Accelerated and instance-optimal policy evaluation with linear function approximation*

Tianjiao Li†, Guanghui Lan†, and Ashwin Pananjady‡

Abstract. We study the problem of policy evaluation with linear function approximation and present efficient and practical algorithms that come with strong optimality guarantees. We begin by proving lower bounds that establish baselines on both the deterministic error and stochastic error in this problem. In particular, we prove an oracle complexity lower bound on the deterministic error in an instance-dependent norm associated with the stationary distribution of the transition kernel, and use the local asymptotic minimax machinery to prove an instance-dependent lower bound on the stochastic error in the i.i.d. observation model. Existing algorithms fail to match at least one of these lower bounds: To illustrate, we analyze a variance-reduced variant of temporal difference learning, showing in particular that it fails to achieve the oracle complexity lower bound. To remedy this issue, we develop an accelerated, variance-reduced fast temporal difference algorithm (VRFTD) that simultaneously matches both lower bounds and attains a strong notion of instance-optimality. Finally, we extend the VRFTD algorithm to the setting with Markovian observations, and provide instance-dependent convergence results. Our theoretical guarantees of optimality are corroborated by numerical experiments.

Key words. policy evaluation, temporal difference, variance reduction, acceleration, Markovian noise

AMS subject classifications. 62M20, 68Q25, 90C15, 90C60, 93E10

1. Introduction.

Reinforcement learning (RL) problems are generally formulated in terms of Markov decision processes (MDPs). At each time step, the agent observes the current state and subsequently takes an action, which leads to the realization of some reward as well as a transition to the next state according to the underlying, but unknown, stochastic transition function. The eventual goal of the agent is to learn a policy, i.e., a mapping from states to actions, to optimize the reward accrued over time. The setting is a very general one, with applications ranging from engineering to the natural and social sciences; see, e.g., [14, 17] for surveys of RL applications.

A fundamental building block in RL is the problem of *policy evaluation*, in which we are interested in estimating the long-term (discounted) *value* of each state under a fixed policy with sample access to the transition and reward functions. The literature considers three observation models for transition and reward samples, namely the generative model, the so-called "i.i.d." model, and the Markovian noise model[1]. Furthermore, in modern applications with large state spaces, it is common to seek an approximation to the true value function within the span of a small number of basis functions, a setting that is commonly known as *linear function approximation*. In the canonical setting of the problem, one is interested in using

*Submitted to the editors May 24, 2023.

†Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 (tli432@gatech.edu, george.lan@isye.gatech.edu).

‡Industrial & Systems Engineering and Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 (ashwinpm@gatech.edu).

[1]Two of these observation models are formally discussed in Section 2.

random observations to compute an approximate value function within the subspace, with the distance between the true value function and its approximation being measured according to an instance-dependent "weighted ℓ_2-norm" that depends on the stationary distribution of the transition kernel.

It is common to use stochastic approximation (SA) algorithms to solve the policy evaluation problem in the setting described above. Given the iterative nature of these algorithms, their convergence rates can be decomposed into two types of error: a *deterministic error* that measures how fast the algorithm converges to its fixed point even in the absence of noise, and a *stochastic error* that measures the contribution of the noise. Either of these errors could dominate in practice. While the stochastic error is typically larger in noisy problems, the deterministic error of the algorithm can dominate, for example, in settings with multiple processors. In particular, the collection and use of multiple samples/trajectories in parallel can reduce the stochastic error considerably.

Loosely speaking, the deterministic error is measured in terms of the *oracle complexity* of the algorithm and the stochastic error is measured in terms of the *sample complexity*. The eventual goal of algorithm design is to develop practical algorithms that have optimal oracle and sample complexities. Ideally, these optimality guarantees should be *instance-specific*, in that they depend explicitly on the problem at hand and enable us to draw distinctions between the performance profiles of different algorithms.

With this context in hand, let us briefly discuss the state of the art relevant to characterizing the complexity of policy evaluation and related problems. Classical work by Nemirovsky [29, 30] established oracle complexity lower bounds for solving linear operator equations in ℓ_2-norm. However, these results do not extend to the more specific policy evaluation setting under the weighted ℓ_2-norm. On the other hand, Khamaru et al. [15] recently provided an instance-specific analysis of the sample complexity of policy evaluation under the ℓ_∞-norm, focusing on the generative observation model without function approximation. Mou et al. [27] studied the broader problem of solving projected fixed point equations, with a focus on characterizing the error incurred due to projection onto a subspace. By virtue of studying the more general problem, the lower bounds on the statistical error proved in this paper are not specific enough to capture the policy evaluation setting with function approximation. Concurrent work by Mou et al. [28] studied SA methods for solving linear fixed point equations with Markovian samples and established a non-asymptotic, instance-dependent lower bound. Given this state of affairs, the central question that motivates this paper is the following:

What are the optimal oracle and sample complexities of policy evaluation in weighted ℓ_2-norm with linear function approximation, and do existing algorithms achieve these bounds?

1.1. Related work. There is a large literature on stochastic approximation for policy evaluation. The most popular stochastic iterative algorithm used for policy evaluation is temporal difference (TD) learning; see Dann et al. [6] for a survey. The TD learning algorithm was first introduced by Sutton [40], and convergence guarantees for TD have been proven in both asymptotic and non-asymptotic settings. Asymptotic convergence of TD with linear function approximation was established in Tsitsiklis and Van Roy [43], and other classical asymptotic guarantees include those due to Borkar and co-authors [4, 3]. The vanilla TD algorithm can also be combined with the iterate averaging technique, and the asymptotic

convergence of this algorithm was shown by Tadic [42], who extended the convergence results for solving noisy linear systems [35].

While asymptotic convergence results offer a proof-of-concept, the algorithm is often run in large-scale applications with relatively small sample sizes. The first results proving finite time convergence under i.i.d. setting were proposed by Sutton et al. [41] and later extended by Lakshminarayanan and Szepesvári [21]. Finite-time analysis of TD learning under Markovian noise was carried out by Bhandari et al. [1], where the authors employed nonsmooth analysis to a variant of TD learning, by requiring projections at each iteration onto a pre-specified ball. A consequence of the nonsmooth approach is that there is no obvious way of benefiting from the variance reduction effect of parallel computing. In recent work, a subset of the current authors [20] provided an improved analysis of vanilla TD algorithm that overcomes this hurdle. There are also several other notable finite sample analyses of policy evaluation in various settings, e.g., [39, 5, 10, 24, 26], and statistical lower bounds have also been shown for offline reinforcement learning with linear function approximation, e.g., [47, 51].

While some of these analyses are sharp, to our knowledge, vanilla TD learning is not known to attain the optimal oracle complexity and instance-dependent sample complexity. Li et al. [24] recently proved that TD learning achieves the minimax lower bound on stochastic error up to logarithmic factors, but their analysis is not instance dependent. Recent work by Kotsalis et al. [20] presented two new algorithms, the conditional temporal difference (CTD) and the fast temporal difference (FTD) learning, where FTD exhibits an accelerated rate in deterministic error. However, these algorithms fail to capture the correct stochastic error in the policy evaluation problem, and the bounds can be shown to be suboptimal for policy evaluation both in an instance-dependent sense, and in the worst-case over natural problem classes.

During the past decade, there has been a flurry of parallel work in stochastic optimization on developing first-order methods with variance-reduction; early examples include IAG [2], SAG [37], SVRG [13, 49], and SAGA [7]. There are several papers that apply variance reduction to reinforcement learning, e.g., [8, 34, 45]. Recent work in policy evaluation has shown that variance reduction techniques can also be applied to algorithms of the TD-type [18, 44, 50, 15]. Among these, the paper [15] is motivated by the desire to draw distinctions between RL algorithms with similar worst-case performance and follows a line of work deriving instance-dependent bounds on the stochastic error in policy evaluation [33, 25]. Specifically, the results of [15] capture the optimal instance-dependent stochastic error in the ℓ_∞-norm. However, the optimal sample complexity is not achieved in [15] since the algorithm requires $\mathcal{O}\{1/(1-\gamma)^3\}$ samples in each epoch.

1.2. Contributions and organization. Towards answering the question posed at the end of Section 1, we make three distinct contributions:

- **Lower bounds.** We construct a worst-case instance that shows an oracle complexity lower bound of order $\Omega\{(1-\gamma)^{-1} \cdot \log(1/\epsilon)\}$ for any iterative method whose iterates lie within the linear span of the initial point v_0 and subsequent temporal differences, to converge to ϵ-error in weighted ℓ_2-norm. We also prove a lower bound on sample complexity using the classical local minimax theorem [11, 22, 23] to provide an instance-specific baseline for algorithm design.

- **Algorithm design in the i.i.d. setting.** We start by applying the variance reduction technique to the classical TD algorithm, showing that the resulting variance-reduced temporal difference (VRTD) algorithm nearly matches the optimal stochastic error, but the analysis suggests a suboptimal deterministic error. This motivates us to further improve the VRTD algorithm with the stochastic operator extrapolation (SOE) device [19]. We provide a sharp analysis of our new algorithm—termed variance-reduced fast temporal difference (VRFTD)—showing that it achieves a convergence rate nearly matching both the deterministic error lower bound (for well-conditioned feature matrices) and the stochastic error lower bound.
- **Extension to the Markovian setting.** We extend the VRFTD algorithm to the Markovian setting by introducing a burn-in period during sample collection. We show that the resulting algorithm also achieves similarly fast convergence, with a dominating stochastic error term that matches the instance-dependent lower bound proved in [28] and a deterministic error that matches that of the i.i.d. setting up to a multiplicative factor of the mixing time. In particular, in the so-called realizable case when the approximation error caused by linear function approximation is 0 (e.g., in the tabular setting), the leading Markovian stochastic error term is equal to the i.i.d. stochastic error term, indicating that the additional dependence on mixing time only appears in terms whose dependence on the final tolerance ϵ is weak.

The rest of this paper is organized as follows. In Section 2, we formally present the problem setting. The three aforementioned main contributions are presented in Sections 3—5. In Section 6, we provide numerical experiments that corroborate our optimality guarantees. The proofs of our main results are postponed to Section 7, and auxiliary results are collected in the supplementary materials.

1.3. Notation.

For a positive integer n, we define $[n] := \{1, 2, ..., n\}$. We let $\mathbf{1}$ denote the all-ones vector in $\mathbb{R}^D$. We let e_j denote the j-th standard basis vector in $\mathbb{R}^D$. Let $\mathbb{I}_S : X \to \{0, 1\}$ denote the indicator function of the subset $S \subseteq X$. Given a vector $x \in \mathbb{R}^m$, denote its i-th entry by[2] $x_{(i)}$. Let $\|x\|_1 := \sum_{i=1}^m |x_{(i)}|$, $\|x\|_2 := \sqrt{\sum_{t=1}^m x_{(i)}^2}$ and $\|x\|_\infty := \max_{i\in[m]} |x_{(i)}|$ denote the ℓ_1, ℓ_2 and ℓ_∞-norms respectively. Given a matrix A, denote its (i, j)-th entry by $P_{i,j}$. Let $\|A\|_2$ denote the spectral norm of matrix A. We let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of a square matrix A, respectively. For a symmetric positive definite matrix A, define the inner product $\langle x, y\rangle_A := x^\top Ay$ and the associated norm $\|x\|_A := \sqrt{x^\top Ax}$. We refer to $\|x\|_A$ as the ℓ_A-norm of x.

2. Background and problem setting.

In this section, we formally introduce Markov reward processes (MRPs) and the (discounted) policy evaluation problem. We also define linear function approximation of the value function, and present the concrete observation models that we study.

