PER-ETD: A POLYNOMIALLY EFFICIENT EMPHATIC TEMPORAL DIFFERENCE LEARNING METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Emphatic temporal difference (ETD) (Sutton et al., 2016) learning is a successful method to conduct the off-policy value function estimation with function approximation. Although the stability and asymptotic behavior of ETD have been established shortly after its birth, the finite-time analysis of such an algorithm still remains vacant. In this work, we propose a new variant of ETD, called PER-ETD (i.e., **PE**riodically **R**estarted-ETD) and show that PER-ETD (either PER-ETD(0) or PER-ETD(λ)) converges to the same fixed point of ETD but with only polynomial sample complexity (whereas ETD takes exponential sample complexity). Effectively, PER-ETD alternates the update between the emphatic weights and the function approximation parameter. The restart of the update of the follow-on trace periodically effectively reduces the variance. More importantly, PER-ETD features a logarithmically increasing period length with the number of iterations so that it attains the same value function estimation accuracy as ETD under the polynomial sample complexity. Experimentally, we demonstrate that PER-ETD can converge in the case that neither TD nor ETD converges. We observe that there does exist a choice of the period length for PER-ETD, which attains the best tradeoff between the variance and bias errors. And counter-intuitively, PER-ETD(λ) can have a larger bias error than PER-ETD(0), which later is explained through the fixed points of PER-ETD(λ).

1 INTRODUCTION

As a major value function valuation algorithm, temporal difference (TD) learning (Sutton, 1988; Dayan, 1992; Dayan & Sejnowski, 1994; Jaakkola et al., 1994; Tsitsiklis & Van Roy, 1997; Dalal et al., 2018a; Bhandari et al., 2018) has been applied widely to developing various planning and policy optimization approaches in reinforcement learning. Although TD performs successfully in on-policy settings, where an agent can interact with environments with the target policy, TD with function approximation can perform poorly and diverge under off-policy settings if the agent has access only to history data sampled by a behavior policy, as pointed out by Baird (1995); Tsitsiklis & Van Roy (1997); Mahmood et al. (2015). To address such an issue, the gradient temporal-difference (GTD) (Sutton et al., 2008) and least-squares temporal difference (LSTD) (Yu, 2010) algorithms have been proposed, which have been shown to converge in off-policy settings. However, since GTD and LSTD still apply the objective function based on the behavior policy, their converging points can be largely biased from the true value function due to the distribution mismatch between the target and the behavior policies, even when the function approximation class is large (Kolter, 2011), making the evaluation of GTD and LSTD unreliable.

In order to provide more accurate evaluation, Sutton et al. (2016) proposed a novel emphatic temporal difference (ETD) algorithm, which introduces the *follow-on trace* to eliminate the distribution mismatch. Ziwei: Sutton et al. (2016); Mahmood et al. (2015) showed the stability of the ETD through the negative definiteness of the key matrix, [what exactly did they show? smaller bias, close to ground truth?] and Yu (2015) provided the asymptotic convergence guarantee for ETD. Further, ETD has been demonstrated to achieve successful performance on many tasks (Ghiassian et al., 2017; Ni, 2021). However, although ETD can reduce the large bias error and offer accurate evaluation, it often suffers from the large variance due to the follow-on trace estimation over a large time-horizon (Hallak et al., 2016). Consequently, the convergence of ETD can be very slow and unstable. In fact, careful analysis can show that the variance of ETD can grow exponentially fast with

the number of iterations so that ETD requires **exponentially** large sample/computational complexity to converge. A natural idea to reduce the variance is to update the emphatic variable only over a finite-time horizon, but such an approach likely introduces a large estimation bias error, which seems to suggest that an exponential sample size is needed to attain an accurate value function estimation of ETD. This thus poses the following intriguing questions:

- Can we design a new variant of ETD, which overcomes its large variance issue and at the same time approaches to the same accurate evaluation as ETD under only **polynomial** sample complexity?
- If so, can we theoretically capture the desired performance guarantee for the proposed algorithm via a finite-time analysis of its convergence?

In this work, we provide affirmative answers to both questions.

1.1 Main Contributions

We propose a new variant of ETD, called PER-ETD (i.e., **PE**riodically **R**estarted-ETD), in which for each update of the value function parameter we restart the iteration for the follow-on trace and update it only for b times (where we call b as the **period length**). Effectively, PER-ETD alternates the update between the emphatic weights and the function approximation parameter. The restart of the update of the follow-on trace periodically effectively reduces the variance. More importantly, PER-ETD features a logarithmically increasing period length with the number of iterations so that it attains the same value function estimation accuracy as ETD under the polynomial sample complexity.

We provide the finite-time analysis for the proposed PER-ETD algorithm, and show that PER-ETD (either PER-ETD(0) or PER-ETD(λ)) converges to the same fixed point of ETD but with only polynomial sample complexity (whereas ETD takes exponential sample complexity). Our analysis features the following key insights: (a) the period length plays the role of trading off between the variance (of emphatic variable) and bias error (with respect to the fixed point of ETD), and its optimal choice of logarithmical increase with the number of iterations achieves the best tradeoff and attains the polynomial sample efficiency; (b) our analysis captures how the mismatch between the behavior and target policies affects the convergence rate of PER-ETD. Interestingly, as long as the mismatch is below a certain threshold, then PER-ETD achieves the same convergence rate as the on-policy TD algorithm; and once the mismatch is above the threshold, then the converge rate of PER-ETD gradually decays as the level of mismatch increases. [Any other points we want to mention?]

