Optimal and instance-dependent oracle inequalities for policy evaluation

Wenlong Mou [1] **Ashwin Pananjady** [2 3] **Martin J. Wainwright** [1 4]

Abstract

Linear fixed point equations in Hilbert spaces naturally arise from the policy evaluation problem in reinforcement learning. We study methods that use a collection of random observations to compute approximate solutions by searching over a known low-dimensional subspace of the Hilbert space. First, we prove an instance-dependent upper bound on the mean-squared error for a linear stochastic approximation scheme that exploits Polyak–Ruppert averaging. This bound consists of two terms: an approximation error term with an instance-dependent approximation factor, and a statistical error term that captures the instance-specific complexity of the noise when projected onto the low-dimensional subspace. Using information-theoretic methods, we also establish lower bounds showing that the approximation factor cannot be improved, again in an instance-dependent sense. A concrete consequence of our characterization is that the optimal approximation factor in this problem can be much larger than a universal constant. We show how our results precisely characterize the error of a class of temporal difference learning methods for the policy evaluation problem with linear function approximation, establishing their optimality.

1. Introduction

Estimating the value function of a Markov reward process (MRP) from data is a fundamental task in reinforcement learning (RL) and approximate dynamic programming (ADP). The estimation problem involves solving the Bellman equation, and when the underlying MRP is time-homogeneous, this problem becomes that of solving a *linear fixed-point equation* in a given Hilbert space $\mathbb{X}$:

$$v = Lv + b, \tag{1}$$

For example, the Bellman equation in a discounted MRP is an instantiation of Eq (1) with $L = \gamma P$ and $b = r$, where γ is the discount factor, P is the Markov transition kernel, and r is the reward function. The Hilbert space $\mathbb{X}$ is usually chosen to be $\mathbb{L}^2(\mathcal{S}, \xi)$, where $\mathcal{S}$ is the state space of the MRP, and ξ is a suitably chosen probability measure on $\mathcal{S}$. See the discussion and examples in the following section for a more detailed formulation of this problem.

For most practical problems, the cardinality of the state space $D := |\mathcal{S}|$ is very large, or even infinite. This poses statistical and computational hurdles to learning the value function. The amount of data available in practice is usually far from sufficient to estimate the transition kernel and reward accurately. This motivates the method of function approximation, the workhorse of modern RL algorithms. Value learning methods with function approximation attempt to approximate the solution to the Bellman equation (1) using a given class of functions in the Hilbert space $\mathbb{X}$.

A commonly-used value learning method is *linear function approximation*. This method chooses a subspace $\mathbb{S}$ of the Hilbert space, of dimension $d \ll D$, and searches for solutions within this subspace. This paper treats the general problem in which n observations $\{(L_i, b_i)\}_{i=1}^n$ are drawn i.i.d. from some distribution with mean (L, b). Letting v^* denote the solution to the fixed point equation (1), our goal is to use these observations in order to produce an estimate $\widehat{v}_n$ of v^* that satisfies an *oracle inequality* of the form

$$\mathbb{E}\|\widehat{v}_n - v^*\|^2 \leq \alpha \cdot \inf_{v \in \mathbb{S}} \|v - v^*\|^2 + \varepsilon_n. \tag{2}$$

Here we use $\|\cdot\|$ to denote the Hilbert norm associated with $\mathbb{X}$. The three terms appearing on the RHS of inequality (2) all have concrete interpretations. The term

$$\mathcal{A}(\mathbb{S}, v^*) := \inf_{v \in \mathbb{S}} \|v - v^*\|^2 \tag{3}$$

defines the *approximation error*; this is the error incurred by an oracle procedure that knows the fixed point v^* in advance and aims to output the best approximation to v^* within the subspace $\mathbb{S}$. The term α is the *approximation factor*, which indicates how poorly the estimator $\widehat{v}_n$ performs at carrying

[1]Department of EECS, University of California, Berkeley [2]School of ECE, Georgia Institute of Technology [3]School of Industrial and Systems Engineering, Georgia Institute of Technology [4]Department of Statistics, University of California, Berkeley. Correspondence to: Wenlong Mou <wmou@eecs.berkeley.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

out the aforementioned approximation; note that $\alpha \geq 1$ by definition, and it is most desirable for α to be as small as possible. The final term ε_n is a proxy for the *statistical error* incurred due to our stochastic observation model; indeed, one expects that as the sample size n goes to infinity, this error should tend to zero for any reasonable estimator, indicating consistent estimation when $v^* \in \mathbb{S}$. More generally, we would like our estimator to also have as small a statistical error as possible in terms of the other parameters that define the problem instance. In an ideal world, we would like both desiderata to hold simultaneously: the approximation factor should be as close to one as possible while the statistical error stays as small as possible. Indeed, such a "best-of-both-worlds" guarantee can indeed be obtained in many canonical statistical problems, and "sharp" oracle inequalities—meaning ones in which the approximation factor is equal to 1—are known (Rakhlin et al., 2017; Dalalyan & Salmon, 2012; Tsybakov, 2004)

A standard approach in value learning with linear function approximation is the family of temporal difference (TD) methods (Sutton, 1988; Boyan, 2002). Borrowing ideas from Galerkin approximation in differential equations (Galerkin, 1915), the method uses the solution to the projected linear equation as a proxy to the original one. In particular, letting $\Pi_{\mathbb{S}}$ denote the orthogonal projection onto this subspace, the method seeks (approximate) solutions to the *projected fixed point equation*

$$v = \Pi_{\mathbb{S}}\big(Lv + b\big). \tag{4}$$

The projected equation can be solved in time that depends only on the dimension of the subspace d, and when $v^* \in \mathbb{S}$, the statistical error $\epsilon_n = O(d/n)$ can be achieved via stochastic approximation procedures (Bhandari et al., 2018). It is also known that the projected fixed-point approach enjoys a bounded approximation factor under natural assumptions. In particular, Tsitsiklis & Van Roy (1997) show that if the operator L is $\gamma_{\max}$-contractive in the norm $\|\cdot\|$, then the (deterministic) solution $\bar{v}$ to the projected fixed point equation (4) satisfies the bound

$$\|\bar{v} - v^*\|^2 \leq \frac{1}{1-\gamma_{\max}^2} \inf_{v \in \mathbb{S}} \|v - v^*\|^2. \tag{5}$$

The finite nature of the approximation factor notwithstanding, it should be noted that the bound (5) has a potentially large approximation factor that can be quite far from unity (as would be the case for a "sharp" oracle inequality). Indeed, and as will be discussed shortly, the approximation factor for discounted MRPs grows with the effective *horizon* of the problem, which can often be quite large.

The primary motivating question for our work is whether or not this bound can be improved, and if so, to what extent. Our work is also driven by the complementary question of whether a sharp bound can be obtained on the statistical error of an estimator that, unlike $\bar{v}$, has access only to the samples $\{(L_i, b_i)\}_{i=1}^n$. In particular, we would like the statistical error ε_n to depend on some notion of complexity of our specific instance.

2. Problem Setup

Let us fix some orthogonal basis $\{\phi_j\}_{j \geq 1}$ of the full space $\mathbb{X}$ such that $\mathbb{S} = \text{span}\{\phi_1, \ldots, \phi_d\}$. In terms of this basis, we can define the projection operator $\Phi_d : \mathbb{X} \to \mathbb{R}^d$ via $\Phi_d(x) := \big(\langle x, \phi_j \rangle\big)_{j=1}^d$. Using these operators, we can define the *projected operator* associated with L—namely

$$M := \Phi_d L \Phi_d^*. \tag{6}$$

Note that M is simply a d-dimensional matrix, one which describes the action of L on $\mathbb{S}$ according to the basis that we have chosen. As we will see in the main theorems, our results do not depend on the specific choice of the orthonormal basis, but it is convenient to use a given one, as we have done here. Consider the quantity $\kappa(M) := \frac{1}{2}\lambda_{\max}\big(M + M^\top\big)$, corresponding to the maximal eigenvalue of the symmetrized version of M. We assume $\kappa(M)$ throughout, which guarantees the existence and uniquess of the solution to Eq (4).

Stochastic observation model: As noted in the introduction, this paper focuses on an observation model in which we observe i.i.d. random pairs (L_i, b_i) for $i = 1, \ldots, n$ that are unbiased estimates of the pair (L, b) so that

$$\mathbb{E}[L_i] = L, \quad \text{and} \quad \mathbb{E}[b_i] = b. \tag{7}$$

In addition to this unbiasedness, we also assume that our observations satisfy a certain second-moment bound.

Assumption 1 (Second-moment bound in projected space) *There exist scalars $\sigma_L, \sigma_b > 0$ such that for any unit-norm vector $u \in \mathbb{S}$ and any basis vector in $\{\phi_j\}_{j=1}^d$ we have the bounds*

$$\mathbb{E}\langle \phi_j, (L_i - L)u \rangle^2 \leq \sigma_L^2 \|u\|^2, \quad \textit{and} \tag{8a}$$

$$\mathbb{E}\langle \phi_j, b_i - b \rangle^2 \leq \sigma_b^2. \tag{8b}$$

In words, Assumption 1 guarantees that the random variable obtained by projecting the "noise" onto any of the basis functions $\phi_1, \ldots, \phi_d$ in the subspace $\mathbb{S}$ has bounded second moment. In Section 3.2, we show that this assumption is satisfied under mild conditions for MRP and associated feature vectors.

2.1. Policy evaluation in discounted MRPs

Now we introduce the problem of estimating the long-term, discounted value function in a Markov reward process. We

will introduce basic setup and notations, and show that LSTD method is a special case of the projected fixed point equation (4).

Consider a Markov chain on a state space $\mathcal{S}$ and an ergodic transition kernel $P : \mathcal{S}\times\mathcal{S} \to \mathbb{R}$, with stationary distribution ξ. The associated (discounted) Markov reward process is given by introducing a reward function $r : \mathcal{S} \to \mathbb{R}$, and discount factor $\gamma \in (0, 1)$. The goal of the policy evaluation problem to estimate the value function, which is given by the solution to the Bellman equation $v^* = \gamma P v^* + r$. We define $\mathbb{X}$ to be the Hilbert space $\mathbb{L}^2(\mathcal{S}, \xi)$.

We consider the i.i.d. observation model in this paper. For each $i = 1, 2, \cdots, n$, suppose that we observe an independent tuple $(s_i, s_i^+, R_i(s_i))$, such that

$$s_i \sim \xi,\ s_i^+ \sim P(s_i, \cdot),\ \text{and}\ \mathbb{E}[R_i(s_i)|s_i] = r(s_i). \quad (9)$$

The i-th observation (L_i, b_i) is then obtained by plugging in these observations to compute unbiased estimates of P and r, respectively.

In the setting with linear function approximation, we search for the solution to the Bellman equation within a linear subspace spanned by a number of given basis functions. In particular, consider a set $\{\psi_1, \psi_2, \cdots, \psi_d\}$ of basis functions in $\mathbb{X}$, and suppose that they are linearly independent on the support of ξ. We are interested in projections onto the subspace $\mathbb{S} = \text{span}(\psi_1, \ldots, \psi_d)$, and in solving the population-level projected fixed point equation (4), which takes the form$\bar{v} = \Pi_{\mathbb{S}}(\gamma P\bar{v} + r)$. By writing $\bar{v}(s) = \psi(s)^\top\bar{\vartheta}$ for a vector of coefficients $\bar{\vartheta} \in \mathbb{R}^d$, the projected fixed-point equation can be equivalently written in terms of the coefficient vector $\bar{\vartheta}$ as

$$\mathbb{E}_{s\sim\xi}[\psi(s)\psi(s)^\top]\bar{\vartheta} = \gamma\mathbb{E}_{s\sim\xi}\left[\mathbb{E}_{s^+\sim P(s,\cdot)}[\psi(s)\psi(s^+)^\top]\right]\bar{\vartheta} + \mathbb{E}_{s\sim\xi}[r(s)\psi(s)]. \quad (10)$$

Equation (10) is the population relation underlying the canonical *least squares temporal difference* (LSTD) learning method (Bradtke & Barto, 1996; Boyan, 2002).

3. Main results and their consequences

3.1. Upper bounds

In this section, we describe a standard stochastic approximation scheme for the problem based on combining ordinary stochastic updates with Polyak–Ruppert averaging (Polyak, 1990; Polyak & Juditsky, 1992; Ruppert, 1988). In particular, given an oracle that provides observations (L_i, b_i), consider the stochastic recursion parameterized by a positive stepsize η:

$$v_{t+1} = (1-\eta)v_t + \eta\Pi_{\mathbb{S}}\big(L_{t+1}v_t + b_{t+1}\big), \quad \text{for } t = 1, 2, \ldots. \quad (11a)$$

For a given sample size $n \geq 2$, our final estimate $\widehat{v}_n$ is given by taking the average of these iterates from time n_0 to n—that is

$$\widehat{v}_n := \frac{1}{n-n_0}\sum_{t=n_0+1}^{n} v_t. \quad (11b)$$

Here the "burn-in" time n_0 is an integer parameter to be specified.

The stochastic approximation procedure (11) is defined in the entire space $\mathbb{X}$; note that it can be equivalently written as iterates in the projected space $\mathbb{R}^d$, via the recursion

$$\vartheta_{t+1} = (1-\eta)\vartheta_t + \eta(\Phi_d L_{t+1}\Phi_d^*\vartheta_t + \Phi_d b_{t+1}). \quad (12)$$

The original iterates can be recovered by applying the adjoint operator—that is, $v_t = \Phi_d^*\vartheta_t$ for $t = 1, 2, \ldots$.

Having introduced the algorithm itself, we are now ready to provide a guarantee on its error. Two matrices play a key role in the statement of our upper bound. The first is the d-dimensional matrix $M := \Phi_d L\Phi_d^*$. We show that the mean-squared error is upper bounded by the approximation error $\inf_{v\in\mathbb{S}}\|v - v^*\|^2$ along with a pre-factor of the form

$$\alpha(M, s) = 1 + \lambda_{\max}\Big((I-M)^{-1}(s^2\, I_d - MM^T)(I-M)^{-T}\Big),$$

for $s = \|L\|_{\text{op}}$. Our bounds also involve the quantity $\kappa(M) = \frac{1}{2}\lambda_{\max}\big(M + M^T\big)$, which we abbreviate by κ when the underlying matrix M is clear from the context.

The second matrix is a covariance matrix, capturing the noise structure of our observations, given by

$$\Sigma^* := \text{cov}\left(\Phi_d(b_1 - b)\right) + \text{cov}\left(\Phi_d(L_1 - L)\bar{v}\right).$$

This matrix, along with the constants (σ_L, σ_b) from Assumption 1, arise in the definition of two additional error terms, namely

$$\mathcal{E}_n(M, \Sigma^*) := \frac{\text{trace}\left((I-M)^{-1}\Sigma^*(I-M)^{-\top}\right)}{n},$$

$$\mathcal{H}_n(\sigma_L, \sigma_b, \bar{v}) := \frac{\sigma_L}{(1-\kappa)^3}\left(\frac{d}{n}\right)^{\frac{3}{2}}\left(\|\bar{v}\|^2\sigma_L^2 + \sigma_b^2\right).$$

As suggested by our notation, the error $\mathcal{H}_n(\sigma_L, \sigma_b, \bar{v})$ is a higher-order term, decaying as $n^{-3/2}$ in the sample size, whereas the quantity $\mathcal{E}_n(M, \Sigma^*)$ is the dominant source of statistical error. With this notation, we have the following:

Theorem 1 *Suppose that we are given n i.i.d. observations $\{(L_i, b_i)\}_{i=1}^n$ that satisfy the noise conditions in Assumption 1. Then there are universal constants (c_0, c) such that for any sample size $n \geq \frac{c_0\sigma_L^2 d}{(1-\kappa)^2}\log^2\left(\frac{\|v_0-\bar{v}\|^2 d}{1-\kappa}\right)$, then running the algorithm (11) with*

$$\text{stepsize } \eta = \frac{1}{c_0\sigma_L\sqrt{dn}}, \quad \text{and burn-in period } n_0 = n/2$$

yields an estimate $\widehat{v}_n$ such that

$$\mathbb{E}\|\widehat{v}_n - v^*\|^2 \leq (1+\omega)\cdot\alpha(M, \|L\|_{\mathbb{X}}) \inf_{v\in\mathbb{S}} \|v - v^*\|^2 + c\left(1+\tfrac{1}{\omega}\right)\cdot\left\{\mathcal{E}_n(M,\Sigma^*) + \mathcal{H}_n(\sigma_L, \sigma_b, \bar{v})\right\}, \tag{13}$$

valid for any $\omega > 0$.