2.1. Markov reward process and policy evaluation.

An MRP is described by a tuple $(\mathcal{S}, \mathsf{P}, R, \gamma)$, where $\mathcal{S} = [D]$ denotes the state space, P is the transition kernel, R is the reward function and $\gamma \in (0, 1)$ is the discount factor. At each iteration, the system moves from the

[2] In situations in which there is no ambiguity, we also use x_i to denote the i-th coordinate of a vector x.

current state $s \in \mathcal{S}$ to some state $s' \in \mathcal{S}$ with probability $\mathsf{P}(s'|s)$, while the agent realizes the reward $R(s, s')$. We denote by $r(s) := \sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s) R(s, s')$ the expected instantaneous reward generated at state s. Let P denote the transition probability matrix having (i, j)-th entry $P_{i,j} = \mathsf{P}(j|i)$. The reward R can also be written in matrix form, i.e., $R_{i,j} = R(i, j)$. The value function specifies the infinite-horizon discounted reward as a function of the initial state:

$$v^*(s) := \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t R(s_t, s_{t+1}) | s_0 = s\big].$$

In the case where the number of states is finite and equal to D, both the expected reward function r and the value function v^* are D-dimensional vectors of reals. The value function is given by the solution to the Bellman equation

$$v^* = \gamma P v^* + r. \tag{2.1}$$

Throughout this paper, we assume that the Markov chain is aperiodic, ergodic and that there exists a unique stationary distribution $\pi := (\pi_1, ..., \pi_D)$ with strictly positive entries, satisfying $\pi P = \pi$. Let $\Pi := \mathrm{diag}(\pi_1, ..., \pi_D)$ denote a $D \times D$ diagonal matrix whose non-zero elements are given by the entries of the stationary distribution.

2.2. Linear function approximation. In modern applications with large state spaces, it is common to seek approximate solutions to the Bellman equation (2.1), and the standard approach is to choose a d-dimensional subspace $\mathbb{S}$ for the purposes of approximation. In particular, one chooses $\mathbb{S} := \mathrm{span}\{\psi_1, ..., \psi_d\}$ for d linearly independent basis vectors $\psi_1, \ldots, \psi_d$. For each state $s \in [D]$ we let $\psi(s) := [\psi_1(s), \psi_2(s), ..., \psi_d(s)]^\top$ denote its feature vector. Letting $\Pi_{\mathbb{S}}$ denote the projection onto the subspace with respect to the $\|\cdot\|_\Pi$-norm, define $\bar{v}$ as the solution to the *projected fixed point equation*

$$\bar{v} = \Pi_{\mathbb{S}}(\gamma P \bar{v} + r). \tag{2.2}$$

It is convenient to write this projection in matrix notation. Let $\Psi := [\psi_1, \psi_2, ..., \psi_d]^\top$, and for $v^\diamondsuit$ in $\mathbb{S}$, use $\theta^\diamondsuit$ to denote its corresponding parameterization in $\mathbb{R}^d$, e.g., $\Psi^\top \theta' = v'$. With this shorthand, equation (2.2) can be equivalently written as

$$\Psi \Pi \Psi^\top \bar{\theta} = \Psi \Pi \gamma P \Psi^\top \bar{\theta} + \Psi \Pi r. \tag{2.3}$$

It is convenient in the analysis to have access to an orthonormal basis spanning the projected space $\mathbb{S}$. Define the matrix $B \in \mathbb{R}^{d \times d}$ by letting $B_{i,j} := \langle \psi_i, \psi_j \rangle_\Pi$ for each i, j, and let

$$\Phi := [\phi_1, \phi_2, ..., \phi_d]^\top = B^{-\frac{1}{2}} \Psi.$$

By construction, the vectors $\phi_1, \ldots, \phi_d$ satisfy $\langle \phi_i, \phi_j \rangle_\Pi = \mathbb{I}(i = j)$. Next, define the scalars

$$\beta := \lambda_{\max}(B), \quad \text{and } \mu := \lambda_{\min}(B), \tag{2.4}$$

so that β/μ is the condition number of the covariance matrix of the features. Finally, let

$$M := \gamma \Phi \Pi P \Phi^\top$$

denote the d-dimensional matrix that describes the action of γP on the projected space $\mathbb{S}$.

2.3. Observation models and problem statement. We start by introducing the i.i.d. observation model, in which we have access to a black box or simulator that generates samples from the transition kernel and reward functions. In particular, we observe independent tuples $\xi_i = (s_i, s_i', R(s_i, s_i'))$, such that

$$s_i \sim \omega, \quad s_i' \sim P(\cdot|s_i),$$

where $\omega := (\omega_1, ..., \omega_D)$ is a distribution with strictly positive entries, and we use the shorthand $\Omega := \mathrm{diag}([\omega_1, ..., \omega_D])$. A natural and popular choice is $\omega = \pi$, in which case the i.i.d. model is meant to approximate the stationary Markov chain.

In the Markovian noise model, we assume that all of our observations come from a single trajectory of a Markov chain. Precisely, the sequence of states $\{s_0, s_1 ...., \}$ generated by the MRP is a time-homogeneous Markov chain, with $s_0 \sim \pi$. The tuple $\xi_t = (s_t, s_{t+1}, R(s_t, s_{t+1}))$ is observed at each time t. The highly correlated nature of these observations renders algorithm design and analysis in the Markovian setting more challenging than in the i.i.d. setting.

Our goal in both cases is to use the observations to generate an estimator $\widehat{v}_n$ of v^* which satisfies an *oracle inequality* of the form

$$\mathbb{E}\|\widehat{v}_n - v^*\|_\Pi^2 \leq \mathcal{O}(1)\|\bar{v} - v^*\|_\Pi^2 + \delta_n \|v_0 - \bar{v}\|_\Pi^2 + \epsilon_{n,\sigma}, \tag{2.5}$$

where v_0 is the initial iterate of the algorithm. The three terms appearing on the RHS of inequality (2.5) all have concrete interpretations. The first term $\|\bar{v} - v^*\|_\Pi^2$ characterizes the *approximation error* incurred by the linear function approximation. As a point of the background, we recall the following instance-dependent upper bound on the approximation error due to Mou et al. [27]:

$$\|\bar{v} - v^*\|_\Pi^2 \leq \mathcal{A}(M, \gamma) \inf_{v \in S} \|v - v^*\|_\Pi^2, \tag{2.6}$$

where $\mathcal{A}(M, \gamma) = 1 + \lambda_{\max}\left((I - M)^{-1}(\gamma^2 I_d - MM^\top)(I - M)^{-\top}\right)$. See Mou et al. [27] for a proof, alongside guarantees of information-theoretic optimality.

This work focuses on sharply analyzing the last two terms on the RHS of inequality (2.5), both of which have concrete operational interpretations. The term $\delta_n \|v_0 - \bar{v}\|_\Pi^2$ is the *deterministic error*, which characterizes the convergence of the iterative algorithm in the purely deterministic setting. Specifically, the term δ_n, which should tend to zero as the number of iterations (or oracle calls) n goes to infinity, quantifies how fast the discrepancy between the initialization v_0 and the approximate solution $\bar{v}$ diminishes by running the iterative algorithm. The third term $\epsilon_{n,\sigma}$ is the *stochastic error*, which is incurred due to the stochastic observation model. Here we use the notation σ as a placeholder for the "noise level" in the observed samples. One should expect the stochastic error $\epsilon_{n,\sigma}$ to go to zero as n goes to infinity or as σ goes to zero. Several previous works mix the deterministic error with stochastic error in their guarantees (see, e.g., [1, 27]). However, the key benefit of separating the deterministic error from the stochastic error is that it allows a clean understanding of situations in which either the observations have low noise or parallel implementation may be available. In these cases, the deterministic error dominates the overall convergence rate of the algorithm, and so having algorithms that attain the optimal deterministic error is a key desideratum.

Having precisely defined the deterministic and stochastic errors, we are now in a position to present our first set of results on lower bounds for both of these terms.

3. Lower bounds in weighted ℓ_2-norm.

We study the oracle complexity lower bound on deterministic error in Section 3.1 and the instance-specific stochastic error lower bound in Section 3.2.

3.1. Oracle complexity lower bound on deterministic error.

It is well-known that a linear rate can be achieved for the deterministic policy evaluation problem, and the convergence rate is highly dependent on the effective horizon $(1-\gamma)^{-1}$ [36]. Accordingly, our goal in this section is to prove an oracle complexity lower bound in terms of $(1-\gamma)^{-1}$, which can be done even in the tabular setting in which the subspace is all of $\mathbb{R}^D$. The following assumption on the oracle captures algorithms in the temporal difference learning family.

Assumption 1 (Amenable iterative method). *An* amenable *iterative method $\mathcal{M}$ generates a sequence of iterates v_k such that*

$$v_k \in v_0 + \operatorname{span}\{G(v_0), G(v_1), ..., G(v_{k-1})\}, \quad k \geq 1, \tag{3.1}$$

where $G(v) = (I - \gamma P)v - r$.

Noting that $G(v)$ is precisely the temporal difference operator applied at the point v, an amenable algorithm is one whose iterates are always in the linear span of the initial point v_0 and subsequent temporal differences. The linear span assumption is commonly used in proving oracle complexity lower bounds [31, 32], and as such, nearly all the algorithms in the temporal difference family can be shown to be amenable. The sole exceptions that we are aware of occur in cases where there are projections involved in the algorithm, e.g., [1]. However, in policy evaluation problems with unbounded feasible region $\mathbb{R}^D$, projection steps are often unnatural and vanilla TD algorithms are able to attain similar performance (see, e.g., [20]). The following theorem provides an oracle complexity lower bound for policy evaluation problem under the ℓ_Π-norm for amenable algorithms.

Theorem 3.1. *Fix a constant $\gamma > \frac{1}{2}$ and an initialization v_0. There exists a transition kernel P and an expected reward vector r such that any iterative method $\mathcal{M}$ satisfying Assumption 1 produces iterates $\{v_k\}_{k\geq 1}$ satisfying the following. If (D, k) satisfies $\frac{1-(2\gamma-1)^{2D-2k}}{1-(2\gamma-1)^{2D}} \geq \frac{1}{2}$, then*

$$\|v_k - v^*\|_\Pi^2 \geq \tfrac{1}{2}(2\gamma - 1)^{2k}\|v_0 - v^*\|_\Pi^2, \tag{3.2}$$

where v^* is the solution of equation (2.1).

See Section 7.1 for the proof of this theorem.

Noting that $2\gamma - 1 = 1 - 2(1-\gamma)$, Theorem 3.1 shows an oracle complexity lower bound $\mathcal{O}\{\frac{1}{1-\gamma}\log(\frac{\|v_0-v^*\|_\Pi^2}{\epsilon})\}$ for finding a solution $\widehat{v} \in \mathbb{R}^m$ such that $\|\widehat{v} - v^*\|_\Pi^2 \leq \epsilon$. It should be noted that the metric (i.e., the ℓ_Π-norm) used in Theorem 3.1 depends on the problem instance through the stationary distribution of the transition kernel P. Such an instance-dependent metric makes the construction of our worst-case instance non-standard and challenging.

On a related note, it is instructive to recall that classical oracle complexity bounds for solving linear operator equations [29, 30] allow the conjugate operator to be queried within the oracle, making the method class wider than the class of amenable algorithms captured in Assumption 1. On the one hand, the conjugate operator is not natural for solving a policy

evaluation problem under stochastic settings since the vector $(I-\gamma P)^\top v$ is hard to estimate with transition and reward samples. On the other hand, our construction used in proving Theorem 3.1 naturally extends to this wider method class, and we provide an even stronger deterministic error lower bound than Theorem 3.1 in Appendix SM1.

3.2. Instance-specific lower bound on stochastic error. We now turn our attention to proving lower bounds on the instance-specific sample complexity under the i.i.d. observation model introduced in Section 2.3. We assume that the feature matrix Ψ is fixed and known, and let $\vartheta = (\omega, P, R)$ denote an individual problem instance parameterized by the initial state distribution ω, transition kernel P, and reward function R. Note at this juncture that we do not require that $\omega = \pi$; this is akin to the so-called *off-policy* situation in which the sampling (or behavior) policy may differ from the policy that we are interested in evaluating. Our result will apply in this general case; but given that the initial state is drawn from the distribution ω, it is convenient to consider solving the projected fixed point equation with respect to the $\|\cdot\|_\Omega$-norm (cf. Eq. (2.3)), written as

$$\Psi\Omega\Psi^\top\theta = \Psi\Omega\gamma P\Psi^\top\theta + \Psi\Omega r.$$

Use the function $\bar{\theta}(\vartheta) := (\Psi\Omega\Psi^\top - \Psi\Omega\gamma P\Psi^\top)^{-1}\Psi\Omega r$ to denote the target of interest.

