Appendix: Proofs

Appendix A: Analysis of Projected TD(0) under Markov chain sampling model

In this section, we complete the proof of Theorem 3. The first subsection restates the theorem, as well as the two key lemmas from Section 8 that underly the proof. The second subsection contains a proof of Theorem 3. Finally, Subsection A.3 contains the proof of a technical result, Lemma 10, which was omitted from the main text but we need for the proof.

A.1. Restatement of the theorem and key lemmas from the main text

THEOREM 3 *Suppose the Projected TD algorithm is applied with parameter $R \geq \|\theta^*\|_2$ under the Markov chain observation model with Assumption 1. Set $G = (r_{\max} + 2R)$. Then the following claims hold.*

(a) With a constant step-size sequence $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left(9 + 12\tau^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\gamma)}.$$

(b) With a constant step-size sequence $\alpha_0 = \cdots = \alpha_T < 1/(2\omega(1-\gamma))$,

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_T}\|_D^2\right] \leq \left(e^{-2\alpha_0(1-\gamma)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{G^2\left(9 + 12\tau^{\text{mix}}(\alpha_0)\right)}{2(1-\gamma)\omega}\right).$$

(c) With a decaying step-size sequence $\alpha_t = 1/(\omega(t+1)(1-\gamma))$ for all $t \in \mathbb{N}_0$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{G^2\left(9 + 24\tau^{\text{mix}}(\alpha_T)\right)}{T(1-\gamma)^2\omega}\left(1 + \log T\right),$$

The key to our proof is the following lemmas, which were established in Section 8. Recall the definition of the semi-gradient error $\zeta_t(\theta) \equiv \left(g_t(\theta) - \bar{g}(\theta)\right)^\top(\theta - \theta^*)$.

LEMMA 8 *With probability 1, for every $t \in \mathbb{N}_0$,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.$$

LEMMA 11 *Consider a non-increasing step-size sequence, $\alpha_0 \geq \alpha_1 \ldots \geq \alpha_T$. Fix any $t < T$, and set $t^* \equiv \max\{0, t - \tau^{\text{mix}}(\alpha_T)\}$. Then,*

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

The following bound also holds:

$$\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq 6G^2\sum_{i=0}^{t-1}\alpha_i.$$

A.2. Proof of Theorem 3.

We now complete the proof of Theorem 3 by directly leveraging Lemmas 8 and 11.

Proof of Theorem 3. From Lemma 8, we have

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha_t(1-\gamma)\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] + 2\alpha_t\mathbb{E}\left[\zeta_t(\theta_t)\right] + \alpha_t^2 G^2. \quad \text{(EC.1)}$$

Proof of part (a): We first show the analysis for a constant step-size and iterate averaging. Considering $\alpha_t = \alpha_0 = 1/\sqrt{T}$ in Equation (EC.1), rearranging terms and summing from $t=0$ to $t=T-1$, we get

$$2\alpha_0(1-\gamma)\sum_{t=0}^{T-1}\mathbb{E}\left[\|V_{\theta^*}-V_{\theta_t}\|_D^2\right] \leq \sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|\theta^*-\theta_t\|_2^2\right]-\mathbb{E}\left[\|\theta^*-\theta_{t+1}\|_2^2\right]\right)+G^2+2\alpha_0\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right].$$

Using Lemma 11 (in which $\alpha_{t^*}=\alpha_0$ in this case) and simplifying, we find

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|V_{\theta^*}-V_{\theta_t}\|_D^2\right] &\leq \frac{\|\theta^*-\theta_0\|_2^2+G^2}{2\alpha_0(1-\gamma)}+\frac{T\cdot 2G^2(2+3\tau^{\text{mix}}(1/\sqrt{T}))\alpha_0}{(1-\gamma)}\\ &= \frac{\sqrt{T}\left(\|\theta^*-\theta_0\|_2^2+G^2\right)}{2(1-\gamma)}+\frac{\sqrt{T}\cdot 2G^2(2+3\tau^{\text{mix}}(1/\sqrt{T}))}{(1-\gamma)}.\end{aligned}$$

This gives us our desired result,

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

Proof of part (b): The proof is analogous to part (b) of Theorem 2. Consider a constant step-size of $\alpha_0 < 1/(2\omega(1-\gamma))$. Starting with Equation (EC.1) and applying Lemma 1, which showed $\|V_{\theta^*}-V_\theta\|_D^2 \geq \omega\|\theta^*-\theta\|_2^2$ for all θ, we get

$$\begin{aligned}\mathbb{E}\left[\|\theta^*-\theta_{t+1}\|_2^2\right] &\leq (1-2\alpha_0(1-\gamma)\omega)\,\mathbb{E}\left[\|\theta^*-\theta_t\|_2^2\right]+\alpha_0^2G^2+2\alpha_0\mathbb{E}\left[\zeta_t(\theta_t)\right]\\ &\leq (1-2\alpha_0(1-\gamma)\omega)\,\mathbb{E}\left[\|\theta^*-\theta_t\|_2^2\right]+\alpha_0^2G^2\left(9+12\tau^{\text{mix}}(\alpha_0)\right),\end{aligned}$$

where we used Lemma 11 to go to the second inequality. Iterating over this inequality gives us our final result. For any $T\in\mathbb{N}_0$,

$$\begin{aligned}\mathbb{E}\left[\|\theta^*-\theta_T\|_2^2\right] &\leq (1-2\alpha_0(1-\gamma)\omega)^T\|\theta^*-\theta_0\|_2^2+\alpha_0^2G^2\left(9+12\tau^{\text{mix}}(\alpha_0)\right)\sum_{t=0}^{\infty}(1-2\alpha_0(1-\gamma)\omega)^t\\ &\leq \left(e^{-2\alpha_0(1-\gamma)\omega T}\right)\|\theta^*-\theta_0\|_2^2+\frac{\alpha_0G^2\left(9+12\tau^{\text{mix}}(\alpha_0)\right)}{2(1-\gamma)\omega}.\end{aligned}$$

Final inequality follows by solving the geometric series and using that $(1-2\alpha_0(1-\gamma)\omega) \leq e^{-2\alpha_0(1-\gamma)\omega}$ along with Lemma 1.

Proof of part (c): We now show the analysis for a linearly decaying step-size using Equation (EC.1) as our starting point. We again use Lemma 1, which showed $\|V_{\theta^*}-V_\theta\|_D^2 \geq \omega\|\theta^*-\theta\|_2^2$ for all θ, to get,

$$\mathbb{E}\left[\|V_{\theta^*}-V_{\theta_t}\|_D^2\right] \leq \frac{1}{(1-\gamma)\alpha_t}\left((1-(1-\gamma)\omega\alpha_t)\mathbb{E}\left[\|\theta^*-\theta_t\|_2^2\right]-\mathbb{E}\left[\|\theta^*-\theta_{t+1}\|_2^2\right]+\alpha_t^2G^2\right)+\frac{2}{(1-\gamma)}\mathbb{E}\left[\zeta_t(\theta_t)\right].$$

Consider a decaying step-size $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$, simplify and sum from $t=0$ to $T-1$ to get

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\|V_{\theta^*}-V_{\theta_t}\|_D^2\right] \leq \underbrace{-\omega T\mathbb{E}\left[\|\theta^*-\theta_T\|_2^2\right]}_{<0}+\frac{G^2}{\omega(1-\gamma)^2}\sum_{t=0}^{T-1}\frac{1}{t+1}+\frac{2}{(1-\gamma)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right]. \quad \text{(EC.2)}$$

To simplify notation, for the remainder of the proof put $\tau = \tau^{\text{mix}}(\alpha_T)$. We can decompose the sum of semi-gradient errors as

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t)\right] = \sum_{t=0}^{\tau} \mathbb{E}\left[\zeta_t(\theta_t)\right] + \sum_{t=\tau+1}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t)\right]. \tag{EC.3}$$

We will upper bound each term. In each case we use that, since $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$,

$$\sum_{t=0}^{T-1} \alpha_t = \frac{1}{\omega(1-\gamma)} \sum_{t=0}^{T-1} \frac{1}{(t+1)} \leq \frac{1+\log T}{\omega(1-\gamma)}.$$

