A. Proofs of Theorems 1 and 2

In this section, we present the detailed proof of the main results of our paper, i.e., Theorem 1 and 2. We begin with the proof of Theorem 1.

Before the proof, we introduce some necessary notation. Let ϕ be the feature vector of a random state generated according to the stationary distribution π. In other words, $\phi = \phi(s_k)$ with probability π_{s_k}. Let ϕ' be the feature vector of the next state s' and let $r = r(s,s')$. Thus ϕ and ϕ' are random vectors and r is a random variable. As shown in Equation (2) of Bhandari et al. (2018), $\bar{g}(\theta)$ can be written as

$$\bar{g}(\theta) = E\left[\phi r\right] + E\left[\phi(\gamma\phi' - \phi)^T\right]\theta.$$

With these notation in place, we begin the proof of Theorem 1.

Proof of Theorem 1. Recall that θ^* is the unique vector with $\bar{g}(\theta^*) = 0$ (see Lemma 6 in (Tsitsiklis & Van Roy, 1997)). Consider

$$\bar{g}(\theta) = \bar{g}(\theta) - \bar{g}(\theta^*) = E\left[\phi(\gamma\phi' - \phi)^T\right](\theta - \theta^*). \quad (13)$$

To conclude that $\bar{g}(\theta)$ is a splitting of the gradient for a quadratic form $f(\theta)$, we need to calculate the gradient of $f(\theta)$. Let us begin with the Dirichlet norm and perform the following sequence of manipulations:

$$\begin{aligned}
&\|V_\theta - V_{\theta^*}\|^2_{\text{Dir}} \\
=&\frac{1}{2}\sum_{s,s'\in\mathcal{S}} \pi(s)P(s,s')\left[V_{\theta^*}(s) - V_\theta(s) - V_{\theta^*}(s') + V_\theta(s')\right]^2 \\
=&\frac{1}{2}\sum_{s,s'\in\mathcal{S}} \pi(s)P(s,s')\left[\left(V_{\theta^*}(s) - V_\theta(s)\right)^2 + \left(V_{\theta^*}(s') - V_\theta(s')\right)^2\right] \\
&- \sum_{s,s'\in\mathcal{S}} \pi(s)P(s,s')\left(V_{\theta^*}(s) - V_\theta(s)\right)\left(V_{\theta^*}(s') - V_\theta(s')\right) \\
=&\frac{1}{2}\sum_{s\in\mathcal{S}} \pi(s)\left(\sum_{s'\in\mathcal{S}} P(s,s')\right)\left(V_{\theta^*}(s) - V_\theta(s)\right)^2 \\
&+ \frac{1}{2}\sum_{s'\in\mathcal{S}}\left(\sum_{s\in\mathcal{S}} \pi(s)P(s,s')\right)\left(V_{\theta^*}(s') - V_\theta(s')\right)^2 \\
&- \sum_{s,s'\in\mathcal{S}} \pi(s)P(s,s')(\theta - \theta^*)^T\phi(s)\phi(s')^T(\theta - \theta^*) \\
=&\frac{1}{2}\sum_{s\in\mathcal{S}} \pi(s)\left(V_\theta(s) - V_{\theta^*}(s)\right)^2 + \frac{1}{2}\sum_{s'\in\mathcal{S}} \pi(s')\left(V_\theta(s') - V_{\theta^*}(s')\right)^2 \\
&- (\theta - \theta^*)^T E\left[\phi\phi'^T\right](\theta - \theta^*) \\
=&\|V_\theta - V_{\theta^*}\|^2_D - (\theta - \theta^*)^T E\left[\phi\phi'^T\right](\theta - \theta^*). \quad (14)
\end{aligned}$$

In the above sequence of equations, the first equality is just the definition of Dirichlet seminorm; the second equality follows by expanding the square; the third equality follows by interchanging sums and the definition of V_θ; the fourth equality uses that π is a stationary distribution of P, as well as the definition of ϕ and ϕ'; and the final equality uses the definition of the $||\cdot||_D$ norm.

Our next step is to use the identity we have just derived to rearrange the definition of $\|V_\theta - V_{\theta^*}\|^2_D$:

$$\begin{aligned}
&\|V_\theta - V_{\theta^*}\|^2_D = (V_\theta - V_{\theta^*})^T D(V_\theta - V_{\theta^*}) \\
=&(\theta - \theta^*)^T\Phi^T D\Phi(\theta - \theta^*) \\
=&(\theta - \theta^*)^T\sum_{s\in\mathcal{S}} \pi(s)\phi(s)\phi(s)^T(\theta - \theta^*) \\
=&(\theta - \theta^*)^T E[\phi\phi^T](\theta - \theta^*). \quad (15)
\end{aligned}$$

We now use these identities to write down a new expression for the function $f(\theta)$:

$$\begin{aligned}
f(\theta) =&(1-\gamma)||V_\theta - V_{\theta^*}||^2_{\text{D}} + \gamma||V_\theta - V_{\theta^*}||^2_{\text{Dir}} \\
=&(1-\gamma)||V_\theta - V_{\theta^*}||^2_{\text{D}} \\
&+ \gamma\left(||V_\theta - V_{\theta^*}||^2_{\text{D}} - (\theta - \theta^*)^T E\left[\phi\phi'^T\right](\theta - \theta^*)\right) \\
=&||V_\theta - V_{\theta^*}||^2_{\text{D}} - \gamma(\theta - \theta^*)^T E\left[\phi\phi'^T\right](\theta - \theta^*) \\
=&(\theta - \theta^*)^T E[\phi\phi^T](\theta - \theta^*) \\
&- \gamma(\theta - \theta^*)^T E\left[\phi\phi'^T\right](\theta - \theta^*) \\
=&(\theta - \theta^*)^T E\left[\phi(\phi - \gamma\phi')^T\right](\theta - \theta^*).
\end{aligned}$$

In the above sequence of equations, the first equality is just the definition of $f(\theta)$; the second equality is obtained by plugging in Eq. (14); the third equality is obtained by cancellation of terms; the fourth equality is obtained by plugging in Eq. (15); and the last step follows by merging the two terms together.

As a consequence of writing $f(\theta)$ this way, we can write down a new expression for the gradient of $f(\theta)$ directly:

$$\nabla f(\theta) = \left(E\left[\phi(\phi - \gamma\phi')^T\right] + E\left[(\phi - \gamma\phi')\phi^T\right]\right)(\theta - \theta^*). \quad (16)$$

Combining Equations (13) and (16), it is immediately that $-\bar{g}(\theta)$ is a splitting of $\nabla f(\theta)$. ■

We next turn to the proof of Theorem 2. Before beginning the proof, we introduce some notation.