In order to state our result, we require some additional notation. Fix an instance $\vartheta = (\omega, P, R)$, and for any $\epsilon > 0$, define an ϵ-neighborhood of problem instances by

$$\mathfrak{N}(\vartheta;\epsilon) := \{\vartheta' = (\omega', P', R') : \|\omega-\omega'\|_2 + \|P-P'\|_F + \|R-R'\|_F \le \epsilon\}.$$

Define the matrix $\widetilde{B} \in \mathbb{R}^{d\times d}$ by $\widetilde{B}_{i,j} := \langle \psi_i, \psi_j\rangle_\Omega$ for $i,j \in [d]$. Thus $\widetilde{B}$ satisfies

$$\widetilde{B}^{-\frac{1}{2}}\Psi\Omega\Psi\widetilde{B}^{-\frac{1}{2}} = I_d. \tag{3.3}$$

Adopting the $\ell_{\widetilde{B}}$-norm as our loss function, define the following local asymptotic minimax risk [11, 22]:

$$\mathfrak{M}(\vartheta) := \lim_{c\to\infty}\lim_{N\to\infty}\inf_{\widehat{\theta}_N}\sup_{\vartheta'\in\mathfrak{N}(\vartheta;c/\sqrt{N})} N\cdot\mathbb{E}_{\vartheta'}\left[\left\|\widehat{\theta}_N - \bar{\theta}(\vartheta')\right\|_{\widetilde{B}}^2\right]. \tag{3.4}$$

The infimum in Eq. (3.4) is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N observations drawn according to the i.i.d. observation model. In contrast to the global minimax risk—which takes a supremum of the risk over all the problem instances within a reasonable class—the local minimax risk $\mathfrak{M}(\vartheta)$ looks for the hardest alternative in a small neighborhood of the instance ϑ with diameter $c/\sqrt{N}$. To capture the hardest local alternative (in an asymptotic sense) it suffices to take the diameter of the neighborhood to be of the order $1/\sqrt{N}$. Invoking Eq. (3.3) yields the equivalent definition

$$\mathfrak{M}(\vartheta) = \lim_{c\to\infty}\lim_{N\to\infty}\inf_{\widehat{\theta}_N}\sup_{\vartheta'\in\mathfrak{N}(\vartheta;c/\sqrt{N})} N\cdot\mathbb{E}_{\vartheta'}\left[\left\|\Psi^\top\widehat{\theta}_N - \Psi^\top\bar{\theta}(\vartheta')\right\|_{\Omega}^2\right]. \tag{3.5}$$

The following proposition characterizes the local asymptotic risk $\mathfrak{M}(\vartheta)$ explicitly.

Proposition 3.2. *Consider the i.i.d. observation model with the initial state drawn from the distribution ω. Let $Z \in \mathbb{R}^d$ be a multivariate Gaussian*

$$Z \sim \mathcal{N}\big(0, (I_d - \widetilde{M})^{-1}\widetilde{\Sigma}(I_d - \widetilde{M})^{-T}\big),$$

where $\widetilde{\Sigma} := \operatorname{cov}\big[\widetilde{B}^{-\frac{1}{2}}\big(\langle\psi(s) - \gamma\psi(s'), \bar{\theta}\rangle - R(s,s')\big)\psi(s)\big]$ *and* $\widetilde{M} := \gamma\widetilde{B}^{-\frac{1}{2}}\Psi\Omega P\Psi\widetilde{B}^{-\frac{1}{2}}$. *Then we have*

$$\mathfrak{M}(\vartheta) = \mathbb{E}[\|Z\|_2^2] = \operatorname{trace}\Big\{(I_d - \widetilde{M})^{-1}\widetilde{\Sigma}(I_d - \widetilde{M})^{-T}\big)\Big\}. \tag{3.6}$$

See Section 7.2 for the proof of this theorem.

A few comments are in order. First, it should be noted that this lower bound is distinct from the asymptotic minimax lower bound shown in Khamaru et al. [15], in which a generative observation model is assumed (where we observe transitions from all D initial states) and there is no function approximation. Consequently, our choice of a problem instance of interest is $\vartheta = (\omega, P, R)$ rather than (P, R) in [15]. Second, and on a related note, it is important that ω be unknown and included in the set of parameters ϑ; if in contrast ω is known a priori, then the local asymptotic minimax risk differs from the characterization (3.6). Finally, we note that Mou et al. [27] provide non-asymptotic, instance-dependent lower bounds on stochastic error for solving projected fixed-point equations using the Bayesian Cramér–Rao bound. However, these lower bounds do not directly apply here, since the family of hardest local alternatives constructed in [27] may not be valid instances in the policy evaluation setting.

Let us now specialize Proposition 3.2 by taking $\omega = \pi$, where π is the stationary distribution of the transition kernel P. Denote by $\vartheta_\pi := (\pi, P, R)$ the instance of interest. Let

$$\bar{\Sigma}_{\mathsf{iid}} := \operatorname{cov}\big[B^{-\frac{1}{2}}\big(\langle\psi(s) - \gamma\psi(s'), \bar{\theta}\rangle - R(s,s')\big)\psi(s)\big] \quad \text{for } s \sim \pi \text{ and } s' \sim \mathsf{P}(\cdot|s).$$

Applying Proposition 3.2, the local asymptotic minimax risk (3.5) under this setting is then given by

$$\begin{aligned}\lim_{c\to\infty}\lim_{N\to\infty}\inf_{\widehat{\theta}_N}\sup_{\vartheta' \in \mathfrak{N}(\vartheta_\pi; c/\sqrt{N})} & N \cdot \mathbb{E}_{\vartheta'}\left[\left\|\Psi^\top\widehat{\theta}_N - \Psi^\top\bar{\theta}(\vartheta')\right\|_\Pi^2\right] \\ &= \operatorname{trace}\big\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-T}\big)\big\}. \end{aligned} \tag{3.7}$$

Taking stock, we have proved two lower bounds (3.2) and (3.7) on the deterministic and stochastic errors in ℓ_Π-norm under the i.i.d. observation model $s \sim \pi$ and $s' \sim \mathsf{P}(\cdot|s)$. Given these baselines, it is natural to ask whether there is a practical iterative algorithm in the TD family that can achieve both lower bounds, which is the main focus of Section 4.

4. Algorithms for policy evaluation in the i.i.d. setting. Taking both lower bounds proved in Section 3 as our baseline, we now turn our attention to the question of algorithm design. In this section, we assume the i.i.d. observation model introduced in Section 2.3 with $\omega = \pi$. In order to state the results clearly, we require some additional notation. For $\theta \in \mathbb{R}^d$, we define the deterministic operator for solving equation (2.3) as

$$g(\theta) = \Psi\Pi(\Psi^\top\theta - r - \gamma P\Psi^\top\theta); \tag{4.1}$$

note that $\bar{\theta}$ is the solution to $g(\theta) = 0$ The corresponding stochastic operator calculated from sample ξ_i is defined as

$$\widetilde{g}(\theta, \xi_i) = \big(\langle \psi(s_i), \theta\rangle - R(s_i, s_i') - \gamma\langle \psi(s_i'), \theta\rangle\big)\, \psi(s_i); \tag{4.2}$$

note that $\mathbb{E}_{s_i\sim\pi, s_i'\sim \mathsf{P}(\cdot|s_i)}[\widetilde{g}(\theta, \xi_i)] = g(\theta)$. To characterize the "variance" of the stochastic operator under the i.i.d. observation model, we make the following assumption:

Assumption 2. *There exists a constant $\varsigma \geq 0$ such that for every $\theta, \theta' \in \mathbb{R}^d$,*

$$\mathbb{E}\|\widetilde{g}(\theta, \xi) - \widetilde{g}(\theta', \xi) - \big(g(\theta) - g(\theta')\big)\|_2^2 \leq \varsigma^2 \|v - v'\|_\Pi^2, \tag{4.3}$$

where $v = \Psi^\top \theta$ and $v' = \Psi^\top \theta'$.

In words, instead of bounding the "variance" of the stochastic operator directly as in [20], Assumption 2 guarantees that the variance of the difference between stochastic operators with different variables θ under the same data ξ is upper bounded by the distance between the variables. This assumption is critical for implementing the variance-reduction techniques and capturing the instance-dependent stochastic error at the approximate solution $\bar{\theta}$. Clearly, the parameter ς^2 is bounded provided the features $\psi(s)$ are bounded, and provides a natural measure of "noise" in the problem. Accordingly, we make use of Assumption 2 throughout Sections 4 and 5.

We are now ready to present our algorithms. We start with a variance-reduced version of the TD algorithm that captures the instance-specific stochastic error lower bound but fails to achieve the oracle complexity lower bound on deterministic error. To remedy this issue, we develop an accelerated variance-reduced TD algorithm that matches both lower bounds proved in Section 3.

4.1. A warm-up algorithm: variance-reduced temporal difference learning. Variance-reduced temporal difference learning (VRTD) solves the policy evaluation problem using epochs. With a slight ambiguity of notation, we let v_t and its corresponding parameterization θ_t denote the iterates generated within each epoch, and let v^0 and its corresponding parameterization θ^0 denote the initialization of the algorithm. At the beginning of each epoch k, the algorithm uses N_k samples to compute an averaged stochastic operator $\widehat{g}$ and evaluates it at a point $\widetilde{\theta}$, where $\widetilde{\theta}$ should be understood as the best current approximation of the optimal solution. The vector $\widehat{g}(\widetilde{\theta})$ is used to recenter the updates in each epoch.

Algorithm 4.1 Variance-reduced Temporal Difference Algorithm under i.i.d observations

Input: $\theta^0 = \widehat{\theta}_0 \in \mathbb{R}^d$, $\eta > 0$, $\{\zeta_t\}_{t=1}^T \geq 0$ and $\{N_k\}_{k=1}^K \subset \mathbb{Z}_+$.
for $k = 1, \ldots, K$ **do**
Set $\theta_1 = \widetilde{\theta} = \widehat{\theta}_{k-1}$. Collect N_k samples $\xi_i^k = (s_i, s_i', R(s_i, s_i'))$ from the i.i.d. model.
Calculate $\widehat{g}(\widetilde{\theta}) = \frac{1}{N_k}\sum_{i=1}^{N_k} \widetilde{g}(\widetilde{\theta}, \xi_i^k)$.
for $t = 1, \ldots, T$ **do**
Collect a sample $\xi_t = (s_t, s_t', R(s_t, s_t'))$ from the i.i.d. observation model and compute

$$\theta_{t+1} = \theta_t - \eta\left(\widetilde{g}(\theta_t, \xi_t) - \widetilde{g}(\widetilde{\theta}, \xi_t) + \widehat{g}(\widetilde{\theta})\right). \tag{4.4}$$

end for
Output of the epoch:

$$\widehat{\theta}_k = \frac{\sum_{t=1}^{T+1} \zeta_t \theta_t}{\sum_{t=1}^{T+1} \zeta_t}. \tag{4.5}$$

end for

Note that this algorithm is distinct from previous instantiations of variance-reduced temporal difference algorithms [50, 15], since the output of each epoch (4.5) is a weighted average of the iterates. The following theorem provides a convergence guarantee on the VRTD algorithm.

Theorem 4.1. *Consider the i.i.d. observation model with the initial state drawn from the distribution π. Fix the total number of epochs K and a positive integer N. Assume that for each epoch $k \in [K]$, the parameters η, N_k and T satisfy*

$$\eta \leq \min\left\{\tfrac{1-\gamma}{2\beta(1+\gamma)^2}, \tfrac{1-\gamma}{32\varsigma^2}\right\}, \quad T \geq \tfrac{32}{\mu(1-\gamma)\eta}, \text{ and } \quad N_k \geq \left\{\tfrac{38\varsigma^2}{\mu(1-\gamma)^2}, (\tfrac{3}{4})^{K-k}N\right\}.$$

Set the output of the epoch to be $\widehat{v}_k := \frac{\sum_{t=1}^T \eta(1-\gamma)v_t + (1/\beta)v_{T+1}}{T\eta(1-\gamma)+(1/\beta)}$. Then for each $\delta > 0$, we have

$$\begin{aligned}\mathbb{E}[\|\widehat{v}_K - v^*\|_\Pi^2] &\leq (1+\delta)\mathcal{A}(M,\gamma)\inf_{v\in S}\|v - v^*\|_\Pi^2 \\ &\quad + (1+\tfrac{1}{\delta})\left[\tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \tfrac{15}{N}\mathrm{trace}\left((I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\right)\right].\end{aligned} \tag{4.6}$$

See Section 7.3 for the detailed proof of this theorem.