Combining this with Lemma 11 gives,

$$\sum_{t=0}^{\tau} \mathbb{E}\left[\zeta_t(\theta_t)\right] \leq \sum_{t=0}^{\tau} \left(6G^2 \sum_{i=0}^{t-1} \alpha_i\right) \leq \tau \left(6G^2 \sum_{i=0}^{T-1} \alpha_i\right) \leq \frac{6G^2\tau}{\omega(1-\gamma)}(1+\log T).$$

Similarly, using Lemma 11, we have

$$\sum_{t=\tau+1}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t)\right] \leq 2G^2\left(2+3\tau\right) \sum_{t=\tau+1}^{T-1} \alpha_{t-\tau} \leq 2G^2\left(2+3\tau\right) \sum_{t=1}^{T-1} \alpha_t \leq \frac{2G^2\left(2+3\tau\right)}{\omega(1-\gamma)}\left(1+\log T\right).$$

Combining the two parts, we get

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t)\right] \leq \frac{4G^2\left(1+3\tau\right)}{\omega(1-\gamma)}(1+\log T).$$

Using this in conjunction with Equation (EC.2) we get,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|V_{\theta_t} - V_{\theta^*}\right\|_D^2\right] \leq \frac{G^2}{\omega T(1-\gamma)^2}\left(1+\log T\right) + \frac{2}{T(1-\gamma)}\sum_{t=0}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t)\right].$$

Simplifying and substituting $\tau = \tau^{\text{mix}}(\alpha_T)$, we get

$$\begin{aligned}\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] &\leq \frac{G^2}{\omega T(1-\gamma)^2}\left(1+\log T\right) + \frac{8G^2\left(1+3\tau^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\gamma)^2}(1+\log T) \\ &\leq \frac{G^2\left(9+24\tau^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\gamma)^2}\left(1+\log T\right).\end{aligned}$$

which gives us the desired result. □

We remark that Equation (EC.2) also gives us a convergence rate of $\mathcal{O}(\log T/T)$ for the iterate θ_T itself (and hence on the value function V_{θ_T}) but the bound degrades by a factor of ω. In particular, we have

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \mathbb{E}\left[\left\|\theta^* - \theta_T\right\|_2^2\right] \leq \frac{G^2\left(9+24\tau^{\text{mix}}(\alpha_T)\right)}{\omega^2 T(1-\gamma)^2}\left(1+\log T\right). \tag{EC.4}$$

where the first inequality follows using Lemma 1.

A.3. Proof of Lemma 10

LEMMA 10 *With probability 1,*

$$|\zeta_t(\theta)| \leq 2G^2 \qquad \textit{for all } \theta \in \Theta_R$$

and

$$|\zeta_t(\theta) - \zeta_t(\theta')| \leq 6G \left\| (\theta - \theta') \right\|_2 \qquad \textit{for all } \theta, \theta' \in \Theta_R.$$

Proof of Lemma 10. The first claim follows from a simple argument using Lemma 6.

$$|\zeta_t(\theta)| = \left| (g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*) \right| \leq (\|g_t(\theta)\|_2 + \|\bar{g}(\theta)\|_2)(\|\theta\|_2 + \|\theta^*\|_2) \leq 4GR \leq 2G^2,$$

where the first inequality follows from the triangle inequality and the Cauchy-Schwartz inequality, and the final inequality uses that $R \leq G/2$ by definition of $G = r_{\max} + 2R$.

To establish the second claim, consider the following inequality for any vectors (a_1, b_1, a_2, b_2):

$$\left| a_1^\top b_1 - a_2^\top b_2 \right| = \left| a_1^\top (b_1 - b_2) + b_2^\top (a_1 - a_2) \right| \leq \|a_1\| \|b_1 - b_2\| + \|b_2\| \|a_1 - a_2\|.$$

This follows as a direct application of Cauchy-Schwartz. It implies that for any $\theta, \theta' \in \Theta_R$,

$$\begin{aligned}
|\zeta_t(\theta) - \zeta_t(\theta')| &= \left| (g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*) - (g_t(\theta') - \bar{g}(\theta'))^\top (\theta' - \theta^*) \right| \\
&\leq \|g_t(\theta) - \bar{g}(\theta)\|_2 \|\theta - \theta'\|_2 + \|\theta' - \theta^*\|_2 \|(g_t(\theta) - \bar{g}(\theta)) - (g_t(\theta') - \bar{g}(\theta'))\|_2 \\
&\leq 2G \|\theta - \theta'\|_2 + 2R \left(\|g_t(\theta) - g_t(\theta')\|_2 + \|\bar{g}(\theta) - \bar{g}(\theta')\|_2 \right) \\
&\leq 2G \|\theta - \theta'\|_2 + 8R \|\theta - \theta'\|_2 \\
&\leq 6G \|\theta - \theta'\|_2.
\end{aligned}$$

where we used that $R \leq G/2$ by the definition of G. We also used that both $g_t(\cdot)$ and $\bar{g}(\cdot)$ are 2-Lipschitz functions which is easy to see. Starting with $g_t(\theta) = (r_t + \gamma \phi(s_t')^\top \theta - \phi(s_t)^\top \theta)\phi(s_t)$, consider

$$\begin{aligned}
\|g_t(\theta) - g_t(\theta')\|_2 &= \left\| \phi(s_t) \left(\gamma \phi(s_t') - \phi(s_t) \right)^\top (\theta - \theta') \right\|_2 \\
&\leq \|\phi(s_t)\|_2 \|(\gamma \phi(s_t') - \phi(s_t))\|_2 \|(\theta - \theta')\|_2 \\
&\leq 2 \|(\theta - \theta')\|_2.
\end{aligned}$$

Similarly, following Equation (2), we have $\|\bar{g}(\theta) - \bar{g}(\theta')\|_2 = \left\| \mathbb{E}\left[\phi \left(\gamma \phi' - \phi \right)^\top (\theta - \theta') \right] \right\|_2$, where $\phi = \phi(s)$ is the feature vector of a random initial state $s \sim \pi$, $\phi' = \phi(s')$ is the feature vector of a random next state drawn according to $s' \sim \mathcal{P}(\cdot \mid s)$. Therefore,

$$\|\bar{g}(\theta) - \bar{g}(\theta')\|_2 \leq \left\| \phi \left(\gamma \phi' - \phi \right)^\top (\theta - \theta') \right\| \leq 2 \|(\theta - \theta')\|_2. \quad \square$$

Appendix B: Analysis of Projected TD(λ) under Markov chain sampling model

In this section, we give a detailed proof of the convergence bounds presented in Theorem 4. Subsection B.1 details our proof strategy along with key lemmas which come together in Subsection B.2 to establish the results. We begin by providing mathematical expressions for TD(λ) updates.