The leading statistical error term $\mathcal{E}_n(M, \Sigma^*)$ matches the Cramér-Rao lower bound, and is known to be asymptotically optimal. A more in-depth discussion of the approximation factor $\alpha(M, \|L\|_{\mathbb{X}})$, including comparison to prior works and useful upper bounds in particular settings, can be found in Appendix B. A useful upper bound is that $\alpha(M, s) \leq \frac{2}{1-\kappa(M)}$ if $s \leq 1$.

3.2. Consequences to value function estimation

Now we turn to the consequence of the general oracle inequality in Theorem 1 to the specific model of value function estimation in Section 2.1. Recall the i.i.d. observation model (9). Also recall the equivalent form of the projected fixed point equation (10), and note that the population-level operator L satisfies the norm bound

$$\|L\|_{\mathbb{X}} = \gamma \cdot \sup_{\|v\|\leq 1} \|Pv\| \leq \gamma.$$

since ξ is the stationary measure of the transition kernel P.

Under the setup of Section 2.1, the temporal difference method is a variant of SA iterates (11a). The proof of Theorem 1 can be then applied to this case. To state the result, we define the matrix $B \in \mathbb{R}^{d\times d}$ by $B_{ij} := \langle \psi_i, \psi_j\rangle$ for $i, j \in [d]$, and define the following quantities:

$$\begin{aligned} M &:= \gamma B^{-1/2}\mathbb{E}_\xi[\psi(s)\psi(s^+)^\top]B^{-1/2}, \\ \Sigma_L &:= \mathrm{cov}_\xi\left[B^{-1/2}\psi(s)\left(\psi(s) - \gamma\psi(s^+)\right)^\top \bar{\vartheta}\right], \\ \Sigma_b &:= \mathrm{cov}_\xi\left[R(s)B^{-1/2}\psi(s)\right]. \end{aligned}$$

It can be shown that the averaged TD iterates satisfy the bound (13) with an approximation factor $\alpha(M, \gamma)$, a statistical error term $\mathcal{E}_n(M, \Sigma_L + \Sigma_b)$, and a high-order term depending on suitable moment assumptions and the condition number of B. See Appendix D for a complete statement and the proof of this corollary.

In the worst case, the approximation factor $\alpha(M, \gamma)$ scales as $\frac{1}{1-\gamma^2}$, recovering the classical result (5), but more generally gives a more fine-grained characterization of the approximation factor depending on the one-step auto-covariance matrix for the feature vectors.

3.3. Minimax lower bound on the approximation factor

Now we turn to the minimax lower bound for the value function estimation problem in MRPs, under an i.i.d. observation model from the stationary distribution.

To set the stage, we say that a Markov reward process (P, γ, r) and associated basis functions $\{\psi_j\}_{j=1}^d$ are in the *canonical set-up* if the following conditions hold:

- The stationary distribution ξ of P exists and is unique.
- The reward function and its observations are uniformly bounded. In particular, we have $\|r\|_\infty \leq 1$, and $\|R\|_\infty \leq 1$ almost surely.
- The basis functions are orthonormal, i.e., $\mathbb{E}_\xi[\psi(s)\psi(s)^\top] = I_d$.

The three conditions are standard assumptions in Markov reward processes.

Now given scalars $\nu \in (0, 1]$ and $\gamma \in (0, 1)$, integer $D > 0$ and scalar $\delta \in (0, 1/2)$, we consider the following class of MRPs and associated feature vectors:

$$\mathbb{C}_{\mathsf{MRP}}(\nu, \gamma, D, \delta) := \left\{ (P, \gamma, r, \psi) \;\middle|\; \begin{array}{c} (P, \gamma, r, \psi) \text{ is in the canonical setup,} \\ |\mathcal{S}| = D, \quad \mathcal{A}(\mathbb{S}, v^*) \leq \delta^2, \\ \kappa\left(\mathbb{E}_\xi[\psi(s)\psi(s^+)^\top]\right) \leq \nu. \end{array} \right\}.$$

Note that under the canonical set-up, we have $M = \gamma\mathbb{E}_\xi[\psi(s)\psi(s^+)^\top]$, and consequently, a problem instance in the class $\mathbb{C}_{\mathsf{MRP}}(\nu, \gamma, D, \delta)$ satisfies $\kappa(M) \leq \nu\gamma$ in the set-up of previous section. The condition $\kappa\left(\mathbb{E}_\xi[\psi(s)\psi(s^+)^\top]\right) \leq \nu$ can be seen as a "mixing" condition in the projected space: when ν is bounded away from 1, the feature vector cannot have too large a correlation with its next-step transition in any direction.

We have the following minimax lower bound for this class, where we use the shorthand $\mathbb{C}_{\mathsf{MRP}} \equiv \mathbb{C}_{\mathsf{MRP}}(\nu, \gamma, D, \delta)$ for convenience.

Theorem 2 *There are universal positive constants (c, c_1) such that if $D \geq c_1(n^2 + d)$, then for all scalars $\nu \in (0, 1]$ and $\gamma \in (0, 1)$, we have*

$$\inf_{\widehat{v}_n \in \widehat{\mathcal{V}}_{\mathbb{X}}} \sup_{(P,\gamma,r,\psi)\in\mathbb{C}_{\mathsf{MRP}}} \|\widehat{v}_n - v^*\|^2 \geq \frac{c}{1-\nu\gamma}\delta^2 \wedge 1. \tag{14}$$

In conjunction with the results from previous section, we can conclude that the TD algorithm for policy evaluation with linear function approximation attains the minimax-optimal approximation factor over the class $\mathbb{C}_{\mathsf{MRP}}$, up to universal constants. It is also worth noting that Theorem 2 also shows that the worst-case upper bound (5) due to Tsitsiklis & Van Roy (1997) is indeed sharp up to a universal constant.

Observe that Theorem 2 requires the state space size D to be larger than n^2. As mentioned in the introduction, we should not expect any non-trivial approximation factor when $n \geq D$, but this leaves open the regime $n \ll D \ll n^2$. Finally, it is worth noticing that Theorem 2 holds true only for the i.i.d. observation models. If we are given the entire trajectory of the Markov reward process, the approximation factor can be made arbitrarily close to 1, using TD(λ) methods (Tsitsiklis & Van Roy, 1997). The trade-off inherent to the Markov observation model is left for our companion paper.

References

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.

Bellec, P. C. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*, volume 22. Springer Science & Business Media, 2012.

Bertsekas, D. P. Temporal difference methods for general projected equations. *IEEE Transactions on Automatic Control*, 56(9):2128–2139, 2011.

Bertsekas, D. P. Proximal algorithms and temporal differences for large linear systems: extrapolation, approximation, and simulation. *arXiv preprint arXiv:1610.05427*, 2016.

Bertsekas, D. P. *Reinforcement Learning and Optimal Control*. Athena Scientific Belmont, MA, 2019.

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

Borkar, V. S. *Stochastic Approximation: a Dynamical Systems Viewpoint*, volume 48. Springer, 2009.

Bousquet, O., Kane, D., and Moran, S. The optimal approximation factor in density estimation. In *Conference on Learning Theory*, pp. 318–341, 2019.

Boyan, J. A. Technical update: Least-squares temporal difference learning. *Machine learning*, 49(2-3):233–246, 2002.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.

Brenner, S. and Scott, R. *The Mathematical Theory of Finite Element Methods*, volume 15. Springer Science & Business Media, 2007.

Bunea, F., Tsybakov, A., and Wegkamp, M. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007a.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation for Gaussian regression. *The Annals of Statistics*, 35 (4):1674–1697, 2007b.

Céa, J. Approximation variationnelle des problèmes aux limites. In *Annales de l'institut Fourier*, volume 14, pp. 345–444, 1964.

Chan, S. O., Diakonikolas, I., Servedio, R. A., and Sun, X. Near-optimal density estimation in near-linear time using variable-width histograms. In *Advances in Neural Information Processing Systems*, pp. 1844–1852, 2014.

Dalalyan, A. S. and Salmon, J. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

Dalalyan, A. S. and Sebbar, M. Optimal Kullback–Leibler aggregation in mixture density estimation by maximum likelihood. *Mathematical Statistics and Learning*, 1(1): 1–35, 2018.

Duchi, J. and Ruan, F. Asymptotic optimality in stochastic optimization. *arXiv preprint arXiv:1612.05612*, 2016.

Fletcher, C. A. Computational Galerkin methods. In *Computational galerkin methods*, pp. 72–85. Springer, 1984.

Galerkin, B. G. Series solution of some problems of elastic equilibrium of rods and plates. *Vestnik inzhenerov i tekhnikov*, 19(7):897–908, 1915.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. Is temporal difference learning optimal? An instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

Klopp, O., Tsybakov, A. B., and Verzelen, N. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Koltchinskii, V. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pp. 1008–1014, 2000.

Lai, T. L. Stochastic approximation. *The Annals of Statistics*, 31(2):391–406, 2003.

Lakshminarayanan, C. and Szepesvári, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355, 2018.

Li, C. J., Mou, W., Wainwright, M. J., and Jordan, M. I. ROOT-SGD: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. *arXiv preprint arXiv:2008.12690*, 2020.

Massart, P. *Concentration Inequalities and Model Selection*, volume 6. Springer, 2007.

Massart, P. and Nédélec, É. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 2947–2997, 2020.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (May):815–857, 2008.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Pananjady, A. and Wainwright, M. J. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67 (1):566–585, 2021.

Polyak, B. T. A new method of stochastic approximation type. *Automat. i Telemekh*, 7(98-107):2, 1990.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Polydorides, N., Wang, M., and Bertsekas, D. P. Approximate solution of large-scale linear inverse problems with Monte Carlo simulation. *Lab. for Information and Decision Systems Report, MIT*, 2009.

Polydorides, N., Wang, M., and Bertsekas, D. P. A quasi Monte Carlo method for large-scale inverse problems. In *Monte Carlo and Quasi-Monte Carlo Methods 2010*, pp. 623–637. Springer, 2012.

Rakhlin, A., Sridharan, K., and Tsybakov, A. B. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23 (2):789–824, 2017.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Rummery, G. A. and Niranjan, M. On-line Q-learning using connectionist systems. Technical report, Cambridge University Engineering Department, 1994.

Ruppert, D. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Scherrer, B. Should one compute the temporal difference fix point or minimize the Bellman residual? The unified oblique projection view. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 959–966, 2010.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 993–1000, 2009.

Szepesvári, C. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2010.

Tsitsiklis, J. N. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1075–1081, 1997.

Tsitsiklis, J. N. and Van Roy, B. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, 1999.

Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Van Roy, B. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006.

Wainwright, M. J. Stochastic approximation with cone-contractive operators: Sharper ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019a.

Wainwright, M. J. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019b.

Watkins, C. J. and Dayan, P. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Yatracos, Y. G. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, pp. 768–774, 1985.

Yu, H. and Bertsekas, D. P. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329, 2010.

Zhu, B., Jiao, J., and Tse, D. Deconstructing generative adversarial networks. *IEEE Transactions on Information Theory*, 2020.

Appendix

A. Additional related works

Our paper touches on various lines of related work, including oracle inequalities for statistical estimation, stochastic approximation and its application to reinforcement learning, and projected linear equation methods. We provide a brief discussion of these connections here.

Oracle inequalities: There is a large literature on misspecified statistical models and oracle inequalities (e.g., see the monographs (Massart, 2007; Koltchinskii, 2011) for overviews). Oracle inequalities in the context of penalized empirical risk minimization (ERM) are quite well-understood (e.g., (Bartlett et al., 2005; Koltchinskii, 2006; Massart & Nédélec, 2006)). Typically, the resulting approximation factor is exactly 1 or arbitrarily close to 1, and the statistical error term depends on the localized Rademacher complexity or metric entropy of this function class. Aggregation methods have been developed in order to obtain *sharp* oracle inequalities with approximation factor exactly 1 (e.g. (Tsybakov, 2004; Bunea et al., 2007b; Dalalyan & Salmon, 2012; Rakhlin et al., 2017)). Sharp oracle inequalities are now available in a variety of settings including for sparse linear models (Bunea et al., 2007a), density estimation (Dalalyan & Sebbar, 2018), graphon estimation (Klopp et al., 2017), and shape-constrained estimation (Bellec, 2018). As previously noted, our setting differs qualitatively from the ERM setting, in that as shown in this paper, sharp oracle inequalities are no longer possible. There is another related line of work on oracle inequalities of density estimation. Yatracos (1985) showed an oracle inequality with the non-standard approximation factor 3, and with a statistical error term depending on the metric entropy. This non-unit approximation factor was later shown to be optimal for the class of one-dimensional piecewise constant densities (Chan et al., 2014; Bousquet et al., 2019; Zhu et al., 2020). The approximation factor lower bound in these papers and our work both make use of the birthday paradox to establish information-theoretic lower bounds.

Stochastic approximation: Stochastic approximation algorithms for linear and nonlinear fixed-point equations have played a central role in large-scale machine learning and statistics (Robbins & Monro, 1951; Lai, 2003; Nemirovski et al., 2009). See the books (Benveniste et al., 2012; Borkar, 2009) for a comprehensive survey of the classical methods of analysis. The seminal works by Polyak (1990); Polyak & Juditsky (1992); Ruppert (1988) propose taking the average of the stochastic approximation iterates, which stabilizes the algorithm and achieves a Gaussian limiting distribution. This asymptotic result is also known to achieve the local asymptotic minimax lower bound (Duchi & Ruan, 2016). Non-asymptotic guarantees matching this asymptotic behavior have also been established for stochastic approximation algorithms and their variance-reduced variants (Moulines & Bach, 2011; Khamaru et al., 2020; Mou et al., 2020; Li et al., 2020).

Stochastic approximation is also a fundamental building block for reinforcement learning algorithms, wherein the method is used to produce an iterative, online solution to the Bellman equation from data; see the books (Szepesvári, 2010; Bertsekas, 2019) for a survey. Such approaches include temporal difference (TD) methods (Sutton, 1988) for the policy evaluation problem and the Q-learning algorithm (Watkins & Dayan, 1992) for policy optimization. Variants of these algorithms also abound, including LSTD (Boyan, 2002), SARSA (Rummery & Niranjan, 1994), actor-critic algorithms (Konda & Tsitsiklis, 2000), and gradient TD methods (Sutton et al., 2009). The analysis of these methods has received significant attention in the literature, ranging from asymptotic guarantees (e.g., (Bradtke & Barto, 1996; Tsitsiklis & Van Roy, 1997; 1999)) to more fine-grained finite-sample bounds (e.g., (Bhandari et al., 2018; Srikant & Ying, 2019; Lakshminarayanan & Szepesvári, 2018; Pananjady & Wainwright, 2021; Wainwright, 2019a;b)). Our work contributes to this literature by establishing finite-sample upper bounds for temporal difference methods with Polyak–Ruppert averaging, as applied to the policy evaluation problem with linear function approximation.