The first term in the bound (4.6) is the approximation error term alluded to previously; let us extract the deterministic and stochastic errors from the remaining terms. The number of epochs required by the VRTD method to find a solution $\widehat{v} \in \mathbb{R}^D$, such that $\mathbb{E}[\|\widehat{v} - \bar{v}\|_\Pi^2] \leq \epsilon$ is bounded by $\mathcal{O}\big\{\log(\|v^0 - \bar{v}\|_\Pi^2/\epsilon)\big\}$. The total number of samples used is $\sum_{k=1}^K (T + N_k)$, which is of the order

$$\underbrace{\tfrac{\beta}{(1-\gamma)^2\mu}\log(\tfrac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon})}_{\text{deterministic error}} + \underbrace{\tfrac{\varsigma^2}{(1-\gamma)^2\mu}\log(\tfrac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon}) + \tfrac{\mathrm{trace}\left((I_d-M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d-M)^{-\top}\right)}{\epsilon}}_{\text{stochastic error}}. \tag{4.7}$$

A few comments on the upper bound provided in Theorem 4.1 are in order.

Comparing the upper and lower bounds. We first focus on the stochastic error in the bound (4.7). The VRTD algorithm requires at least $\mathcal{O}\{\frac{\varsigma^2}{(1-\gamma)^2\mu}\}$ samples in each epoch, which accounts for the first term. Note that with noisy observations, it is necessary to have $\mathcal{O}\{\frac{1}{(1-\gamma)^2}\}$ samples in order to obtain an estimate of the value function within $\mathcal{O}(1)$ error, so this higher-order term is natural. The dominating stochastic error term is the last term, and matches the lower bound in equation (3.7). Therefore the VRTD algorithm is instance-optimal in terms of its stochastic error.

Next, we turn our attention to the deterministic error, noticing that the dependence on $\frac{1}{1-\gamma}$ is quadratic. Comparing with the oracle complexity lower bound proved in Theorem 3.1, this quadratic dependence is suboptimal. This shortcoming motivates us to develop an accelerated algorithm in the next subsection.

Comparing with related work. To our knowledge, the only work using variance reduction that captures the correct instance-specific stochastic error is that of Khamaru et al. [15], which showed that the VRTD algorithm can match the lower bound on stochastic error in ℓ_∞-norm. However, their guarantees require $\mathcal{O}\{\frac{1}{(1-\gamma)^3}\}$ samples in each epoch to compute the recentered update, and this sample size is suboptimal. In addition, the deterministic error proved in this paper is of the order $\frac{1}{(1-\gamma)^2\epsilon}$ which is also suboptimal.[3] The work of Mou et al. [27] provided an analysis for the Polyak–Ruppert averaged temporal difference learning algorithm with linear function approximation in the weighted ℓ_2-norm, showing that the dominant stochastic error term matches the stochastic lower bound proved in Proposition 3.2. However, the sample complexity is suboptimal due to the presence of higher-order terms (see [15] for simulations demonstrating this suboptimality), as is the oracle complexity.

4.2. Variance-reduced fast temporal temporal difference algorithm.

Motivated by the suboptimality of VRTD in its oracle complexity, we now present a variance-reduced "fast" temporal difference (VRFTD) algorithm, which incorporates the idea of operator extrapolation introduced in [19]. This serves to accelerate the algorithm, and our analysis of VRFTD shows a convergence rate matching both the deterministic and stochastic error lower bounds.

The VRFTD algorithm is formally presented in Algorithm 4.2, and we introduce the basic idea of the algorithm below. First, it utilizes the idea of recentering updates from VRTD with the operator $\widehat{g}(\widetilde{\theta})$ used in each epoch. Second, in terms of iterate updating in the inner loop, it involves an inner mini-batch that generates the averaged operator $\widetilde{g}_t$, which allows the algorithm to be run with a much larger stepsize. Finally, each iteration within an epoch involves an operator extrapolation step (4.8). This is crucial to achieving the optimal deterministic error (cf. the VRTD update (4.4)).

The following theorem establishes a convergence rate for the VRFTD algorithm.

Theorem 4.2. *Fix the total number of epochs K and a positive integer N. Assume that for each epoch, the parameters η, λ, m, N_k, T satisfy*

(4.10)
$$\eta \le \tfrac{1}{4\beta(1+\gamma)},\ \lambda = 1,\ T \ge \tfrac{32}{\mu(1-\gamma)\eta},\ m \ge \max\left\{1, \tfrac{256\eta\varsigma^2}{1-\gamma}\right\} \text{ and } N_k \ge \max\left\{\tfrac{56\varsigma^2}{\mu(1-\gamma)^2}, (\tfrac{3}{4})^{K-k}N\right\}.$$

[3] In more detail, a family of such deterministic error guarantees is possible to extract from the paper. The dependence on ϵ can be improved but the dependence on $(1-\gamma)^{-1}$ is at least quadratic.

Algorithm 4.2 Variance-reduced Fast Temporal Difference Algorithm under i.i.d. observations

Input: $\theta^0 = \widehat{\theta}_0 \in \mathbb{R}^d$, $\eta > 0$, $\lambda \geq 0$, $\{\zeta_t\}_{t=1}^T \geq 0$ and nonnegative integers m, $\{N_k\}_{k=1}^K$.
for $k = 1, \ldots, K$ **do**
Set $\theta_0 = \theta_1 = \widetilde{\theta} = \widehat{\theta}_{k-1}$. Collect N_k sample tuples $\xi_i^k = (s_i, s_i', R(s_i, s_i'))$ from the i.i.d. observation model.
Calculate $\widehat{g}(\widetilde{\theta}) = \frac{1}{N_k}\sum_{i=1}^{N_k} \widetilde{g}(\widetilde{\theta}, \xi_i^k)$.
for $t = 1, \ldots, T$ **do**
Collect m sample tuples $\xi_j^t = (s_j, s_j', R(s_j, s_j'))$ from the i.i.d. observation model.
Calculate $\widetilde{g}_t(\cdot) = \frac{1}{m}\sum_{j=1}^m \widetilde{g}(\cdot, \xi_j^t)$.
Denote $\widetilde{F}_t(\theta_t) = \widetilde{g}_t(\theta_t) - \widetilde{g}_t(\widetilde{\theta}) + \widehat{g}(\widetilde{\theta})$. Set $\widetilde{F}_0(\theta_0) = \widetilde{F}_1(\theta_1)$. Let

$$\theta_{t+1} = \theta_t - \eta\left[\widetilde{F}_t(\theta_t) + \lambda\left(\widetilde{F}_t(\theta_t) - \widetilde{F}_{t-1}(\theta_{t-1})\right)\right]. \tag{4.8}$$

end for
Output of the epoch:

$$\widehat{\theta}_k = \frac{\sum_{t=1}^{T+1}\zeta_t\theta_t}{\sum_{t=1}^{T+1}\zeta_t}. \tag{4.9}$$

end for

Set the output of this epoch to be $\widehat{v}_k := \frac{\sum_{t=2}^{T+1} v_t}{T}$. *Then for* $\delta > 0$,

$$\mathbb{E}[\|\widehat{v}_K - v^*\|_\Pi^2] \leq (1+\delta)\mathcal{A}(M,\gamma)\inf_{v\in S}\|v - v^*\|_\Pi^2 + (1+\tfrac{1}{\delta})\left[\tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \tfrac{15}{N}\mathrm{trace}\left((I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\right)\right].$$

See Section 7.4 for the detailed proof of this theorem.

In view of Theorem 4.2, the number of epochs required by the VRFTD method to find a solution $\widehat{v} \in \mathbb{R}^D$, such that $\mathbb{E}[\|\widehat{v} - \bar{v}\|_\Pi^2] \leq \epsilon$ is bounded by $\mathcal{O}\{\log(\|v^0 - \bar{v}\|_\Pi^2/\epsilon)\}$. The total number of samples used is $\sum_{k=1}^K (mT + N_k)$, which is bounded on the order

$$\underbrace{\tfrac{\beta}{(1-\gamma)\mu}\log(\tfrac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon})}_{\text{deterministic error}} + \underbrace{\tfrac{\varsigma^2}{(1-\gamma)^2\mu}\log(\tfrac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon}) + \tfrac{\mathrm{trace}\left((I_d-M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d-M)^{-\top}\right)}{\epsilon}}_{\text{stochastic error}}. \tag{4.11}$$

Similar to the VRTD algorithm, the VRFTD algorithm achieves optimal sample complexity in terms of stochastic error. For the deterministic error, the dependence on $1/(1-\gamma)$ matches the oracle complexity lower bound proved in Theorem 3.1. Note that the term β/μ is the condition number of the feature matrix in ℓ_Π-norm. Therefore, for "well-conditioned" feature matrices, the VRFTD algorithm achieves optimal oracle complexity[4]. In summary, VRFTD

[4]It is an interesting open problem to prove an oracle complexity lower bound for policy evaluation with linear function approximation having linear dependence on β/μ.

is an accelerated and instance-optimal policy evaluation algorithm, and answers the central question posed in this paper.

5. Algorithm for policy evaluation in the Markovian setting. Finally, we extend the VRFTD algorithm to the Markovian setting, noting in passing that such an extension is also possible for the VRTD algorithm. The challenge of Markovian noise stems from the presence of dependent data that leads to biased samples. To control the bias caused by correlation, we need a standard ergodicity assumption on the underlying Markov chain.

Assumption 3. *There exist constants $C_P > 0$ and $\rho \in (0,1)$ such that*

$$\max_{s\in S} \|\mathbb{P}(s_t = \cdot | s_0 = s) - \pi\|_\infty \leq C_P \cdot \rho^t \quad \text{for all } t \in \mathbb{Z}_+. \tag{5.1}$$

In other words, with the following definition of mixing time

$$t_{\mathsf{mix}} := \inf\{t \in \mathbb{Z}_+ \mid \max_{s\in S} \|\mathbb{P}(s_t = \cdot | s_0 = s) - \pi\|_\infty \leq 1/4\},$$

Assumption 3 guarantees that the mixing time is bounded as[5] $t_{\mathsf{mix}} \leq \frac{\log(4C_P)}{\log(1/\rho)}$.

In order to overcome the difficulty caused by highly-correlated data, we introduce a burn-in period for sample collection. For instance, to compute the operator $\widehat{g}$ defined in Algorithm 4.2, we collect N_k successive samples and only use the last $N_k - n_0$ of them. With this method, we are able to reduce the bias induced by Markovian samples and achieve the desired variance reduction properties. The following two lemmas make this quantitative.

Lemma 5.1. *For every $t, \tau \in \mathbb{Z}_+$, with probability 1,*

$$\|\mathbb{E}[\widetilde{g}(\bar{\theta}, \xi_{t+\tau})|\mathcal{F}_t] - g(\bar{\theta})\|_2 \leq C_M \cdot \rho^\tau \|\bar{v} - v^*\|_\Pi. \tag{5.2}$$

where $C_M := \frac{C_P}{\sqrt{\min_{i\in[D]} \pi_i}} \|\Psi\|_2 \|I - \gamma P\|_2$ and $\mathcal{F}_t := [\xi_1, ..., \xi_t]$.

See Appendix SM3.1 for a proof of this lemma. In words, Lemma 5.1 provides an upper bound on the bias of the stochastic operator at the solution $\bar{\theta}$ in terms of the approximation error, and the bound decays exponentially with τ.

Lemma 5.2. *For every $t, \tau \in \mathbb{Z}_+$ and $\theta, \theta' \in \mathbb{R}^d$, with probability 1,*

$$\|\mathbb{E}[\widetilde{g}(\theta, \xi_{t+\tau})|\mathcal{F}_t] - \mathbb{E}[\widetilde{g}(\theta', \xi_{t+\tau})|\mathcal{F}_t] - [g(\theta) - g(\theta')]\|_2 \leq C_M \cdot \rho^\tau \|v - v'\|_\Pi. \tag{5.3}$$

See Appendix SM3.2 for a proof of this lemma. In contrast to Lemma 5.1, Lemma 5.2 provides an upper bound on the bias of the *difference* of stochastic operators, which allows us to get rid of any dependence on the approximation error. We are now ready to formally state the VRFTD algorithm in the Markovian noise setting in Algorithm 5.1.

[5]Note that while our choice of the constant 1/4 in the definition is arbitrary, there is no additional dependence on ϵ when accounting for the mixing time, unlike in the assumptions made by [1, Eq. (21)].