Stationary distribution of TD(λ) updates. Recall that the projected TD(λ) update at time t is given by:

$$\theta_{t+1} = \Pi_{2,R}\left(\theta_t + \alpha_t x_t(\theta_t, z_{0:t})\right)$$

where $\Pi_{2,R}(\cdot)$ denotes the projection operator onto a norm ball of radius $R < \infty$ and $x_t(\theta_t, z_{0:t})$ is the update direction. Let us now give explicit mathematical expressions for $x_t(\theta, z_{0:t})$ and its steady-state mean $\bar{x}(\theta)$. Note that these are analogous to the expressions for the negative semi-gradient $g_t(\theta)$ and its steady-state expectation $\bar{g}(\theta)$ for TD(0). At time t, as a general function of (non-random) θ and the tuple $\mathcal{O}_t = (s_t, r_t, s_t')$ along with the eligibility trace term $z_{0:t}$, we have

$$x_t(\theta, z_{0:t}) = \left(r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right) z_{0:t} = \delta_t(\theta) z_{0:t} \quad \forall\, \theta \in \mathbb{R}^d.$$

The asymptotic convergence of TD(λ) is closely related to the expected value of $x_t(\theta, z_{0:t})$ under the steady-state behavior of $(O_t, z_{0:t})$,

$$\bar{x}(\theta) = \lim_{t\to\infty} \mathbb{E}\left[\delta_t(\theta) z_{0:t}\right].$$

Rather than take this limit, it will be helpful in our analysis to think of an equivalent *backward view* by constructing a stationary process with mean $\bar{x}(\theta)$. Consider a stationary sequence of states $(\ldots, s_{-1}, s_0, s_1, \ldots)$ and set $z_{-\infty:t} = \sum_{k=0}^{\infty}(\gamma\lambda)^k\phi(s_{t-k})$. Then the sequence $(x_0(\theta, z_{-\infty:0}), x_1(\theta, z_{-\infty:1}), \ldots)$ is stationary, and we have

$$\bar{x}(\theta) = \mathbb{E}\left[\delta_t(\theta) z_{-\infty:t}\right]. \tag{EC.5}$$

It should be emphasized that $\bar{x}(\theta)$ and the states $(\ldots, s_{-2}, s_{-1})$ are introduced only for the purposes of our analysis and are never used by the algorithm itself. However, this turns out to be quite useful as it is easy to show (Van Roy 1998) that

$$\bar{x}(\theta) = \Phi^\top D\left(T_\mu^{(\lambda)}\Phi\theta - \Phi\theta\right), \tag{EC.6}$$

where Φ is the feature matrix and $(T_\mu^{(\lambda)}\Phi\theta - \Phi\theta)$ denotes the Bellman error defined with respect to the Bellman operator $T_\mu^{(\lambda)}(\cdot)$, corresponding to a policy μ. Careful readers will notice the stark similarity between Equation (EC.6) and Equation (3). Exploiting the property that $\Pi_D T_\mu^\lambda(\cdot)$ is also a contraction operator, one can easily show a result equivalent to Lemma 3, thus quantifying the progress we make by taking steps in the direction of $\bar{x}(\theta)$. The rest of our proof essentially shows how to control for the observation noise, i.e. the fact that we use $x_t(\theta, z_{0:t})$ rather than $\bar{x}(\theta)$ to make updates. To remind the readers of the results, we first restate Theorem 4 below.

THEOREM 4 *Suppose the Projected TD(λ) algorithm is applied with parameter $R \geq \|\theta^*\|_2$ under the Markov chain observation model with Assumption 1. Set $B = \frac{(r_{\max}+2R)}{(1-\gamma\lambda)}$. Then the following claims hold.*

(a) With a constant step-size $\alpha_t = \alpha_0 = 1/\sqrt{T}$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2\left(13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\kappa)}.$$

(b) With a constant step-size $\alpha_t = \alpha_0 < 1/(2\omega(1-\kappa))$ and $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_T}\|_D^2\right] \leq \left(e^{-2\alpha_0(1-\kappa)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)}{2(1-\kappa)\omega}\right).$$

(c) With a decaying step-size $\alpha_t = 1/(\omega(t+1)(1-\kappa))$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{B^2\left(13 + 52\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{T(1-\kappa)^2\omega}\left(1 + \log T\right).$$

B.1. Proof strategy and key lemmas

We now describe our proof strategy and give key lemmas used to establish Theorem 4. Throughout, we consider the Markov chain observation model with Assumption 1 and study the Projected TD (λ) algorithm applied with parameter $R \geq \|\theta^*\|_2$ and step-size sequence $(\alpha_0, \ldots, \alpha_T)$. To simplify our exposition, we introduce some notation below.

Notation: We specify the notation used throughout this section. Define the set $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, so $\theta_t \in \Theta_R$ for each t because of the algorithm's projection step. Next, we generically define $z_{l:t} = \sum_{k=0}^{t-l} (\gamma\lambda)^k \phi(s_{t-k})$ for any lower limit $l \leq t$. Thus, $z_{l:t}$ denotes the eligibility trace as a function of the states $(s_l, \ldots, s_t)$. Next, we define $\zeta_t(\theta, z_{l:t})$ as a general function of θ and $z_{l:t}$,

$$\zeta_t(\theta, z_{l:t}) = \left(\delta_t(\theta) z_{l:t} - \bar{x}(\theta)\right)^\top (\theta - \theta^*). \tag{EC.7}$$

Here, the subscript t in ζ_t encodes the dependence on the tuple $O_t = (s_t, r_t, s_t')$ which is used to compute the Bellman error, $\delta_t(\cdot)$ at time t. Finally, we set $B := (r_{\max} + 2R)/(1 - \gamma\lambda)$ which implies $B > 2R$, a fact we use many times in our proofs to simplify constant terms. As a reminder, note that our bounds depend on the mixing time, which we defined in Section 9 as

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

where $\tau^{\text{MC}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}$ and $\tau^{\text{Algo}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid (\gamma\lambda)^t \leq \epsilon\}$.

Proof outline: The analysis for TD(λ) can be broadly divided into three parts and closely mimics the steps used to prove TD(0) results.

1. As a first step, we do an error decomposition, similar to the result shown in Lemma 8. This is enabled by two key lemmas, which are analogues of Lemma 3 and Lemma 6 for Projected TD(0). The first one spells out a clear relationship of how the updates following $\bar{x}(\theta)$ point in the descent direction of $\|\theta^* - \theta\|_2^2$ while the second one upper bounds the norm of the update direction, $x_t(\theta, z_{0:t})$, by the constant B (as defined above).
2. The error decomposition that we obtain from Step 1 can be stated as:

$$\mathbb{E}[\|\theta^* - \theta_{t+1}\|_2^2] \leq \mathbb{E}[\|\theta^* - \theta_t\|_2^2] - 2\alpha_t(1-\kappa)\mathbb{E}[\|V_{\theta^*} - V_{\theta_t}\|_D^2] + 2\alpha_t \mathbb{E}[\zeta_t(\theta_t, z_{0:t})] + \alpha_t^2 B^2.$$

   In the second step, we establish an upper bound on the bias term, $\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]$, which is the main challenge in our proof. Recall that the dependent nature of the state transitions may result in strong coupling between the tuples $\mathcal{O}_{t-1}$ and $\mathcal{O}_t$ under the Markov chain observation model. Therefore, this bias in update direction can potentially be non-zero. Presence of the eligibility trace term, $z_{0:t}$, which is a function of the entire history of states, $(s_0, \ldots, s_t)$, further complicates the analysis by introducing subtle dependencies.

   To control for this, we use information-theoretic techniques shown in Lemma 9 which exploit the geometric ergodicity of the MDP, along with the geometric weighting of state features in the eligibility trace term. Our result essentially shows that the bias scales the noise in update direction by a factor of the mixing time. Mathematically, for a constant step-size α, we show that $\mathbb{E}\left[\alpha\zeta_t(\theta_t, z_{0:t})\right] \approx B^2(6 + 12\tau_\lambda^{\text{mix}}(\alpha))\alpha^2$. We show a similar result for decaying step-sizes as well.
3. In the final step, we combine the error decomposition from Step 1 and the bound on the bias from Step 2, to establish finite time bounds on the performance of Projected TD(λ) for different step-size choices. We closely mimic the analysis of Nemirovski et al. (2009) for a constant, aggressive step-size of $(1/\sqrt{T})$ and the proof ideas of Lacoste-Julien et al. (2012) for decaying step-sizes.

B.1.1. Error decomposition under Projected TD(λ) We first state two important lemmas below which enable the error decomposition shown in Lemma EC.3.