The operator $T^{(\lambda)}$ is defined as:

$$\begin{aligned}
&\left(T^{(\lambda)}J\right)(s) \\
=&(1-\lambda)\sum_{m=0}^{\infty}\lambda^m E\left[\sum_{t=0}^{m}\gamma^t r(s_t, s_{t+1}) + \gamma^{m+1}J(s_{m+1})|s_0 = s\right] \quad (17)
\end{aligned}$$

for vectors $J \in \mathbb{R}^n$. The expectation is taken over sample paths taken by following actions according to policy μ;

recalling that this results in the transition matrix P, we can write this as

$$\begin{aligned} &T^{(\lambda)}J \\ =&(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}\gamma^t P^t R+(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}P^{m+1}J. \end{aligned} \tag{18}$$

We next devise new notation that is analogous to the TD(0) case. Let us denote the quantity $\delta_t z_t$ by $x(\theta_t, z_t)$ and its steady-state mean by $\bar{x}(\theta)$. It is known that

$$\bar{x}(\theta) = \Phi^T D\left(T^{(\lambda)}(\Phi\theta)-\Phi\theta\right), \tag{19}$$

see Lemma 8 of Tsitsiklis & Van Roy (1997); it also shown there that TD(λ) converges to a unique fixed point of a certain Bellman equation which we'll denote by θ^*_λ, and which satisfies

$$\bar{x}(\theta^*_\lambda) = 0. \tag{20}$$

With these preliminaries in place, we can begin the proof.

Proof of Theorem 2. By the properties of $T^{(\lambda)}$ and $\bar{x}(\theta)$ given in Equations (19) and (17) respectively, we have

$$\begin{aligned} \bar{x}(\theta) =&\bar{x}(\theta)-\bar{x}(\theta^*_\lambda) \\ =&\Phi^T D\left(T^{(\lambda)}(\Phi\theta)-\Phi\theta\right)-\Phi^T D\left(T^{(\lambda)}\left(\Phi\theta^*_\lambda\right)-\Phi\theta^*_\lambda\right) \\ =&\Phi^T D\left(T^{(\lambda)}(\Phi\theta)-T^{(\lambda)}(\Phi\theta^*_\lambda)-\Phi(\theta-\theta^*_\lambda)\right) \\ =&\left[(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\Phi^T DP^{m+1}\Phi-\Phi^T D\Phi\right](\theta-\theta^*_\lambda), \end{aligned} \tag{21}$$

where the last line used Eq. (18).

Our next step is to derive a convenient expression for $f^{(\lambda)}(\theta)$. We begin by finding a clean expression for the Dirichlet form that appears in the definition of $f^{(\lambda)}(\theta)$:

$$\begin{aligned} &\|V_\theta - V_{\theta^*_\lambda}\|^2_{\mathrm{Dir},m+1} \\ =&\frac{1}{2}\sum_{s,s'\in\mathcal{S}}\pi_s P^{m+1}(s,s')(V_\theta(s)-V_{\theta^*_\lambda}(s)-V_\theta(s')+V_{\theta^*_\lambda}(s'))^2 \\ =&\frac{1}{2}\sum_{s,s'\in\mathcal{S}}\pi_s P^{m+1}(s,s')\left[(V_\theta(s)-V_{\theta^*_\lambda}(s))^2+(V_\theta(s')-V_{\theta^*_\lambda}(s'))^2\right] \\ &-\sum_{s,s'\in\mathcal{S}}\pi_s P^{m+1}(s,s')\left(V_\theta(s)-V_{\theta^*_\lambda}(s)\right)\left(V_\theta(s')-V_{\theta^*_\lambda}(s')\right) \\ =&\frac{1}{2}\sum_{s\in\mathcal{S}}\pi_s\left(\sum_{s'\in\mathcal{S}}P^{m+1}(s,s')\right)(V_\theta(s)-V_{\theta^*_\lambda}(s))^2 \\ &+\frac{1}{2}\sum_{s'\in\mathcal{S}}\left(\sum_{s\in\mathcal{S}}\pi_s P^{m+1}(s,s')\right)(V_\theta(s')-V_{\theta^*_\lambda}(s'))^2 \\ &-\sum_{s,s'\in\mathcal{S}}\pi_s P^{m+1}(s,s')\left(V_\theta(s)-V_{\theta^*_\lambda}(s)\right)\left(V_\theta(s')-V_{\theta^*_\lambda}(s')\right) \\ =&\frac{1}{2}\sum_{s\in\mathcal{S}}\pi_s(V_\theta(s)-V_{\theta^*_\lambda}(s))^2+\frac{1}{2}\sum_{s'\in\mathcal{S}}\pi_{s'}(V_\theta(s')-V_{\theta^*_\lambda}(s'))^2 \\ &-\sum_{s,s'\in\mathcal{S}}\pi_s P^{m+1}(s,s')\left(V_\theta(s)-V_{\theta^*_\lambda}(s)\right)\left(V_\theta(s')-V_{\theta^*_\lambda}(s')\right) \\ =&\sum_{s\in\mathcal{S}}\pi_s(V_\theta(s)-V_{\theta^*_\lambda}(s))^2 \\ &-\sum_{s\in\mathcal{S}}\pi_s\left(V_\theta(s)-V_{\theta^*_\lambda}(s)\right)\sum_{s'\in\mathcal{S}}P^{m+1}(s,s')\left(V_\theta(s')-V_{\theta^*_\lambda}(s')\right) \\ =&(\theta-\theta^*_\lambda)^T\left(\Phi^T D\Phi-\Phi^T DP^{m+1}\Phi\right)(\theta-\theta^*_\lambda). \end{aligned} \tag{22}$$

In the above sequence of equations, the first equality follows by the definition of the $m+1$-Dirichlet norm; the second equality follows by expanding the square; the third equality follows by interchanging the order of summations; the fourth equality uses that any power of a stochastic matrix is stochastic, and the $\pi P^{m+1}=\pi$; the fifth equality combines terms and rearranges the order of summation; and the last line uses the definition $V_\theta = \Phi\theta$.

We'll also make use of the obvious identity

$$\|V_\theta - V_{\theta^*_\lambda}\|^2_D = (\theta-\theta^*_\lambda)^T\Phi^T D\Phi(\theta-\theta^*_\lambda). \tag{23}$$

Putting all these together, we can express the function $f^{(\lambda)}(\theta)$ as:

$$\begin{aligned} &f^{(\lambda)}(\theta) \\ =&(1-\gamma\kappa)||V_\theta - V_{\theta^*_\lambda}||^2_D \\ &+(1-\lambda)\sum_{m=0}^{+\infty}\lambda^m\gamma^{m+1}||V_\theta - V_{\theta^*_\lambda}||^2_{\mathrm{Dir,m+1}} \\ =&(\theta-\theta^*_\lambda)^T\left[(1-\gamma\kappa)\Phi^T D\Phi\right. \\ &\left.+(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\Phi^T D(I-P^{m+1})\Phi\right](\theta-\theta^*_\lambda) \\ =&(\theta-\theta^*_\lambda)^T\left[\left((1-\gamma\kappa)+(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\right)\Phi^T D\Phi\right. \\ &\left.-(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\Phi^T DP^{m+1}\Phi\right](\theta-\theta^*_\lambda) \\ =&(\theta-\theta^*_\lambda)^T\left[\left((1-\gamma\kappa)+\gamma\frac{1-\lambda}{1-\gamma\lambda}\right)\Phi^T D\Phi\right. \\ &\left.-(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\Phi^T DP^{m+1}\Phi\right](\theta-\theta^*_\lambda) \\ =&(\theta-\theta^*_\lambda)^T\left[\Phi^T D\Phi\right. \\ &\left.-(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\Phi^T DP^{m+1}\Phi\right](\theta-\theta^*_\lambda). \end{aligned}$$

In the above sequence of equations, the first equality is from the definition of the function $f^{(\lambda)}(\theta)$; the second line comes from plugging in Eq. (23) and Eq. (22); the third equality

from breaking the sum in the second term into two pieces, one of which is then absorbed into the first term; the fourth equality follows by using the sum of a geometric series; and the last equality by the definition of κ from the theorem statement, which, recall, is $\kappa = (1-\lambda)/(1-\gamma\lambda)$.