Projected methods for linear equations: Galerkin (1915) first proposed the method of approximating the solution to a linear PDE by solving the projected equation in a finite-dimensional subspace. This method later became a cornerstone of finite-element methods in numerical methods for PDEs; see the books (Fletcher, 1984; Brenner & Scott, 2007) for a comprehensive survey. A fundamental tool used in the analysis of Galerkin methods is Céa's lemma (Céa, 1964), which corresponds to a special case of the approximation factor upper bounds that we establish. As mentioned before, projected linear equations were also considered independently by Tsitsiklis & Van Roy (1997) in the context of reinforcement learning; they established the worst-case upper bound (5) on the approximation factor under contractivity assumptions. These contraction-based bounds were further extended to the analysis of Q-learning in optimal stopping problems (Tsitsiklis & Van Roy, 1999). The connection between the Galerkin method and TD methods was discovered by Yu & Bertsekas (2010); Bertsekas (2011), and the former paper shows an instance-dependent upper bound on the approximation factor. This analysis

was later applied to Monte–Carlo methods for solving linear inverse problems (Polydorides et al., 2009; 2012).

We note that the Bellman equation can be written in infinitely many equivalent ways—by using powers of the transition kernel and via the formalism of resolvents—leading to a continuous family of projected equations indexed by a scalar parameter λ (see, e.g., Section 5.5 of Bertsekas (2019)). Some of these forms can be specifically leveraged in other observation models; for instance, by observing the trajectory of the Markov chain instead of i.i.d. samples, it becomes possible to obtain unbiased observations for integer powers of the transition kernel. This makes it possible to efficiently estimate the solution to the projected linear equation for various values of λ, and underlies the family of TD(λ) methods (Sutton, 1988; Boyan, 2002). Indeed, Tsitsiklis & Van Roy (1997) also showed that the worst-case approximation factor in equation (5) can be improved by using larger values of λ. Based on this observation, a line of work has studied the trade-off between approximation error and estimation measure in model selection for reinforcement learning problems (Bertsekas, 2016; Scherrer, 2010; Munos & Szepesvári, 2008; Van Roy, 2006). However, unlike this body of work, our focus in the current paper is on studying the i.i.d. observation model; we postpone a detailed investigation of the Markov setting to a companion paper.

B. Detailed discussion of the approximation error

As mentioned in the introduction, upper bounds on the approximation factor have received significant attention in the literature, and it is interesting to compare our bounds.

B.1. Past results

In the case where $\gamma_{\max} := \|L\|_{\mathbb{X}} < 1$, the approximation-factor bound (5) was established by Tsitsiklis & Van Roy (1997), via the following argument. Letting $\widetilde{v} := \Pi_{\mathbb{S}}(Lv^* + b)$, we have

$$\begin{aligned}\|\bar{v} - v^*\|^2 \overset{(i)}{=} \|\bar{v} - \widetilde{v}\|^2 + \|\widetilde{v} - v^*\|^2 &= \|\Pi_{\mathbb{S}}(L\bar{v} + b) - \Pi_{\mathbb{S}}(Lv^* + b)\|^2 + \|\widetilde{v} - v^*\|^2 \\ &\overset{(ii)}{\leq} \|L\bar{v} - Lv^*\|^2 + \|\widetilde{v} - v^*\|^2 \\ &\overset{(iii)}{\leq} \gamma_{\max}^2 \|\bar{v} - v^*\|^2 + \|\widetilde{v} - v^*\|^2. \end{aligned} \tag{15}$$

Step (i) uses Pythagorean theorem; step (ii) follows from the non-expansiveness of the projection operator; and step (iii) makes use of the contraction property of the operator L. Note that by definition, we have $\alpha(M, \|L\|_{\mathbb{X}}) \leq (1 - \|L\|_{\mathbb{X}})^{-2}$, and so the approximation factor in Theorem 1 recovers the bound (5) in the worst case. In general, however, the factor $\alpha(M, \|L\|_{\mathbb{X}})$ can be significantly smaller.

Yu & Bertsekas (2010) derived two fine-grained approximation factor upper; in terms of our notation, their bounds take the form

$$\begin{aligned}\alpha_{\mathsf{YB}}^{(1)} &:= 1 + \|L\|_{\mathbb{X}}^2 \cdot \lambda_{\max}\left((I - M)^{-1}(I - M)^{-\top}\right), \\ \alpha_{\mathsf{YB}}^{(2)} &:= 1 + \|(I - \Pi_{\mathbb{S}} L)^{-1} \Pi_{\mathbb{S}} L \Pi_{\mathbb{S}^\perp}\|_{\mathbb{X}}^2.\end{aligned}$$

It is clear from the definition that $\alpha(M, \|L\|_{\mathbb{X}}) \leq \alpha_{\mathsf{YB}}^{(1)}$, but $\alpha(M, \|L\|_{\mathbb{X}})$ can often provide an improved bound. This improvement is indeed significant, as will be shown shortly in Lemma 1. On the other hand, the term $\alpha_{\mathsf{YB}}^{(2)}$ is never larger than $\alpha(M, \|L\|_{\mathbb{X}})$, and is indeed the smallest possible bound that depends only on L and *not* b. However, as pointed out by Yu and Bertsekas, the value of $\alpha_{\mathsf{YB}}^{(2)}$ is not easily accessible in practice, since it depends on the precise behavior of the operator L over the orthogonal complement $\mathbb{S}^\perp$. Thus, estimating the quantity $\alpha_{\mathsf{YB}}^{(2)}$ requires $O(D)$ samples. In contrast, the term $\alpha(M, \|L\|_{\mathbb{X}})$ depends only on the projected operator M and the operator norm $\|L\|_{\mathbb{X}}$. The former can be easily estimated using d samples and at smaller computational cost, while the latter is usually known a priori. The discussion about LSTD methods in Section 3.2 fleshes out these distinctions.

B.2. Some useful bounds on $\alpha(M, \|L\|_{\mathbb{X}})$

We conclude our discussion of the approximation factor with some bounds that can be derived under different assumptions on the operator L and its projected version M. The following lemma is useful in understanding the behavior of the approximation factor as a function of the contractivity properties of the operator L; this is particularly useful in analyzing convergence rates in numerical PDEs.

Lemma 1 *Consider a projected matrix $M \in \mathbb{R}^{d\times d}$ such that $(I-M)$ is invertible and $\kappa(M) < 1$.*

(a) For any $s > 0$, we have the bound

$$\alpha(M,s) \leq 1 + \|(I-M)^{-1}\|_{op}^2 \cdot s^2 \leq 1 + \frac{s^2}{(1-\kappa(M))^2}. \tag{16a}$$

(b) For $s \in [0,1]$, we have

$$\alpha(M,s) \leq 1 + 2\|(I-M)^{-1}\|_{op} \leq 1 + \frac{2}{1-\kappa(M)}. \tag{16b}$$

See Appendix B.2.1 for the proof of this lemma.

A second special case, also useful, is when the matrix M is symmetric, a setting that appears in least-squares regression, value function estimation in reversible Markov chains, and self-adjoint elliptic operators. The optimal approximation factor $\alpha(M, \gamma_{\max})$ can be explicitly computed in such cases.

Lemma 2 *Suppose that M is symmetric with eigenvalues $\{\lambda_j(M)\}_{j=1}^d$ such that $\lambda_{max}(M) < 1$. Then for any $s > 0$, we have*

$$\alpha(M,s) = 1 + \max_{j=1,\ldots,d} \frac{s^2 - \lambda_j^2}{(1-\lambda_j)^2}. \tag{17}$$

See Appendix B.2.2 for the proof of this lemma.

In the study of Galerkin approximation methods for differential equations, the bound of the form (a) in Lemma 1 is known as Céa's lemma (Céa, 1964), which plays a central role in the convergence rate analysis of the Galerkin methods for numerical differential equations. However, the instance-dependent approximation factor $\alpha(M, \|L\|_{\mathbb{X}})$ can often be much smaller: the global coercive parameter needed in Céa's estimate is replaced by the bounds on the behavior of the operator L in the finite-dimensional subspace. The part (b) in Lemma 1 generalizes Céa's energy estimate from the symmetric positive-definite case to the general non-expansive setting.

B.2.1. Proof of Lemma 1

Recall that

$$\alpha(M,s) = 1 + \lambda_{\max}\left((I-M)^{-1}(s^2 I_d - MM^\top)(I-M)^{-\top}\right). \tag{18}$$

In the following, we prove upper bounds for the two different cases separately.

Bounds in the general case: By assumption, we have $\|M\|_{\text{op}} \leq s$, and consequently,

$$0 \preceq s^2 I_d - MM^\top \preceq s^2.$$

Thus, we have the sequence of implications

$$\begin{aligned}
\alpha(M,s) - 1 &= \lambda_{\max}\left((I-M)^{-1}(s^2 I - MM^\top)(I-M)^{-\top}\right) \\
&= \|(I-M)^{-1}(s^2 I - MM^\top)(I-M)^{-\top}\|_{\text{op}} \\
&\leq \|(I-M)^{-1}\|_{\text{op}} \cdot \|s^2 I_d - MM^\top\|_{\text{op}} \cdot \|(I-M)^{-1}\|_{\text{op}} \\
&\leq \|(I-M)^{-1}\|_{\text{op}}^2 \cdot s^2,
\end{aligned}$$

which proves the bound.

Bounds under non-expansive condition: When $s \leq 1$, we have

$$s^2 I - MM^\top \preceq I - MM^\top = \frac{1}{2}(I-M)(I+M^\top) + \frac{1}{2}(I+M)(I-M^\top).$$

Consequently, we have the chain of bounds

$$\begin{aligned}
\alpha(M,s) - 1 &\leq \lambda_{\max}\left((I-M)^{-1}(I-MM^\top)(I-M)^{-\top}\right) \\
&= \frac{1}{2}\lambda_{\max}\left((I+M)^\top(I-M^\top)^{-1} + (I-M)^{-1}(I+M)\right) \\
&\leq \frac{1}{2}|||(I+M)^\top(I-M^\top)^{-1}|||_{\text{op}} + \frac{1}{2}|||(I-M)^{-1}(I+M)|||_{\text{op}} \\
&\leq |||(I-M)^{-1}|||_{\text{op}} + |||(I-M^\top)^{-1}|||_{\text{op}} \\
&= 2|||(I-M)^{-1}|||_{\text{op}}.
\end{aligned}$$

Finally, we note that if $\kappa(M) < 1$, then for any $u \in \mathbb{R}^d$, we have

$$(1-\kappa(M))\|u\|_2^2 \leq \langle (I-M)u,\, u\rangle \leq \|(I-M)u\|_2 \cdot \|u\|_2.$$

Consequently, we have $|||(I-M)^{-1}|||_{\text{op}} \leq \frac{1}{1-\kappa(M)}$, which completes the proof of this lemma.

B.2.2. Proof of Lemma 2

Once again, recall the definition

$$\alpha(M,s) = 1 + \lambda_{\max}\left((I-M)^{-1}(s^2 I_d - MM^\top)(I-M)^{-\top}\right).$$

Since M is symmetric, let $M = P\Lambda P^\top$ be its eigen-decomposition, where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$, and note that

$$\begin{aligned}
\alpha(M,s) &= 1 + \lambda_{\max}\left(P(I-\Lambda)^{-1}(s^2 - \Lambda^2)(I-\Lambda)^{-1}P^\top\right) \\
&= 1 + \lambda_{\max}\left((I-\Lambda)^{-2}(s^2-\Lambda^2)\right) \\
&= 1 + \max_{1\leq i\leq d}\left(\frac{\gamma_{\max}^2 - \lambda_i^2}{(1-\lambda_i)^2}\right),
\end{aligned}$$

which completes the proof.

C. Proof of Theorem 1

We divide the proof into two parts, corresponding to the two components in the mean-squared error of the estimator $\widehat{v}_n$. The first term is the *approximation error* $\|\bar{v} - v^*\|^2$ that arises from the difference between the exact solution v^* to the original fixed point equation, and the exact solution $\bar{v}$ to the projected set of equations. The second term is the *estimation error* $\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2$, measuring the difficulty of estimating $\bar{v}$ on the basis of n noisy samples.

In particular, under the conditions of the theorem, we prove that the approximation error is upper bounded as

$$\|\bar{v} - v^*\|^2 \leq \alpha(M, |||L|||_{\mathbb{X}}) \inf_{v\in\mathbb{S}} \|v - v^*\|^2, \tag{19a}$$

whereas the estimation error is bounded as

$$\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 \leq c\frac{\operatorname{trace}\left((I-M)^{-1}\Sigma^*(I-M)^{-\top}\right)}{n} + c\frac{\sigma_L}{(1-\kappa)^3}\left(\frac{d}{n}\right)^{3/2}\left(\|\bar{v}\|^2\sigma_L^2 + \sigma_b^2\right). \tag{19b}$$

Given these two inequalities, it is straightforward to prove the bound (13) stated in the theorem. By expanding the square, we have

$$\begin{aligned}
\mathbb{E}\|\widehat{v}_n - v^*\|^2 &= \mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 + \|\bar{v} - v^*\|^2 + 2\mathbb{E}\langle \widehat{v}_n - \bar{v},\, \bar{v} - v^*\rangle \\
&\overset{(i)}{\leq} \mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 + \|\bar{v} - v^*\|^2 + 2\sqrt{\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 \cdot \|\bar{v} - v^*\|^2} \\
&\overset{(ii)}{\leq} \mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 + \|\bar{v} - v^*\|^2 + \tfrac{1}{\omega}\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 + \omega\|\bar{v} - v^*\|^2 \\
&= (1+\omega)\|\bar{v} - v^*\|^2 + (1+\tfrac{1}{\omega})\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2
\end{aligned}$$

where step (i) follows from the Cauchy–Schwarz inequality; and step (ii) follows from the arithmetic-geometric mean inequality, and is valid for any $\omega > 0$. Substituting the bounds from equations (19a) and (19b) yields the claim of the theorem.

The remainder of our argument is devoted to the proofs of the bounds (19a) and (19b).

C.1. Proof of approximation error bound (19a)

We begin with some decomposition relations for vectors and operators. Note that $\mathbb{S}$ is a finite-dimensional subspace, and therefore is closed. We use

$$\mathbb{S}^{\perp} := \{u \in \mathbb{X} \mid \langle u, v\rangle = 0 \mid \text{for all } v \in \mathbb{S}.\}$$

to denote its orthogonal complement. The pair $(\mathbb{S}, \mathbb{S}^{\perp})$ forms a direct product decomposition of $\mathbb{X}$, and the projection operators satisfy $\Pi_{\mathbb{S}} + \Pi_{\mathbb{S}^{\perp}} = I$. Also define the operators $L_{\mathbb{S},\mathbb{S}} = \Pi_{\mathbb{S}} L \Pi_{\mathbb{S}}$ and $L_{\mathbb{S},\perp} = \Pi_{\mathbb{S}} L \Pi_{\mathbb{S}^{\perp}}$. With this notation, our proof can be broken down into two auxiliary lemmas, which we state here:

Lemma 3 *The error $\|\bar{v} - v^*\|$ between the projected fixed point $\bar{v}$ and the original fixed point v^* is bounded as*

$$\|\bar{v} - v^*\|^2 \leq \left(1 + \|(I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp}\|_{\mathbb{X}}^2\right) \inf_{v \in \mathbb{S}} \|v - v^*\|^2. \tag{20}$$

Lemma 4 *Under the set-up above, we have*

$$\|(I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp}\|_{\mathbb{X}}^2 \leq \lambda_{max}\Big((I_d - M)^{-1}\Big(\|L\|_{\mathbb{X}}^2 I_d - MM^{\top}\Big)(I_d - M)^{-\top}\Big).$$

The claimed bound (19a) on the approximation error follows by combining these two lemmas, and recalling our definition of $\alpha(M, L)$. We now prove these two lemmas in turn.