LEMMA EC.1 **(Tsitsiklis and Van Roy (1997))**. *Let V_{θ^*} be the unique fixed point of $\Pi_D T_\mu^{(\lambda)}(\cdot)$ i.e. $V_{\theta^*} = \Pi T_\mu^{(\lambda)} V_{\theta^*}$. For any $\theta \in \mathbb{R}^d$,*

$$(\theta^* - \theta)^\top \bar{x}(\theta) \geq (1-\kappa)\|V_{\theta^*} - V_\theta\|_D^2.$$

Proof of Lemma EC.1. We use the definition of $\bar{x}(\theta) = \langle \Phi^\top, T_\mu^{(\lambda)}\Phi\theta - \Phi\theta\rangle_D$ as shown in Equation (EC.6) along with the fact that $\Pi_D T_\mu^{(\lambda)}(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus κ. See Appendix C for a complete proof. □

LEMMA EC.2. *For all $\theta \in \Theta_R$, $\|x_t(\theta, z_{0:t})\|_2 \leq B$ with probability 1. Additionally, $\|\bar{x}(\theta)\|_2 \leq B$.*

Proof of Lemma EC.2. See Subsection B.3 for a complete proof. □

The above two lemmas can be easily combined to establish a recursion for the error under projected TD(λ) that holds for each sample path.

LEMMA EC.3. *With probability 1, for every $t \in \mathbb{N}_0$,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\kappa)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 B^2.$$

Proof of Lemma EC.3. The Projected TD(λ) algorithm updates the parameter as: $\theta_{t+1} = \Pi_{2,R}[\theta_t + \alpha_t x_t(\theta_t, z_{0:t})]\ \forall\, t \in \mathbb{N}_0$. This implies,

$$\begin{aligned}
\|\theta^* - \theta_{t+1}\|_2^2 &= \|\theta^* - \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta, z_{0:t}))\|_2^2 \\
&= \|\Pi_{2,R}(\theta^*) - \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))\|_2^2 \\
&\leq \|\theta^* - \theta_t - \alpha_t x_t(\theta_t, z_{0:t})\|_2^2 \\
&= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t x_t(\theta_t, z_{0:t})^\top(\theta^* - \theta_t) + \alpha_t^2\|x_t(\theta_t, z_{0:t})\|_2^2 \\
&\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t x_t(\theta_t, z_{0:t})^\top(\theta^* - \theta_t) + \alpha_t^2 B^2 \\
&= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t \bar{x}(\theta_t)^\top(\theta^* - \theta_t) + 2\alpha_t\zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 G^2. \\
&\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\kappa)\|V_{\theta^*} - V_\theta\|_D^2 + 2\alpha_t\zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 B^2.
\end{aligned}$$

The first inequality used that orthogonal projection operators onto a convex set are non-expansive, the second used Lemma EC.2 together with the fact $\|\theta_t\|_2 \leq R$ due to projection, and the third used Lemma EC.1. Note that we used $\zeta_t(\theta_t, z_{0:t})$ to simplify the notation for the error in the update direction. Recall the definition of the error function from Equation (EC.7) which implies,

$$\zeta_t(\theta_t, z_{0:t}) = (\delta_t(\theta_t)z_{0:t} - \bar{x}(\theta_t))^\top(\theta_t - \theta^*) = (x_t(\theta_t, z_{0:t}) - \bar{x}(\theta_t))^\top(\theta_t - \theta^*).$$

B.1.2. Upper bound on the bias in update direction. We give an upper bound on the expected error in the update direction, $\mathbb{E}[\zeta_t(\theta_t, z_{0:t})]$, which as explained above, is the key challenge for our analysis. For this, we first establish some basic regularity properties of the error function $\zeta_t(\cdot,\cdot)$ in Lemma EC.4 below. In particular, part (a) shows boundedness, part (b) shows that it is Lipschitz in the first argument and part (c) bounds the error due to truncation of the eligibility trace. Recall that $z_{l:t}$ denotes the eligibility trace as a function of the states $(s_l, \ldots, s_t)$.

LEMMA EC.4. *Consider any $l \leq t$ and any $\theta, \theta' \in \Theta_R$. With probability 1,*

(a) $|\zeta_t(\theta, z_{l:t})| \leq 2B^2$.

(b) $|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| \leq 6B\left\|(\theta - \theta')\right\|_2$.

(c) The following two bounds also hold,

$$\begin{aligned}
&|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau \text{ for all } \tau \leq t, \\
&|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| \leq B^2(\gamma\lambda)^t.
\end{aligned}$$

Proof of Lemma EC.4. We essentially use the uniform bound on $x_t(\theta, z_{0:t})$ and $\bar{x}(\theta)$ as stated in Lemma EC.2 to show this result. See Subsection B.3 for a detailed proof. □

Lemma EC.4 can be combined with Lemma 9 to give an upper bound on the bias term, $\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]$, as shown below.

LEMMA EC.5. *Consider a non-increasing step-size sequence, $\alpha_0 \geq \alpha_1 \ldots \geq \alpha_T$. Then the following hold.*

(a) For $2\tau_\lambda^{\text{mix}}(\alpha_T) < t \leq T$,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T))\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)}.$$

(b) For $0 \leq t \leq 2\tau_\lambda^{\text{mix}}(\alpha_T)$,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2\left(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T)\right)\alpha_0 + B^2(\gamma\lambda)^t.$$

(c) For all $t \in \mathbb{N}_0$,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t$$

Proof of EC.5. We proceed in two cases below. Throughout the proof, results from Lemma EC.4 are applied using the fact that $\theta_t \in \Theta_R$, because of the algorithm's projection step.

Case (a): Let $t > 2\tau$ and consider the following decomposition for all $\tau \in \{0, 1, \ldots, t/2\}$. We show an upper bound on each of the three terms separately.

$$\begin{aligned}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq |\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| \\ + |\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]|.\end{aligned}$$

Step 1: Use regularity properties of the error function to bound first two terms.

We relate $\zeta_t(\theta_t, z_{0:t})$ and $\zeta_t(\theta_{t-\tau}, z_{0:t})$ using the Lipschitz property shown in part (b) of Lemma EC.4 to get,

$$|\zeta_t(\theta_t, z_{0:t}) - \zeta_t(\theta_{t-2\tau}, z_{0:t})| = 6B\|\theta_t - \theta_{t-2\tau}\|_2 \leq 6B^2 \sum_{i=t-2\tau}^{t-1} \alpha_i. \tag{EC.8}$$

Taking expectations on both sides gives us the desired bound on the first term. The last inequality used the norm bound on update direction as shown in Lemma EC.2 to simplify,

$$\|\theta_t - \theta_{t-2\tau}\|_2 \leq \sum_{i=t-2\tau}^{t-1} \|\Pi_{2,R}\left(\theta_{i+1} + \alpha_i x_i(\theta_i, z_{0:i})\right) - \theta_i\|_2 \leq \sum_{i=t-2\tau}^{t-1} \alpha_i \|x_i(\theta_i, z_{0:i})\|_2 \leq B \sum_{i=t-2\tau}^{t-1} \alpha_i.$$

Similarly, by part (c) of Lemma EC.4, we have a bound on the second term.

$$|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| \leq B^2(\gamma\lambda)^\tau. \tag{EC.9}$$

Step 2: Use information-theoretic arguments to upper bound $\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]$.

We will essentially use Lemma 9 to upper bound $\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]$. We first introduce some notation to highlight subtle dependency issues. Note that $\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})$ is a function of $(\theta_{t-2\tau}, s_{t-\tau}, \ldots, s_{t-1}, O_t)$. To simplify, let $Y_{t-\tau:t} = (s_{t-\tau}, \ldots, s_{t-1}, O_t)$. Define,

$$f(\theta_{t-2\tau}, Y_{t-\tau:t}) := \zeta_t(\theta_{t-2\tau}, z_{t-\tau:t}).$$

Consider random variables $\theta'_{t-2\tau}$ and $Y'_{t-\tau:t}$ drawn independently from the marginal distributions of $\theta_{t-2\tau}$ and $Y_{t-\tau:t}$, so $\mathbb{P}(\theta'_{t-2\tau} = \cdot, Y'_{t-\tau:t} = \cdot) = \mathbb{P}(\theta_{t-\tau} = \cdot) \otimes \mathbb{P}(Y_{t-\tau:t} = \cdot)$. By Lemma EC.4 we have that $|f(\theta, Y_{t-\tau:t})| \leq 2B^2$ for all $\theta \in \Theta_R$ with probability 1. As

$$\theta_{t-2\tau} \to s_{t-2\tau} \to s_{t-\tau} \to s_t \to O_t$$

form a Markov chain, a direct application of Lemma 9 gives us:

$$\left|\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})] - \mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]\right| \leq 4B^2 m\rho^\tau. \tag{EC.10}$$