By comparing the expression for $f^{(\lambda)}(\theta)$ we have just derived to Eq. (21), it is immediate that $-\bar{x}(\theta)$ is a splitting of the gradient of $f^{(\lambda)}(\theta)$. ■

B. Proof of Corollary 2

We will find it convenient to use several observations made in (Bhandari et al., 2018). First, Lemma 6 of that paper says that, under the assumptions of Corollary 2, we have that

$$\|g_t(\theta_t)\|_2 \leq G = r_{\max} + 2R_\theta. \tag{24}$$

This holds with probability one; note, however, that because the number of states and actions is finite, this just means one takes the maximum over all states and actions to obtain this upper bound.

A second lemma from (Bhandari et al., 2018) deals with a measure of "gradient bias," the quantity $\zeta_t(\theta) = (\bar{g}(\theta) - g_t(\theta))^T(\theta^* - \theta)$. As should be unsurprising, what matters in the analysis is not the natural measure of gradient bias, e.g., $\bar{g}(\theta) - g_t(\theta)$, but rather how the angle with the direction to the optimal solution is affected, which is precisely what is measured by $\zeta_t(\theta)$. We have the following upper bound.

Lemma 1 (Lemma 11 in (Bhandari et al., 2018))**.** *Consider a non-increasing step-size sequence, $\alpha_0 \geq \alpha_1 \geq \cdots \geq \alpha_T$. Fix any $t < T$, and set $t^* = \max\{0, t - \tau^{\text{mix}}(\alpha_T)\}$. Then*

$$E[\zeta_t(\theta_t)] \leq G^2\left(4 + 6\tau^{\text{mix}}(\alpha_T)\right)\alpha_{t^*}.$$

With these preliminaries in place, we are now ready to prove the corollary. The proof follows the steps of (Sun et al., 2018) to analyze Markov gradient descent, using the fact that the gradient splitting has the same inner product with the direction to the optimal solution as the gradient.

Proof of Corollary 2. From the projected TD(0) recursion, for any t,

$$\begin{aligned}
&\|\theta^* - \theta_{t+1}\|_2^2 \\
=&\|\theta^* - \text{Proj}_\Theta(\theta_t + \alpha_t g_t(\theta_t))\|_2^2 \\
\leq&\|\theta^* - \theta_t - \alpha_t g_t(\theta_t)\|_2^2 \\
=&\|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^T(\theta^* - \theta_t) + \alpha_t^2\|g_t(\theta_t)\|_2^2 \\
=&\|\theta^* - \theta_t\|_2^2 - 2\alpha_t\left[\bar{g}(\theta_t)^T - (\bar{g}(\theta_t) - g_t(\theta_t))^T\right](\theta^* - \theta_t) \\
&+ \alpha_t^2\|g_t(\theta_t)\|_2^2 \\
\leq&\|\theta^* - \theta_t\|_2^2 - 2\alpha_t\bar{g}(\theta_t)^T(\theta^* - \theta_t) + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.
\end{aligned}$$

In the above sequence of equations, all the equalities are just rearrangements of terms; whereas the first inequality follows that the projection onto a convex set does not increase distance, while the second inequality follows by Eq. (24).

Next we use Corollary 1, rearrange terms, and sum from $t = 0$ to $t = T-1$:

$$\begin{aligned}
&\sum_{t=0}^{T-1} 2\alpha_t E\left[(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + \gamma\|V_{\theta^*} - V_{\theta_t}\|_{\text{Dir}}^2\right] \\
\leq&\sum_{t=0}^{T-1}\left(E\left[\|\theta^* - \theta_t\|_2^2\right] - E\left[\|\theta^* - \theta_{t+1}\|_2^2\right]\right) + \sum_{t=0}^{T-1}\alpha_t^2 G^2 \\
&+ \sum_{t=0}^{T-1} 2\alpha_t E\left[\zeta_t(\theta_t)\right] \\
=&\left(\|\theta^* - \theta_0\|_2^2 - E\left[\|\theta^* - \theta_T\|_2^2\right]\right) + \sum_{t=0}^{T-1}\alpha_t^2 G^2 \\
&+ \sum_{t=0}^{T-1} 2\alpha_t E\left[\zeta_t(\theta_t)\right] \\
\leq&\|\theta^* - \theta_0\|_2^2 + \sum_{t=0}^{T-1} 2\alpha_t E\left[\zeta_t(\theta_t)\right] + \sum_{t=0}^{T-1}\alpha_t^2 G^2.
\end{aligned}$$

Now plugging in the step-sizes $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$, it is immediate that

$$\begin{aligned}
&\sum_{t=0}^{T-1} E\left[(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + \gamma\|V_{\theta^*} - V_{\theta_t}\|_{\text{Dir}}^2\right] \\
\leq&\frac{\sqrt{T}}{2}\left(\|\theta^* - \theta_0\|_2^2 + G^2\right) + \sum_{t=0}^{T-1} E\left[\zeta_t(\theta_t)\right].
\end{aligned}$$

Using Lemma 1, have that

$$\begin{aligned}
\sum_{t=0}^{T-1} E\left[\zeta_t(\theta_t)\right] &\leq \sum_{t=0}^{T-1} G^2\left(4 + 6\tau^{\text{mix}}(\alpha_T)\right)\alpha_{t^*} \\
&= \sqrt{T}G^2\left(4 + 6\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right).
\end{aligned}$$

Putting all this together and using the convexity of the function $f(\theta)$, we can bound the error at the average iterate as:

$$\begin{aligned}
&E\left[(1-\gamma)\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 + \gamma\|V_{\theta^*} - V_{\bar{\theta}_T}\|_{\text{Dir}}^2\right] \\
\leq&\frac{1}{T}\sum_{t=0}^{T-1} E\left[(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + \gamma\|V_{\theta^*} - V_{\theta_t}\|_{\text{Dir}}^2\right] \\
\leq&\frac{\|\theta^* - \theta_0\|_2^2 + G^2}{2\sqrt{T}} + \frac{G^2\left(4 + 6\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)}{\sqrt{T}} \\
=&\frac{\|\theta^* - \theta_0\|_2^2 + G^2\left(9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)}{2\sqrt{T}}.
\end{aligned}$$

■

C. Proof of Corollary 3

Before starting the proof, we will need a collection of definitions, observations, and preliminary lemmas. We organize these into subheadings below.