C.1.1. Proof of Lemma 3

For any vector $v \in \mathbb{X}$, we perform the orthogonal decomposition $v = v_{\mathbb{S}} + v_{\perp}$, where $v_{\mathbb{S}} := \Pi_{\mathbb{S}}(v)$ is a member of the set $\mathbb{S}$, and $v_{\perp} := \Pi_{\mathbb{S}^{\perp},\xi}$ is a member of the set $\mathbb{S}^{\perp}$. With this notation, the operator L can be decomposed as

$$L = (\Pi_{\mathbb{S}} + \Pi_{\mathbb{S}^{\perp}}) L (\Pi_{\mathbb{S}} + \Pi_{\mathbb{S}^{\perp}}) \;=\; \underbrace{\Pi_{\mathbb{S}} L \Pi_{\mathbb{S}}}_{=:L_{\mathbb{S},\mathbb{S}}} + \underbrace{\Pi_{\mathbb{S}} L \Pi_{\mathbb{S}^{\perp}}}_{=:L_{\mathbb{S},\perp}} + \underbrace{\Pi_{\mathbb{S}^{\perp}} L \Pi_{\mathbb{S}}}_{=:L_{\perp,\mathbb{S}}} + \underbrace{\Pi_{\mathbb{S}^{\perp}} L \Pi_{\mathbb{S}^{\perp}}}_{=:L_{\perp,\perp}}.$$

The four operators $L_{\mathbb{S},\mathbb{S}}, L_{\mathbb{S},\perp}, L_{\perp,\mathbb{S}}, L_{\perp,\perp}$ defined in the equation above are also bounded linear operators. By the properties of projection operators, we note that $L_{\mathbb{S},\mathbb{S}}$ and $L_{\perp,\mathbb{S}}$ both map each element of $\mathbb{S}^{\perp}$ to 0, and $L_{\mathbb{S},\perp}$ and $L_{\perp,\perp}$ both map each element of $\mathbb{S}$ to 0.

Decomposing the target vector v^* in an analogous manner yields the two components

$$\widetilde{v} := \Pi_{\mathbb{S}}(v^*), \quad \text{and} \quad v^{\perp} := v^* - \widetilde{v}.$$

The fixed point equation $v^* = Lv^* + b$ can then be written using $\mathbb{S}$ and its orthogonal complement as

$$\widetilde{v} \overset{(a)}{=} L_{\mathbb{S},\mathbb{S}}\widetilde{v} + L_{\mathbb{S},\perp} v^{\perp} + b_{\mathbb{S}}, \quad \text{and} \quad v^{\perp} \overset{(b)}{=} L_{\perp,\mathbb{S}}\widetilde{v} + L_{\perp,\perp} v^{\perp} + b_{\perp}. \tag{21}$$

For the projected solution $\bar{v}$, we have the defining equation

$$\bar{v} = L_{\mathbb{S},\mathbb{S}}\bar{v} + b_{\mathbb{S}}. \tag{22}$$

Subtracting equation (21)(a) from equation (22) yields

$$(I - L_{\mathbb{S},\mathbb{S}})(\widetilde{v} - \bar{v}) = L_{\mathbb{S},\perp} v^{\perp}.$$

Recall the quantity $M = \Phi_d L \Phi_d^*$, and our assumption that $\kappa(M) = \frac{1}{2}\lambda_{\max}(M + M^T) < 1$. This condition implies that $I - L_{\mathbb{S},\mathbb{S}}$ is invertible on the subspace $\mathbb{S}$. Since this operator also maps each element of $\mathbb{S}^{\perp}$ to itself, it is invertible on all of $\mathbb{X}$, and we have $\widetilde{v} - \bar{v} = (I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp} v^{\perp}$.

Applying the Pythagorean theorem then yields

$$\|\bar{v} - v^*\|^2 \;=\; \|\bar{v} - \widetilde{v}\|^2 + \|\widetilde{v} - v^*\|^2 = \|(I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp} v^{\perp}\|^2 + \|v^{\perp}\|^2 \\ \leq \left(1 + \|\!|(I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp}|\!\|_{\mathbb{X}}^2\right) \cdot \|v^{\perp}\|^2, \tag{23}$$

as claimed.

C.1.2. Proof of Lemma 4

By the definition of operator norm for any vector $v \in \mathbb{X}$ such that $\|v\| = 1$, we have

$$|\!|\!|L|\!|\!|_{\mathbb{X}}^2 \geq \|Lv\|^2 = \|L_{\mathbb{S},\mathbb{S}} v_{\mathbb{S}} + L_{\mathbb{S},\perp} v_{\perp}\|^2 + \|L_{\perp,\mathbb{S}} v_{\mathbb{S}} + L_{\perp,\perp} v_{\perp}\|^2 \geq \|L_{\mathbb{S},\mathbb{S}} v_{\mathbb{S}} + L_{\mathbb{S},\perp} v_{\perp}\|^2.$$

Noting the fact that $L_{\mathbb{S},\mathbb{S}} v_{\perp} = 0 = L_{\mathbb{S},\perp} v_{\mathbb{S}}$, we have the following norm bound on the linear operator $L_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp}$:

$$|\!|\!|L_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp}|\!|\!|_{\mathbb{X}} = \sup_{\|v\|=1} \|(L_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp}) v\| \\ = \sup_{\|v\|=1} \|L_{\mathbb{S},\mathbb{S}} v_{\mathbb{S}} + L_{\mathbb{S},\perp} v_{\perp}\| \leq |\!|\!|L|\!|\!|_{\mathbb{X}}.$$

By definition, the operator $L^*_{\mathbb{S},\perp} = \Pi_{\mathbb{S}^\perp} L^* \Pi_{\mathbb{S}}$ maps any vector to $\mathbb{S}^\perp$, and the operator $L_{\mathbb{S},\mathbb{S}}$ maps any element of $\mathbb{S}^\perp$ to 0. Therefore, we have the identity $L_{\mathbb{S},\mathbb{S}} L^*_{\mathbb{S},\perp} = 0$. A similar argument yields that $L_{\mathbb{S},\perp} L^*_{\mathbb{S},\mathbb{S}} = 0$. Consequently, we have

$$|\!|\!|L|\!|\!|_{\mathbb{X}}^2 \geq |\!|\!|L_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp}|\!|\!|_{\mathbb{X}}^2 = |\!|\!|(L_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp})(L_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp})^*|\!|\!|_{\mathbb{X}} \\ = |\!|\!| \underbrace{L_{\mathbb{S},\mathbb{S}} L^*_{\mathbb{S},\mathbb{S}} + L_{\mathbb{S},\perp} L^*_{\mathbb{S},\perp}}_{=:G} |\!|\!|_{\mathbb{X}}. \tag{24}$$

Note that the operator G can be expressed as $G = \Pi_{\mathbb{S}} \left(L \Pi_{\mathbb{S}} L^* + L \Pi_{\mathbb{S}^\perp} L^*\right) \Pi_{\mathbb{S}}$. From this representation, we see that:

- For any vector $x \in \mathbb{X}$, we have $Gx \in \mathbb{S}$.
- For any vector $y \in \mathbb{S}^\perp$, we have $Gy = 0$.

Consequently, there exists a matrix $\widetilde{G} \in \mathbb{R}^{d \times d}$, such that $G = \Phi_d^* \widetilde{G} \Phi_d$. Since G is a positive semi-definite operator, the matrix $\widetilde{G}$ is positive semi-definite. Equation (24) implies that

$$\lambda_{\max}(\widetilde{G}) = |\!|\!|\widetilde{G}|\!|\!|_{\text{op}} = |\!|\!|G|\!|\!|_{\mathbb{X}} \leq |\!|\!|L|\!|\!|_{\mathbb{X}}^2. \tag{25}$$

Now defining $\tau := |\!|\!|(I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp}|\!|\!|_{\mathbb{X}}$, note that

$$\tau^2 = |\!|\!| \underbrace{(I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp} L^*_{\mathbb{S},\perp} (I - L^*_{\mathbb{S},\mathbb{S}})^{-1}}_{=:H} |\!|\!|_{\mathbb{X}}. \tag{26}$$

Moreover, the operator H is self-adjoint, and we have the following properties:

- The operator $L_{\mathbb{S},\perp}$ maps any vector to $\mathbb{S}$, and $(I - L_{\mathbb{S},\mathbb{S}})^{-1}$ maps $\mathbb{S}$ to itself. Consequently, for any $x \in \mathbb{X}$, the vector $Hx = (I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp} \left(L^*_{\mathbb{S},\perp} (I - L^*_{\mathbb{S},\mathbb{S}})^{-1}\right) x$ is a member of the set $\mathbb{S}$.
- The operator $L^*_{\mathbb{S},\perp} = \Pi_{\mathbb{S}^\perp} L^* \Pi_{\mathbb{S}}$ maps any vector from $\mathbb{S}^\perp$ to 0. Consequently, for any $y \in \mathbb{S}^\perp$, we have $Hy = (I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp} \left(L^*_{\mathbb{S},\perp} (I - L^*_{\mathbb{S},\mathbb{S}})^{-1}\right) y = 0$.

Owing to the facts above, there exists a matrix $\widetilde{H} \in \mathbb{R}^{d \times d}$, such that $H = \Phi_d^* \widetilde{H} \Phi_d$. Since the operator H is positive semi-definite, so is the matrix $\widetilde{H}$. Consequently, by equation (26), we obtain the identity $\tau^2 = |\!|\!|H|\!|\!|_{\mathbb{X}} = |\!|\!|\widetilde{H}|\!|\!|_{\text{op}} = \lambda_{\max}(H)$. In particular, letting $u \in \mathbb{S}^{d-1}$ be a maximal eigenvector of $\widetilde{H}$, we have

$$\widetilde{H} \succeq \tau^2 u u^\top. \tag{27}$$

Since $M = \Phi_d L_{\mathbb{S},\mathbb{S}} \Phi_d^*$ by definition, combining the above matrix inequalities (25) and (27), we arrive at the bound:

$$\begin{aligned}
&\|\!|L|\!\|_{\mathbb{X}}^2 I_d \succeq \widetilde{G} \\
&= \Phi_d \left(L_{\mathbb{S},\mathbb{S}} L_{\mathbb{S},\mathbb{S}}^* + L_{\mathbb{S},\perp} L_{\mathbb{S},\perp}^* \right) \Phi_d^* \\
&= \Phi_d L_{\mathbb{S},\mathbb{S}} L_{\mathbb{S},\mathbb{S}}^* \Phi_d^* + (\Phi_d (I - L_{\mathbb{S},\mathbb{S}}) \Phi_d^*) \cdot \left(\Phi_d (I - L_{\mathbb{S},\mathbb{S}})^{-1} L_{\mathbb{S},\perp} L_{\mathbb{S},\perp}^* (I - L_{\mathbb{S},\mathbb{S}}^*)^{-1} \Phi_d^* \right) \cdot \left(\Phi_d (I - L_{\mathbb{S},\mathbb{S}}^*) \Phi_d^* \right) \\
&= M M^\top + (I - M) \widetilde{H} (I - M^\top) \\
&\succeq M M^\top + \tau^2 (I - M) u u^\top (I - M^\top).
\end{aligned}$$

Re-arranging and noting that $u \in \mathbb{S}^{d-1}$, we arrive at the inequality

$$\tau^2 \leq u^\top \left[(I - M)^{-1} (\|\!|L|\!\|_{\mathbb{X}}^2 I_d - M M^\top) (I - M)^{-\top} \right] u \;\leq\; \lambda_{\max} \Big((I - M)^{-1} (\|\!|L|\!\|_{\mathbb{X}}^2 I_d - M M^\top) (I - M)^{-\top} \Big),$$

which completes the proof of Lemma 4.

C.2. Proof of estimation error bound (19b)

We now turn to the proof of our claimed bound on the estimation error. Our analysis relies on two auxiliary lemmas. The first lemma provides bounds on the mean-squared error of the standard iterates $\{v_t\}_{t \geq 0}$—that is, without the averaging step:

Lemma 5 *Suppose that the noise conditions in Assumption 1 hold. Then for any stepsize $\eta \in \left(0, \frac{1-\kappa}{4\sigma_L^2 d + 1 + \|\!|L|\!\|_{\mathbb{X}}^2}\right)$, we have the bound*

$$\mathbb{E}\|v_t - \bar{v}\|^2 \leq e^{-(1-\kappa)\eta t/2} \mathbb{E}\|v_0 - \bar{v}\|^2 + \frac{8\eta}{1-\kappa} (\|\bar{v}\|^2 \sigma_L^2 d + \sigma_b^2 d) \qquad \text{valid for } t = 1, 2, \ldots. \tag{28}$$

See Section C.2.1 for the proof of this claim.

Our second lemma provides a bound on the PR-averaged estimate $\widehat{v}_n$ based on n observations in terms of a covariance term, along with the error of the non-averaged sequences $\{v_t\}_{t \geq 1}$:

Lemma 6 *Under the setup above, we have the bound*

$$\begin{aligned}
\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 \leq \frac{6}{n - n_0} \operatorname{trace} \left((I - M)^{-1} \Sigma^* (I - M)^{-\top} \right) \\
+ \frac{6}{(n - n_0)^2} \sum_{t = n_0}^{n} \mathbb{E}\|(I - M)^{-1} \Phi_d (L_{t+1} - L)(v_t - \bar{v})\|_2^2 + \frac{3\mathbb{E}\|v_n - v_{n_0}\|^2}{\eta^2 (n - n_0)^2 (1 - \kappa)^2}.
\end{aligned} \tag{29}$$

See Section C.2.2 for the proof of this claim.