Experimentally, we demonstrate that PER-ETD can converge in the case that neither TD nor ETD converges. Further, our experiments provide the following two interesting observations. (a) There does exist a choice of the period length for PER-ETD, which attains the best tradeoff between the variance and bias errors. Below such a choice, the bias error is large so that value function estimation is not accurate, and above it the variance error is large so that the convergence is unstable. (b) Under a small period length b, PER-ETD(λ) can have larger bias error than PER-ETD(0), which is somewhat counter-intuitive, but in fact reasonable. This is because under a small period length b, the fixed point of PER-ETD(1) can be distant from the fixed point of ETD. Thus, as λ increases, the fixed point of PER-ETD(λ) gets closer to that of PER-ETD(1) rather than the fixed point of ETD. This suggests that if we adopt a small period length b in practice, then PER-ETD(0) will offer a better value function estimation than PER-ETD(λ).

1.2 Related Works

Finite-sample analysis of TD learning and GTD: Dalal et al. (2018a) studied the TD(0) algorithm with i.i.d. samples. They gave both an $\mathcal{O}\left(\frac{1}{T^\sigma}\right)$, $0 < \sigma < 1$, convergence guarantee in expectation and polynomial convergence rate with high probability. In the meanwhile, Bhandari et al. (2018) provided an extended analysis which covers both i.i.d. samples case and Markovian samples in a rate of $\tilde{\mathcal{O}}\left(\frac{1}{T}\right)$ convergence in expectation. Lakshminarayanan & Szepesvari (2018) investigated the linear stochastic approximation (SA) with constant stepsize of which TD(0) and GTD could be treated as special cases. They have proven, with i.i.d. samples, linear SA converges with a rate of $\mathcal{O}(\frac{1}{T})$ depending on the variance and $\mathcal{O}(\frac{1}{T^2})$ depending on the initial distance from the fixed point, which match the lower bounds they developed. Recently, Kotsalis et al. (2020) analyzed the TD learning algorithm with on-policy samples through the lens of generalized monotone variational

inequality. They incorporated the TD learning with periodic update and stochastic operator extrapolation technique, which provides huge performance improvement when discounting fact γ near to 1. For the GTD algorithms, Dalal et al. (2018b); Gupta et al. (2019); Wang et al. (2018) both provided the finite-time analysis of GTD and its variants.

Emphatic Temporal Difference Learning: The emphatic temporal difference learning was originally proposed by Sutton et al. (2016). In this seminal work, Sutton et al. (2016) introduced the follow-on trace to overcome the distribution mismatch between the behavior policy and the target policy. Later, Yu (2015) provided the asymptotic convergence guarantee. Hallak et al. (2016) investigated the variance of the follow-on trace and found out that it may has unbounded variance. They proposed a general framework of ETD with a variable decay rate, $\beta \in (0, 1)$, of follow-on trace instead of γ in the original design. At the cost of larger approximation error of the fixed point, the variance of follow-on trace is bounded with small β. Imani et al. (2018) developed a new policy gradient theorem, where the emphatic weighted is used to correct the distribution shift. Zhang et al. (2020b) provided a new variant of ETD, where the emphatic weights are estimated through function approximation. Van Hasselt et al. (2018); Jiang et al. (2021) studied the ETD algorithm with deep neural function class.

Comparison to concurrent work: During our preparation of this paper, a concurrent work (Zhang & Whiteson, 2021) was posted on arXiv, which proposed a similar algorithm of truncating the update of emphatic variable to reduce the variance of ETD. However, our theoretical guarantee is very different from that given in (Zhang & Whiteson, 2021). Our paper shows that PER-ETD converges to the fixed point of the original ETD with the infinite updates of emphatic variable, whereas (Zhang & Whiteson, 2021) shows that their algorithm converges only to the fixed point determined by the truncated ETD and its distance to the fixed point of ETD is truncation-length dependent. Further, Zhang & Whiteson (2021) obtained the convergence through a fixed truncation-length n and hidden the exponentially exploded terms in the big-O notation. Their bound will lose effectiveness if n is scalable with T. Our analysis provides the guidance on the period length in order to attain arbitrarily small bias error to the fixed point of ETD as well as the dependence of the convergence rate on the mismatch between behavior and target polices, whereas (Zhang & Whiteson, 2021) does not capture these properties.