The Dirichlet Laplacian. Let $L = (L(i,j))_{n\times n}$ be a symmetric matrix in $\mathbb{R}^{n\times n}$ defined as

$$L(i,j) = \begin{cases} -(1/2)\left(\pi_i P(i,j) + \pi_j P(j,i)\right) & \text{if } i \neq j \\ \sum_{i'\neq i} |L(i,i')| & \text{if } i = j \end{cases}.$$

It is immediate that the diagonal elements of L are positive and its rows sum to zero.

Furthermore, it can be shown that for any vector x, we have that $\|x\|^2_{\text{Dir}} = x^T L x$. Indeed:

$$\begin{aligned} x^T L x &= \sum_{i=1}^{n} \left[\sum_{j\neq i} -\frac{1}{2}\left(\pi_i P(i,j) + \pi_j P(j,i)\right) x(i)x(j) \right. \\ &\quad \left. + \left(\sum_{j\neq i} \frac{1}{2}\left(\pi_i P(i,j) + \pi_j P(j,i)\right) \right) x(i)^2 \right] \\ &= \sum_{i<j} \frac{1}{2}\left(\pi_i P(i,j) + \pi_j P(j,i)\right)\left(x(i) - x(j)\right)^2 \\ &= \frac{1}{2} \sum_{i,j\in[n]} \frac{1}{2}\left(\pi_i P(i,j) + \pi_j P(j,i)\right)\left(x(i) - x(j)\right)^2 \\ &= \frac{1}{2} \sum_{i,j\in[n]} \pi_i P(i,j)(x(i) - x(j))^2 = \|x\|^2_{\text{Dir}}. \end{aligned}$$

Connection to the reversed chain. We remark that the matrix L is connected to the so-called "additive reversibilization" of the matrix P, which we explain next. For a stochastic matrix P with stationary distribution π, it is natural to define the matrix P^* as

$$[P^*]_{ij} = \frac{\pi(j)}{\pi(i)} P_{ji}.$$

It is possible to verify that the matrix P^* has the same stationary distribution as the matrix P (see Aldous & Fill (1995)). Intuitively, the equality

$$\pi(i)[P^*]_{ij} = \pi(j)P_{ji},$$

means that it is natural to interpret P^* as the "reversed" chain of P: for all pairs i, j, the link from i to j is traversed as often under the stationary distribution in P^* as the link from j to i in P.

It can then be shown that the matrix $Q = (P + P^*)/2$ is reversible (see (Aldous & Fill, 1995)); this matrix is called the "additive reversibilization" of the matrix P. It is easy to see that $Q = I - D^{-1}L$; indeed, both the left-hand side and the right-hand side have the same off-diagonal entries and have rows that sum to one. Because Q is reversible, its spectrum is real.

The matrix $D^{-1}L$ is clearly similar to the symmetric matrix $D^{-1/2}LD^{-1/2}$ and thus has a real spectrum, with all the eigenvalues nonnegative. Moreover, $D^{-1}L$ has an eigenvalue of zero as $D^{-1}L\mathbf{1} = 0$. As a consequence of these two observations, if we denote by $r(P)$ the spectral gap of the matrix Q, then we have

$$r(P) = \frac{1}{1-\lambda_2(Q)} = \frac{1}{\lambda_{n-1}(D^{-1}L)}, \tag{25}$$

where $\lambda_{n-1}(D^{-1}L)$ is the second smallest eigenvalue of $D^{-1}L$.

Equivalence of norms on $\mathbf{1}^\perp$. We will need to pass between the $||\cdot||_D$ norm and the $||\cdot||_{\text{Dir}}$ norm. To that end, we have the following lemma.

Lemma 2. *For any x with $\langle x, \mathbf{1}\rangle_D = 0$, we have that*

$$\|x\|^2_D \leq r(P)\|x\|^2_{\text{Dir}}.$$

Proof. Indeed,

$$\min_{\langle x,\mathbf{1}\rangle_D=0} \frac{\|x\|^2_{\text{Dir}}}{\|x\|^2_D} = \min_{\langle x,\mathbf{1}\rangle_D=0} \frac{x^T L x}{\langle x,x\rangle_D} = \min_{\langle x,\mathbf{1}\rangle_D=0} \frac{\langle x, D^{-1}Lx\rangle_D}{\langle x,x\rangle_D}.$$

We next observe that the matrix $D^{-1}L$ is self adjoint in the $\langle\cdot,\cdot\rangle_D$ inner product:

$$\langle x, D^{-1}Ly\rangle_D = x^T L y = \langle D^{-1}Lx, y\rangle_D.$$

Since the smallest eigenvalue of $D^{-1}L$ is zero with associated eigenvector of $\mathbf{1}$, by the Rayleigh-Ritz theorem we have

$$\min_{\langle x,\mathbf{1}\rangle_D=0} \frac{\langle x, D^{-1}Lx\rangle_D}{\langle x,x\rangle_D} = \lambda_{n-1}(D^{-1}L).$$

Putting it all together, we obtain

$$\frac{\|x\|^2_{\text{Dir}}}{\|x\|^2_D} \geq \lambda_{n-1}(D^{-1}L) = r(P)^{-1},$$

where the last step used Eq. (25). This completes the proof. ■

Error in mean estimation.

Recall that we set $\hat{V}_T$ be an estimate for the mean of value function in Algorithm 1. Our next lemma upper bounds the error in the estimate $\hat{V}_T$.

Lemma 3. *Suppose that $\hat{V}_T$ is generated by Algorithm 1 and $\bar{V} = \pi^T V$ denote the mean of value function. Let $t_0 = \max\left\{t \in \mathbb{N} | t_0 \leq 2\tau^{\text{mix}}\left(\frac{1}{2(t_0+1)}\right)\right\}$. Then, for $t > t_0$, we have*

$$E\left[\left(\hat{V}_t - \bar{V}\right)^2\right] \leq O\left(\frac{r^2_{\max}\tau^{\text{mix}}\left(\frac{1}{2(t+1)}\right)}{(1-\gamma)^2 t}\right).$$

Proof. By the definition of $\hat{V}_t$ and $\bar{A}_t$ given in Algorithm 1, we can write the recursion:

$$\begin{aligned}\hat{V}_t =& \frac{\bar{A}_t}{1-\gamma} = \frac{1}{1-\gamma}\left[\bar{A}_{t-1} + \frac{1}{t+1}(r_t - \bar{A}_{t-1})\right] \\ =& \hat{V}_{t-1} + \frac{1}{t+1}\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right).\end{aligned}$$