We also have the following bound for all fixed $\theta \in \Theta_R$. Using $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta) z_{-\infty:t}]$, we get

$$\mathbb{E}[f(\theta, Y_{t-\tau:t})] = \left(\mathbb{E}[\delta_t(\theta) z_{t-\tau:t}] - \bar{x}(\theta)\right)^\top (\theta - \theta^*) \leq \left|\left(\delta_t(\theta) z_{-\infty:t-\tau}\right)^\top (\theta - \theta^*)\right| \leq B^2(\gamma\lambda)^\tau \tag{EC.11}$$

Combining the above with Equation (EC.10), we get

$$\begin{aligned}
|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| &= |\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})]| \\
&\leq \left|\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})] - \mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]\right| + \left|\mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]\right| \\
&\leq 4B^2 m\rho^\tau + \left|\mathbb{E}[\mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})|\theta'_{t-2\tau}]]\right| \\
&\leq 4B^2 m\rho^\tau + B^2(\gamma\lambda)^\tau.
\end{aligned} \tag{EC.12}$$

Step 3. Combine terms to show part (a) of our claim.

Taking $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ and combining Equations (EC.8), (EC.9) and (EC.12) establishes the first claim.

$$\begin{aligned}
\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2 \sum_{i=t-2\tau}^{t-1} \alpha_i + 4B^2 m\rho^\tau + 2B^2(\gamma\lambda)^\tau &\leq 12B^2 \tau_\lambda^{\text{mix}}(\alpha_T)\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)} + 6B^2\alpha_T \\
&\leq 6B^2(1 + 2\tau^{\text{mix}}(\alpha_T))\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)}.
\end{aligned}$$

Here we used that letting $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ implies: $\max\{m\rho^\tau, (\gamma\lambda)^\tau\} \leq \alpha_T$. Two additional facts which we also use follow from a non-increasing step-size sequence, $\sum_{i=t-2\tau}^{t-1} \alpha_i \leq 2\tau\alpha_{t-2\tau}$ and $\alpha_T \leq \alpha_{t-2\tau}$.

Case (b) and (c): Consider the following decomposition for all $t \in \mathbb{N}_0$,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq |\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] - \mathbb{E}[\zeta_t(\theta_0, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]|.$$

Step 1: Use regularity properties of the error function to upper bound the first two terms.

Using parts (b), (c) of Lemma EC.4 and following the arguments shown in Step 1, 2 of case (a) above, we get

$$|\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] - \mathbb{E}[\zeta_t(\theta_0, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]| \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t. \tag{EC.13}$$

Step 2: Characterizing $\mathbb{E}[\zeta_t(\theta, z_{-\infty:t})]$ for any fixed (non-random) θ.

Recall the definition of $\bar{x}(\theta)$ from Equation (EC.5). For any fixed θ, we have $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta) z_{-\infty:t}]$. Therefore,

$$\mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})] = \left(\mathbb{E}[\delta_t(\theta_0) z_{-\infty:t}] - \bar{x}(\theta_0)\right)^\top (\theta_0 - \theta^*) = 0. \tag{EC.14}$$

Step 3. Combine terms to show parts (b), (c) of our claim.

Combining Equations (EC.13) and (EC.14) establishes part (c) which states,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t \quad \forall\, t \in \mathbb{N}_0.$$

We establish part (b) by using that the step-size sequence is non-increasing which implies: $\sum_{i=0}^{t-1} \alpha_i \leq t\alpha_0$. For all $t \leq 2\tau_\lambda^{\text{mix}}(\alpha_T)$, we have the following loose upper bound.

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2 t\alpha_0 + B^2(\gamma\lambda)^t \leq 6B^2\left(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T)\right)\right)\alpha_0 + B^2(\gamma\lambda)^t. \quad \square$$

B.2. Proof of Theorem 4

In this subsection, we establish convergence bounds for Projected TD(λ) as stated in Theorem 4 using Lemmas EC.3 and EC.5. From Lemma EC.3 we have,

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha_t(1-\kappa)\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] + 2\alpha_t\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] + \alpha_t^2 B^2. \quad \text{(EC.15)}$$

Equation (EC.15) will be used as a starting point for analyzing different step-size choices.

Proof of part (a): Fix a constant step-size of $\alpha_0 = \ldots = \alpha_t = 1/\sqrt{T}$ in Equation (EC.15), rearrange terms and sum from $t=0$ to $t=T-1$, we get

$$2\alpha_0(1-\kappa)\sum_{t=0}^{T-1}\mathbb{E}\|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \sum_{t=0}^{T-1}\left(\mathbb{E}\|\theta^* - \theta_t\|_2^2 - \mathbb{E}\|\theta^* - \theta_{t+1}\|_2^2\right) + B^2 + 2\alpha_0\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right].$$

Using Lemma EC.5 where $\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)} = \alpha_0$ along with the fact that $(\gamma\lambda) < 1$, we simplify to get

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\|V_{\theta^*} - V_{\theta_t}\|_D^2 &\leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2}{2\alpha_0(1-\kappa)} + \frac{6B^2T(1+2\tau_\lambda^{\text{mix}}(1/\sqrt{T}))\alpha_0}{(1-\kappa)} + \frac{1}{(1-\kappa)}\sum_{t=0}^{2\tau_\lambda^{\text{mix}}(\frac{1}{\sqrt{T}})} B^2(\gamma\lambda)^t \\ &\leq \frac{\sqrt{T}\left(\|\theta^* - \theta_0\|_2^2 + B^2\right)}{2(1-\kappa)} + \frac{6B^2\sqrt{T}(1+2\tau_\lambda^{\text{mix}}(1/\sqrt{T}))}{(1-\kappa)} + \frac{2B^2\tau_\lambda^{\text{mix}}(1/\sqrt{T})}{(1-\kappa)}.\end{aligned}$$

Adding these terms, we conclude

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2\left(13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\kappa)}.$$

Proof of part (b): For a constant step-size of $\alpha_0 < 1/(2\omega(1-\kappa))$, we show that the expected distance between the iterate θ_T and the TD(λ) limit point, θ^* converges at an exponential rate below some level that depends on the choice of step-size and λ. Starting with Equation (EC.15) and applying Lemma 1 which shows that $\|V_{\theta^*} - V_\theta\|_D^2 \geq w\|\theta^* - \theta\|_2^2$ for any θ, we have that for all $t > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\begin{aligned}\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] &\leq (1 - 2\alpha_0(1-\kappa)\omega)\,\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + \alpha_0^2 B^2 + 2\alpha_0\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \\ &\leq (1 - 2\alpha_0(1-\kappa)\omega)\,\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + \alpha_0^2 B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right),\end{aligned}$$

where we used part (a) of Lemma EC.5 for the second inequality. Iterating over it gives us our final result. For any $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\begin{aligned}\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] &\leq (1 - 2\alpha_0(1-\kappa)\omega)^T\|\theta^* - \theta_0\|_2^2 + \alpha_0^2 B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)\sum_{t=0}^{\infty}(1 - 2\alpha_0(1-\kappa)\omega)^t \\ &\leq \left(e^{-2\alpha_0(1-\kappa)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \frac{B^2\alpha_0\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)}{2(1-\kappa)\omega}.\end{aligned}$$

Final inequality follows by solving the geometric series and using that $(1 - 2\alpha_0(1-\kappa)\omega) \leq e^{-2\alpha_0(1-\kappa)\omega}$ along with Lemma 1.