2 Background and Preliminaries

2.1 Markov Decision Process

We consider the infinite-horizon Markov decision process (MDP) defined by the five tuple $(\mathcal{S}, \mathcal{A}, r, \mathsf{P}, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces respectively, which are both assumed to be finite sets, $r : \mathcal{S} \times \mathcal{A} \to R$ denotes the reward function, $\mathsf{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the transition kernel, where $\Delta(\mathcal{S})$ denotes the probability simplex over the state space $\mathcal{S}$, and $\gamma \in (0, 1)$ is the discount factor. The policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ of an agent maps from the state space to the probability simplex over the action space $\mathcal{A}$, i.e., $\pi(a|s)$ represents the probability of taking the action a under the state s. At any time t, given that the system is at the state s_t, the agent takes an action a_t with the probability $\pi(a_t|s_t)$, and receives a reward $r(s_t, a_t)$. The system then transitions to the next state s_{t+1} at time $t + 1$ with the probability $\mathbb{P}(s_{t+1}|s_t, a_t)$.

For a given policy π, we define the value function corresponding to an initial state $s_0 = s \in \mathcal{S}$ as,

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, \pi\right].$$

Let the cardinality of the state space be $|\mathcal{S}| = N$. Then the value function over the space space can be expressed as the vector $V_\pi = (V_\pi(1), V_\pi(2), \ldots, V_\pi(N))^\top \in \mathbb{R}^N$.

When the state space is large, we approximate the value function V_π via a linear function class as $V_\theta(s) = \phi^\top(s)\theta$, where $\phi(s) \in \mathbb{R}^d$ denotes the feature vector, and $\theta \in \mathbb{R}^d$ denotes the parameter to be learned. We further let $\Phi = [\phi(1), \phi(2), \ldots, \phi(N)]^\top$ denote the feature matrix, and then $V_\theta = \Phi\theta$. We assume that the feature matrix Φ has linearly independent columns and each feature vector has bounded ℓ_1 norm, i.e., $\|\phi(s)\|_2 \leq B_\phi$ for any $s \in \mathcal{S}$.

Algorithm	Error bounds of fixed point	Sample Complexity
GTD(Wang et al., 2018)	✗	$\mathcal{O}\left(\frac{1}{T}\right)$
ETD(Hallak et al., 2016; Yu, 2015)	✓	✗
Truncated ETD(Zhang & Whiteson, 2021)	✓[1]	$\mathcal{O}\left(\frac{\log(T)^3}{T}\right)$
This work	✓	$\mathcal{O}\left(\frac{1}{T^a}\right)$

[1] Zhang & Whiteson (2021) provided a performance guarantee of the changing fixed point ($w_{*,n}$ in their paper) with large enough truncation number n. Theorectically, $w_{*,n}$ converges to the original fixed points of ETD (also the fixed point of this work) when n goes to infinity.

Table 1: Comparison of different TD-type algorithm with off-policy samples

2.2 Temporal Difference (TD) Learning for On-policy Evaluation

In order to evaluate the value function for a given target policy π (i.e., find the linear function approximation parameter θ, the temporal difference (TD) learning can be employed based on a sampling trajectory, which takes the following update rule at each time t:

$$\theta_{t+1} = \theta_t + \eta_t \left(r(s_t, a_t) + \gamma \theta_t^\top \phi(s_{t+1}) - \theta_t^\top \phi(s_t)\right) \phi(s_t), \tag{1}$$

where η_t is the stepsize at time t. The main idea here is to follow the Bellman operation update to approach its fixed point, and the above sampled version update can be viewed as the so-called semi-gradient descent update. If the trajectory is sampled by the target policy π, then the above TD algorithm can be shown to converge to the fixed point solution, where the convergence is guaranteed by the negative definiteness of the so-called *key matrix* $A := \lim_{t\to\infty} \mathbb{E}\left[(\gamma\phi(s_t+1) - \phi(s_t))\phi^\top(s_t)\right]$.

2.3 Emphatic TD (ETD) Learning for Off-policy Evaluation

Consider the off-policy setting, where the goal is still to evaluate the value function for a given target policy π, but the agent has the access only to the off-policy sample trajectories sampled under a behavior policy μ. Namely, at each time t, the probability of taking an action a_t given s_t is $\mu(a_t|s_t)$. Let d_μ denote the stationary distribution of the Markov chain induced by the behavior policy μ, which satisfies $d_\mu^\top = d_\mu^\top P_\pi$. We assume that $d_\mu(s) > 0$ for all states.

The mismatch between the target and behavior policies can be addressed by incorporating the importance sampling factor $\rho(s,a) := \frac{\pi(a|s)}{\mu(a|s)}$ into eq. (1) to adjust the TD learning update direction. However, with such modification, the key matrix A may not be negative definite any more so that the algorithm is not guaranteed to converge.