Equipped with these two lemmas, we can now complete the proof of the claimed bound (19b) on the estimation error. Recalling that $n_0 = n/2$, we see that the first term in the bound (29) matches a term in the bound (19b). As for the remaining two terms in equation (29), the second moment bounds from Assumption 1 combined with the assumption that $\kappa(M) < 1$ imply that

$$\begin{aligned}
\mathbb{E}\|(I - M)^{-1} \Phi_d (L_{t+1} - L)(v_t - \bar{v})\|_2^2 &\leq \frac{1}{(1-\kappa)^2} \mathbb{E}\|\Phi_d (L_{t+1} - L)(v_t - \bar{v})\|_2^2 \\
&\leq \frac{1}{(1-\kappa)^2} \sum_{j=1}^{d} \mathbb{E}\langle \phi_j, (L_{t+1} - L)(v_t - \bar{v}) \rangle^2 \\
&\leq \frac{\sigma_L^2 d \|v_t - \bar{v}\|^2}{(1-\kappa)^2}.
\end{aligned}$$

On the other hand, we can use Lemma 5 to control the third term in the bound (29). We begin by observing that

$$\|v_n - v_{n_0}\|^2 \leq 2\|v_n - \bar{v}\|^2 + 2\|v_{n_0} - \bar{v}\|^2 \leq 4 \sup_{n_0 \leq t \leq n} \mathbb{E}\|v_t - \bar{v}\|^2.$$

If we choose a burn-in time $n_0 > \frac{c_0}{(1-\kappa)\eta} \log\left(\frac{\|v_0 - \bar{v}\|^2 d}{1-\kappa}\right)$, then Lemma 5 ensures that

$$\sup_{n_0 \leq t \leq n} \mathbb{E}\|v_t - \bar{v}\|^2 \leq \frac{16\eta}{1-\kappa}\left(\|\bar{v}\|^2 \sigma_L^2 d + \sigma_b^2 d\right).$$

Finally, taking the step size $\eta = \frac{1}{24\sigma_L\sqrt{dn}}$, recalling that $n_0 = n/2$, and putting together the pieces yields

$$\begin{aligned}
\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 &\leq \frac{12}{n}\operatorname{trace}\left((I-M)^{-1}\Sigma^*(I-M)^{-\top}\right) + \frac{1}{(1-\kappa)^2}\left(\frac{12\sigma_L^2 d}{n} + \frac{48}{\eta^2 n^2}\right)\sup_{n_0 \leq t \leq n}\mathbb{E}\|v_t - \bar{v}\|^2 \\
&\leq \frac{12}{n}\operatorname{trace}\left((I-M)^{-1}\Sigma^*(I-M)^{-\top}\right) + \frac{48\sigma_L}{(1-\kappa)^3}\left(\frac{d}{n}\right)^{3/2}\left(\|\bar{v}\|^2\sigma_L^2 + \sigma_b^2\right),
\end{aligned}$$

as claimed.

It remains to prove our two auxiliary lemmas, which we do in the following subsections.

C.2.1. Proof of Lemma 5

We now prove Lemma 5, which provides a bound on the error of the non-averaged iterates $\{v_t\}_{t \geq 1}$, as defined in equation (11a). Using the form of the update, we expand the mean-squared error to find that

$$\begin{aligned}
\mathbb{E}\|v_{t+1} - \bar{v}\|^2 &= \mathbb{E}\|(I - \eta I + \eta\Pi_{\mathbb{S}} L)(v_t - \bar{v}) + \eta\Pi_{\mathbb{S}}(L_{t+1} - L)v_t + \eta\Pi_{\mathbb{S}}(b_{t+1} - b)\|^2 \\
&\stackrel{(i)}{=} \mathbb{E}\|(I - \eta I + \eta\Pi_{\mathbb{S}} L)(v_t - \bar{v})\|^2 + \eta^2\mathbb{E}\|\Pi_{\mathbb{S}}(L_{t+1} - A)v_t + \Pi_\phi(b_{t+1} - b)\|^2 \\
&\stackrel{(ii)}{\leq} (1 - \eta(1-\kappa))\mathbb{E}\|v_t - \bar{v}\|^2 + 2\eta^2\mathbb{E}\|\Pi_{\mathbb{S}}(L_{t+1} - L)(v_t - \bar{v})\|^2 \\
&\qquad + 2\eta^2\mathbb{E}\|\Pi_{\mathbb{S}}(L_{t+1} - L)\bar{v} + \Pi_{\mathbb{S}}(b_{t+1} - b)\|^2. \qquad (30)
\end{aligned}$$

In step (i), we have made use of the fact that the noise is unbiased, and in step (ii), we have used that for any Δ in the subspace $\mathbb{S}$ and any stepsize $\eta \in \left(0, \frac{1-\kappa}{1+\|L\|_{\mathbb{X}}^2}\right)$, we have

$$\begin{aligned}
\|(I - \eta I + \eta\Pi_{\mathbb{S}} L)\Delta\|^2 &= (1-\eta)^2\|\Delta\|^2 + \eta^2\|\Pi_{\mathbb{S}} L\Delta\|^2 + 2(1-\eta)\eta\langle\Delta, \Pi_{\mathbb{S}} L\Delta\rangle \\
&\leq \left\{1 - 2\eta + \eta^2 + \eta^2\|L\|_{\mathbb{X}}^2 + 2(1-\eta)\eta\kappa\right\}\|\Delta\|^2 \\
&\leq \left(1 - \eta(1-\kappa)\right)\|\Delta\|^2.
\end{aligned}$$

Turning to the second term of equation (30), the moment bounds in Assumption 1 imply that

$$\mathbb{E}\|\Pi_{\mathbb{S}}(L_{t+1} - L)(v_t - \bar{v})\|^2 = \sum_{j=1}^d \mathbb{E}\langle\phi_j, (L_{t+1} - L)(v_t - \bar{v})\rangle^2 \leq \mathbb{E}\|v_t - \bar{v}\|^2\sigma_L^2 d.$$

Finally, the last term of equation (30) is also handled by Assumption 1, whence we obtain

$$\begin{aligned}
\mathbb{E}\|\Pi_{\mathbb{S}}(L_{t+1} - L)\bar{v} + \Pi_{\mathbb{S}}(b_{t+1} - b)\|^2 &\leq 2\sum_{j=1}^d \mathbb{E}\langle\phi_j, (L_{t+1} - L)\bar{v}\rangle^2 \\
&\qquad + 2\sum_{j=1}^d \mathbb{E}\langle\phi_j, b_{t+1} - b\rangle^2 \leq 2\|\bar{v}\|^2\sigma_L^2 d + 2\sigma_b^2 d.
\end{aligned}$$

Putting together the pieces, we see that provided $\eta < \frac{1-\kappa}{4\sigma_L^2 d+1+\|L\|_{\mathbb{X}}^2}$, we have

$$\begin{aligned}\mathbb{E}\|v_{t+1}-\bar{v}\|^2 &\leq (1-\eta(1-\kappa)+2\eta^2\sigma_L^2 d)\mathbb{E}\|v_t-\bar{v}\|^2 + 4\eta^2(\|\bar{v}\|^2\sigma_L^2 d+\sigma_b^2 d)\\ &\leq \left(1-\frac{\eta(1-\kappa)}{2}\right)\mathbb{E}\|v_t-\bar{v}\|^2 + 4\eta^2(\|\bar{v}\|^2\sigma_L^2 d+\sigma_b^2 d).\end{aligned}$$

Finally, rolling out the recursion yields the bound

$$\mathbb{E}\|v_n-\bar{v}\|^2 \leq e^{-(1-\kappa)\eta n/2}\mathbb{E}\|v_0-\bar{v}\|^2 + \frac{8\eta}{1-\kappa}(\|\bar{v}\|^2\sigma_L^2 d+\sigma_b^2 d),$$

which completes the proof.

C.2.2. Proof of Lemma 6

Recall that $\bar{v}$ satisfies the fixed point equation $\bar{v} = \Pi_{\mathbb{S}} L\bar{v} + \Pi_{\mathbb{S}} b$. Using this fact, we can derive the following elementary identity:

$$\begin{aligned}\frac{v_{n_0}-v_n}{\eta(n-n_0)} &= \frac{1}{n-n_0}\sum_{t=n_0}^{n-1}(v_t - \Pi_{\mathbb{S}}L_{t+1}v_t - \Pi_{\mathbb{S}}b_{t+1})\\ &= (I-\Pi_{\mathbb{S}}L)(\widehat{v}_n-\bar{v}) + \frac{1}{n-n_0}\underbrace{\sum_{t=n_0}^{n-1}\Pi_{\mathbb{S}}(L_{t+1}-L)v_t}_{=:\Psi_n^{(1)}} + \frac{1}{n-n_0}\underbrace{\sum_{t=n_0}^{n-1}\Pi_{\mathbb{S}}(b_{t+1}-b)}_{=:\Psi_n^{(2)}}. \quad (31)\end{aligned}$$

Re-arranging terms and applying the Cauchy–Schwarz inequality, we have

$$\|\widehat{v}_n-\bar{v}\|^2 \leq \frac{3}{(n-n_0)^2}\left(\frac{1}{\eta^2}\|(I-\Pi_{\mathbb{S}}L)^{-1}(v_n-v_{n_0})\|^2 + \|(I-\Pi_{\mathbb{S}}L)^{-1}\Psi_n^{(1)}\|^2 + \|(I-\Pi_{\mathbb{S}}L)^{-1}\Psi_n^{(2)}\|^2\right).$$

Note that the quantities $\Psi_n^{(1)}$ and $\Psi_n^{(2)}$ are martingales adapted to the filtration $\mathcal{F}_n := \sigma(\{L_i, b_i\}_{i=1}^n)$, so that

$$\begin{aligned}\mathbb{E}\|\widehat{v}_n-\bar{v}\|^2 \leq \frac{3}{(n-n_0)^2}&\sum_{t=n_0}^{n-1}\mathbb{E}\|(I-\Pi_{\mathbb{S}}L)^{-1}\Pi_{\mathbb{S}}(L_{t+1}-L)v_t\|^2\\ &+ \frac{3}{(n-n_0)^2}\sum_{t=n_0}^{n-1}\mathbb{E}\|(I-\Pi_{\mathbb{S}}A)^{-1}\Pi_{\mathbb{S}}(b_{t+1}-b)\|^2\\ &+ \frac{3}{(n-n_0)^2\eta^2}\mathbb{E}\|(I-\Pi_{\mathbb{S}}L)^{-1}(v_n-v_{n_0})\|^2.\end{aligned}$$

We claim that for any vector $v \in \mathbb{X}$, we have

$$(I-\Pi_{\mathbb{S}}L)^{-1}\Pi_{\mathbb{S}}v = \Phi_d^*\left((I-M)^{-1}\Phi_d v\right). \quad (32)$$

Taking this claim as given for the moment, by applying equation (32) with $v = (L_{t+1}-L)v_t$ and $v = b_{t+1}-b$, we find that

$$\begin{aligned}\mathbb{E}\|(I-\Pi_{\mathbb{S}}L)^{-1}\Pi_{\mathbb{S}}(L_{t+1}-L)v_t\|^2 &= \mathbb{E}\|(I-M)^{-1}\Phi_d(L_{t+1}-L)v_t\|_2^2\\ &\leq 2\mathbb{E}\|(I-M)^{-1}\Phi_d(L_{t+1}-L)\bar{v}\|_2^2 + 2\mathbb{E}\|(I-M)^{-1}\Phi_d(L_{t+1}-L)(v_t-\bar{v})\|_2^2,\end{aligned}$$

and

$$\mathbb{E}\|(I-L)^{-1}\Pi_{\mathbb{S}}(b_{t+1}-b)\|^2 = \mathbb{E}\|(I-M)^{-1}\Phi_d(b_{t+1}-b)\|_2^2.$$

Putting together the pieces, we obtain

$$\begin{aligned}\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 \leq \frac{3}{n-n_0}\,\mathrm{trace}\left((I-M)^{-1}\cdot\mathrm{cov}(\Phi_d(b_1-b))\cdot(I-M)^{-\top}\right)\\ + \frac{6}{n-n_0}\,\mathrm{trace}\left((I-M)^{-1}\cdot\mathrm{cov}(\Phi_d(L_1-L)\bar{v})\cdot(I-M)^{-\top}\right)\\ + \frac{6}{(n-n_0)^2}\sum_{t=n_0}^{n}\mathbb{E}\|(I-M)^{-1}\Phi_d(L_{t+1}-L)(v_t-\bar{v})\|_2^2 + \frac{3\mathbb{E}\|v_n - v_{n_0}\|^2}{\eta^2(n-n_0)^2(1-\kappa)^2},\end{aligned}$$

as claimed.

It remains to prove the identity (32).

Proof of claim (32)**:** Note that for any vector $v \in \mathbb{X}$, the vector $z := (I - \Pi_{\mathbb{S}}L)^{-1}\Pi_{\mathbb{S}}v$ is a member of $\mathbb{S}$, since $z = \Pi_{\mathbb{S}}Lz + \Pi_{\mathbb{S}}v$. Furthermore, since $\{\phi_j\}_{j=1}^d$ is a standard basis for $\mathbb{S}$, we have $z = \Pi_{\mathbb{S}}z = \Phi_d^*\Phi_d z$, and consequently,

$$\Phi_d z = \Phi_d L z + \Phi_d v = (\Phi_d L \Phi_d^*)\Phi_d z + \Phi_d v = M\Phi_d z + \Phi_d v.$$

Since the matrix M is invertible, we have $\Phi_d z = (I_d - M)^{-1}\Phi_d v$. Consequently, we have the identity $z = \Phi_d^*\Phi_d z = \Phi_d^*(I_d - M)^{-1}\Phi_d v$, which proves the claim.

D. Proof of the corollary for TD learning

We start by formally introducing the assumptions and stating the result. First, we recall that the Polyak-Ruppert-averaged TD(0) algorithm takes the form:

$$\vartheta_{t+1} = \vartheta_t - \eta\left(\psi(s_{t+1})\psi(s_{t+1})^\top\vartheta_t - \gamma\psi(s_{t+1})\psi(s_{t+1}^+)^\top\vartheta_t - R_{t+1}(s_{t+1})\psi(s_{t+1})\right). \tag{33a}$$

The Polyak–Ruppert averaged estimator is then given by the relations

$$\widehat{\vartheta}_n = \frac{2}{n}\sum_{t=n/2}^{n-1}\vartheta_t, \quad \text{and} \quad \widehat{v}_n := \widehat{\vartheta}_n^\top\psi. \tag{33b}$$

Recall our definition of the positive definite matrix B, with $B_{ij} = \langle\psi_i, \psi_j\rangle$. This defines an orthonormal basis given by

$$\begin{bmatrix}\phi_1 & \phi_2 & \cdots & \phi_d\end{bmatrix} := \begin{bmatrix}\psi_1 & \psi_2 & \cdots & \psi_d\end{bmatrix} B^{-1/2}.$$

Let $\beta := \lambda_{\max}(B)$ and $\mu := \lambda_{\min}(B)$, so that β/μ is the condition number of the covariance matrix of the features.

Having set up this transformation, we are now ready to state the implication of our main theorem to the case of LSTD problems. We assume the following fourth-moment condition:

$$\forall u \in \mathbb{S}^{d-1},\ \mathbb{E}_\xi\left(u^\top B^{-1/2}\psi(s)\right)^4 \leq \varsigma^4, \quad \text{and} \quad \mathbb{E}_\xi\left[R^4(s)\right] \leq \varsigma^4. \tag{34}$$

As verified in the proof of Corollary 1 to follow, equation (34) suffices to guarantee that Assumption 1 is satisfied with parameters $(\sigma_L, \sigma_b) = (2\varsigma^2, \varsigma^2/\sqrt{\beta})$.

The following corollary then provides a guarantee on the Polyak–Ruppert averaged TD(0) iterates.

Corollary 1 *Under the set-up above, there are universal positive constants (c, c_0) such that given a sample size $n \geq \frac{c_0\varsigma^4\beta^2 d}{\mu^2(1-\kappa(M))^2}\log^2\left(\frac{\|v_0-\bar{v}\|_2^2\beta d}{\mu(1-\kappa(M))}\right)$, then when the stochastic approximation scheme (33a) is run with step size $\eta = \frac{1}{c_0\varsigma^2\beta\sqrt{dn}}$, then the averaged iterates satisfy the bound*

$$\mathbb{E}\|\widehat{v}_n - v^*\|^2 \leq (1+\omega)\alpha(M,\gamma)\mathcal{A}(\mathbb{S}, v^*) + c\left(1+\frac{1}{\omega}\right)\left[\mathcal{E}_n(M, \Sigma_L+\Sigma_b) + \left(1+\|\bar{v}\|^2\right)\left(\frac{\varsigma^2\beta}{(1-\kappa(M))\mu}\sqrt{\frac{d}{n}}\right)^3\right] \tag{35}$$

for any $\omega > 0$.