We next use this recursion to argue:

$$\begin{aligned}& E\left[\left(\hat{V}_t - \bar{V}\right)^2\right] \\ =& E\left[\left(\hat{V}_{t-1} + \frac{1}{t+1}\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right) - \bar{V}\right)^2\right] \\ =& E\left[\left(\hat{V}_{t-1} - \bar{V}\right)^2 + \frac{1}{(t+1)^2}\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right)^2\right] \\ &+ E\left[\frac{2}{t+1}\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right)\left(\hat{V}_{t-1} - \bar{V}\right)\right] \\ =& E\left[\left(\hat{V}_{t-1} - \bar{V}\right)^2 + \frac{1}{(t+1)^2}\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right)^2\right] \\ &+ E\left[\frac{2}{t+1}\left(\frac{r_t}{1-\gamma} - \bar{V} - \hat{V}_{t-1} + \bar{V}\right)\left(\hat{V}_{t-1} - \bar{V}\right)\right] \\ =& E\left[\left(1 - \frac{2}{t+1}\right)\left(\hat{V}_{t-1} - \bar{V}\right)^2 + \frac{1}{(t+1)^2}\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right)^2\right] \\ &+ E\left[\frac{2}{t+1}\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1} - \bar{V}\right)\right]. \qquad (26)\end{aligned}$$

To bound the second term on the right-hand side of Eq. (26), we will use that, since $r_{\max}$ is the upper bound on absolute values of the rewards, we have that

$$\left(\frac{r_t}{1-\gamma} - \hat{V}_{t-1}\right)^2 \le \left(\frac{r_{\max}}{1-\gamma} + \frac{r_{\max}}{1-\gamma}\right)^2 = \frac{4r_{\max}^2}{(1-\gamma)^2}.$$

We next analyze the third term on the right-hand side of Eq. (26). Let $\tau_t = \tau^{\text{mix}}\left(\frac{1}{2(t+1)}\right)$ so that for any state s'',

$$\sum_{s=1}^{n} |P^{\tau_t}(s'', s) - \pi_s| = 2d_{\text{TV}}(P^{\tau_t}(s'', \cdot), \pi) \le 2m\rho^{\tau_t} \le \frac{1}{t+1}. \qquad (27)$$

We have that

$$\begin{aligned}& E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1} - \bar{V}\right)\right] \\ =& E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1} - \hat{V}_{t-1-\tau_t} + \hat{V}_{t-1-\tau_t} - \bar{V}\right)\right] \\ =& E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1} - \hat{V}_{t-1-\tau_t}\right)\right] \\ &+ E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\right].\end{aligned}$$

We now bound each of the two terms in the last equation separately. For the first term, we have

$$\begin{aligned}& E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1} - \hat{V}_{t-1-\tau_t}\right)\right] \\ \le& \frac{2r_{\max}}{1-\gamma}\sum_{d=t-\tau_t}^{t-1} E\left[\left|\hat{V}_d - \hat{V}_{d-1}\right|\right] \\ =& \frac{2r_{\max}}{1-\gamma}\sum_{d=t-\tau_t}^{t-1}\frac{1}{d+1}E\left[\left|\frac{r_d}{1-\gamma} - \hat{V}_{d-1}\right|\right] \\ \le& \frac{4r_{\max}^2}{(1-\gamma)^2}\sum_{d=t-\tau_t}^{t-1}\frac{1}{d+1} \\ \le& O\left(\frac{\tau_t r_{\max}^2}{(1-\gamma)^2(t+1)}\right),\end{aligned}$$

where the last inequality follows from $t > 2\tau_t$ (which in turn follows from $t \ge t_0$).

For the second term, we denote the following sigma algebra χ^t denote the sigma algebra generated by the information collected by time t, i.e., by the random variables $s_0, r_0, \theta_0, \cdots, s_t, r_t, \theta_t$. We then have that

$$\begin{aligned}& E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\right] \\ =& E\left[E\left[\left(\frac{r_t}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)|\chi^{t-1-\tau_t}\right]\right] \\ =& E\left[\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\right. \\ &\qquad \left.\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)P^{\tau_t}(s_{t-1-\tau_t}, s)\right] \\ =& E\left[\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\right. \\ &\qquad \left.\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\left(P^{\tau_t}(s_{t-1-\tau_t}, s) - \pi_s + \pi_s\right)\right] \\ =& E\left[\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\right. \\ &\qquad \left.\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\left(P^{\tau_t}(s_{t-1-\tau_t}, s) - \pi_s\right)\right] \\ &+ E\left[\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\pi_s\right] \\ =& E\left[\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\right. \\ &\qquad \left.\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\left(P^{\tau_t}(s_{t-1-\tau_t}, s) - \pi_s\right)\right] \\ &+ E\left[\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\right]\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\pi_s \\ =& E\left[\sum_{s=1}^{n}\left(\frac{\sum_{s'=1}^{n} P(s,s')r(s,s')}{1-\gamma} - \bar{V}\right)\right.\end{aligned}$$

$$\left(\hat{V}_{t-1-\tau_t} - \bar{V}\right)\left(P^{\tau_t}(s_{t-1-\tau_t}, s) - \pi_s\right)\Bigg] + 0$$
$$\leq \frac{4r_{\max}^2}{(1-\gamma)^2(t+1)}$$
$$\leq O\left(\frac{r_{\max}^2}{(1-\gamma)^2(t+1)}\right).$$

Here the first equality follows by iterating conditional expectation; the second, third, fourth, and fifth equality is just rearranging terms; the sixth equality follows from Eq. (12); and the next inequality follows from Eq. (27) as well as the fact that all rewards are upper bounded by $r_{\max}$ in absolute value.

Combining all the inequalities, we can conclude that as long as $t > t_0$, we have that

$$E\left[\left(\hat{V}_t - \bar{V}\right)^2\right]$$
$$\leq \left(1 - \frac{2}{t+1}\right) E\left[\left(\hat{V}_{t-1} - \bar{V}\right)^2\right] + O\left(\frac{\tau_t r_{\max}^2}{(1-\gamma)^2(t+1)^2}\right).$$

Let $b_t = O\left(\frac{\tau_t r_{\max}^2}{(1-\gamma)^2}\right)$; then the above equation can be compactly written as

$$E\left[\left(\hat{V}_t - \bar{V}\right)^2\right] \leq \left(1 - \frac{2}{t+1}\right) E\left[\left(\hat{V}_{t-1} - \bar{V}\right)^2\right] + \frac{b_t}{(t+1)^2}.$$