Proof of part (c): Consider a decaying step-size of $\alpha_t = 1/(\omega(t+1)(1-\kappa))$. We start with Equation (EC.15) and use Lemma 1 which showed $\mathbb{E}\left[\|V_{\theta^*} - V_\theta\|_D^2\right] \geq w\mathbb{E}\left[\|\theta^* - \theta\|_2^2\right]$ for all θ to get,

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] \leq \frac{1}{(1-\kappa)\alpha_t}\left((1-(1-\kappa)\omega\alpha_t)\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - \mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] + \alpha_t^2 B^2\right) + \frac{2}{(1-\kappa)}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right].$$

Substituting $\alpha_t = \frac{1}{\omega(t+1)(1-\kappa)}$, simplify and sum from $t=0$ to $T-1$ to get,

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] &\leq -\omega T\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] + \frac{B^2}{\omega(1-\kappa)^2}\sum_{t=0}^{T-1}\frac{1}{t+1} + \frac{2}{(1-\kappa)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \\ &\leq \underbrace{-\omega T\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right]}_{<0} + \frac{B^2(1+\log T)}{\omega(1-\kappa)^2} + \frac{2}{(1-\kappa)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right], \quad \text{(EC.16)}\end{aligned}$$

where we use that $\sum_{t=0}^{T-1}\frac{1}{t+1} \leq (1+\log T)$. To simplify notation, we put $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ for the remainder of the proof. We use Lemma EC.5 to upper bound the total bias, $\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]$ which can be decomposed as:

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] = \sum_{t=0}^{2\tau}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] + \sum_{t=2\tau+1}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]. \quad \text{(EC.17)}$$

First, note that for a decaying step-size $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$ we have

$$\sum_{t=0}^{T-1}\alpha_t = \frac{1}{\omega(1-\gamma)}\sum_{t=0}^{T-1}\frac{1}{(t+1)} \leq \frac{1+\log T}{\omega(1-\gamma)}.$$

We will combine this with Lemma EC.5 to upper bound each term separately. First,

$$\begin{aligned}\sum_{t=0}^{2\tau}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] &\leq \sum_{t=0}^{2\tau}\left(6B^2\sum_{i=0}^{t-1}\alpha_i\right) + \sum_{t=0}^{2\tau}B^2(\gamma\lambda)^t \\ &\leq \frac{6B^2}{\omega(1-\kappa)}\sum_{t=0}^{2\tau}\sum_{i=0}^{T-1}\frac{1}{(i+1)} + 2B^2\tau \leq \frac{14B^2\tau}{\omega(1-\kappa)}(1+\log T),\end{aligned}$$

where we used the fact that $\omega, \kappa, (\gamma\lambda) < 1$. Similarly,

$$\sum_{t=2\tau+1}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2(1+2\tau)\sum_{t=2\tau+1}^{T-1}\alpha_{t-2\tau} \leq 6B^2(1+2\tau)\sum_{t=0}^{T-1}\alpha_t \leq \frac{6B^2(1+2\tau)}{\omega(1-\kappa)}(1+\log T).$$

Combining the two parts, we get

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq \frac{B^2(6+26\tau)}{\omega(1-\kappa)}(1+\log T).$$

Using this in conjunction with Equation (EC.16) we get,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|V_{\theta_t} - V_{\theta^*}\|_D^2\right] \leq \frac{B^2(1+\log T)}{\omega T(1-\kappa)^2} + \frac{2}{T(1-\kappa)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right].$$

Simplifying and putting back $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$, we get our final result.

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{B^2}{\omega T(1-\kappa)^2}(1+\log T) + \frac{2B^2\left(6+26\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\kappa)^2}(1+\log T)$$
$$\leq \frac{B^2\left(13+52\tau^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\kappa)^2}(1+\log T). \quad \square$$

We remark that Equation (EC.16) implies a convergence rate of $\mathcal{O}(\log T/T)$ for the iterate θ_T (and hence the value function V_{θ_T}) itself but the bounds degrade by a factor of ω. In particular, we have

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_T}\|_D^2\right] \leq \mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] \leq \frac{B^2\left(13+52\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{\omega^2 T(1-\kappa)^2}(1+\log T) \tag{EC.18}$$

where the first inequality follows using Lemma 1.

B.3. Proof of supporting lemmas.

In this subsection, we provide standalone proofs of Lemma EC.2 and EC.4 used above.

LEMMA EC.2 *For all $\theta \in \Theta_R$, $\|x_t(\theta, z_{0:t})\|_2 \leq B$ with probability 1. Additionally, $\|\bar{x}(\theta)\|_2 \leq B$.*

Proof of Lemma EC.2. We start with the mathematical expression for $x_t(\theta, z_{0:t})$.

$$x_t(\theta, z_{0:t}) = \delta_t(\theta) z_{0:t} \;\Rightarrow\; \|x_t(\theta, z_{0:t})\|_2 = |\delta_t(\theta)| \|z_{0:t}\|_2.$$

We give an upper bound on both $|\delta_t(\theta)|$ and $\|z_{0:t}\|_2$. Starting with the definition of $\delta_t(\theta)$ and using that $\|\phi(s_t)\|_2 \leq 1 \;\forall\, t$ along with $\|\theta\|_2 \leq R$, we get

$$|\delta_t(\theta)| = \left|r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right| \leq r_{\max} + \|\phi(s_t')\|_2\|\theta\|_2 + \|\phi(s_t)\|_2\|\theta\|_2 \leq (r_{\max} + 2R).$$

Next,

$$\|z_{0:t}\|_2^2 = \left\|\sum_{k=0}^{t}(\gamma\lambda)^k\phi(s_{t-k})\right\|_2^2 \leq \left(\sum_{k=0}^{t}(\gamma\lambda)^k\right)^2 \leq \left(\sum_{k=0}^{\infty}(\gamma\lambda)^k\right)^2 = \frac{1}{(1-\gamma\lambda)^2}.$$

Combining these two implies the first part of our claim.

$$\|x_t(\theta, z_{0:t})\|_2 = |\delta_t(\theta)| \|z_{0:t}\|_2 \leq \frac{(r_{\max}+2R)}{(1-\gamma\lambda)} = B.$$

Note that can easily show an upper bound $\|\delta_t(\theta) z_{l:t}\|_2 \leq B$ for any pair $(\theta, z_{l:t})$ with $l \leq t$. Consider,

$$\|z_{l:t}\|_2^2 \leq \|z_{-\infty:t}\|_2^2 \leq \left(\sum_{k=0}^{\infty}(\gamma\lambda)^k\right)^2 = \frac{1}{(1-\gamma\lambda)^2}$$
$$\Rightarrow\; \|\delta_t(\theta) z_{l:t}\|_2 = |\delta_t(\theta)| \|z_{l:t}\|_2 \leq \frac{(r_{\max}+2R)}{(1-\gamma\lambda)} = B.$$

Taking $l \to -\infty$ implies that $\|\delta_t(\theta) z_{-\infty:t}\|_2 \leq B$. As $\bar{x}(\theta) = \mathbb{E}\left[\delta_t(\theta) z_{-\infty:t}\right]$, we also have a uniform norm bound on the expected updates, $\|\bar{x}(\theta)\|_2 \leq B$, as claimed. $\square$

LEMMA EC.4 *Consider any $l \leq t$ and any $\theta, \theta' \in \Theta_R$. With probability 1,*

(a) $|\zeta_t(\theta, z_{l:t})| \leq 2B^2$.

(b) $|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| \leq 6B\left\|(\theta - \theta')\right\|_2$.

(c) The following two bounds also hold,

$$|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau \text{ for all } \tau \leq t,$$
$$|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| \leq B^2(\gamma\lambda)^t.$$

Proof of Lemma EC.4. Throughout, we use the assumption that basis vectors are normalized i.e. $\|\phi(s_t)\|_2 \leq 1 \;\forall\, t$.