Letting $\theta_t := B^{1/2}\vartheta_t$, the iterates (33a) can be equivalently written as

$$\theta_{t+1} = \theta_t - \eta B \cdot \left(\phi(s_{t+1})\phi(s_{t+1})^\top \theta_t - \gamma\phi(s_{t+1})\phi(s_{t+1}^+)^\top\theta_t + R_{t+1}(s_{t+1})\phi(s_{t+1})\right), \tag{36}$$

and the Polyak–Ruppert averaged iterate is given by $\widehat{\theta}_n := \frac{2}{n}\sum_{t=n/2}^{n-1}\theta_t$. We also define $\bar{\theta} := \Phi_d\bar{v}$, which is the solution to projected linear equations under the orthogonal basis. Clearly, we have $\bar{\theta} = B^{1/2}\bar{\vartheta}$.

We now claim that if $n \geq \frac{c_0\varsigma^4\beta^2}{\mu^2(1-\kappa(M))^2} d\log^2\left(\frac{\|\vartheta_0-\bar{\vartheta}\|_2 d\beta}{\mu(1-\kappa(M))}\right)$, then

$$\|\Phi_d^*\bar{\theta} - v^*\|^2 \leq \alpha(M,\gamma)\mathcal{A}(\mathbb{S}, v^*), \text{ and} \tag{37a}$$

$$\mathbb{E}\|\widehat{\theta}_n - \bar{\theta}\|_2^2 \leq c\mathcal{E}_n(M, \Sigma_L + \Sigma_b) + c\big(1 + \|\bar{v}\|^2\big)\left(\frac{\varsigma^2\beta}{(1-\kappa(M))\mu}\sqrt{\frac{d}{n}}\right)^3. \tag{37b}$$

Taking both inequalities as given for now, we proceed with the proof of this corollary. Combining equation (37a) and equation (37b) via Young's inequality, we arrive at the bound

$$\begin{aligned}
\mathbb{E}\|\widehat{v}_n - v^*\|^2 &\leq (1+\omega)\|\Phi_d^*\bar{\theta} - v^*\|^2 + \left(1 + \frac{1}{\omega}\right)\mathbb{E}\|\widehat{\theta}_n - \bar{\theta}\|_2^2 \\
&\leq (1+\omega)\mathcal{A}(\mathbb{S}, v^*) + c\left(1+\frac{1}{\omega}\right)\left[\mathcal{E}_n(M, \Sigma_L+\Sigma_b) + \big(1+\|\bar{v}\|^2\big)\left(\frac{\varsigma^2\beta}{(1-\kappa(M))\mu}\sqrt{\frac{d}{n}}\right)^3\right],
\end{aligned}$$

which completes the proof of this corollary.

D.1. Proof of equation (37a)

By equation (10) and the definition of $\bar{\theta}$, we have

$$\bar{\theta} = \gamma M\bar{\theta} + \mathbb{E}_\xi[R(s)\phi(s)].$$

It is easy to see that $\Phi_d^*\bar{\theta}$ solves the projected Bellman equation. Note furthermore that the projected linear operator is given by

$$\Phi_d L\Phi_d^* = \gamma\Phi_d P\Phi_d^* = M.$$

Invoking the bound in equation (19a), we complete the proof of this inequality.

D.2. Proof of equation (37b)

Following the proof strategy for the bound (19b), we first show an upper bound on the iterates $\mathbb{E}\|\vartheta_t - \bar{\vartheta}\|^2$ under the non-orthogonal basis $(\psi_j)_{j\in[d]}$, and then use this bound to establish the final estimation error guarantee under $\|\cdot\|$-norm.

Recall the stochastic approximation procedure under the non-orthogonal basis:

$$\vartheta_{t+1} = \vartheta_t - \eta\left(\psi(s_{t+1})\psi(s_{t+1})^\top\vartheta_t - \gamma\psi(s_{t+1})\psi(s_{t+1}^+)^\top\vartheta_t - R_{t+1}(s_{t+1})\psi(s_{t+1})\right).$$

Let $\widetilde{M} := I_d - \frac{1}{\beta}B^{1/2}(I_d - M)B^{1/2}$ and $\widetilde{h} := \frac{1}{\beta}\mathbb{E}[R(s)\psi(s)]$. We can view equation (33a) as a stochastic approximation procedure for solving the linear fixed-point equation $\bar{\vartheta} = \widetilde{M}\bar{\vartheta} + \widetilde{h}$, with stochastic observations

$$\widetilde{M}_t := I_d - \beta^{-1}\left(\psi(s_t)\psi(s_t)^\top - \gamma\psi(s_t)\psi(s_t^+)^\top\right), \quad \text{and} \quad \widetilde{h}_t := \beta^{-1}R(s_t)\psi(s_t).$$

To verify Assumption 1, we note that for $p, q \in \mathbb{S}^{d-1}$, the following bounds directly follows from the condition (34):

$$\begin{aligned}
\mathbb{E}\left(p^\top(\widetilde{M}_t - \widetilde{M})q\right)^2 &\leq 2\beta^{-2}\mathbb{E}\left((p^\top\psi(s_t))\cdot(\psi(s_t)^\top q)\right)^2 + 2\beta^{-2}\mathbb{E}\left((p^\top\psi(s_t))\cdot(\psi(s_t^+)^\top q)\right)^2 \\
&\leq 2\beta^{-2}\sqrt{\mathbb{E}\left(p^\top\psi(s_t)\right)^4\cdot\mathbb{E}\left(\psi(s_t)^\top q\right)^4} + 2\beta^{-2}\sqrt{\mathbb{E}\left(p^\top\psi(s_t)\right)^4\cdot\mathbb{E}\left(\psi(s_t^+)^\top q\right)^4} \\
&\leq \frac{4\varsigma^4}{\beta^2}\|B^{1/2}p\|_2^2\cdot\|B^{1/2}q\|_2^2 \leq 4\varsigma^4,
\end{aligned}$$

and

$$\mathbb{E}\left(p^\top\left(\widetilde{h}_t-\widetilde{h}\right)\right)^2 \le \beta^{-2}\mathbb{E}\left(R(s_t)\cdot p^\top\psi(s_t)\right)^2 \le \beta^{-2}\sqrt{\mathbb{E}\left[R(s_t)^4\right]\cdot\mathbb{E}\left(p^\top B^{1/2}\phi(s_t)\right)^4} \le \varsigma^4/\beta.$$

Consequently, for the stochastic approximation procedure in equation (33a), Assumption 1 is satisfied with $\sigma_L = 2\varsigma^2$ and $\sigma_b = \varsigma^2/\sqrt{\beta}$.

To establish an upper bound on $\kappa(\widetilde{M})$, we note that

$$\begin{aligned}1-\kappa(\widetilde{M}) &= \frac{1}{\beta}\lambda_{\min}\left(B - B^{1/2}\frac{M+M^\top}{2}B^{1/2}\right)\\ &= \frac{1}{\beta}\inf_{u\in\mathbb{S}^{d-1}}(B^{1/2}u)^\top\left(I_d-\frac{M+M^\top}{2}\right)(B^{1/2}u)\\ &\ge \frac{\mu}{\beta}\inf_{u\in\mathbb{S}^{d-1}}u^\top\left(I_d-\frac{M+M^\top}{2}\right)u \ge \frac{\mu}{\beta}\left(1-\kappa(M)\right).\end{aligned}$$

Invoking Lemma 5, for $\eta < \frac{c_0(1-\kappa(M))\mu}{(\varsigma^4 d+1)\beta^2}$, we have

$$\mathbb{E}\|\vartheta_t-\bar{\vartheta}\|_2^2 \le e^{-\frac{\mu}{2}(1-\kappa(M))\eta t}\mathbb{E}\|\vartheta_0-\bar{\vartheta}\|_2^2 + \frac{8\eta\beta}{(1-\kappa(M))\mu}\left(\|\vartheta\|_2^2\varsigma^4 d+\varsigma^4 d/\beta\right). \tag{38}$$

On the other hand, applying Lemma 6 to the stochastic approximation procedure (36) under the orthogonal coordinates, we have the bound

$$\begin{aligned}\mathbb{E}\|\widehat{v}_n-\bar{v}\|^2 \le{}& \frac{6}{n-n_0}\operatorname{trace}\left((I-M)^{-1}\Sigma^*(I-M)^{-\top}\right)\\ &+\frac{6}{(n-n_0)^2}\sum_{t=n_0}^{n}\mathbb{E}\|(I-B^{1/2}\widetilde{M}B^{-1/2})^{-1}B^{1/2}(\widetilde{M}_{t+1}-\widetilde{M})B^{-1/2}(\theta_t-\bar{\theta})\|_2^2\\ &+\frac{3\mathbb{E}\|\left(I_d-B^{1/2}\widetilde{M}B^{-1/2}\right)^{-1}(\theta_n-\theta_{n_0})\|_2^2}{\eta^2\beta^2(n-n_0)^2}.\end{aligned} \tag{39}$$

Straightforward calculation yields

$$\mathbb{E}\|(I-B^{1/2}\widetilde{M}B^{-1/2})^{-1}B^{1/2}(\widetilde{M}_{t+1}-\widetilde{M})B^{-1/2}(\theta_t-\bar{\theta})\|_2^2 = \beta^2\mathbb{E}\|(I-M)^{-1}B^{-1/2}(\widetilde{M}_{t+1}-\widetilde{M})(\vartheta_t-\bar{\vartheta})\|_2^2.$$

For any vector $p\in\mathbb{R}^d$, using condition (34), we note that

$$\begin{aligned}\mathbb{E}\|B^{-1/2}(\widetilde{M}_t-\widetilde{M})p\|_2^2 &\le 2\beta^{-2}\mathbb{E}\|\phi(s_t)\phi(s_t)^\top B^{1/2}p\|_2^2 + 2\beta^{-2}\mathbb{E}\|\phi(s_t)\phi(s_t^+)^\top B^{1/2}p\|_2^2\\ &\le 2\beta^{-2}\sqrt{\mathbb{E}\|\phi(s_t)\|_2^4}\cdot\sqrt{\mathbb{E}\left(\phi(s_t)^\top B^{1/2}p\right)^4} + 2\beta^{-2}\sqrt{\mathbb{E}\|\phi(s_t)\|_2^4}\cdot\sqrt{\mathbb{E}\left(\phi(s_t^+)^\top B^{1/2}p\right)^4}\\ &\le 4\beta^{-1}\varsigma^4 d.\end{aligned}$$

Substituting into the identity above, we obtain

$$\mathbb{E}\|(I-B^{1/2}\widetilde{M}B^{-1/2})^{-1}B^{1/2}(\widetilde{M}_{t+1}-\widetilde{M})B^{-1/2}(\theta_t-\bar{\theta})\|_2^2 \le \frac{4\beta\varsigma^4 d}{\left(1-\kappa(M)\right)^2}\mathbb{E}\|\vartheta_t-\bar{\vartheta}\|_2^2.$$

For the third term in equation (39), we note that

$$\begin{aligned}\mathbb{E}\|\left(I_d-B^{1/2}\widetilde{M}B^{-1/2}\right)^{-1}(\theta_n-\theta_{n_0})\|_2^2 &= \beta^2\mathbb{E}\|(I-M)^{-1}B^{-1/2}(\vartheta_n-\vartheta_{n_0})\|_2^2\\ &\le \frac{2\beta^2}{\mu\left(1-\kappa(M)\right)^2}\left(\mathbb{E}\|\vartheta_n-\bar{\vartheta}\|_2^2+\mathbb{E}\|\vartheta_{n_0}-\bar{\vartheta}\|_2^2\right).\end{aligned}$$

Putting together the pieces and invoking the bound (38), we see that if $n_0 \geq c_0 \frac{1}{\mu\eta(1-\kappa)} \log\big(\frac{d\beta}{\mu(1-\kappa)}\big)$, then

$$\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 \leq 6\mathcal{E}_n(M, \Sigma^*) + \left[\frac{24\beta\varsigma^4 d}{\big(1-\kappa(M)\big)^2 n} + \frac{48}{\mu\big(1-\kappa(M)\big)^2\eta^2 n^2}\right] \cdot \sup_{n_0 \leq t \leq n} \mathbb{E}\|\vartheta_t - \bar{\vartheta}\|_2^2$$
$$\leq 6\mathcal{E}_n(M, \Sigma^*) + c\frac{\beta^3}{\mu^2\big(1-\kappa(M)\big)^3}\left[\frac{\varsigma^4\eta d}{n} + \frac{1}{\eta\beta^2 n^2}\right]\big(\|\bar{\vartheta}\|_2^2\varsigma^4 d + \varsigma^4 d/\beta\big).$$

Now note that $\|\bar{\vartheta}\|_2^2 = \|B^{-1/2}\bar{\theta}\|_2^2 \leq \mu^{-1}\|\bar{v}\|^2$, and so choosing the step size $\eta := \frac{1}{c_0\varsigma^2\beta\sqrt{dn}}$ yields

$$\mathbb{E}\|\widehat{v}_n - \bar{v}\|^2 \leq 6\mathcal{E}_n(M, \Sigma^*) + c\frac{\beta^3\varsigma^6}{\mu^3\big(1-\kappa(M)\big)^3}\left(\frac{d}{n}\right)^{3/2}.$$

This completes the proof of equation (37b), and thus the corollary.

E. Proof of Theorem 2

Letting D and d be integer multiples of four without loss of generality, we denote the state space by $\mathcal{S} = \{1, 2, \cdots, D\}$. We decompose the state space into $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2$, with $\mathcal{S}_0 := \{1, 2, \cdots, 2d\}$, $\mathcal{S}_1 := \{2d+1, \cdots, d + \frac{D}{2}\}$, and $\mathcal{S}_2 := \{d + \frac{D}{2} + 1, \cdots, D\}$. Define the scalars $\rho = \min(\gamma, \nu) \in (0, 1)$ and $\tau := \frac{\delta}{\sqrt{2(1-\rho)}} \wedge 1$.

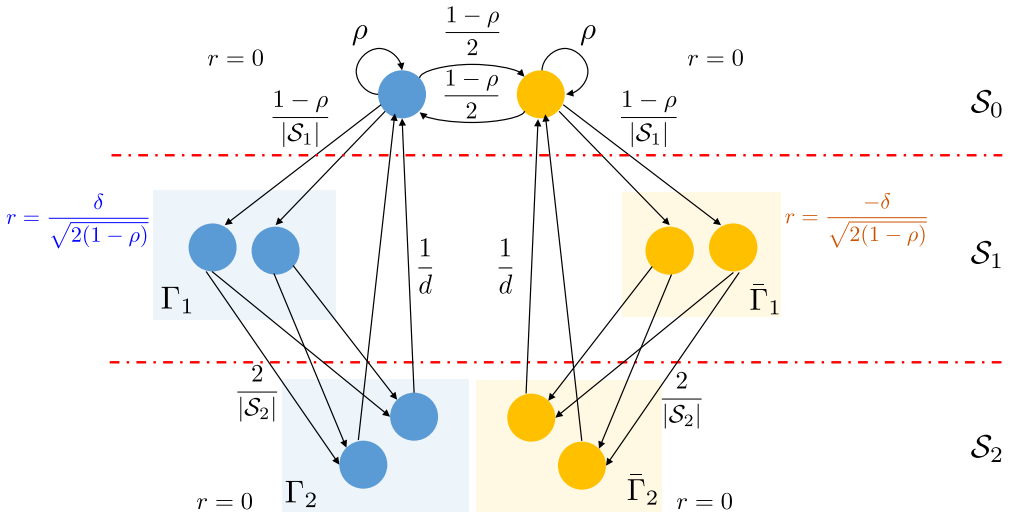

Figure 1. A graphical illustration of the MRP instance constructed above. For this instance, we let $d = 1$, $|\mathcal{S}_1| = 4$ and $|\mathcal{S}_2| = 4$, so that the total number of states is $D = 10$. In the graph, solid rounds stand for states, and arrows stand for the possible transitions. The numbers associated to the arrows stand for the probability of the transitions, and the equations $r = \cdots$ standard for the reward at a state. The sets $\mathcal{S}_0$, $\mathcal{S}_1$ and $\mathcal{S}_2$ are separated by red dotted lines, and the sets Γ_1, $\bar{\Gamma}_1$, Γ_2, and $\bar{\Gamma}_2$ are marked by transparent rectangles. A blue round stands for a state with positive value function, and an orange round stands for a state with negative value function.