Let $C_t = \max\left\{(t_0+1)\left(\hat{V}_{t_0} - \bar{V}\right)^2, b_t\right\}$. We will prove by induction that $t \geq t_0$,

$$E\left[\left(\hat{V}_t - \bar{V}\right)^2\right] \leq \frac{C_t}{t+1}.$$

Indeed, the assertion holds for $t = t_0$. Suppose that the assertion holds at time t, i.e., suppose that $E\left[\left(\hat{V}_t - \bar{V}\right)^2\right] \leq C_t/(t+1)$. Then,

$$E\left[\left(\hat{V}_{t+1} - \bar{V}\right)^2\right] \leq \left(1 - \frac{2}{t+2}\right)\frac{C_t}{t+1} + \frac{b_t}{(t+2)^2}$$
$$= \frac{C_{t+1}}{t+2} + \left(1 - \frac{2}{t+2}\right)\frac{C_t}{t+1} + \frac{b_t}{(t+2)^2} - \frac{C_{t+1}}{t+2}$$
$$= \frac{C_t(t+2)^2 - 2C_t(t+2) + b_t(t+1) - C_{t+1}(t+1)(t+2)}{(t+1)(t+2)^2} + \frac{C_{t+1}}{t+2}$$
$$= \frac{(C_t - C_{t+1})(t+1)(t+2) + (b_t - C_t)(t+1) - C_t}{(t+1)(t+2)^2} + \frac{C_{t+1}}{t+2}$$
$$\leq \frac{C_{t+1}}{t+2},$$

where the last inequality follows because $C_t \leq C_{t+1}$, $b_t \leq C_t$ and $C_t \geq 0$. Therefore, we have that, for $t \geq t_0$,

$$E\left[\left(\hat{V}_t - \bar{V}\right)^2\right] \leq \frac{C_t}{t+1}.$$

Since $\left(\hat{V}_{t_0} - \bar{V}\right)^2 \leq 4\frac{r_{\max}^2}{(1-\gamma)^2}$ with probability one, and by definition $t_0 \leq 2\tau^{\max}\left(\frac{1}{2(t_0+1)}\right)$, we have that $C_t = O\left(\frac{\tau_t r_{\max}^2}{(1-\gamma)^2}\right)$ for $t \geq t_0$; this completes the proof. ■

With all these preliminary lemmas in place, we can now give the main result of this section, the proof of Corollary 3.

Proof of Corollary 3. By the Pythagorean theorem, we have

$$\|V'_T - V\|_D^2 = \|\pi^T V'_T \mathbf{1} - \pi^T V \mathbf{1}\|_D^2 + \|V'_{T,\mathbf{1}^\perp} - V_{\mathbf{1}^\perp}\|_D^2, \quad (28)$$

where $V'_{T,\mathbf{1}^\perp}, V_{\mathbf{1}^\perp}$ are the projections of V'_T, V onto $\mathbf{1}^\perp$ in the $\langle\cdot,\cdot\rangle_D$ inner product.

Recall, that, in Algorithm 1, we defined

$$V'_T = V_{\tilde{\theta}_T} + \mathbf{1}\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right).$$

Therefore,

$$\pi^T V'_T \mathbf{1} = \pi^T V_{\tilde{\theta}_T}\mathbf{1} + \pi^T \mathbf{1}\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\mathbf{1}$$
$$= \pi^T V_{\tilde{\theta}_T}\mathbf{1} + \hat{V}_T \mathbf{1} - \pi^T V_{\tilde{\theta}_T}\mathbf{1}$$
$$= \hat{V}_T \mathbf{1}.$$

Plugging this as well as $\bar{V} = \pi^T V$ into Eq. (28) we obtain:

$$\|V'_T - V\|_D^2 = \|\hat{V}_T \mathbf{1} - \bar{V}\mathbf{1}\|_D^2 + \|V'_{T,\mathbf{1}^\perp} - V_{\mathbf{1}^\perp}\|_D^2. \quad (29)$$

For the first term on the right hand side of Eq. (29), by the definition of the square norm under π, it is immediate that

$$\|\hat{V}_T \mathbf{1} - \bar{V}\mathbf{1}\|_D^2 = \sum_{i=1}^n \pi_i (\hat{V}_T - \bar{V})^2 = \left(\hat{V}_T - \bar{V}\right)^2.$$

For the second term on the right hand side of Eq. (29), we have

$$\|V'_{T,\mathbf{1}^\perp} - V_{\mathbf{1}^\perp}\|_D^2$$
$$= \|V'_{T,\mathbf{1}^\perp} - V_{\theta^*,\mathbf{1}^\perp} + V_{\theta^*,\mathbf{1}^\perp} - V_{\mathbf{1}^\perp}\|_D^2$$
$$\leq 2\|V'_{T,\mathbf{1}^\perp} - V_{\theta^*,\mathbf{1}^\perp}\|_D^2 + 2\|V_{\theta^*,\mathbf{1}^\perp} - V_{\mathbf{1}^\perp}\|_D^2$$
$$\leq 2r(P)\|V'_{T,\mathbf{1}^\perp} - V_{\theta^*,\mathbf{1}^\perp}\|_{\mathrm{Dir}}^2 + 2\|V_{\theta^*} - V\|_D^2$$
$$= 2r(P)\|V_{\tilde{\theta}_T} - V_{\theta^*}\|_{\mathrm{Dir}}^2 + 2\|V_{\theta^*} - V\|_D^2,$$

where the third line follows by the Lemma 2 and the Pythagorean theorem and the fourth line comes from the

observation that $\|\cdot\|_{\text{Dir}}$ does not change when we add a multiple of $\mathbf{1}$.

Combining these results and taking expectation of Eq. (29), we obtain

$$\begin{aligned}
&E\left[\|V'_T - V\|_D^2\right] \\
\leq& E\left[(\hat{V}_T - \bar{V})^2\right] + 2r(P)E\left[\|V_{\tilde{\theta}_T} - V_{\theta^*}\|_{\text{Dir}}^2\right] \\
&+ 2E\left[\|V_{\theta^*} - V\|_D^2\right] \\
\leq& O\left(\frac{\tau^{\text{mix}}\left(\frac{1}{2(T+1)}\right) r_{\max}^2}{(1-\gamma)^2 T}\right) + 2E\left[\|V_{\theta^*} - V\|_D^2\right] \\
&+ r(P)\cdot\frac{\|\theta^* - \theta_0\|_2^2 + \left(9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)G^2}{\gamma\sqrt{T}}, \quad (30)
\end{aligned}$$

where the second inequality follows by Lemma 3 and Eq. (11) from the main text.