In order to address such a divergence issue, the emphatic temporal difference (ETD) algorithm has been proposed by Sutton et al. (2016), where the simple ETD(0) algorithm takes the following update

$$\theta_{t+1} = \theta_t + \eta_t \rho(s_t, a_t) F_t \left(r(s_t, a_t) + \gamma \theta_t^\top \phi(s_{t+1}) - \theta_t^\top \phi(s_t)\right) \phi(s_t). \tag{2}$$

In eq. (2), in addition to the importance sampling factor ρ, a follow-on trace coefficient F_t is introduced as a calibration factor, which is updated as

$$F_t = \gamma \rho(s_{t-1}, a_{t-1}) F_{t-1} + 1, \tag{3}$$

with initialization $F_0 = 1$. With such a follow-on trace factor, the key matrix becomes negative definite, and the ETD algorithm has been shown to converge asymptotically in Yu (2015) to the fixed point

$$\theta^* = \left(\Phi^\top F(I - \gamma P_\pi)\Phi\right)^{-1} \Phi^\top F r_\pi,$$

where $F = \mathrm{diag}(f(1), f(2), \ldots, f(N))$ and $f(i) = d_\mu(i) \lim_{t\to\infty} \mathbb{E}\left[F_t | s_t = i\right]$.

Similarly, the ETD(λ) algorithm can be further derived, which has the following update

$$\theta_{t+1} = \theta_t + \eta_t \rho(s_t, a_t) \left(r(s_t, a_t) + \gamma \theta_t^\top \phi(s_{t+1}) - \theta_t^\top \phi(s_t)\right) e_t,$$

where F_t and e_t are updated as

$$\begin{aligned} F_t &= \gamma\rho(s_{t-1}, a_{t-1})F_{t-1} + 1 \\ M_t &= \lambda + (1-\lambda)F_t \\ e_t &= \gamma\lambda\rho(s_{t-1}, a_{t-1})e_{t-1} + M_t\phi(s_t), \end{aligned}$$

with $F_0 = M_0 = 1$ and $e_0 = \phi(s_0)$. It has been shown that with a diminishing stepsize, the ETD(λ) algorithm converges to the fixed point given by

$$\theta_\lambda^* = \left(\Phi^\top M(I-\gamma\lambda P_\pi)^{-1}(I-\gamma P_\pi)\Phi\theta\right)^{-1}\Phi^\top M(I-\gamma\lambda P_\pi)^{-1}r_\pi,$$

where $M = \mathrm{diag}(m(1), m(2), \ldots, m(N))$ and $m(i) = d_\mu(i)\lim_{t\to\infty}\mathbb{E}\left[M_t|s_t = i\right]$.

2.4 Notations

For the simplicity of expression, we use the shorthand notations in the remaining paragraphs. For a fixed integer b, let $s_t^\tau := s_{t(b+1)+\tau}$, $a_t^\tau := a_{t(b+1)+\tau}$, $\rho_t^\tau = \frac{\pi(a_t^\tau|s_t^\tau)}{\mu(a_t^\tau|s_t^\tau)}$ and $\phi_t^\tau = \phi(s_t^\tau)$. And define the filtration, $\mathcal{F}_t = \sigma\left(s_0, a_0, s_1, a_1, \ldots, s_{t(b+1)+b}, a_{t(b+1)+b}, s_{t(b+1)+b+1}\right)$. We define $r_\pi \in \mathbb{R}^N$, with $r_\pi(s) = \sum_{a\in\mathcal{A}} r(s,a)\pi(a|s)$. $P_\pi \in \mathbb{R}^{N\times N}$, with $P_\pi(s'|s) = \sum_{a\in\mathcal{A}}\pi(a|s)\mathsf{P}(s'|s,a)$.

For a matrix $M \in \mathbb{R}^{N\times N}$, we specify that $M_{(s,\cdot)}$ denotes its sth row and $M_{(\cdot,s)}$ is its sth column.

3 Proposed PER-ETD Algorithms

Drawbacks of ETD: In the original design of ETD, the follow-on trace coefficient F_t is updated throughout the execution of the algorithm. Hence, its variance increases exponentially with the number of iterations, and causes the algorithm to unstable and diverge, as having been observed in Hallak et al. (2016). Here, we propose stable and convergent ETD algorithms with guaranteed global convergence.

Algorithm 1 PER-ETD(0)

1: **Input:** Parameters T, b, and η_t.
2: Initialize: $\theta_0 = 0$.
3: **for** $t = 0, 1, ..., T$ **do**
4: $\quad F$ update: $F_t^{\tau+1} = \gamma\rho_t^\tau F_t^\tau + 1$, where $\tau = 0, 1, \ldots, b-1$ and $F_t^0 = 1$.
5: $\quad \theta$ update: $\theta_{t+1} = \Pi_\Theta\left(\theta_t + \eta_t F_t^b\rho_t^b(r_t^b + \gamma\theta_t^\top\phi_t^{b+1} - \theta_t^\top\phi_t^b)\phi_t^b\right)$
6: **end for**

In order to overcome the divergence issue of ETD, we propose the following ETD algorithm with the follow-on trace updated periodically with restart, which we call as PER-ETD. We first introduce the basic version PER-ETD(0), which is formally described in Algorithm 1. At iteration t, we restart the update for the follow-on trace F and iteratively update it for b times to obtain a good estimate F_t^b. We then define the emphatic update operator at t as,

$$\widehat{\mathcal{T}}_t(\theta) = F_t^b\rho_t^b\phi_t^b(\phi_t^b - \gamma\phi_t^{b+1})^\top\theta - F_t^b\rho_t^b\phi_t^b r_t^b, \tag{4}$$

and update the value function parameter θ_t as

$$\theta_{t+1} = \theta_t - \eta_t\widehat{\mathcal{T}}_t(\theta_t).$$

It can be shown that $\lim_{b\to\infty}\mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta)\middle|\mathcal{F}_{t-1}\right] = \mathcal{T}(\theta)$ defined as

$$\mathcal{T}(\theta) := \left(\Phi^\top F(I-\gamma P_\pi)\Phi\right)\theta - \Phi^\top F r_\pi,$$

which takes a unique fixed point θ^* as the original ETD. Clearly, a large value of the hyperparameter b attains a small bias error, but we need the value of b to be sufficiently small to control the variance of the follow-on trace. Our analysis in the next section will provide a guidance to balance the value of such a hyperparameter in order to attain the target convergence error.