Given a sign $z \in \{-1, 1\}$ and subsets $\Gamma_1 \subseteq \mathcal{S}_1$ and $\Gamma_2 \subseteq \mathcal{S}_2$ such that $|\Gamma_i| = \frac{1}{2}|\mathcal{S}_i|$ for each $i \in \{1, 2\}$, we let $\bar{\Gamma}_i := \mathcal{S}_i \setminus \Gamma_i$ for $i \in \{1, 2\}$. We then construct Markov reward processes $(P^{(\Gamma_1,\Gamma_2,z)}, r^{(\Gamma_1,\Gamma_2,z)})$ and feature vectors $(\psi^{(\Gamma_1,\Gamma_2,z)}(s_i))_{i=1}^D$,

indexed by the tuple (Γ_1, Γ_2, z). Entry (i, j) of the transition matrix is given by

$$P^{(\Gamma_1,\Gamma_2,z)}(i,j) := \begin{cases} \rho & i = j \in \mathcal{S}_0, \\ \frac{1-\rho}{2} & i, j \in \mathcal{S}_0,\ |i-j| = d, \\ \frac{1-\rho}{|\mathcal{S}_1|} & (i,j) \in (\{1,\cdots,d\} \times \Gamma_1) \cup \left(\{d+1,\cdots,2d\} \times \bar{\Gamma}_1\right), \\ \frac{2}{|\mathcal{S}_2|} & (i,j) \in (\Gamma_1 \times \Gamma_2) \cup \left(\bar{\Gamma}_1 \times \bar{\Gamma}_2\right), \\ \frac{1}{d} & (i,j) \in (\Gamma_2 \times \{1,2,\cdots,d\}) \cup \left(\bar{\Gamma}_2 \times \{d+1,\cdots,2d\}\right) \\ 0 & \text{otherwise.} \end{cases} \tag{40a}$$

The reward function at state i is given by

$$r^{(\Gamma_1,\Gamma_2,z)}(i) := \begin{cases} z\tau & i \in \Gamma_1, \\ -z\tau & i \in \bar{\Gamma}_1, \\ 0 & \text{otherwise.} \end{cases} \tag{40b}$$

This MRP is illustrated in Figure 1 for convenience. It remains to specify the feature vectors, and we use the same set of features for each tuple (Γ_1, Γ_2, z). The i-th such feature vector is given by

$$\psi(i) := \begin{cases} \sqrt{\frac{3-\rho}{2}d}e_i & i \in \{1,2,\cdots,d\}, \\ -\sqrt{\frac{3-\rho}{2}d}e_{i-d} & i \in \{d+1,\cdots,2d\}, \\ 0 & \text{otherwise.} \end{cases} \tag{40c}$$

It is easy to see that for any tuple (z, Γ_1, Γ_2), the Markov chain is irreducible and aperiodic, and furthermore, that the stationary distribution of the transition kernel $P^{(\Gamma_1,\Gamma_2,z)}$ is independent of the tuple (Γ_1, Γ_2, z), and given by

$$\xi = \Big[\underbrace{\tfrac{1}{(3-\rho)d} \quad \cdots \quad \tfrac{1}{(3-\rho)d}}_{2d} \quad \underbrace{\tfrac{1-\rho}{(3-\rho)(D-2d)} \quad \cdots \quad \tfrac{1-\rho}{(3-\rho)(D-2d)}}_{D-2d}\Big].$$

Clearly, we have $\mathbb{E}_\xi[\psi(s)\psi(s)^\top] = I_d$ under the stationary distribution. For the projected transition kernel, we have

$$\mathbb{E}[\psi(s)\psi(s^+)^\top] = \frac{3-\rho}{2} \cdot \left(\rho - \frac{1-\rho}{2}\right) I_d \preceq \rho I_d \preceq \nu I_d. \tag{41}$$

Given the discount factor $\gamma \in (0,1)$, let $c_0 := \frac{(1-\rho)/2}{1-\gamma(\rho-(1-\rho)(1-\gamma^2)/2)}$ for convenience. Straightforward calculation then yields that the value function for the problem instance $\left(P^{(\Gamma_1,\Gamma_2,z)}, r^{(\Gamma_1,\Gamma_2,z)}\right)$ at state i is given by

$$v^*_{\Gamma_1,\Gamma_2,z}(i) = \begin{cases} c_0 z\tau & i \in \{1,2,\cdots,d\}, \\ -c_0 z\tau & i \in \{d+1,\cdots,2d\}, \\ (1+\gamma^2 c_0) z\tau & i \in \Gamma_1, \\ -(1+\gamma^2 c_0) z\tau & i \in \bar{\Gamma}_1, \\ \gamma c_0 z\tau & i \in \Gamma_2, \\ -\gamma c_0 z\tau & i \in \bar{\Gamma}_2. \end{cases}$$

For $\rho > 1/2$, we have the bounds

$$c_0 \geq \frac{1}{4} \cdot \frac{1-\rho}{1-\gamma\rho} \geq \frac{1-\rho}{4(1-\rho^2)} \geq \frac{1}{8}, \quad \text{and} \quad c_0 \leq \frac{1-\rho}{1-\gamma\rho} \leq 1.$$

Consequently, we have $|v^*_{\Gamma_1,\Gamma_2,z}(i)| \asymp |v^*_{\Gamma_1,\Gamma_2,z}(j)|$ for each pair (i,j).

Note that by our construction, the subspace $\mathbb{S}$ spanned by the basis functions $\psi(1), \psi(2), \cdots, \psi(2d)$ is given by

$$\mathbb{S} = \left\{v \in \mathbb{L}^2(\mathcal{S}, \xi) : v(s) = 0 \text{ for } s \notin \mathcal{S}_0, \text{ and } v(i+d) = -v(i) \text{ for all } i \in [d]\right\}.$$

Consequently, we have

$$\inf_{v\in\mathbb{S}} \|v - v^*_{\Gamma_1,\Gamma_2,z}\|^2 = \frac{1-\rho}{3-\rho}\cdot\left(\frac{1}{2}(1+\gamma^2 c_0)^2\tau^2 + \frac{1}{2}\gamma^2 c_0^2\tau^2\right) \leq 2(1-\rho)\tau^2 = \delta^2. \tag{42}$$

Putting the equations (41) and (42) together, for any tuple (Γ_1, Γ_2, z), we conclude that the problem instance $(P^{(\Gamma_1,\Gamma_2,z)}, r^{(\Gamma_1,\Gamma_2,z)}, \gamma, \psi^{((\Gamma_1,\Gamma_2,z))})$ belongs to the class $\mathbb{C}_{\mathsf{MRP}}(\nu, \gamma, D, \delta)$.

In order to apply Le Cam's lemma, we define the following mixture distributions for each $z \in \{-1, 1\}$:

$$\mathbb{P}_z^{(n)} := \binom{|\mathcal{S}_1|}{|\mathcal{S}_1|/2}^{-2} \sum_{\substack{\Gamma_1\subseteq\mathcal{S}_1, \Gamma_2\subseteq\mathcal{S}_2 \\ |\Gamma_1|=|\Gamma_2|=\frac{1}{2}|\mathcal{S}_1|}} \mathbb{P}^{\otimes n}_{\Gamma_1,\Gamma_2,z},$$

where $\mathbb{P}_{\Gamma_1,\Gamma_2,z}$ is the law of an observed tuple $(s_i, s_i^+, r(s_i))$ under the MRP $\left(P^{(\Gamma_1,\Gamma_2,z)}, \gamma, r^{(\Gamma_1,\Gamma_2,z)}\right)$, and $\mathbb{P}^{\otimes n}_{\Gamma_1,\Gamma_2,z}$ denotes its n-fold product. Our next result gives a bound on the total variation distance.

Lemma 7 *Under the set-up above, we have* $d_{\mathrm{TV}}\left(\mathbb{P}_1^{(n)}, \mathbb{P}_{-1}^{(n)}\right) \leq \frac{Cn^2}{D-2d}$.

Taking this lemma as given, we now turn to the proof of the proposition. Consider any estimator $\widehat{v}$ for the value function. For any pair Γ_1, Γ_2 and Γ_1', Γ_2', we have

$$\|\widehat{v} - v^*_{\Gamma_1,\Gamma_2,1}\|^2 + \|\widehat{v} - v^*_{\Gamma_1',\Gamma_2',-1}\|^2 \geq \frac{1}{2}\|v^*_{\Gamma_1,\Gamma_2,1} - v^*_{\Gamma_1',\Gamma_2',-1}\|^2 \geq \frac{1}{2}c_0^2\tau^2 \geq \frac{\delta^2}{64(1-\rho)}.$$

Invoking Le Cam's lemma, for $D > 2C(n^2 + d)$, we have

$$\inf_{\widehat{v}_n} \sup_{(P,\gamma,r,\psi)\in\mathbb{C}_{\mathsf{MRP}}} \geq \frac{c}{1-\rho}\delta^2\left(1 - d_{\mathrm{TV}}(\mathbb{P}_1^{(n)}, \mathbb{P}_{-1}^{(n)})\right) \geq \frac{c'}{1-\nu\gamma}\delta^2,$$

which completes the proof. ■

E.1. Proof of Lemma 7

The high-level idea behind the proof is by recursive application of a one-step birthday argument combined with bounds on the bias induced by drawing without replacement.

To prove the lemma, we construct a probability distribution $\mathbb{Q}^{(n)}$ and bound the total variation distance between $\mathbb{Q}^{(n)}$ and $\mathbb{P}_z^{(n)}$ for each $z \in \{-1, 1\}$. In particular, for $k \in [n]$, we let $\mathbb{Q}^{(k)}$ be the law of k independent samples drawn from the following observation model:

- (Initial state:) Generate the state $s_i \sim \xi$.
- (Next state:) If $s_i \in \mathcal{S}_1$, then generate $s_i^+ \sim \mathcal{U}(\mathcal{S}_2)$. If $s_i \in \mathcal{S}_2$, then generate $s_i^+ \sim \mathcal{U}(\mathcal{S}_0)$. On the other hand, if $s_i \in \mathcal{S}_0$, then generate $S \sim \mathcal{U}(\mathcal{S}_1)$ and let[1]

$$s_i^+ = \begin{cases} s_i & \text{w.p. } \rho, \\ (s_i + d) \mod 2d & \text{w.p. } \frac{1-\rho}{2}, \\ S & \text{w.p. } \frac{1-\rho}{2}. \end{cases} \tag{43}$$

- (Reward:) If $s_i \in \mathcal{S}_1$, randomly draw $R_i = \zeta^{(i)} \sim \mathcal{U}(\{-1, 1\})$, and output $\zeta^{(i)}\tau$ as the reward. Otherwise, output the reward $R_i = 0$.

To bound the total variation distance $d_{\mathrm{TV}}(\mathbb{Q}^{(n)}, \mathbb{P}_z^{(n)})$, we use the following recursive relation, which holds for each $k = 0, 1, \cdots, n-1$:

$$d_{\mathrm{TV}}\left(\mathbb{Q}^{(k+1)}, \mathbb{P}_z^{(k+1)}\right) \leq d_{\mathrm{TV}}\left(\mathbb{Q}^{(k)}, \mathbb{P}_z^{(k)}\right) + \sup_{(s_i, s_i^+, R_i)_{i=1}^k} d_{\mathrm{TV}}\left(\mathbb{Q}^{(k+1)}|(s_i, s_i^+, R_i)_{i=1}^k, \mathbb{P}_z^{(k+1)}|(s_i, s_i^+, R_i)_{i=1}^k\right). \tag{44}$$

[1]The expression $a \mod b$ denotes the remainder of a divided by b, when a and b are integers.

Owing to the i.i.d. nature of the sampling model for $\mathbb{Q}^{(k+1)}$, note that we have the equivalence $(s_{k+1}, s_{k+1}^+, R_{k+1})|(s_i, s_i^+, R_i)_{i=1}^k \stackrel{d}{=} (s_{k+1}, s_{k+1}^+, R_{k+1})$.

At this juncture, it is helpful to view the probability distributions $\mathbb{P}_1^{(k)}$ and $\mathbb{P}_{-1}^{(k)}$ via the following two-step sampling procedure: First, for $j \in \{1, 2\}$, sample the subsets $\Gamma_j \subseteq \mathcal{S}_j$ uniformly at random from the collection of all subsets of size $|\mathcal{S}_j|/2$. Then, generate k i.i.d. samples $(s_i, s_i^+, R_i)_{i=1}^k$ according to the observation model (40a)-(40b). Consequently, for the rest of this proof, we view Γ_1 and Γ_2 as *random* sets. With this equivalence at hand, the following technical lemma shows that the posterior distribution of the subsets (Γ_1, Γ_2) conditioned on sampling the tuple $(s_i, s_i^+, R_i)_{i=1}^k$ is very close to the distribution of subsets chosen uniformly at random.

Lemma 8 *There is a universal positive constant c such that for each bit $z \in \{\pm 1\}$ and indices $j \in \{1, 2\}$ and $k \in [n]$, the following statement is true almost surely. For each tuple $(s_i, s_i^+, R_i)_{i=1}^k$ in the support of $\mathbb{P}_z^{(k)}$, the posterior distribution of Γ_j conditioned on $(s_i, s_i^+, R_i)_{i=1}^k \sim \mathbb{P}_z^{(k)}$ satisfies*

$$\max_{s \in \mathcal{S}_j \setminus \cup_{i=1}^k \{s_i, s_i^+\}} \left| \mathbb{P}\left(\Gamma_j \ni s \mid (s_i, s_i^+, R_i)_{i=1}^k\right) - \frac{1}{2} \right| \leq \frac{ck}{D-d}.$$

In words, for any "observable" tuple $(s_i, s_i^+, R_i)_{i=1}^k$ and each state $s \in \mathcal{S}_j \setminus \cup_{i=1}^k \{s_i, s_i^+\}$, the posterior probability of the event $\{\Gamma_j \ni s\}$ conditioned on observing the tuple $(s_i, s_i^+, R_i)_{i=1}^k$ is close to $1/2$ provided $D - d$ is large relative to k. In addition to the sets $\Gamma_j, j = 1, 2$ being close to uniformly random, we also require the following analog of a "birthday-paradox" argument in this setting. For convenience, we let $\mathcal{T}_k := \bigcup_{i=1}^k \{s_i, s_i^+\}$ denote the subset of states seen up until sample k.