Proof of Part (a): We show a uniform norm bound on $\zeta_t(\theta, z_{l:t})\ \forall\ \theta \in \Theta_R$. First consider the following:

$$\begin{aligned}\|\delta_t(\theta)z_{l:t}\|_2 = |\delta_t(\theta)|\|z_{l:t}\|_2 &\leq \left|r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right|\left\|\sum_{k=0}^{t-l}(\gamma\lambda)^k\phi(s_{t-k})\right\|_2 \\ &\leq |r_t + \|\phi(s_t')\|_2\|\theta\|_2 + \|\phi(s_t)\|_2\|\theta\|_2|\left\|\sum_{k=0}^{\infty}(\gamma\lambda)^k\phi(s_{t-k})\right\|_2 \\ &\leq \frac{(r_{\max}+2R)}{(1-\gamma\lambda)} = B.\end{aligned}$$

Using this along with the fact that $\|\theta - \theta^*\|_2 \leq 2R \leq B$ and $\|\bar{x}(\theta)\|_2 \leq B$ for all $\theta \in \Theta_R$, we get

$$\begin{aligned}|\zeta_t(\theta, z_{l:t})| = \left|\left(\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\right)^\top(\theta - \theta^*)\right| &\leq \|\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\|_2\|(\theta-\theta^*)\|_2 \\ &\leq \left(\|\delta_t(\theta)z_{l:t}\|_2 + \|\bar{x}(\theta)\|_2\right)\|(\theta-\theta^*)\|_2 \\ &\leq 2B\|(\theta-\theta^*)\|_2 \leq 2B^2.\end{aligned}$$

Proof of Part (b): To show that $\zeta_t(\cdot, z_{l:t})$ is L-Lipschitz, consider the following inequality for any four vectors (a_1, b_1, a_2, b_2), which follows as a direct application of Cauchy-Schwartz.

$$\left|a_1^\top b_1 - a_2^\top b_2\right| = \left|a_1^\top(b_1 - b_2) + b_2^\top(a_1 - a_2)\right| \leq \|a_1\|_2\|b_1 - b_2\|_2 + \|b_2\|_2\|a_1 - a_2\|_2.$$

This implies,

$$\begin{aligned}|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| &= \left|\left(\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\right)^\top(\theta-\theta^*) - \left(\delta_t(\theta')z_{l:t} - \bar{x}(\theta')\right)^\top(\theta'-\theta^*)\right| \\ &\leq \|\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\|_2\|\theta-\theta'\|_2 + \|\theta'-\theta^*\|_2\|(\delta_t(\theta)z_{l:t} - \bar{x}(\theta)) - (\delta_t(\theta')z_{l:t} - \bar{x}(\theta'))\|_2 \\ &\leq 2B\|\theta-\theta'\|_2 + 2R\Big[\|z_{l:t}\left(\delta_t(\theta) - \delta_t(\theta')\right)\|_2 + \|\bar{x}(\theta) - \bar{x}(\theta')\|_2\Big] \\ &\leq 2B\|\theta-\theta'\|_2 + \frac{8R}{(1-\gamma\lambda)}\|\theta-\theta'\|_2 \\ &\leq 6B\|\theta-\theta'\|_2,\end{aligned}$$

where the last inequality follows as $\frac{R}{1-\gamma\lambda} \leq B/2$ by definition. In the penultimate inequality, we used that $\|z_{l:t}\left(\delta_t(\theta) - \delta_t(\theta')\right)\|_2 \leq \frac{2}{(1-\gamma\lambda)}\|\theta-\theta'\|_2$ which is easy to prove. Consider,

$$\begin{aligned}\|z_{l:t}\left(\delta_t(\theta) - \delta_t(\theta')\right)\|_2 &\leq \|z_{l:t}\|_2|\left(\delta_t(\theta) - \delta_t(\theta')\right)| \\ &\leq \left\|\sum_{k=0}^{\infty}(\gamma\lambda)^k\phi(s_{t-k})\right\|_2|\left(\delta_t(\theta) - \delta_t(\theta')\right)| \\ &\leq \frac{1}{(1-\gamma\lambda)}\left|\left(\gamma\phi(s_t') - \phi(s_t)\right)^\top(\theta-\theta')\right| \\ &\leq \frac{\left(\|\phi(s_t')\|_2 + \|\phi(s_t)\|_2\right)}{(1-\gamma\lambda)}\|\theta-\theta'\|_2 \leq \frac{2}{(1-\gamma\lambda)}\|\theta-\theta'\|_2.\end{aligned}$$

As $\bar{x}(\theta) = \mathbb{E}\left[\delta_t(\theta)z_{-\infty:t}\right]$, this also implies $\|\bar{x}(\theta) - \bar{x}(\theta')\|_2 \leq \frac{2}{(1-\gamma\lambda)}\|\theta-\theta'\|_2$ which completes the proof.

Proof of Part (c): To show that $|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau$ for all $\theta \in \Theta_R$ and $\tau \leq t$, we use that $\|\theta - \theta^*\|_2 \leq 2R \leq B$.

$$\begin{aligned}
|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| &= \left|\left(\delta_t(\theta) z_{0:t} - \delta_t(\theta) z_{t-\tau:t}\right)^\top (\theta - \theta^*)\right| \\
&\leq |\delta_t(\theta)| \|z_{0:t} - z_{t-\tau:t}\|_2 \|\theta - \theta^*\|_2 \\
&\leq \left|r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right| \left\|\sum_{k=\tau}^{\infty} (\gamma\lambda)^k \phi(s_{t-k})\right\|_2 B \\
&\leq |r_t + 2\|\theta\|_2| \cdot \frac{(\gamma\lambda)^\tau}{(1-\gamma\lambda)} \cdot B \\
&\leq B\frac{(r_{\max} + 2R)}{(1-\gamma\lambda)}(\gamma\lambda)^\tau \\
&= B^2(\gamma\lambda)^\tau.
\end{aligned}$$

Similarly,

$$\begin{aligned}
|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| &\leq \left|\delta_t(\theta)(z_{0:t} - z_{-\infty:t})^\top(\theta - \theta^*)\right| \\
&\leq |\delta_t(\theta)| \|z_{0:t} - z_{-\infty:t}\|_2 \|\theta - \theta^*\|_2 \\
&\leq \left|\left(r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right)\right| \left\|\sum_{k=t}^{\infty} (\gamma\lambda)^k \phi(s_{t-k})\right\|_2 B \\
&\leq B\frac{(r_{\max} + 2R)}{(1-\gamma\lambda)}(\gamma\lambda)^t \\
&\leq B^2(\gamma\lambda)^t. \quad \square
\end{aligned}$$

Appendix C: Proofs of Additional Lemmas

In this section, we prove some additional lemmas stated and used in other parts of the paper. We first give a proof of Lemma 7 which gives an upper bound on the projection radius, R.

LEMMA 7 *$\|\theta^*\|_\Sigma \leq \frac{2r_{\max}}{(1-\gamma)^{3/2}}$ and hence $\|\theta^*\|_2 \leq \frac{2r_{\max}}{\sqrt{\omega}(1-\gamma)^{3/2}}$.*

Proof of Lemma 7. Because rewards are uniformly bounded, $|V_\mu(s)| \leq r_{\max}/(1-\gamma)$ for all $s \in \mathcal{S}$. Recall that V_μ denotes the true value function of the Markov reward process. This implies that

$$\|V_\mu\|_D \leq \|V_\mu\|_\infty \leq \frac{r_{\max}}{(1-\gamma)}.$$

Lemma 2 along with simple matrix inequalities enable a simple upper bound on $\|\theta^*\|_2$. We have

$$\|V_{\theta^*} - V_\mu\|_D \leq \frac{1}{\sqrt{1-\gamma^2}}\|V_\mu - \Pi_D V_\mu\|_D \leq \frac{1}{\sqrt{1-\gamma^2}}\|V_\mu\|_D \leq \frac{1}{\sqrt{1-\gamma}}\|V_\mu\|_D,$$

where the penultimate inequality holds by the Pythagorean theorem. By the reverse triangle inequality we have $\big|\|V_{\theta^*}\|_D - \|V_\mu\|_D\big| \leq \|V_{\theta^*} - V_\mu\|_D$. Thus,

$$\|V_{\theta^*}\|_D \leq \|V_{\theta^*} - V_\mu\|_D + \|V_\mu\|_D \leq \frac{2}{\sqrt{1-\gamma}}\|V_\mu\|_D \leq \frac{2}{\sqrt{1-\gamma}}\frac{r_{\max}}{(1-\gamma)}.$$

Recall from Section 2 we have, $\|V_{\theta^*}\|_D = \|\theta^*\|_\Sigma$ which establishes first part of the claim. The second claim uses that $\|\theta^*\|_\Sigma \geq \omega\|\theta^*\|_2$ which follows by Lemma 1. □

Next, we give a combined proof of Lemmas EC.1 and 13 which quantify the progress of the expected updates towards the limit point θ^* for TD(λ) and the Q-function approximation algorithm respectively. These lemmas can be restated more generally as shown below, instead of using the Bellman operators $F(\cdot)$ and $T^{(\lambda)}(\cdot)$.