Algorithm 5.1 Variance-reduced Fast Temporal Difference Algorithm under Markovian noise

Input: $\theta^0 = \widehat{\theta}_0 \in \mathbb{R}^d$, $\eta > 0$, $\lambda \geq 0$, $\{\zeta_t\}_{t=1}^T \geq 0$ and nonnegative integers m, m_0, n_0, $\{N_k\}_{k=1}^K$.

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

Set $\theta_1 = \widetilde{\theta} = \widehat{\theta}_{s-1}$. Collect N_k successive samples $\xi_i^k := (s_i, s_{i+1}, R(s_i, s_{i+1}))$ from the single Markov trajectory. Calculate

$$\widehat{g}(\widetilde{\theta}) = \tfrac{1}{N_k - n_0} \textstyle\sum_{i=n_0+1}^{N_k} \widetilde{g}(\widetilde{\theta}, \xi_i^k). \tag{5.4}$$

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

Collect m successive samples $\widehat{\xi}_j^t := (s_j, s_{j+1}, R(s_j, s_{j+1}))$ from the Markov trajectory. Calculate $\widetilde{g}_t(\cdot) = \frac{1}{m-m_0} \sum_{j=m_0+1}^{m} \widetilde{g}(\cdot, \widehat{\xi}_j^t)$.

Let $\widetilde{F}_t(\theta_t) = \widetilde{g}_t(\theta_t) - \widetilde{g}_t(\widetilde{\theta}) + \widehat{g}(\widetilde{\theta})$ and set $\widetilde{F}_0(\theta_0) = \widetilde{F}_1(\theta_1)$. Let

$$\theta_{t+1} = \theta_t - \eta \left[\widetilde{F}_t(\theta_t) + \lambda \left(\widetilde{F}_t(\theta_t) - \widetilde{F}_{t-1}(\theta_{t-1}) \right) \right]. \tag{5.5}$$

end for

Output of the epoch:

$$\widehat{\theta}_k = \frac{\sum_{t=1}^{T+1} \zeta_t \theta_t}{\sum_{t=1}^{T+1} \zeta_t}. \tag{5.6}$$

end for

Before presenting our main convergence result for the VRFTD algorithm, we first define the matrix $\bar{\Sigma}_{\mathsf{Mkv}}$, which is a covariance matrix analog for the Markovian case (see Mou et al. [28] and references therein). Letting $\{\widetilde{s}_t\}_{t=-\infty}^{\infty}$ define a sequence of samples obtained from a *stationary* Markov trajectory, define

$$\bar{\Sigma}_{\mathsf{Mkv}} := \sum_{t=-\infty}^{\infty} B^{-\frac{1}{2}} \, \mathbb{E}\left[\left(\widetilde{g}(\bar{\theta}, \widetilde{\xi}_t) - g(\bar{\theta}) \right) \left(\widetilde{g}(\bar{\theta}, \widetilde{\xi}_0) - g(\bar{\theta}) \right)^\top \right] B^{-\frac{1}{2}},$$

where $\widetilde{\xi}_t := (\widetilde{s}_t, \widetilde{s}_{t+1}, R(\widetilde{s}_t, \widetilde{s}_{t+1}))$. This matrix is an infinite sum of matrices where one of the summands (when $t = 0$) is the matrix $\bar{\Sigma}_{\mathsf{iid}}$ defined in Eq. (3.2).

Similarly to the i.i.d. setting, the instance-dependent complexity of Markovian linear stochastic approximation was shown in [28] to be governed by the trace of the matrix $(I_d - M)^{-1} \bar{\Sigma}_{\mathsf{Mkv}} (I_d - M)^{-T}$. To interpret this functional, consider the special case in which the approximation error caused by linear function approximation is 0, i.e., $v^* = \bar{v}$. Let $\widetilde{\mathcal{F}_i}$ denote the σ-field generated by samples $\widetilde{\xi}_0, ..., \widetilde{\xi}_i$ and let $\widetilde{\Pi}_j^i := \mathrm{diag}\{[\mathbb{P}(\widetilde{s}_j = 1 | \widetilde{s}_i), ..., \mathbb{P}(\widetilde{s}_j = D | \widetilde{s}_i)]\}$

for $j \geq i$. Then we have

$$\begin{aligned}
&\mathbb{E}\left\langle (I_d - M)^{-1} B^{-\frac{1}{2}} \big(\widetilde{g}(\bar{\theta}, \widetilde{\xi}_0) - g(\bar{\theta})\big), (I_d - M)^{-1} B^{-\frac{1}{2}} \big(\widetilde{g}(\bar{\theta}, \widetilde{\xi}_i) - g(\bar{\theta})\big) \right\rangle \\
&= \mathbb{E}\left\langle (I_d - M)^{-1} B^{-\frac{1}{2}} \big(\widetilde{g}(\bar{\theta}, \widetilde{\xi}_0) - g(\bar{\theta})\big), (I_d - M)^{-1} B^{-\frac{1}{2}} \big(\mathbb{E}[\widetilde{g}(\bar{\theta}, \widetilde{\xi}_i) | \widetilde{\mathcal{F}}_0] - g(\bar{\theta})\big) \right\rangle \\
&= \mathbb{E}\left\langle (I_d - M)^{-1} B^{-\frac{1}{2}} \big(\widetilde{g}(\bar{\theta}, \widetilde{\xi}_0) - g(\bar{\theta})\big), (I_d - M)^{-1} B^{-\frac{1}{2}} \Psi (\widetilde{\Pi}_j^i - \Pi)(\Psi^\top \bar{\theta} - \gamma P \Psi^\top \bar{\theta} - r) \right\rangle = 0,
\end{aligned}$$

where the last equation follows from the fact that $v^* = \bar{v} = \Psi^\top \bar{\theta}$ and the Bellman equation (2.1). Then

$$\operatorname{trace}\left((I_d - M)^{-1} \bar{\Sigma}_{\mathsf{Mkv}} (I_d - M)^{-\top} \right) = \operatorname{trace}\left((I_d - M)^{-1} \bar{\Sigma}_{\mathsf{iid}} (I_d - M)^{-\top} \right).$$

Armed with this intuition, we are now ready to establish the main convergence result for the VRFTD algorithm under Markovian noise. Given the calculation above, we discuss the cases $\bar{v} = v^*$ and $\bar{v} \neq v^*$ separately for clarity.

Theorem 5.3. *Fix the total number of epochs K and a positive integer N. Consider an integer τ satisfying $\rho^\tau \leq \min\{\frac{2(1-\rho)\varsigma}{3C_M}, \frac{2(1-\rho)^2}{5C_M}\}$. Suppose the parameters n_0 and m_0 satisfy*

$$\rho^{n_0} \leq \frac{\min_{i\in[D]} \pi_i}{C_P}, \quad \text{and} \quad \rho^{m_0} \leq \min\left\{ \frac{\min_{i\in[D]} \pi_i}{C_P}, \frac{\sqrt{\mu}\eta\tau\varsigma^2(1-\rho)}{C_M} \right\}. \tag{5.7}$$

Assume that for each epoch $k \in [K]$, the parameters η, λ, m, N_k, T satisfy

$$\begin{aligned}
&\eta \leq \tfrac{1}{4\beta(1+\gamma)}, \quad \lambda = 1, \quad T \geq \tfrac{64}{\mu(1-\gamma)\eta}, \quad m - m_0 \geq \max\left\{1, \tfrac{792\eta(\tau+1)\varsigma^2}{1-\gamma}\right\}, \\
&\rho^{N_k - n_0} \leq \tfrac{\tau(1-\rho)}{5C_M(N_k - n_0)}, \quad \text{and} \quad N_k - n_0 \geq \left\{ \tfrac{206(\tau+1)\varsigma^2}{\mu(1-\gamma)^2}, (\tfrac{3}{4})^{K-k} N \right\}.
\end{aligned} \tag{5.8}$$

Set the output of each epoch to be $\widehat{v}_k := \frac{\sum_{t=2}^{T+1} v_t}{T}$. Then the following results hold.

(a) If $\bar{v} = v^*$, we have

$$\mathbb{E}[\|\widehat{v}_K - v^*\|_\Pi^2] \leq \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \tfrac{30}{N} \cdot \operatorname{trace}\left((I_d - M)^{-1} \bar{\Sigma}_{\mathsf{iid}} (I_d - M)^{-\top} \right). \tag{5.9}$$

(b) If $\bar{v} \neq v^*$, then for any $\delta > 0$ we have

$$\begin{aligned}
\mathbb{E}[\|\widehat{v}_K - v^*\|_\Pi^2] &\leq \left(1 + \delta + \tfrac{18(\tau+1)(1+1/\delta)}{\mu(1-\gamma)^2 N^2}\right) \cdot \mathcal{A}(M, \gamma) \cdot \inf_{v \in S} \|v - v^*\|_\Pi^2 \\
&\quad + (1 + \tfrac{1}{\delta}) \left[\tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \tfrac{30}{N} \cdot \operatorname{trace}\left((I_d - M)^{-1} \bar{\Sigma}_{\mathsf{Mkv}} (I_d - M)^{-\top} \right) + \tfrac{\mathcal{H}}{N^2} \right],
\end{aligned} \tag{5.10}$$

where $\mathcal{H} := (90\tau^2 + 18\tau + 18) \cdot \operatorname{trace}\left((I_d - M)^{-1} \bar{\Sigma}_{\mathsf{iid}} (I_d - M)^{-\top} \right)$.

See Section 7.5 for detailed proofs of Theorem 5.3. Let us now discuss a few aspects of the theorem.

Estimation of mixing time. From the conditions above, e.g., Ineq. (5.7), the parameters τ, n_0, m_0 scale linearly in the mixing time t_{mix} and logarithmically in other problem parameters. As such, only some rough estimation of the mixing time is sufficient, which has been the topic of active research. Nontrivial confidence intervals for the reversible case can be found in Hsu et al. [12]. There are also guarantees in the more challenging and prevalent case when the underlying Markov chain is non-reversible [48].

Sample complexities. We first consider the case when $\bar{v} = v^*$. In view of Ineq. (5.9) in Theorem 5.3, the total number of samples required by the VRFTD method to find a solution $\widehat{v} \in \mathbb{R}^D$, such that $\mathbb{E}[\|\widehat{v} - v^*\|_\Pi^2] \le \epsilon$ is $\sum_{k=1}^K (mT + N_k)$, which is bounded on the order[6]

$$\underbrace{\frac{t_{\mathsf{mix}}\beta}{(1-\gamma)\mu}\log\Big(\frac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon}\Big)}_{\text{deterministic error}} + \underbrace{\frac{t_{\mathsf{mix}}\varsigma^2}{(1-\gamma)^2\mu}\log\Big(\frac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon}\Big) + \frac{\mathrm{trace}\Big((I_d-M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d-M)^{-\top}\Big)}{\epsilon}}_{\text{stochastic error}}, \tag{5.11}$$

where the mixing time only enters along with terms that scale logarithmically in $1/\epsilon$. The phenomenon that the mixing time does not enter multiplicatively with the leading-order stochastic error term was also noticed by Li et al. [24] for vanilla TD learning, but as mentioned before, this algorithm does not attain the correct instance-dependent stochastic error.

When $\bar{v} \neq v^*$, the Markovian setting has a biased stochastic operator at optimal solution $\bar{v}$, and a larger approximation error $\|\bar{v} - v^*\|_\Pi^2$ caused by linear function approximation enlarges the bias of the stochastic operator and consequently enlarges the dependence on the approximation error in Eq. (5.10). Therefore, a natural stopping criterion for the Markovian setting is to find a solution $\widehat{v} \in \mathbb{R}^D$ satisfying $\mathbb{E}[\|\widehat{v} - v^*\|_\Pi^2] \le c\|\bar{v} - v^*\|_\Pi^2 + \epsilon$ for some absolute constant $c > 0$. From Ineq. (5.10), the total number of required samples $\sum_{k=1}^K (mT + N_k)$ is bounded on the order

$$\underbrace{\frac{t_{\mathsf{mix}}\beta}{(1-\gamma)\mu}\log\Big(\frac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon}\Big)}_{\text{deterministic error}} + \underbrace{\frac{t_{\mathsf{mix}}\varsigma^2\log\big(\frac{\|v^0-\bar{v}\|_\Pi^2}{\epsilon}\big)}{(1-\gamma)^2\mu} + \frac{\sqrt{\mathcal{H}}}{\sqrt{\epsilon}} + \frac{\mathrm{trace}\Big((I_d-M)^{-1}\bar{\Sigma}_{\mathsf{Mkv}}(I_d-M)^{-\top}\Big)}{\epsilon}}_{\text{stochastic error}}. \tag{5.12}$$

Note that, in this bound, the leading stochastic error matches the lower bound proved in [28], which can depend on the mixing time, but is generally smaller than the product of the i.i.d. stochastic error and the mixing time. These results show a delicate difference between how the mixing time of the Markov chain enters the bound depending on whether the function approximation is exact or not. It should be noted that, the stronger stopping criterion, i.e., finding $\widehat{v}$ to satisfy $\mathbb{E}[\|\widehat{v} - \bar{v}\|_\Pi^2] \le \epsilon$, can also be applied in this setting. We can generate a similar sample complexity by enlarging the constants τ and N_k by an additive factor of $\log(\|v^* - \bar{v}\|_\Pi^2)$. However, given that the approximation error is unavoidable and generally unknown, there is marginal benefit to using this stronger stopping criterion.