Algorithm 2 PER-ETD(λ)

Input: Parameters T, b, and η_t.
Initialize: $\theta_0 = 0$.
for $t = 0, 1, \ldots, T$ **do**
  Set $F_t^0 = M_t^0 = 1$ and $e_t^0 = \phi_t^0$
  for $\tau = 1, \ldots, b$ **do**
    $F_t^\tau = \rho_t^{\tau-1}\gamma F_t^{\tau-1} + 1$
    $M_t^\tau = \lambda + (1-\lambda)F_t^\tau$
    $e_t^\tau = \gamma\lambda\rho_t^{\tau-1}e_t^{\tau-1} + M_t^\tau\phi_t^\tau$
  end for
  θ update: $\theta_{t+1} = \theta_t + \eta_t\rho_t^b\left(r_t^b + \gamma\theta_t^\top\phi_t^{b+1} - \theta_t^\top\phi_t^b\right)e_t^b$
end for

We then extend PER-ETD(0) to the PER-ETD(λ) algorithm (see Algorithm 2), which incorporates the eligible trace. Specifically, at each iteration t, we restart the update for the follow-on trace F_t and iteratively update it together with M_t and the eligible trace e_t for b times to obtain a good estimate e_t^b. We then define the emphatic update operator at t as:

$$\widehat{\mathcal{T}}_t^\lambda(\theta) = \rho_t^b e_t^b\left(\phi_t^b - \gamma\phi_t^{b+1}\right)^\top\theta - \rho_t^b r_t^b e_t^b, \tag{5}$$

and then update the value function parameter θ_t as

$$\theta_{t+1} = \theta_t - \eta_t\widehat{\mathcal{T}}_t^\lambda(\theta_t).$$

It can be shown that $\lim_{b\to\infty}\mathbb{E}\left[T_t^\lambda(\theta)\middle|\mathcal{F}_{t-1}\right] = \mathcal{T}^\lambda(\theta)$ defined as

$$\mathcal{T}^\lambda(\theta) = \Phi^\top M(I-\gamma\lambda P_\pi)^{-1}(I-\gamma P_\pi)\Phi\theta - \Phi^\top M(I-\gamma\lambda P_\pi)^{-1}r_\pi,$$

which takes a unique fixed point θ_λ^* as the original ETD.

4 Finite-Time Analysis of PER-ETD Algorithms

In this section, we show that both PER-ETD(0) and PER-ETD(λ)) converge to the globally optimal fixed point and characterize the finite-time convergence rate to attain a target convergence error.

4.1 Technical Assumptions

We take the following standard assumptions for analyzing the TD type algorithms in the literature (Jiang et al., 2021; Zhang & Whiteson, 2021; Yu, 2015).

Assumption 1 (Coverage of behavior policy)**.** *For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the behavior policy μ satisfies $\mu(a|s) > 0$ as long as $\pi(a|s) > 0$.*

Assumption 2. *The Markov chain induced by the behavior policy μ is irreducible and recurrent.*

The following lemma on the geometric ergodicity has been established.

Lemma 1 (Geometric ergodicity)**.** *(Levin & Peres, 2017, Thm. 4.9) Suppose Assumption 2 holds. Then the Markov chain induced by the behavior policy μ has a unique stationary distribution d_μ over the state space $\mathcal{S}$. Moreover, the Markov chain is uniformly geometric ergodic, i.e., there exists constants $C_M \geq 0$ and $0 < \chi < 1$ such that the state distribution $d_{\mu,t}(s) = \mathbb{P}(s_t = s|s_0)$ after t transitions satisfies*

$$\|d_{\mu,t} - d_\mu\|_1 \leq C_M\chi^t,$$

for every initial state $s_0 \in \mathcal{S}$.

4.2 Finite-time Analysis of PER-ETD(0)

In PER-ETD(0), the update of the value function parameter is fully determined by the empirical emphatic operator $\widehat{\mathcal{T}}_t(\theta)$ defined in eq. (4). Thus, we first characterize the bias and variance errors of $\widehat{\mathcal{T}}_t(\theta)$, which serve the central role for establishing the convergence rate for PER-ETD(0).