On the other hand,

$$\begin{aligned}
&E\left[\|V'_T - V\|_D^2\right] \\
=& E\left[\|V_{\tilde{\theta}_T} - V\|_D^2\right] + E\left[\left\|\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\mathbf{1}\right\|_D^2\right] \\
&- 2E\left[\left(\pi^T V - \pi^T V_{\tilde{\theta}_T}\right)\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\right] \\
=& E\left[\|V_{\tilde{\theta}_T} - V\|_D^2\right] + E\left[\left\|\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\mathbf{1}\right\|_D^2\right] \\
&- 2E\left[\left(\bar{V} - \pi^T V_{\tilde{\theta}_T}\right)\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\right] \\
=& E\left[\|V_{\tilde{\theta}_T} - V\|_D^2\right] + E\left[\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)^2\right] \\
&- 2E\left[\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T} + \bar{V} - \hat{V}_T\right)\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\right] \\
=& E\left[\|V_{\tilde{\theta}_T} - V\|_D^2\right] - E\left[\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)^2\right] \\
&+ 2E\left[\left(\hat{V}_T - \bar{V}\right)\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)\right] \\
\leq& E\left[\|V_{\tilde{\theta}_T} - V\|_D^2\right] - E\left[\left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)^2\right] \\
&+ E\left[\left(\hat{V}_T - \bar{V}\right)^2 + \left(\hat{V}_T - \pi^T V_{\tilde{\theta}_T}\right)^2\right] \\
=& E\left[\|V_{\tilde{\theta}_T} - V\|_D^2\right] + E\left[\left(\hat{V}_T - \bar{V}\right)^2\right] \\
\leq& 2E\left[\|V_{\tilde{\theta}_T} - V_{\theta^*}\|_D^2\right] + 2E\left[\|V_{\theta^*} - V\|_D^2\right] + E\left[\left(\hat{V}_T - \bar{V}\right)^2\right] \\
\leq& \frac{2\left[\|\theta^* - \theta_0\|_2^2 + \left(9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)G^2\right]}{(1-\gamma)\sqrt{T}} \\
&+ 2E\left[\|V_{\theta^*} - V\|_D^2\right] + O\left(\frac{r_{\max}^2\tau^{\text{mix}}\left(\frac{1}{2(T+1)}\right)}{(1-\gamma)^2 T}\right). \quad (31)
\end{aligned}$$

Here the first four equalities come from rearranging; the next inequality comes from the identity $2ab \leq a^2 + b^2$); the next equality comes from cancellation; the next inequality uses $||u+v||_D^2 \leq 2||u||_D^2 + 2||v||_D^2$; and the final inequality uses Corollary 2 and Lemma 3.

We have just derived two bounds on $E[||V'_T - V||_D^2]$, one in Eq. (30) and one in Eq. (31). We could, of course, take the minimum of these two bounds. We then obtain:

$$\begin{aligned}
&E\left[\|V'_T - V\|_D^2\right] \\
\leq& 2E\left[\|V_{\theta^*} - V\|_D^2\right] + O\left(\frac{\tau^{\text{mix}}\left(\frac{1}{2(T+1)}\right) r_{\max}^2}{(1-\gamma)^2 T}\right) \\
&+ \min\left\{ r(P)\cdot\frac{\|\theta^* - \theta_0\|_2^2 + \left(9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)G^2}{\gamma\sqrt{T}}, \right. \\
&\left. \frac{2\|\theta^* - \theta_0\|_2^2 + 2\left(9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)G^2}{(1-\gamma)\sqrt{T}} \right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E\left[\|V'_T - V\|_D^2\right] \\
\leq& 2E\left[\|V_{\theta^*} - V\|_D^2\right] + O\left(\frac{\tau^{\text{mix}}\left(\frac{1}{T+1}\right) r_{\max}^2}{(1-\gamma)^2 T}\right) \\
&+ \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left[1 + \tau^{\text{mix}}(1/\sqrt{T})\right]}{\sqrt{T}}\cdot\min\left\{\frac{r(P)}{\gamma}, \frac{2}{1-\gamma}\right\},
\end{aligned}$$

and the proof is complete. ■

D. Error Bound for TD with Eligibility Traces

We now analyze the performance of projected TD(λ) which updates as

$$\theta_{t+1} = \text{Proj}_{\Theta_\lambda}(\theta_t + \alpha_t\delta_t\hat{z}_t), \quad (32)$$

where we now use

$$z_t = \sum_{k=0}^{t}(\gamma\lambda)^k\phi(s_{t-k}).$$

We remark that this is an abuse of notation, as previously z_t was defined with the sum starting at negative infinity, rather than zero; however, in this section, we will assume that the sum starts at zero. The consequence of this modification of notation is that Theorem 4 does not imply that $-E[z_t]$ is the gradient splitting of an appropriately defined function anymore, as now one needs to account for the error term coming from the beginning of the sum.

We assume Θ_λ is a convex set containing the optimal solution θ_λ^*. We will further assume that the norm of every element in Θ_λ is at most R_λ. Recall that

We begin by introducing some notation. Much of our analysis follows (Bhandari et al., 2018) with some deviations where we appeal to Theorem 4, and the notation bellow is mostly identical to what is used in that paper. First, recall that we denote the quantity $\delta_t z_t$ by $x(\theta_t, z_t)$. We define $\zeta_t(\theta_t, z_t)$ as a random variable which can be thought of as a measure of the bias that TD(λ) has in estimation of the gradient:

$$\zeta_t(\theta_t, z_t) = (\bar{x}(\theta_t) - \delta_t z_t)^T(\theta_\lambda^* - \theta_t).$$

Analogously to the TD(0) case, what turns out to matter for our analysis is not so much the bias per se, but the inner product of the bias with the direction of the optimal solution as in the definition of $\zeta_t(\theta_t, z_t)$.

We will next need an upper bound on how big $||x(\theta, z_t)||_2$ can get. Since under Assumption 2, we have that $||\phi(s)||_2 \leq 1$ for all s, we have that

$$\|z_t\|_2 \leq \frac{1}{1-\gamma\lambda}.$$

Furthermore, we have that

$$|\delta_t| = \left|r(s,s') + \gamma\phi(s')^T\theta_t - \phi(s)^T\theta_t\right| \leq r_{\max} + 2R_\lambda,$$

where we used $|r(s,s')| \leq r_{\max}$ as well as Cauchy-Schwarz. Putting the last two equations together, we obtain

$$\|x(\theta, z_t)\|_2 \leq \frac{r_{\max} + 2R_\lambda}{1-\gamma\lambda} := G_\lambda. \tag{33}$$

Compared to the result for TD(0), the bound depends on a slightly different definition of the mixing time that takes into account the geometric weighting in the eligibility trace. Define

$$\tau_\lambda^{\text{mix}}(\varepsilon) = \max\{\tau^{\text{MC}}(\varepsilon), \tau^{\text{Algo}}(\varepsilon)\},$$

where

$$\tau^{\text{MC}}(\varepsilon) = \min\left\{t \in \mathbb{N}_0 | m\rho^t \leq \varepsilon\right\},$$
$$\tau^{\text{Algo}}(\varepsilon) = \min\left\{t \in \mathbb{N}_0 | (\gamma\lambda)^t \leq \varepsilon\right\}.$$

The main result of this section is the following corollary of Theorem 2.

Corollary 4. *Suppose Assumptions 1-2 hold. Suppose further that $(\theta_t)_{t\geq 0}$ is generated by the Projected TD(λ) algorithm of Eq. (32) with θ_λ^* belonging to the convex set Θ_λ and step-sizes $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$. Then*

$$E\left[f^{(\lambda)}(\theta)\right] \leq \frac{\|\theta_\lambda^* - \theta_0\|_2^2 + G_\lambda^2\left[14 + 28\tau_\lambda^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2\sqrt{T}},$$

where the function $f^{(\lambda)}(\theta)$ was defined in Theorem 2.