Mou et al. [28] established convergence guarantees for TD with averaging in the Markovian noise setting, showing a similar leading order stochastic error term but without accelerating the deterministic error. Besides improving on the deterministic error, Theorem 5.3 also guarantees that the higher-order terms on stochastic error are smaller than those proved in [28].

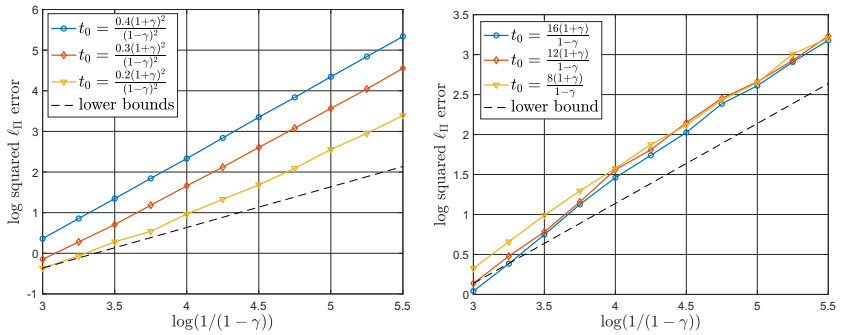
6. Numerical experiments.

In this section, we report numerical experiments for both VRTD and VRFTD, comparing them against temporal difference learning (TD), conditional temporal difference learning (CTD), and fast temporal difference learning (FTD) [19, 20]. To generate a comprehensive performance profile, we conduct experiments under both the i.i.d. and Markovian noise models.

[6] Note that we omit the logarithmic dependence on the problem parameters, e.g., μ, β, γ. Same for the following complexity.

6.1. The i.i.d. setting: A simple two-state construction. We consider a family of two-state MRPs inspired by the construction of Duan et al. [9]. For a discount factor $\gamma \in (\frac{1}{2}, 1)$, the transition kernel P and reward vector r are given by $P = \begin{bmatrix} \frac{2\gamma-1}{\gamma} & \frac{1-\gamma}{\gamma} \\ \frac{1-\gamma}{\gamma} & \frac{2\gamma-1}{\gamma} \end{bmatrix}$ and $r = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Clearly the transition kernel is symmetric, thus the stationary distribution is $\pi = [0.5, 0.5]$. For simplicity, we choose the feature matrix $\Psi = \mathrm{diag}([\sqrt{2}, \sqrt{2}])$, which forms an orthonormal basis under ℓ_Π-norm. Assuming the i.i.d. model in which $s_i \sim \pi$ and $s_i' \sim P(\cdot|s_i)$, it can be shown via simple calculation that the stochastic error term is given by

$$\mathrm{trace}\big((I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\big) = \tfrac{40}{81} \cdot \tfrac{2\gamma-1}{(1-\gamma)^3}. \tag{6.1}$$



(a) TD algorithm (b) FTD algorithm

(c) VRTD algorithm (d) VRFTD algorithm

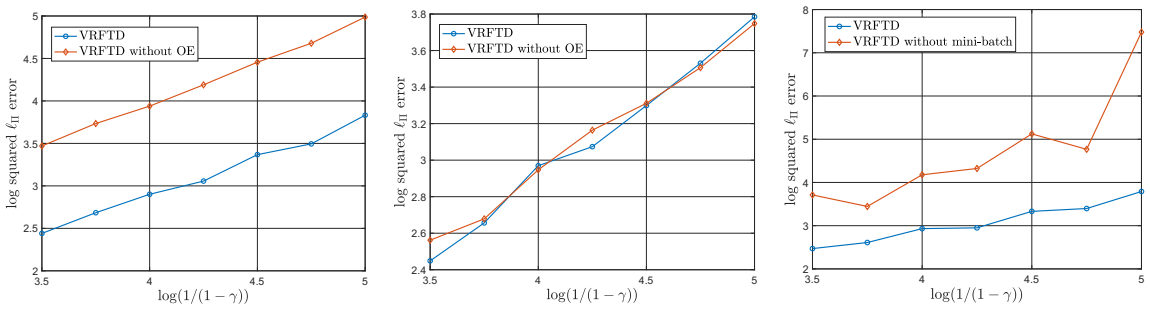
Figure 1. Log-log plots of the squared ℓ_Π-norm error versus $1/(1-\gamma)$. Logarithms are to the natural base. The number of samples used in each single experiment is $N = \lceil 5/(1-\gamma)^2 \rceil$. Each point in the plot is an average of 1000 independent trials. The slope of the lower bound is 1.

Instance-optimality. We generate a range of MRPs with different values of discount factor γ and run the four aforementioned algorithms on each MRP. In order to test the robustness of our results, we simulate various step-sizes for each algorithm. To be fair in our comparison, we also include a simulation of the best-tuned stepsize for each algorithm. We plot the prediction from the lower bound (6.1) as well.

From subplots (a) and (b) of Figure 1, it is clear that the vanilla TD and FTD algorithms with diminishing stepsizes [20] do not achieve the lower bound calculated in equation (6.1).

On the other hand, sub-plots (c) and (d) show that the VRTD and VRFTD algorithms achieve the lower bound (6.1), and that these behaviors are robust to the choice of stepsize parameters. However, given their epoch-wise nature, the outputs of variance-reduced algorithms are more volatile than TD and FTD. Another interesting observation is that the accelerated algorithms—FTD and VRFTD—are less sensitive to stepsize parameters. Our next set of experiments explores this further.

Ablation analysis of VRFTD. Notice that VRFTD includes two new ingredients when compared with VRTD: mini-batching and operator extrapolation (OE). We now perform an ablation analysis to disentangle the contribution of both ingredients. We generate a range of MRPs with different values of γ and run the experimental and control groups on each MRP.



(a) Ablation study for OE (I) (b) Ablation study for OE (II) (c) Ablation study for mini-batching

Figure 2. Log-log plots of the squared ℓ_Π-norm error versus $1/(1-\gamma)$. Logarithms are to the natural base. The number of samples used in each single experiment is $N = \lceil 5/(1-\gamma)^2 \rceil$. Each point in the plot is an average of 1000 independent trials.

In the first experiments, we ran the experimental group with OE steps and the control group without OE steps. We first ran both groups with the stepsize policies suggested by the theoretical analysis (see subplot (a) of Figure 2). The results indicate that the experimental group significantly outperforms the control group. However, this performance difference can largely be attributed to the conservative stepsize for the control group, as prescribed by the theory. To be more fair to both algorithms, we further fine-tuned the stepsize parameters of both algorithms and obtained subplot (b), where the two algorithms exhibit similar convergence rates. Taking stock, the first set of experiments shows that the analysis and stepsize policy of the VRFTD (with OE steps) serves as a better theoretical guideline for practical applications.

To demonstrate the benefits of mini-batching in the inner loop, we ran a second experiment with two groups with and without mini-batching. Note that we keep the stepsizes and the total number of samples the same for both groups (which means that the control group without mini-batching has larger epoch lengths). From subplot (c), we can see that the performance difference is significant, showing that without mini-batching, the algorithm exhibits instability when run with aggressive stepsize policies.

6.2. The Markovian setting: 2D Grid World. Our experiments under the Markovian noise model are conducted on the 2D Grid World environment. This is a classical problem in reinforcement learning with finite state and action spaces. An agent realizes a positive reward when reaching a predetermined goal and negative ones when going through "traps". The dimension of the state space is set to be $D = 400$, among which we assign a goal state (with reward $r = 1$) and 30 traps (with reward $r = -0.2$). The transition kernel is fixed as follows: With probability 0.95, the agent moves in a direction that points towards the goal and with probability 0.05 in a random direction. Our goal is to compute the value function v^*—for each possible initial state of the agent. We also incorporate linear function approximation in these experiments. Specifically, we generate random features with dimension $d = 50$ to estimate the $D = 400$ dimensional value function in this problem.

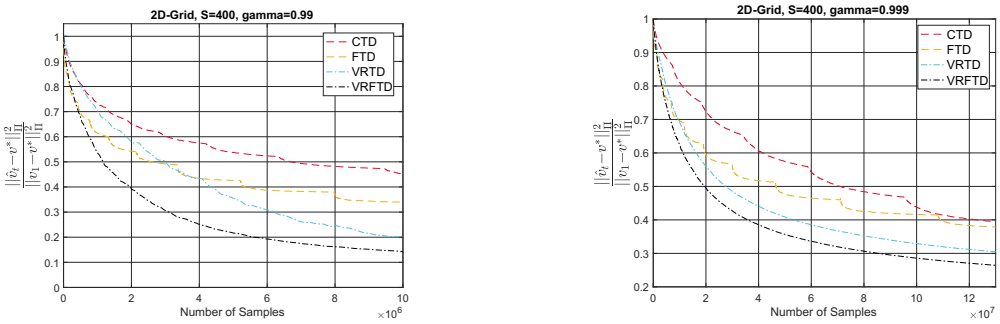


Figure 3. Comparison of the algorithms for the 2D-Grid world example. From left to right γ is set to 0.99, and 0.999 respectively. In the y-axis we report ratios in terms of the Euclidean norm $\|\cdot\|_\Pi$.

We test the performance of four algorithms, with the discount factor γ set to 0.99 and 0.999. Figure 3 plots the normalized error in ℓ_Π-norm against the length of the trajectory. In both experiments, the VRFTD algorithm exhibits the fastest convergence to the true value function, thereby corroborating our theoretical results. Note that in both experiments, the estimation errors do not converge to zero, because there is a nontrivial error incurred by linear function approximation. Another salient takeaway is the following: Closer to the optimal solution v^*, the variance-reduced algorithms (VRTD/VRFTD) achieve faster convergence rate compared to their counterparts that do not employ variance reduction.

7. Proofs. In this section, we provide the proofs of our main results. The proof of other auxiliary results are collected in the supplementary material.

7.1. Proof of Theorem 3.1. First, it is clear that the methods of this type are invariant to a simultaneous shift of variables. The sequence of iterates for solving $G(v) = 0$ starting from v_0 is just a shift of the sequence generated for solving $G(v + v_0) = 0$ starting from the origin. Therefore, without loss of generality, we assume $v_0 = 0$.