Proposition 1 (Bias bound)**.** *Suppose Assumptions 1 and 2 hold. Then we have*

$$\mathbb{E}\left\|\mathcal{T}(\theta_t) - \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t)\middle|\mathcal{F}_{t-1}\right]\right\|_2 \leq C_b\left(B_\phi\|\theta_t - \theta^*\|_2 + \epsilon_{approx}\right)\xi^b,$$

where $\epsilon_{approx} = \|\Phi\theta^* - V_\pi\|_\infty$ is the function approximation error of the fixed point, $C_b > 0$ is a constant, and $\xi = \max\{\gamma, \chi\} < 1$.

Proposition 1 characterizes the conditional expectation of the bias error of the empirical emphatic operator $\widehat{\mathcal{T}}_t(\theta)$. Since $\xi = \max\{\gamma, \chi\} < 1$, such a bias error decays exponentially fast as b increases.

Proposition 2 (Variance bound)**.** *Suppose Assumptions 1 and 2. Then we have*

$$\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2\middle|\mathcal{F}_{t-1}\right] \leq \sigma^2, \quad \textit{where} \quad \sigma^2 = \begin{cases} \mathcal{O}(1), & \textit{if} \quad \gamma^2\rho_{max} < 1, \\ \mathcal{O}(b), & \textit{if} \quad \gamma^2\rho_{max} = 1, \\ \mathcal{O}\left((\gamma^2\rho_{max})^b\right), & \textit{if} \quad \gamma^2\rho_{max} > 1, \end{cases} \tag{6}$$

where $\mathcal{O}(\cdot)$ is with respect to the scaling of b.

Proposition 2 captures the variance bound of the empirical emphatic operator. It can be seen that if the distribution mismatch is large (i.e., $\gamma^2\rho_{max} > 1$), the variance bound grows exponentially large as b increases, which is consistent with the finding in Hallak et al. (2016). However, as we show below, as long as b is controlled to grow only logarithmically with the number of iterations, such a variance error will grow only polynomially with the number of iterations, and at the same time, the bias error can still be controlled to delay polynomially, so that the overall convergence of PER-ETD can be guaranteed.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Consider PER-ETD(0) specified in Algorithm 1. Let the step-size $\eta_t = \frac{2}{\mu_0(t+t_0)}$, where $t_0 = \frac{8L_0^2}{\mu_0^2}$, and let $b \geq \left\lceil \frac{\log(\mu_0) - \log(5C_bB_\phi)}{\log(\xi)} \right\rceil$. Let the projected set $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B_\theta\right\}$ with $\theta^* \in \Theta$. Then output θ_T falls into the following two cases.*

(a) If $\gamma^2\rho_{max} \leq 1$, then $\mathbb{E}\left[\|\theta_T - \theta^*\|_2^2\right] \leq \tilde{\mathcal{O}}\left(\frac{1}{T}\right)$.

(b) If $\gamma^2\rho_{max} > 1$, then $\mathbb{E}\left[\|\theta_T - \theta^*\|_2^2\right] \leq \mathcal{O}\left(\frac{1}{T^a}\right)$, where $a = \frac{1}{\log_{1/\xi}(\gamma^2\rho_{max})+1} < 1$.

Thus, PER-ETD(0) attains an ϵ-accurate solution with $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$ samples if $\gamma^2\rho_{max} \leq 1$, and with $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^{1/a}}\right)$ samples if $\gamma^2\rho_{max} \geq 1$.

Theorem 1 also captures how the convergence rate depends on the mismatch between the behavior and target policies via the parameter $\rho_{max} \geq 1$ into two cases. (a) If $\gamma^2\rho_{max} \leq 1$, i.e., the mismatch between the behavior and target policies is less than a threshold, then Theorem 1 indicates that PER-ETD(0) converges at the rate of $\tilde{\mathcal{O}}\left(\frac{1}{T}\right)$, which is the same as that for the on-policy TD learning Bhandari et al. (2018). (b) If $\gamma^2\rho_{max} \geq 1$, i.e., the mismatch between the behavior and target policies is above the threshold, then Theorem 1 indicates that PER-ETD(0) converges at a rate of $\tilde{\mathcal{O}}\left(\frac{1}{T^a}\right)$, which becomes shower (i.e., a becomes smaller) as the mismatch becomes more serious (i.e., ρ_{max} gets larger).

Theorem 1 also indicates that although there is the update of the emphatic coefficient is truncated to be length b, PER-ETD(0) still converges to the desired fixed point θ^* determined by the infinite-length emphatic variable. This benefits from the value of the length b that we choose as we explain below.

We next explain how to obtain the convergence rate in Theorem 1 by controlling the value of b, which will also illustrate how the value of b can trade off between the bias and variance errors. The main proof of Theorem 1 shows that the output θ_T of PER-ETD(0) satisfies the following convergence rate:

$$\mathbb{E}\left[\|\theta_T - \theta^*\|_2^2\right] \leq \mathcal{O}\left(\frac{\|\theta_0 - \theta^*\|_2^2}{T^2}\right) + \underbrace{\mathcal{O}\left(\frac{\sigma^2}{T}\right)}_{\text{variance}} + \underbrace{\mathcal{O}\left(\frac{\xi^{2b}}{T}\right) + \mathcal{O}\left(\xi^b\right)}_{\text{bias}}. \tag{7}$$

If $\gamma^2\rho_{max} \leq 1$, then σ^2 in the variance term in eq. (7) takes the form $\sigma^2 \leq \mathcal{O}(b)$ as given in eq. (6), which increases at most linearly fast with b. Then we set $b = \left\lceil \frac{\log(\mu_0)-\log(5C_bB_\phi)}{\log(\xi)} \right\rceil + \frac{\log T}{\log \xi}$ so that both the variance and the bias terms in eq. (7) achieve the same order of $\mathcal{O}\left(\frac{1}{T}\right)$, which dominates the overall convergence.