LEMMA EC.6. *Consider a linear function approximation such that $J_\theta = \Phi\theta$. Let $\Pi_D H(\cdot)$ be a contraction with respect to $\|\cdot\|_D$ with modulus γ and let J_{θ^*} be the unique fixed point of $\Pi_D H(\cdot)$, i.e. $J_{\theta^*} = \Pi_D H J_{\theta^*}$. Define $\bar{g}(\theta) = \Phi^\top D\,(H\Phi\theta - \Phi\theta)$ for all $\theta \in \mathbb{R}^d$ to be the expected update. Then,*

$$\bar{g}(\theta)^\top(\theta^* - \theta) \geq (1-\gamma)\|J_{\theta^*} - J_\theta\|_D^2.$$

Proof of Lemma EC.6. We have

$$\begin{aligned}
(\theta^* - \theta)^\top \bar{g}(\theta) &= (\theta^* - \theta)^\top \Phi^\top D\,(H\Phi\theta - \Phi\theta) \\
&= \langle \Phi(\theta^* - \theta), (H\Phi\theta - \Phi\theta)\rangle_D \\
&= \langle \Pi_D\Phi(\theta^* - \theta), (H\Phi\theta - \Phi\theta)\rangle_D && \text{(EC.19)} \\
&= \langle \Phi(\theta^* - \theta), \Pi_D\,(H\Phi\theta - \Phi\theta)\rangle_D && \text{(EC.20)} \\
&= \langle \Phi(\theta^* - \theta), \Pi_D H\Phi\theta - \Phi\theta\rangle_D \\
&= \langle \Phi(\theta^* - \theta), \Pi_D H\Phi\theta - \Phi\theta^* + \Phi\theta^* - \Phi\theta\rangle_D \\
&= \|\Phi(\theta^* - \theta)\|_D^2 - \langle \Phi(\theta^* - \theta), \Phi\theta^* - \Pi_D H\Phi\theta\rangle_D \\
&\geq \|\Phi(\theta^* - \theta)\|_D^2 - \|\Phi(\theta^* - \theta)\|_D \cdot \|\Pi_D H\Phi\theta - \Phi\theta^*\|_D \\
&\geq \|\Phi(\theta^* - \theta)\|_D^2 - \gamma \cdot \|\Phi(\theta^* - \theta)\|_D^2 && \text{(EC.21)} \\
&= (1-\gamma)\cdot\|\Phi(\theta^* - \theta)\|_D^2 = (1-\gamma)\cdot\|J_{\theta^*} - J_\theta\|_D^2,
\end{aligned}$$

where in going to Equation (EC.19), we used that $\forall\, \mathbf{x} \in \text{Span}(\Phi)$, we have $\Pi_D\,\mathbf{x} = \mathbf{x}$. In Equation (EC.20), we used that the projection matrix Π_D is symmetric. In going to Equation (EC.21), we used that that $\Pi_D H(\cdot)$ is a contraction operator with modulus γ with $\Phi\theta^*$ as its fixed point, which implies that $\|\Pi_D H\Phi\theta - \Phi\theta^*\|_D = \|\Pi_D H\Phi\theta - \Pi_D H\Phi\theta^*\|_D \leq \gamma\|\Phi\theta - \Phi\theta^*\|_D$. □

Finally, we restate and prove Lemmas 14 and 15 used in Section 10 to analyze Q-learning for optimal stopping problems under a linear approximation model.

LEMMA 14 *For any fixed* $\theta \in \mathbb{R}^d$, $\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] \leq 2\sigma^2 + 8\|Q_\theta - Q_{\theta^*}\|_D^2$ *where* $\sigma^2 = \mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right]$.

Proof of Lemma 14. We use notation and proof strategy mirroring the proof of Lemma 5. Set $\phi = \phi(s_t)$, $\phi' = \phi(s_t')$ and $U' = U(s')$. Define $\xi = (\theta^* - \theta)^\top \phi$ and $\xi' = (\theta^* - \theta)^\top \phi'$. By stationarity ξ and ξ' have the same marginal distribution and $\mathbb{E}[\xi^2] = \|Q_{\theta^*} - Q_\theta\|_D^2$. Using the formula for $g_t(\theta)$ in Equation (33), we have

$$\begin{aligned}
\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] &\leq 2\mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right] + 2\mathbb{E}\left[\|g_t(\theta) - g_t(\theta^*)\|_2^2\right] \\
&= 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\phi^\top(\theta^* - \theta) - \gamma\left[\max\left(U', {\phi'}^\top\theta^*\right) - \max\left(U', {\phi'}^\top\theta\right)\right]\right)\right\|_2^2\right] \\
&\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\left|\phi^\top(\theta^* - \theta)\right| + \gamma\left|\max\left(U', {\phi'}^\top\theta^*\right) - \max\left(U', {\phi'}^\top\theta\right)\right|\right)\right\|_2^2\right] \\
&\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\left|\phi^\top(\theta^* - \theta)\right| + \gamma\left|{\phi'}^\top(\theta^* - \theta)\right|\right)\right\|_2^2\right] \qquad \text{(EC.22)} \\
&\leq 2\sigma^2 + 2\mathbb{E}[|\xi + \gamma\xi'|^2] \\
&\leq 2\sigma^2 + 4\left(\mathbb{E}\left[|\xi|^2\right] + \gamma^2\mathbb{E}\left[|\xi'|^2\right]\right) \\
&= 2\sigma^2 + 4(1+\gamma^2)\|Q_\theta - Q_{\theta^*}\|_D^2 \leq 2\sigma^2 + 8\|Q_\theta - Q_{\theta^*}\|_D^2,
\end{aligned}$$

where we used the assumption that features are normalized so that $\|\phi\|_2^2 \leq 1$ almost surely. Additionally, in going to Equation (EC.22), we used that $|\max(c_1, c_3) - \max(c_2, c_3)| \leq |c_1 - c_2|$ for any scalars c_1, c_2 and c_3. □

LEMMA 15 *Define* $G = (r_{\max} + 2R)$. *With probability 1,* $\|g_t(\theta)\|_2 \leq G$ *for all* $\theta \in \Theta_R$.

Proof of Lemma 15. We start with the mathematical expression for the semi-gradient,

$$g_t(\theta) = \left(u(s_t) + \gamma \max\left\{U(s_t'), \phi(s_t')^\top\theta\right\} - \phi(s_t)^\top\theta\right)\phi(s_t).$$

As $r_{\max} \leq R$, we have: $\max\left\{U(s_t'), \phi(s_t')^\top\theta\right\} \leq \max\left\{U(s_t'), \|\phi(s_t')\|_2\|\theta\|_2\right\} \leq R$. Then,

$$\begin{aligned}
\|g_t(\theta)\|_2^2 &= \left(u(s_t) + \gamma \max\left\{U(s_t'), \phi(s_t')^\top\theta\right\} - \phi(s_t)^\top\theta\right)^2 \|\phi(s_t)\|^2 \\
&\leq \left(r_{\max} + \gamma R - \phi(s_t)^\top\theta\right)^2 \\
&\leq \left(r_{\max} + \gamma R + \|\phi(s_t)\|_2\|\theta\|_2\right)^2 \leq \left(r_{\max} + 2R\right)^2 = G^2.
\end{aligned}$$

We used here that the basis vectors are normalized, $\|\phi(s_t)\|_2 \leq 1$ for all t. □