Now let us construct a specific instance (P, r) to show the lower bound. Consider the

$D \times D$ matrix

$$P := \begin{bmatrix} \frac{1}{2\gamma} & 0 & 0 & \dots & 0 & 1-\frac{1}{2\gamma} \\ 1-\frac{1}{2\gamma} & \frac{1}{2\gamma} & 0 & \dots & 0 & 0 \\ 0 & 1-\frac{1}{2\gamma} & \frac{1}{2\gamma} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1-\frac{1}{2\gamma} & \frac{1}{2\gamma} \end{bmatrix}. \tag{7.1}$$

Also define the D-dimensional vector

$$r := [\gamma - \tfrac{1}{2} + (\tfrac{1}{2} - \gamma)(2\gamma - 1)^D,\ 0,\ 0, \dots,\ 0,\ 0]^\top \tag{7.2}$$

The matrix $I - \gamma P$ is square and full rank, and it is straightforward to verify that the unique solution of the linear equation $(I - \gamma P)v^* = r$ is

$$(v^*)_{(i)} = (2\gamma - 1)^i \quad \text{for all } i \in [D]. \tag{7.3}$$

It is also easy to see that the Markov chain induced by the transition kernel P is irreducible and aperiodic. From the cyclical behavior of the Markov chain, we have that the stationary distribution is uniform, i.e.,

$$\Pi = [\tfrac{1}{D},\ \tfrac{1}{D},\ \tfrac{1}{D}, \dots,\ \tfrac{1}{D}]^\top.$$

Thus, we obtain

$$\|v_0 - v^*\|_\Pi^2 = \tfrac{1}{D}\textstyle\sum_{i=1}^{D}(2\gamma - 1)^{2i} = \frac{(2\gamma-1)^2[1-(2\gamma-1)^{2D}]}{D[1-(2\gamma-1)^2]}. \tag{7.4}$$

Let $\mathbb{R}^{k,D} := \{v \in \mathbb{R}^D \mid v_{(i)} = 0 \text{ for all } k+1 \le i \le D\}$ denote the set of all D-dimensional vectors lying in the span of the first k standard basis vectors. Since all entries of $I - \gamma P$ below its subdiagonal are equal to 0 and all entries of r except for its first are equal to 0, we conclude that $v_k \in \mathbb{R}^{k,D}$. Therefore,

$$\|v_k - v^*\|_\Pi^2 \ge \tfrac{1}{D}\textstyle\sum_{i=k+1}^{D}(2\gamma - 1)^{2i} = \frac{(2\gamma-1)^{2k+2}[1-(2\gamma-1)^{2D-2k}]}{D[1-(2\gamma-1)^2]}.$$

If $D \gg k$ is such that $\frac{1-(2\gamma-1)^{2D-2k}}{1-(2\gamma-1)^{2D}} \ge \frac{1}{2}$, then we conclude that

$$\|v_k - v^*\|_\Pi^2 \ge \tfrac{1}{2}(2\gamma - 1)^{2k}\|v_0 - v^*\|_\Pi^2,$$

as desired.

7.2. Proof of Proposition 3.2. For the reader's convenience, we begin by stating a version of the Hájek-Le Cam local asymptotic minimax theorem.

Theorem 7.1. *Let $\{\mathbb{P}_\vartheta\}_{\vartheta \in \Theta}$ denote a family of parametric models, quadratically mean differentiable with Fisher information matrix $J_{\vartheta'}$. Fix some parameter $\vartheta \in \Theta$, and consider a*

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

$$\lim_{c\to\infty} \lim_{N\to\infty} \inf_{\widehat{h}_N} \sup_{\vartheta' : \|\vartheta'-\vartheta\|_2 \le c/\sqrt{N}} \mathbb{E}_{\vartheta'} \left[L\big(\sqrt{N}(\widehat{h}_N - h(\vartheta'))\big) \right] = \mathbb{E}[L(Z)], \tag{7.5}$$

where the infimum is taken over all estimators $\widehat{h}_N$ that are measurable functions of N i.i.d. data points drawn from $\mathbb{P}_{\vartheta'}$, and the expectation is taken over a multivariate Gaussian

$$Z \sim \mathcal{N}\left(0, \nabla h(\vartheta)^\top J_\vartheta^\dagger \, \nabla h(\vartheta)\right).$$

In our model, we have $\vartheta = (\omega, P, R)$ and $h(\vartheta) = (\Psi\Omega\Psi^\top - \Psi\Omega\gamma P\Psi^\top)^{-1}\Psi\Omega r$. We set the loss function L to be $\|\cdot\|_{\widetilde{B}}^2$. Invoking Theorem 7.1 yields

$$\mathfrak{M}(\vartheta) = \mathbb{E}\left[\|Z\|_{\widetilde{B}}^2\right] \quad \text{where} \quad Z \sim \mathcal{N}\left(0, \nabla h(\vartheta)^\top J_\vartheta^\dagger \, \nabla h(\vartheta)\right). \tag{7.6}$$

The covariance is explicitly computed in the following lemma.

Lemma 7.2. *We have*

$$\nabla h(\vartheta)^\top J_\vartheta^\dagger \, \nabla h(\vartheta) = (\Psi\Omega\Psi^\top - \gamma\Psi\Omega P\Psi^\top)^{-1}\Sigma \, (\Psi\Omega\Psi^\top - \gamma\Psi\Omega P\Psi^\top)^{-T}, \tag{7.7}$$

where $\Sigma = \operatorname{cov}\left[\big(\langle\psi(s) - \gamma\psi(s'), \bar{\theta}\rangle - R(s,s')\big)\psi(s)\right]$ for $s \sim \omega$, $s' \sim \mathbb{P}(\cdot|s)$.

The proof of this lemma is straightforward but involves some lengthy calculations, which we defer to Appendix SM3.3.

Recall the definition of $\widetilde{M}$ and $\widetilde{\Sigma}$ in the statement of Proposition 3.2. By substituting equality (7.7) into (7.6) and invoking the relation $\widetilde{B}^{-\frac{1}{2}}\Psi\Omega\Psi^\top\widetilde{B}^{-\frac{1}{2}} = I_d$, we obtain

$$\widetilde{B}^{\frac{1}{2}} Z \sim \mathcal{N}\left(0, (I_d - \widetilde{M})^{-1}\widetilde{\Sigma}(I_d - \widetilde{M})^{-T})\right),$$

which completes the proof of Proposition 3.2.

7.3. Proof of Theorem 4.1. Let $\underline{\theta}$ satisfy $g(\underline{\theta}) - g(\widetilde{\theta}) + \widehat{g}(\widetilde{\theta}) = 0$ and $\underline{v} = \Psi^\top\underline{\theta}$. Recalling the definition of $\bar{v}$ from Eq. (2.2), the following lemma provides a bound for $\|\underline{v} - \bar{v}\|_\Pi^2$. This bound is also valid for the VRFTD algorithm in the i.i.d. setting.

Lemma 7.3. *Consider a single epoch with index $k \in [K]$. We have*

$$\mathbb{E}[\|\underline{v} - \bar{v}\|_\Pi^2] \le \tfrac{2}{N_k}\operatorname{trace}\left((I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\right) + \tfrac{2\varsigma^2}{N_k(1-\gamma)^2\mu}\mathbb{E}\|\widetilde{v} - \bar{v}\|_\Pi^2. \tag{7.8}$$

See Appendix SM3.4 for the proof of this lemma.

Given Lemma 7.3, we can derive the following proposition which characterizes the progress of the VRTD algorithm in a single epoch.

Proposition 7.4. *Consider a single epoch with index $k \in [K]$. Suppose that the parameters η, N_k and T satisfy*

$$\eta \le \min\{\tfrac{(1-\gamma)}{2\beta(1+\gamma)^2}, \tfrac{1-\gamma}{32\varsigma^2}\}, \quad T \ge \tfrac{32}{\mu(1-\gamma)\eta}, \quad \text{and} \quad N_k \ge \tfrac{38\varsigma^2}{\mu(1-\gamma)^2}. \tag{7.9}$$

Set the output of this epoch to be $\widehat{v}_k := \frac{\sum_{t=1}^T \eta(1-\gamma)v_t + (1/\beta)v_{T+1}}{T\eta(1-\gamma)+(1/\beta)}$. Then we have

$$\mathbb{E}[\|\widehat{v}_k - \bar{v}\|_\Pi^2] \le \tfrac{1}{2}\mathbb{E}[\|\widehat{v}_{k-1} - \bar{v}\|_\Pi^2] + \tfrac{5}{N_k}\operatorname{trace}\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\}. \tag{7.10}$$

See Appendix SM2.2 for the proof of this proposition. The basic idea of the proof is first providing an upper bound on the term $\mathbb{E}\|\widehat{v}_k - \underline{v}\|_\Pi^2$ and then combining it with Lemma 7.3.

Taking Proposition 7.4 as given for the moment, let us complete the proof of the theorem. The main idea is to bound the approximation error term $\|\bar{v} - v^*\|_\Pi^2$ separately from the term $\|\widehat{v}_K - \bar{v}\|_\Pi^2$. To bound $\|\bar{v} - v^*\|_\Pi^2$, we use the instance-dependent upper bound in Ineq. (2.6). We bound the term $\|\widehat{v}_K - \bar{v}\|_\Pi^2$ by using Ineq. (7.10) as follows

$$\begin{aligned}\mathbb{E}\|\widehat{v}_K - \bar{v}\|_\Pi^2 &\le \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \textstyle\sum_{k=1}^K \frac{5}{2^{K-k}N_k} \cdot \mathrm{trace}\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\} \\ &\overset{(i)}{=} \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \textstyle\sum_{k=1}^K (\frac{2}{3})^{K-k}\frac{5}{N} \cdot \mathrm{trace}\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\} \\ &\le \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \tfrac{15}{N} \cdot \mathrm{trace}\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\}. \qquad (7.11)\end{aligned}$$

Here, step (i) from the condition that $N_k \ge (\frac{3}{4})^{K-k}N$ for all $k \in [K]$. To conclude, we use Young's inequality to obtain

$$\begin{aligned}\mathbb{E}\|\widehat{v}_K - v^*\|_\Pi^2 &\le (1+\delta)\cdot\mathbb{E}\|\bar{v} - v^*\|_\Pi^2 + (1+\tfrac{1}{\delta})\cdot\mathbb{E}\|\widehat{v}_K - \bar{v}\|_\Pi^2 \\ &\le (1+\delta)\cdot\mathcal{A}(M,\gamma)\cdot\inf_{v\in S}\|v - v^*\|_\Pi^2 + (1+\tfrac{1}{\delta})\cdot\mathbb{E}\|\widehat{v}_K - \bar{v}\|_\Pi^2,\end{aligned}$$

which completes the proof.

7.4. Proof of Theorem 4.2. The structure of the proof is similar to the analysis of VRTD in Section 7.3. We first state a proposition that characterizes the progress of the VRFTD algorithm in a single epoch.

Proposition 7.5. *Assume that for each epoch $k \in [K]$, the parameter η, λ, m, N_k and T satisfy*

$$\eta \le \tfrac{1}{4\beta(1+\gamma)},\ \lambda = 1,\ T \ge \tfrac{32}{\mu(1-\gamma)\eta},\ m \ge \max\{1, \tfrac{256\eta\varsigma^2}{1-\gamma}\},\ \text{and } N_k \ge \tfrac{56\varsigma^2}{\mu(1-\gamma)^2}. \qquad (7.12)$$

Set the output of this epoch to be $\widehat{v}_k := \frac{\sum_{t=2}^{T+1} v_t}{T}$. Then we have

$$\mathbb{E}[\|\widehat{v}_k - \bar{v}\|_\Pi^2] \le \tfrac{1}{2}\mathbb{E}[\|\widehat{v}_{k-1} - \bar{v}\|_\Pi^2] + \tfrac{5}{N_k}\mathrm{trace}\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\}. \qquad (7.13)$$

See Appendix SM2.3 for the proof of this proposition.

Taking Proposition 7.5 as given, the proof of Theorem 4.2 follows exactly as the proof of Theorem 4.1 in Section 7.3.

7.5. Proof of Theorem 5.3. The structure of this proof is similar to proofs of Theorems 4.1 and 4.2: We first derive a bound for a single epoch, and then apply it recursively to obtain the eventual convergence result. The following proposition characterizes the progress in a single epoch $k \in [K]$.