If $\gamma^2\rho_{max} > 1$, then σ^2 in the variance term in eq. (7) takes the form $\sigma^2 = \mathcal{O}\left((\gamma^2\rho_{max})^b\right)$ as given in eq. (6), Now, we need to set b as $b = \left\lceil \frac{\log(\mu_0)-\log(5C_bB_\phi)}{\log(\xi)} \right\rceil + \frac{\log(T)}{\log(\gamma^2\rho_{max})+\log(1/\xi)}$, where the increase with $\log T$ has a smaller coefficient than the previous case, so that both the variance and the bias terms in eq. (7) achieve the same order of $\mathcal{O}\left(\frac{1}{T^a}\right)$. Thus, such a choice of b balances the exponentially increasing variance and exponentially decay bias to attain a common convergence rate.

4.3 Finite-time Analysis of PER-ETD(λ)

In PER-ETD(λ), the update of the value function parameter is fully determined by the empirical emphatic operator $\widehat{\mathcal{T}}_t^\lambda(\theta)$ defined in eq. (5). Thus, we first obtain the bias and variance errors of $\widehat{\mathcal{T}}_t^\lambda(\theta)$, which facilitate the analysis of the convergence rate for PER-ETD(λ).

Proposition 3. *Suppose Assumptions 1 and 2 hold. Then we have*

$$\left\|\mathbb{E}\left[\widehat{\mathcal{T}}_t^\lambda(\theta_t)\middle|\mathcal{F}_{t-1}\right] - \mathcal{T}^\lambda(\theta_t)\right\|_2 \leq C_{b,\lambda}\left(B_\phi\|\theta_t - \theta_\lambda^*\|_2 + \epsilon_{approx}\right)\xi^b,$$

where $\epsilon_{approx} = \|\Phi\theta_\lambda^* - V_\pi\|_\infty$ is the function approximation error, $C_{b,\lambda}$ is a constant given a fixed λ, and $\xi = \max\{\chi, \gamma\} < 1$.

The above proposition shows that the bias error of the empirical emphatic operator $\widehat{\mathcal{T}}_t^\lambda(\theta)$ in PER-ETD(λ) decays exponentially fast as b increases, because $\xi = \max\{\gamma, \chi\} < 1$.

Proposition 4. *Suppose Assumptions 1 and 2 hold. Then we have*

$$\mathbb{E}\Big[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2\Big|\mathcal{F}_{t-1}\Big] \leq \sigma_\lambda^2, \quad \textit{where} \quad \sigma_\lambda^2 = \begin{cases}\mathcal{O}(b), & \textit{if} \quad \rho_{max} = 1,\\ \mathcal{O}\left(\rho_{max}^b\right), & \textit{if} \quad \rho_{max} > 1,\end{cases} \tag{8}$$

where $\mathcal{O}(\cdot)$ is with respect to the scaling of b.

Compared to Proposition 2 for ETD(0), Proposition 4 indicates that ETD(λ) has a larger variance, which increases exponentially with b right after the mismatch between the behavior and target policies occurs (i.e., $\rho_{max} > 1$). This is due to the fact that the eligible trace e_t^b carries the historical information and is less stable than ϕ_t^b.

Theorem 2. *Suppose Assumptions 1 and 2 hold. Consider PER-ETD(λ) specified in Algorithm 2. Let the step-size $\eta_t = \frac{2}{\mu_\lambda(t+t_\lambda)}$, where $t_\lambda = \frac{8L_\lambda^2}{\mu_\lambda^2}$, and let $b \geq \left\lceil \frac{\log(\mu_\lambda)-\log(5C_{b,\lambda}B_\phi)}{\log(\xi)} \right\rceil$. Let the projected set $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B_\theta\right\}$ with $\theta_\lambda^* \in \Theta$. Then output θ_T falls into the following two cases.*

(a) If $\rho_{max} = 1$, then $\mathbb{E}\left[\|\theta_T - \theta_\lambda^*\|_2^2\right] \leq \tilde{\mathcal{O}}\left(\frac{1}{T}\right)$.

(b) If $\rho_{max} > 1$, then $\mathbb{E}\left[\|\theta_T - \theta_\lambda^*\|_2^2\right] \leq \mathcal{O}\left(\frac{1}{T^{a_\lambda}}\right)$, where $a_\lambda = \frac{1}{\log_{1/\xi}(\rho_{max})+1} < 1$.