Proof. We begin with the standard recursion for the distance to the limit:

$$\begin{aligned}
&\|\theta_\lambda^* - \theta_{t+1}\|_2^2 \\
=&\|\theta_\lambda^* - \text{Proj}_{\Theta_\lambda}(\theta_t + \alpha_t\delta_t z_t)\|_2^2 \\
\leq&\|\theta_\lambda^* - \theta_t - \alpha_t\delta_t z_t\|_2^2 \\
=&\|\theta_\lambda^* - \theta_t\|_2^2 - 2\alpha_t\delta_t z_t^T(\theta_\lambda^* - \theta_t) + \alpha_t^2\|\delta_t z_t\|_2^2 \\
=&\|\theta_\lambda^* - \theta_t\|_2^2 - 2\alpha_t\left(\bar{x}(\theta_t)^T - \left(\bar{x}(\theta_t)^T - \delta_t z_t^T\right)\right)(\theta_\lambda^* - \theta_t) \\
&+ \alpha_t^2\|\delta_t z_t\|_2^2 \\
=&\|\theta_\lambda^* - \theta_t\|_2^2 - 2\alpha_t\left(\bar{x}(\theta_t) - \bar{x}(\theta_\lambda^*)\right)^T(\theta_\lambda^* - \theta_t) \\
&+ 2\alpha_t\zeta_t(\theta_t, z_t) + \alpha_t^2\|x(\theta_t, z_t)\|_2^2 \\
=&\|\theta_\lambda^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma\kappa)\|V_{\theta_t} - V_{\theta_\lambda^*}\|_D^2 \\
&- 2\alpha_t(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V_{\theta_t} - V_{\theta_\lambda^*}\|_{\text{Dir},m+1}^2 \\
&+ 2\alpha_t\zeta_t(\theta_t, z_t) + \alpha_t^2\|x(\theta_t, z_t)\|_2^2 \\
\leq&\|\theta_\lambda^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma\kappa)\|V_{\theta_t} - V_{\theta_\lambda^*}\|_D^2 \\
&- 2\alpha_t(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V_{\theta_t} - V_{\theta_\lambda^*}\|_{\text{Dir},m+1}^2 \\
&+ 2\alpha_t\zeta_t(\theta_t, z_t) + \alpha_t^2 G_\lambda^2.
\end{aligned}$$

In the sequence of equations above the first inequality follows that the projection onto a convex set does not increase distance; the remaining equalities are rearrangements, using the quantity $\bar{x}(\theta)$ defined in Eq. (19), that $\bar{x}(\theta_\lambda^*) = 0$ from Eq. (20), and Proposition 1; and the final inequality used Eq. (33).

We next take expectations, rearrange terms, and sum:

$$\begin{aligned}
&\sum_{t=0}^{T-1} 2\alpha_t E\left[(1-\gamma\kappa)\|V_{\theta_t} - V_{\theta_\lambda^*}\|_D^2\right] \\
&+ \sum_{t=0}^{T-1} 2\alpha_t E\left[(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V_{\theta_t} - V_{\theta_\lambda^*}\|_{\text{Dir},m+1}^2\right] \\
\leq&\sum_{t=0}^{T-1}\left(E[\|\theta_\lambda^* - \theta_t\|_2^2] - E[\|\theta_\lambda^* - \theta_{t+1}\|_2^2]\right) \\
&+ \sum_{t=0}^{T-1} 2\alpha_t E[\zeta_t(\theta_t, z_t)] + \sum_{t=0}^{T-1}\alpha_t^2 G_\lambda^2 \\
=&\left(\|\theta_\lambda^* - \theta_0\|_2^2 - E[\|\theta_\lambda^* - \theta_T\|_2^2]\right) \\
&+ \sum_{t=0}^{T-1} 2\alpha_t E[\zeta_t(\theta_t, z_t)] + \sum_{t=0}^{T-1}\alpha_t^2 G_\lambda^2 \\
\leq&\|\theta_\lambda^* - \theta_0\|_2^2 + \sum_{t=0}^{T-1} 2\alpha_t E[\zeta_t(\theta_t, z_t)] + \sum_{t=0}^{T-1}\alpha_t^2 G_\lambda^2.
\end{aligned}$$

Plugging in the step-sizes $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$, we obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1} E\left[(1-\gamma\kappa)\|V_{\theta_t} - V_{\theta_\lambda^*}\|_D^2\right] \\
&+ \sum_{t=0}^{T-1} E\left[(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V_{\theta_t} - V_{\theta_\lambda^*}\|_{\text{Dir},m+1}^2\right] \\
\leq&\frac{\sqrt{T}}{2}\left(\|\theta_\lambda^* - \theta_0\|_2^2 + G_\lambda^2\right) + \sum_{t=0}^{T-1} E[\zeta_t(\theta_t, z_t)].
\end{aligned}$$

Using Lemma 20 in (Bhandari et al., 2018), we have that

$$\sum_{t=0}^{T-1} E[\zeta_t(\theta_t, z_t)]$$

$$\leq 6\sqrt{T}\left(1+2\tau_\lambda^{\text{mix}}(\alpha_T)\right)G_\lambda^2+\sum_{t=0}^{2\tau_\lambda^{\text{mix}}(\alpha_T)}(\gamma\lambda)^t G_\lambda^2$$
$$\leq 6\sqrt{T}\left(1+2\tau_\lambda^{\text{mix}}(\alpha_T)\right)G_\lambda^2+\left(2\tau_\lambda^{\text{mix}}(\alpha_T)+1\right)G_\lambda^2.$$

Combining with convexity, we get

$$\begin{aligned}
&E\left[(1-\gamma\kappa)\|V_{\theta_\lambda^*}-V_{\bar{\theta}_T}\|_D^2\right]\\
&\quad+E\left[(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V_{\theta_\lambda^*}-V_{\bar{\theta}_T}\|_{\text{Dir},m+1}^2\right]\\
\leq&\frac{1}{T}\sum_{t=0}^{T-1}E\left[(1-\gamma\kappa)\|V_{\theta_t}-V_{\theta_\lambda^*}\|_D^2\right]\\
&\quad+\frac{1}{T}\sum_{t=0}^{T-1}E\left[(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V_{\theta_t}-V_{\theta_\lambda^*}\|_{\text{Dir},m+1}^2\right]\\
\leq&\frac{\|\theta_\lambda^*-\theta_0\|_2^2+G_\lambda^2}{2\sqrt{T}}\\
&\quad+\frac{6\sqrt{T}\left(1+2\tau_\lambda^{\text{mix}}(\alpha_T)\right)G_\lambda^2+\left(2\tau_\lambda^{\text{mix}}(\alpha_T)+1\right)G_\lambda^2}{T}\\
\leq&\frac{\|\theta_\lambda^*-\theta_0\|_2^2+G_\lambda^2\left(14+28\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right)}{2\sqrt{T}}.
\end{aligned}$$

■