Lemma 9 *There is a universal positive constant c such that for each $k \in [n]$ and each distribution $\mathbb{M}^{(k+1)} \in \left\{\mathbb{P}_{-1}^{(k+1)}, \mathbb{P}_{-1}^{(k+1)}, \mathbb{Q}^{(k+1)}\right\}$, the following statement holds almost surely. For each tuple $(s_i, s_i^+, R_i)_{i=1}^{k+1}$ in the support of $\mathbb{M}^{(k+1)}$, the probability the tuple of states $\left\{s_{k+1}, s_{k+1}^+\right\}$ conditioned on $(s_i, s_i^+, R_i)_{i=1}^k \sim \mathbb{M}^{(k)}$ satisfies*

$$\mathbb{P}\Big(\underbrace{\left\{s_{k+1}, s_{k+1}^+\right\} \cap \mathcal{T}_k \cap (\mathcal{S}_1 \cup \mathcal{S}_2) \neq \varnothing}_{:= \mathscr{E}_{k+1}^{(1)}} \mid (s_i, s_i^+, R_i)_{i=1}^k \Big) \leq \frac{ck}{D-d}. \tag{45}$$

In words, Lemma 9 ensures that if $D - d$ is large relative to k, then the states seen in sample $k + 1$ are different from those seen up until that point (provided we only count states in the set $\mathcal{S}_1 \cup \mathcal{S}_2$). Lemmas 8 and 9 are both proved at the end of this section; we take them as given for the rest of this proof.

Now consider tuples $(s_{k+1}, s_{k+1}^+, R_{k+1}) \sim \mathbb{P}_z^{(k+1)}|(s_i, s_i^+, R_i)_{i=1}^k$ and $(\widetilde{s}_{k+1}, \widetilde{s}_{k+1}^+, \widetilde{R}_{k+1}) \sim \mathbb{Q}^{(k+1)}|(s_i, s_i^+, R_i)_{i=1}^k$; we will now construct a coupling between these two tuples in order to show that the total variation between between the respective laws is small. First, note that under both $\mathbb{P}_z^{(k+1)}$ and $\mathbb{Q}^{(k+1)}$, the initial state is drawn from the stationary distribution, i.e., $s_{k+1}, \widetilde{s}_{k+1} \sim \xi$, regardless of the sequence $(s_i, s_i^+, R_i)_{i=1}^k$. We can therefore couple the two conditional laws together so that $s_{k+1} = \widetilde{s}_{k+1}$ almost surely. To construct the coupling for the rest, we consider the following three cases:

Coupling on the event $s_{k+1} \in \mathcal{S}_0$: We begin by coupling the reward random variables; we have $R_{k+1} = \widetilde{R}_{k+1} = 0$ under both conditional distributions, so this component of the distribution can be coupled trivially. Next, we couple the next state: By construction of the observation models (40a) and (43), we have

$$\mathbb{P}\big(s_{k+1}^+ = s_{k+1} | s_{k+1}\big) = \mathbb{P}\big(\widetilde{s}_{k+1}^+ = \widetilde{s}_{k+1} | \widetilde{s}_{k+1}\big) = \rho, \quad \text{and}$$

$$\mathbb{P}\big(s_{k+1}^+ = s_{k+1} + d \mod 2d \mid s_{k+1}\big) = \mathbb{P}\big(\widetilde{s}_{k+1}^+ = \widetilde{s}_{k+1} + d \mod 2d \mid \widetilde{s}_{k+1}\big) = \frac{1-\rho}{2},$$

and so these two components of the distribution can be coupled trivially. It remains to handle the case where $s_{k+1} \in \mathcal{S}_0$ and $s_{k+1}^+ \in \mathcal{S}_1$. By the symmetry of elements within set $\mathcal{S}_1$, we note that on the event $\big(\mathscr{E}_{k+1}^{(1)}\big)^C$, both random variables $\widetilde{s}_{k+1}^+$ and s_{k+1}^+ are uniformly distributed on the set $\mathcal{S}_1 \setminus \mathcal{T}_k$. Consequently, on the event $\big(\mathscr{E}_{k+1}^{(1)}\big)^C$, we can couple the conditional laws so that $s_{k+1}^+ = \widetilde{s}_{k+1}^+$ almost surely.

Coupling on the event $s_{k+1} \in \mathcal{S}_1$: As before, we begin by coupling the rewards, but first, note that on the event $\left(\mathscr{E}_{k+1}^{(1)}\right)^C$, we have $s_{k+1} \in \mathcal{S}_1 \setminus \mathcal{T}_k$. Invoking Lemma 8, under $\mathbb{P}_z^{(k)}$ and conditionally on the value of s_{k+1}, we have the bound

$$\left| \mathbb{P}\left(s_{k+1} \in \Gamma_1 \mid (s_i, s_i^+, R_i)_{i=1}^k\right) - \frac{1}{2} \right| \leq \frac{ck}{D-d}.$$

Now the reward function (40b) satisfies $r(s) = z\tau$ for $s \in \Gamma_1$ and $r(s) = -z\tau$ for $s \in \bar{\Gamma}_1$. On the other hand, under $\mathbb{Q}^{(k+1)}$, the reward $\widetilde{R}_{k+1}$ takes value of τ and $-\tau$, each with probability half. Consequently, there exists a coupling between R_{k+1} and $\widetilde{R}_{k+1}$, such that

$$\mathbb{P}\Big(\underbrace{R_{k+1} \neq \widetilde{R}_{k+1},\ s_{k+1} \in \mathcal{S}_1}_{:=\mathscr{E}_{k+1}^{(2)}} \mid (s_i, s_i^+, R_i)_{i=1}^k \Big) \leq \frac{ck}{D-d}.$$

Next, we construct the coupling for next-step transition conditionally on the current step. By the symmetry of elements within set $\mathcal{S}_2$, we note that under $\left(\mathscr{E}_{k+1}^{(1)}\right)^C$, both random variables $\widetilde{s}_{k+1}^+$ and s_{k+1}^+ are uniformly distributed on the set $\mathcal{S}_2 \setminus \mathcal{T}_k$. Consequently, on the event $\left(\mathscr{E}_{k+1}^{(1)}\right)^C$, we can couple the conditional laws so that $s_{k+1}^+ = \widetilde{s}_{k+1}^+$ almost surely.

Coupling on the event $s_{k+1} \in \mathcal{S}_2$: In this case, we have $R_{k+1} = \widetilde{R}_{k+1} = 0$ under both conditional distributions, so this coupling is once again trivial. It remains to construct a coupling between next-step transitions s_{k+1}^+ and $\widetilde{s}_{k+1}^+$. On the event $\left(\mathscr{E}_{k+1}^{(1)}\right)^C$, we have $s_{k+1} \in \mathcal{S}_2 \setminus \mathcal{T}_k$. Under $\mathbb{P}_z^{(k)}$ and conditionally on the value of s_{k+1}, Lemma 8 leads to the bound

$$\left| \mathbb{P}\left(s_{k+1} \in \Gamma_2 \mid (s_i, s_i^+, R_i)_{i=1}^k\right) - \frac{1}{2} \right| \leq \frac{ck}{D-d}.$$

By definition, under $\mathbb{P}_z^{(n)}$, we have that $s_{k+1}^+ \sim \mathcal{U}(\{1, 2, \cdots, d\})$ when $s_{k+1} \in \Gamma_2$, and $s_{k+1}^+ \sim \mathcal{U}(\{d+1, \cdots, 2d\})$ when $s_{k+1} \in \bar{\Gamma}_2$. Under $\mathbb{Q}^{(n)}$, we have $\widetilde{s}_{k+1}^+ \sim \mathcal{U}(\{1, 2, \cdots, 2d\})$. Consequently, there exists a coupling such that

$$\mathbb{P}\Big(\underbrace{s_{k+1}^+ \neq \widetilde{s}_{k+1}^+,\ s_{k+1} \in \mathcal{S}_2}_{:=\mathscr{E}_{k+1}^{(3)}} \mid (s_i, s_i^+, R_i)_{i=1}^k \Big) \leq \frac{ck}{D-d}.$$

Putting together our bounds from the three cases, note that for any sequence $(s_i, s_i^+, R_i)_{i=1}^k$ on the support of $\mathbb{Q}^{(k)}$ and $\mathbb{P}_z^{(k)}$, we almost surely have

$$\begin{aligned} &d_{\mathrm{TV}}\left(\mathcal{L}\left[(s_{k+1}, s_{k+1}^+, R_{k+1}) \mid (s_i, s_i^+, R_i)_{i=1}^k \right], \mathcal{L}\left[(\widetilde{s}_{k+1}, \widetilde{s}_{k+1}^+, \widetilde{R}_{k+1}) \mid (s_i, s_i^+, R_i)_{i=1}^k \right] \right) \\ &\leq \sum_{j=1}^3 \mathbb{P}\left(\mathscr{E}_{k+1}^{(j)} \mid (s_i, s_i^+, R_i)_{i=1}^k \right) \leq \frac{c'k}{D-d}, \end{aligned}$$

where the final inequality follows from applying Lemma 9. Substituting into the recursion (44), we conclude that for any $z \in \{-1, 1\}$, we have

$$d_{\mathrm{TV}}\left(\mathbb{Q}^{(n)}, \mathbb{P}_z^{(n)} \right) \leq \sum_{k=0}^{n-1} \sum_{j=1}^3 \sup_{(s_i, s_i^+, R_i)_{i=1}^k} \mathbb{P}\left(\mathscr{E}_{k+1}^{(j)} \mid (s_i, s_i^+, R_i)_{i=1}^k \right) \leq \frac{c'n^2}{D-d},$$

which completes the proof of this lemma. ■

It remains to prove the two helper lemmas.

E.1.1. Proof of Lemma 8

Given $z \in \{\pm 1\}$, we define the sets

$$Z_1 := \{s_i : i \in [k], s_i \in \mathcal{S}_1, R_i = z\tau\}, \quad \bar{Z}_1 := \left(\{s_i\}_{i\in[k]} \cap \mathcal{S}_1\right) \setminus Z_1, \quad \text{and}$$
$$Z_2 := \left\{s_i : i \in [k], s_i \in \mathcal{S}_2, s_i^+ \in [d]\right\}, \quad \bar{Z}_2 := \left(\{s_i\}_{i\in[k]} \cap \mathcal{S}_2\right) \setminus Z_2$$

By the reward model (40b) in our construction, for any valid pair of subsets (Γ_1, Γ_2), under the law $\mathbb{P}^{\otimes k}_{\Gamma_1,\Gamma_2,z}$, the observations $(s_i, s_i^+, R_i)_{i=1}^k$ have positive probability if and only if $Z_1 \subseteq \Gamma_1$ and $\Gamma_1 \cap \bar{Z}_1 = \varnothing$. Furthermore, by the symmetry between the elements in Γ_1, for any Γ_1 such that $Z_1 \subseteq \Gamma_1$ and $\Gamma_1 \cap \bar{Z}_1 = \varnothing$, the probability of observing $(s_i, s_i^+, R_i)_{i=1}^k$ under $\mathbb{P}^{\otimes k}_{\Gamma_1,\Gamma_2,z}$ is independent of the choice of Γ_1. Consequently, the probability under the mixture distribution $\mathbb{P}_z^{(k)}$ can be calculated as

$$\begin{aligned}
\mathbb{P}\left(\Gamma_1 \ni s \mid (s_i, s_i^+, R_i)_{i=1}^k\right) &= \sum_{\substack{|\Gamma'|=|\mathcal{S}_1|/2 \\ s\in\Gamma'}} \frac{\mathbb{P}\left((s_i, s_i^+, R_i)_{i=1}^k \mid \Gamma_1 = \Gamma'\right) \cdot \mathbb{P}(\Gamma_1 = \Gamma')}{\mathbb{P}\left((s_i, s_i^+, R_i)_{i=1}^k\right)} \\
&= \frac{\left|\left\{\Gamma' \subseteq \mathcal{S}_1 :\ |\Gamma'| = \frac{1}{2}|\mathcal{S}_1|,\ Z_1 \subseteq \Gamma',\ \bar{Z}_1 \cap \Gamma' = \varnothing,\ s \in \Gamma'\right\}\right|}{\left|\left\{\Gamma' \subseteq \mathcal{S}_1 :\ |\Gamma'| = \frac{1}{2}|\mathcal{S}'|,\ Z_1 \subseteq \Gamma'\ \bar{Z}_1 \cap \Gamma' = \varnothing\right\}\right|} \\
&= \binom{|\mathcal{S}_1| - |Z_1| - |\bar{Z}_1|}{|\mathcal{S}_1|/2 - |Z_1|}^{-1} \binom{|\mathcal{S}_1| - |Z_1| - |\bar{Z}_1| - 1}{|\mathcal{S}_1|/2 - |Z_1| - 1} \\
&= \frac{|\mathcal{S}_1|/2 - |Z_1|}{|\mathcal{S}_1| - |Z_1| - |\bar{Z}_1|}.
\end{aligned}$$

By definition, we have $|Z_1| + |\bar{Z}_1| \le k$, and $|\mathcal{S}_1| = \frac{D-2d}{2}$. For $D \ge d + 8k$, this yields

$$\left|\mathbb{P}\left(\Gamma_1 \ni s \mid (s_i, s_i^+, R_i)_{i=1}^k\right) - \frac{1}{2}\right| \le \frac{4k}{D-2d} \le \frac{8k}{D-d}.$$

Similarly, by the transition model (40a) in our construction, for any $\Gamma_2 \subseteq \mathcal{S}_2$ with $|\Gamma_2| = \frac{1}{2}|\mathcal{S}_2|$, under the law $\mathbb{P}^{\otimes k}_{\Gamma_1,\Gamma_2,z}$, the observations $(s_i, s_i^+, R_i)_{i=1}^k$ have positive probability if and only if $Z_2 \subseteq \Gamma_2$ and $\Gamma_2 \cap \bar{Z}_2 = \varnothing$. Following exactly the same calculation as above, we arrive at the bound

$$\left|\mathbb{P}\left(\Gamma_2 \ni s \mid (s_i, s_i^+, R_i)_{i=1}^k\right) - \frac{1}{2}\right| \le \frac{8k}{D-d},$$

as desired. ■

E.1.2. Proof of Lemma 9

Under the conditional distribution $\mathbb{M}^{(k+1)}|(s_i, s_i^+, R_i)_{i=1}^k$, for each $s \in \mathcal{S}_1 \cup \mathcal{S}_2$, we have

$$\mathbb{P}\left(s_{k+1} = s\right) \le \frac{2}{|\mathcal{S}_1|}, \quad \text{and} \quad \mathbb{P}\left(s_{k+1}^+ = s\right) \le \frac{2}{|\mathcal{S}_1|}.$$

Applying a union bound, we arrive at the inequality

$$\begin{aligned}
&\mathbb{P}\left(\mathscr{E}_{k+1}^{(1)} \mid (s_i, s_i^+, R_i)_{i=1}^k\right) \\
&\le \sum_{\substack{i\in[k] \\ s_i\in\mathcal{S}_1\cup\mathcal{S}_2}} \left(\mathbb{P}\left(s_{k+1} = s_i\right) + \mathbb{P}\left(s_{k+1}^+ = s_i\right)\right) + \sum_{\substack{i\in[k] \\ s_i\in\mathcal{S}_1\cup\mathcal{S}_2}} \left(\mathbb{P}\left(s_{k+1} = s_i^+\right) + \mathbb{P}\left(s_{k+1}^+ = s_i^+\right)\right) \\
&\le \frac{8k}{|\mathcal{S}_1|} \le \frac{32k}{D-d},
\end{aligned}$$

which completes the proof. ■