department of Statistics, 1987.

[19] D. L. Donoho and R. C. Liu, *Geometrizing rates of convergence II*, Annals of Statistics, 19 (1991), pp. 633–667.

[20] J. C. Duchi and F. Ruan, *The right complexity measure in locally private estimation: It is not the Fisher information*, arXiv preprint arXiv:1806.05756, (2018).

[21] R. Durrett, *Essentials of Stochastic Processes*, Springer, 1999.

[22] W. Feller, *An Introduction to Probability Theory and its Applications: Volume II*, John Wiley and Sons, New York, 1966.

[23] J. Hájek, *Local asymptotic minimax and admissibility in estimation*, in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1972, pp. 175–194.

[24] T. Jaakkola, M. I. Jordan, and S. P. Singh, *Convergence of stochastic iterative dynamic programming algorithms*, in Advances in Neural Information Processing Systems, 1994, pp. 703–710.

[25] N. Jiang and A. Agarwal, *Open problem: The dependence of sample complexity lower bounds on planning horizon*, in Proceedings of the Conference On Learning Theory, 2018, pp. 3395–3398.

[26] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems, 2013, pp. 315–323.

[27] M. Kearns and S. Singh, *Finite-sample convergence rates for Q-learning and indirect algorithms*, in Advances in Neural Information Processing Systems, 1999.

[28] N. Korda and P. La, *On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence*, in International Conference on Machine Learning, 2015, pp. 626–634.

[29] C. Lakshminarayanan and C. Szepesvari, *Linear stochastic approximation: How far does constant step-size and iterate averaging go?*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2018, pp. 1347–1355.
[30] T. Lattimore and M. Hutter, *Near-optimal PAC bounds for discounted MDPs*, Theoretical Computer Science, 558 (2014), pp. 125–143.
[31] L. Le Cam, *Limits of experiments*, in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1972, pp. 245–261.
[32] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*, Springer, 2000.
[33] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, *Breaking the sample size barrier in model-based reinforcement learning with a generative model*, arXiv preprint arXiv:2005.12900, (2020).
[34] O.-A. Maillard, T. A. Mann, and S. Mannor, *How hard is my MDP? The distribution-norm to the rescue*, in Advances in Neural Information Processing Systems, 2014, pp. 1835–1843.
[35] W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, *On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration*, arXiv preprint arXiv:2004.04719, (2020).
[36] W. Mou, A. Pananjady, and M. J. Wainwright, *Optimal oracle inequalities for solving projected fixed-point equations*, arXiv preprint arXiv:2012.05299, (2020).
[37] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal of Optimization, 19 (2009), pp. 1574–1609.
[38] A. Pananjady and M. J. Wainwright, *Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning*, IEEE Transactions on Information Theory, 67 (2021), pp. 566–585.
[39] B. T. Polyak, *New stochastic approximation type procedures*, Avtomat. Telemekh., (1990), pp. 98–107.
[40] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control and Optimization, 30 (1992), pp. 838–855.
[41] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics, (1951), pp. 400–407.
[42] D. Ruppert, *Efficient estimators from a slowly convergent Robbins-Monro process*, Tech. Report 781, Cornell University, 1988.
[43] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye, *Near-optimal time and sample complexities for solving Markov decision processes with a generative model*, in Advances in Neural Information Processing Systems, 2018, pp. 5186–5196.
[44] C. Stein, *Efficient nonparametric testing and estimation*, in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1956, pp. 187–195.
[45] R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine Learning, 3 (1988), pp. 9–44.
[46] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2nd ed., 2018.
[47] V. B. Tadic, *On the almost sure rate of convergence of linear stochastic approximation algorithms*, IEEE Transactions on Information Theory, 50 (2004), pp. 401–409.
[48] J. N. Tsitsiklis and B. Van Roy, *Analysis of temporal-difference learning with function approximation*, in Advances in Neural Information Processing Systems, 1997, pp. 1075–1081.
[49] Y. Z. Tsypkin and B. T. Polyak, *Attainable accuracy of adaptation algorithms*, Doklady Akademii Nauk, 218 (1974), pp. 532–535.
[50] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.
[51] H.-T. Wai, M. Hong, Z. Yang, Z. Wang, and K. Tang, *Variance reduced policy evaluation with smooth function approximation*, in Advances in Neural Information Processing Systems, 2019, pp. 5776–5787.
[52] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press, 2019.
[53] M. J. Wainwright, *Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning*, arXiv preprint arXiv:1905.06265, (2019).
[54] M. J. Wainwright, *Variance-reduced Q-learning is minimax optimal*, arXiv preprint arXiv:1906.04697, (2019).
[55] T. Xu, Z. Wang, Y. Zhou, and Y. Liang, *Reanalysis of variance reduced temporal difference learning*,

arXiv preprint arXiv:2001.01898, (2020).
[56] A. Zanette and E. Brunskill, *Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds*, arXiv preprint arXiv:1901.00210, (2019).