Proposition 7.6. *Consider a single epoch with index $k \in [K]$. Consider an integer τ satisfying $\rho^\tau \le \min\{\frac{2(1-\rho)\varsigma}{3C_M}, \frac{2(1-\rho)^2}{5C_M}\}$. Suppose the parameters N_k, n_0 and m_0 satisfy*

$$\rho^{N_k - n_0} \le \tfrac{\tau(1-\rho)}{5C_M(N_k - n_0)},\ \rho^{n_0} \le \tfrac{\min_{i\in[D]}\pi_i}{C_P},\ \text{and}\ \ \rho^{m_0} \le \min\left\{\tfrac{\min_{i\in[D]}\pi_i}{C_P}, \tfrac{\sqrt{\mu}\eta\tau\varsigma^2(1-\rho)}{C_M}\right\}. \qquad (7.14)$$

Suppose that the parameter η, λ, m, N_k and T satisfy

(7.15)
$$\eta \le \tfrac{1}{4\beta(1+\gamma)}, \quad \lambda = 1, \quad T \ge \tfrac{64}{\mu(1-\gamma)\eta}, \quad m - m_0 \ge \max\{1, \tfrac{792\eta(\tau+1)\varsigma^2}{1-\gamma}\}, \text{ and } N_k - n_0 \ge \tfrac{206(\tau+1)\varsigma^2}{\mu(1-\gamma)^2}.$$

Set the output of this epoch to be $\widehat{v}_k := \frac{\sum_{t=2}^{T+1} v_t}{T}$. Then we have the following results:

(a) If $v^* = \bar{v}$, we have

$$\mathbb{E}\|\widehat{v}_k - v^*\|_\Pi^2 \le \tfrac{1}{2}\mathbb{E}\|\widehat{v}_{k-1} - \bar{v}\|_\Pi^2 + \tfrac{10\cdot\mathrm{trace}\{(I_d-M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d-M)^{-\top}\}}{N_k - N_0}. \tag{7.16}$$

(b) If $v^* \neq \bar{v}$, we have

$$\mathbb{E}\|\widehat{v}_k - v^*\|_\Pi^2 \le \tfrac{1}{2}\mathbb{E}\|\widehat{v}_{k-1} - \bar{v}\|_\Pi^2 + \tfrac{10\cdot\mathrm{trace}\{(I_d-M)^{-1}\bar{\Sigma}_{\mathsf{Mkv}}(I_d-M)^{-\top}\}}{N_k - N_0} + \tfrac{\widetilde{\mathcal{H}}}{(N_k-n_0)^2}, \tag{7.17}$$

where $\widetilde{\mathcal{H}} := \frac{2(\tau+1)}{(1-\gamma)^2\mu}\|\bar{v} - v^*\|_\Pi^2 + (10\tau^2 + 2\tau + 2)\cdot\mathrm{trace}\{(I_d - M)^{-1}\bar{\Sigma}_{\mathsf{iid}}(I_d - M)^{-\top}\}.$

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

Taking Proposition 7.6 as given for the moment, let us complete the proof of the theorem. First, consider the case when $\bar{v} \neq v^*$. Recursively using Ineq. (7.17) yields

$$\begin{aligned}
\mathbb{E}\|\widehat{v}_K - \bar{v}\|_\Pi^2 &\le \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \textstyle\sum_{k=1}^K \left(\tfrac{10\cdot\mathrm{trace}\{(I_d-M)^{-1}\bar{\Sigma}_{\mathsf{Mkv}}(I_d-M)^{-\top}\}}{2^{K-k}(N_k-n_0)} + \tfrac{\widetilde{\mathcal{H}}}{2^{K-k}(N_k-n_0)^2}\right) \\
&\overset{(i)}{\le} \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \textstyle\sum_{k=1}^K (\tfrac{2}{3})^{K-k}\tfrac{10\cdot\mathrm{trace}\{(I_d-M)^{-1}\bar{\Sigma}_{\mathsf{Mkv}}(I_d-M)^{-\top}\}}{N} + \sum_{k=1}^K (\tfrac{8}{9})^{K-k}\tfrac{\widetilde{\mathcal{H}}}{N^2} \\
&\le \tfrac{1}{2^K}\|v^0 - \bar{v}\|_\Pi^2 + \tfrac{30\cdot\mathrm{trace}\{(I_d-M)^{-1}\bar{\Sigma}_{\mathsf{Mkv}}(I_d-M)^{-\top}\}}{N} + \tfrac{9\widetilde{\mathcal{H}}}{N^2}, \qquad (7.18)
\end{aligned}$$

where step (i) follows from the condition $N_k - n_0 \ge (\frac{3}{4})^{K-k}N$. The proof of the case when $\bar{v} = v^*$ follows from the same derivation.

8. Discussion. In this paper, we investigated the problem of policy evaluation with linear function approximation, making three contributions. First, we proved lower bounds on both deterministic error and stochastic error. With these lower bounds in hand, we presented an analysis of a variance-reduced variant of temporal difference algorithm (VRTD) in the i.i.d. observation model and showed that it fails to match the oracle complexity lower bound on the deterministic error. In order to remedy this difficulty, we developed an optimal variance-reduced fast temporal difference algorithm (VRFTD) that nearly matches both lower bounds simultaneously. Finally, we extended the VRFTD algorithm to the Markovian setting and provided instance-dependent convergence results. The leading stochastic error matches the instance-dependent lower bound for Markovian linear stochastic approximation [28], and the deterministic error matches the i.i.d. setting up to a multiplicative factor proportional to the mixing time of the chain. Our theoretical guarantees were corroborated with numerical experiments in both the i.i.d. and Markovian settings, showing that the VRFTD algorithm enjoys several advantages over the prior state-of-the-art.

Our work leaves open severaal salient future directions; let us mention two. First, our oracle complexity lower bound is proved in the tabular setting. On the other hand, our upper bounds on the deterministic error indicate that with linear function approximation, we pay

a multiplicative factor depending on the condition number of the feature matrix. It would be interesting to see if an oracle complexity lower bound can be proved under linear function approximation, and whether the linear dependence on the condition number in our bounds is optimal. Second, and more broadly, note that our analysis relies heavily on the linear structure of the problem. However, there are many problems in the reinforcement learning literature that have nonlinear structures, e.g., the policy optimization problem involving the Bellman optimality operator. An interesting direction for future work is to understand problems with nonlinear structure from an instance-specific point of view and develop efficient algorithms to capture the optimal deterministic and stochastic errors. For instance, variance reduction has been applied to the policy optimization problem under the generative model [38, 46] and some instance-dependent bounds are known (e.g., for variants of Q-learning [16]). It is an important open question to develop acceleration schemes for such algorithms in a fashion similar to our paper, while extending the results to the more realistic Markovian setting.

Acknowledgments. TL and GL were supported in part by Office of Naval Research grant N00014-20-1-2089. TL and AP were supported in part by the National Science Foundation grant CCF-2107455, and are thankful to the Simons Institute for the Theory of Computing for their hospitality when part of this work was performed.

REFERENCES

[1] J. Bhandari, D. Russo, and R. Singal, *A finite time analysis of temporal difference learning with linear function approximation*, in Conference on learning theory, PMLR, 2018, pp. 1691–1692.
[2] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM Journal on Optimization, 18 (2007), pp. 29–51.
[3] V. S. Borkar, *Stochastic approximation: A dynamical systems viewpoint*, vol. 48, Springer, 2009.
[4] 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.
[5] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam, *A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants*, arXiv preprint arXiv:2102.01567, (2021).
[6] C. Dann, G. Neumann, and J. Peters, *Policy evaluation with temporal differences: A survey and comparison*, Journal of Machine Learning Research, 15 (2014), pp. 809–883.
[7] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Advances in neural information processing systems, 2014, pp. 1646–1654.
[8] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou, *Stochastic variance reduction methods for policy evaluation*, in International Conference on Machine Learning, PMLR, 2017, pp. 1049–1058.
[9] Y. Duan, M. Wang, and M. J. Wainwright, *Optimal policy evaluation using kernel-based temporal difference methods*, arXiv preprint arXiv:2109.12002, (2021).
[10] A. Durmus, E. Moulines, A. Naumov, S. Samsonov, and H. T. Wai, *On the stability of random matrix product with Markovian noise: Application to linear stochastic approximation and TD learning*, arXiv preprint arXiv:2102.00185, (2021).
[11] J. Hájek, *Local asymptotic minimax and admissibility in estimation*, in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics, University of California Press, 1972, pp. 175–194.
[12] D. Hsu, A. Kontorovich, D. A. Levin, Y. Peres, C. Szepesvári, and G. Wolfer, *Mixing time estimation in reversible Markov chains from a single sample path*, The Annals of Applied Probability, 29 (2019), pp. 2439–2480.
[13] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*,

in Advances in Neural Information Processing Systems, vol. 26, 2013, pp. 315–323.
[14] L. P. Kaelbling, M. L. Littman, and A. W. Moore, *Reinforcement learning: A survey*, Journal of artificial intelligence research, 4 (1996), pp. 237–285.
[15] K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan, *Is temporal difference learning optimal? An instance-dependent analysis*, SIAM Journal on Mathematics of Data Science, 3 (2021), pp. 1013–1040.
[16] K. Khamaru, E. Xia, M. J. Wainwright, and M. I. Jordan, *Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning*, arXiv preprint arXiv:2106.14352, (2021).
[17] J. Kober, J. A. Bagnell, and J. Peters, *Reinforcement learning in robotics: A survey*, The International Journal of Robotics Research, 32 (2013), pp. 1238–1274.
[18] 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, PMLR, 2015, pp. 626–634.
[19] G. Kotsalis, G. Lan, and T. Li, *Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation*, arXiv preprint arXiv:2011.02987, (2020).
[20] G. Kotsalis, G. Lan, and T. Li, *Simple and optimal methods for stochastic variational inequalities, II: Markovian noise and policy evaluation in reinforcement learning*, arXiv preprint arXiv:2011.08434, (2020).
[21] C. Lakshminarayanan and C. Szepesvári, *Linear stochastic approximation: How far does constant step-size and iterate averaging go?*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1347–1355.
[22] L. Le Cam, *Limits of experiments*, in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics, University of California Press, 1972, pp. 245–282.
[23] L. Le Cam and G. L. Yang, *Asymptotics in statistics: Some basic concepts*, Springer Science & Business Media, 2000.
[24] G. Li, C. Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, *Is q-learning minimax optimal? a tight sample complexity analysis*, arXiv preprint arXiv:2102.06548, (2021).
[25] 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*, in Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 12861–12872.
[26] Y. Min, T. Wang, D. Zhou, and Q. Gu, *Variance-aware off-policy evaluation with linear function approximation*, Advances in neural information processing systems, 34 (2021), pp. 7598–7610.
[27] W. Mou, A. Pananjady, and M. J. Wainwright, *Optimal oracle inequalities for solving projected fixed-point equations*, arXiv preprint arXiv:2012.05299, (2020).
[28] W. Mou, A. Pananjady, M. J. Wainwright, and P. L. Bartlett, *Optimal and instance-dependent guarantees for Markovian linear stochastic approximation*, arXiv preprint arXiv:2112.12770, (2021).
[29] A. S. Nemirovsky, *On optimality of Krylov's information when solving linear operator equations*, Journal of Complexity, 7 (1991), pp. 121–130.
[30] A. S. Nemirovsky, *Information-based complexity of linear operator equations*, Journal of Complexity, 8 (1992), pp. 153–175.
[31] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2003.
[32] Y. Ouyang and Y. Xu, *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*, Mathematical Programming, 185 (2021), pp. 1–35.
[33] 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.
[34] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli, *Stochastic variance-reduced policy gradient*, in International conference on machine learning, PMLR, 2018, pp. 4026–4035.
[35] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM journal on control and optimization, 30 (1992), pp. 838–855.
[36] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014.
[37] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*,

Mathematical Programming, 162 (2017), pp. 83–112.
[38] A. Sidford, M. Wang, X. Wu, L. F. Yang, and Y. Ye, *Near-optimal time and sample complexities for solving Markov decision processes with a generative model*, in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 5192–5202.
[39] R. Srikant and L. Ying, *Finite-time error bounds for linear stochastic approximation and TD learning*, in Conference on Learning Theory, PMLR, 2019, pp. 2803–2830.
[40] R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine learning, 3 (1988), pp. 9–44.
[41] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, *Fast gradient-descent methods for temporal-difference learning with linear function approximation*, in International Conference on Machine Learning, 2009, pp. 993–1000.
[42] 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.
[43] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE transactions on automatic control, 42 (1997), pp. 674–690.
[44] 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, vol. 32, 2019, pp. 5784–5795.
[45] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong, *Multi-agent reinforcement learning via double averaging primal-dual optimization*, Advances in Neural Information Processing Systems, 31 (2018).
[46] M. J. Wainwright, *Variance-reduced Q-learning is minimax optimal*, arXiv preprint arXiv:1906.04697, (2019).
[47] R. Wang, D. P. Foster, and S. M. Kakade, *What are the statistical limits of offline rl with linear function approximation?*, arXiv preprint arXiv:2010.11895, (2020).
[48] G. Wolfer and A. Kontorovich, *Estimating the mixing time of ergodic Markov chains*, in Conference on Learning Theory, PMLR, 2019, pp. 3120–3159.
[49] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM Journal on Optimization, 24 (2014), pp. 2057–2075.
[50] T. Xu, Z. Wang, Y. Zhou, and Y. Liang, *Reanalysis of variance reduced temporal difference learning*, arXiv preprint arXiv:2001.01898, (2020).
[51] A. Zanette, *Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl*, in International Conference on Machine Learning, PMLR, 2021, pp. 12287–12297.