Thus, PER-ETD(λ) attains an ϵ-accurate solution with $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$ samples if $\rho_{max} = 1$, and with $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^{1/a_\lambda}}\right)$ samples if $\rho_{max} \geq 1$.

Theorem 2 indicates that PER-ETD(λ) also converges to the desired fixed point θ_λ^* determined by the infinite-length emphatic variable. Further, PER-ETD(λ) converges at the same rate $\tilde{\mathcal{O}}\left(\frac{1}{T}\right)$ as PER-ETD(0) and on-policy TD if the behavior and target policies are the same (i.e., $\rho_{max} = 1$). Once the mismatch occurs (i.e., $\rho_{max} > 1$), PER-ETD(λ) converges at the rate of $\tilde{\mathcal{O}}\left(\frac{1}{T^{a_\lambda}}\right)$ which is slower than PER-ETD(0) because $a_\lambda < a$.

We next further explain how the value of b trades off between the bias and variance errors. The main proof of Theorem 2 shows that the output θ_T of PER-ETD(λ) satisfies the following convergence rate:

$$\mathbb{E}\left[\|\theta_T - \theta_\lambda^*\|_2^2\right] \leq \mathcal{O}\left(\frac{\|\theta_0 - \theta_\lambda^*\|_2^2}{T^2}\right) + \underbrace{\mathcal{O}\left(\frac{\sigma_\lambda^2}{T}\right)}_{\text{variance}} + \underbrace{\mathcal{O}\left(\frac{\xi^{2b}}{T}\right) + \mathcal{O}\left(\xi^b\right)}_{\text{bias}}. \tag{9}$$

If $\rho_{max} = 1$, then σ_λ^2 in the variance term in eq. (9) takes the form $\sigma_\lambda^2 \leq \mathcal{O}(b)$ as given in eq. (8), which increases at most linearly fast with b. Then we set $b = \left\lceil \frac{\log(\mu_0) - \log(5C_b B_\phi)}{\log(\xi)} \right\rceil + \frac{\log T}{\log \xi}$ so that both the variance and the bias terms in eq. (9) achieve the same order of $\mathcal{O}\left(\frac{1}{T}\right)$, which dominates the overall convergence.

If $\rho_{max} > 1$, then σ_λ^2 in the variance term in eq. (9) takes the form $\sigma^2 = \mathcal{O}\left(\rho_{max}^b\right)$ as given in eq. (8), Now, we need to set b as $b = \left\lceil \frac{\log(\mu_0) - \log(5C_b B_\phi)}{\log(\xi)} \right\rceil + \frac{\log(T)}{\log(\rho_{max}) + \log(1/\xi)}$, where the increase with $\log T$ has a smaller coefficient than the previous case, so that both the variance and the bias terms in eq. (9) achieve the same order of $\mathcal{O}\left(\frac{1}{T^{a_\lambda}}\right)$. Thus, such a choice of b balances the exponentially increasing variance and exponentially decay bias to attain a common convergence rate.

[add a remark about finite b, how λ affects the performance]

Ziwei: With the aid of eligible trace, both TD learning and ETD learning achieve tighter error bounds (Sutton & Barto, 2018; Hallak et al., 2016). Intuitively, this phenomenon would also appear in our PER-ETD. However, the performance for increasing λ may be reversed for a small fixed periodic length b. On the one hand, increasing λ, we have a better fixed point θ_λ. On the other hand, for both PER-ETD(0) and PER-ETD(λ), there exist biases exponentially decaying with the periodic length b. Compared with the performance gain of λ, the biases are significant when b is small. Such biases may finally reverse the performance of increasing λ. Because the biases are decreasing with b very fast, for large enough b, we can always expect better performance with larger λ. We refer the readers to the experiment for more details.

5 Experiment

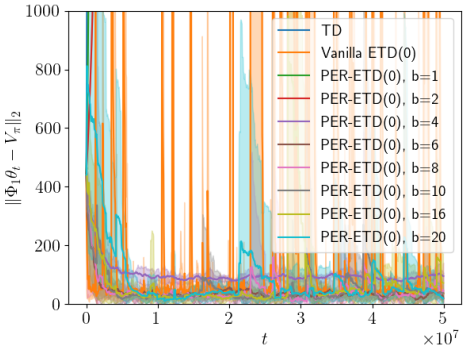
We conduct the computer simulations to verify the following two points. (1). The periodic parameter b could balance the bias and variance. (2). For the PER-ETD(λ), increasing λ may provide better convergence.

We take our experiment on the BAIRD counter-example. There exist 7 states and 2 actions. For all states, when taking the first action, the environment would transition current state to states 1 to 6 following uniform distribution and return 0 reward, and when taking the second action, the environment would transition current state to state 7 with probability 1 and return a reward 1. We pick the target policy as $\pi(0|s) = 0.1$ and $\pi(1|s) = 0.9$ for all $s = 1, 2, \ldots, 7$; And the behavior policy $\mu(0|s) = 6/7$ and $\mu(1|s) = 1/7$ for all $s = 1, 2, \ldots, 7$. We adopt constant learning rate for both PER-ETD(0) and PER-ETD(λ) and every plot is an average of 20 random initializations.

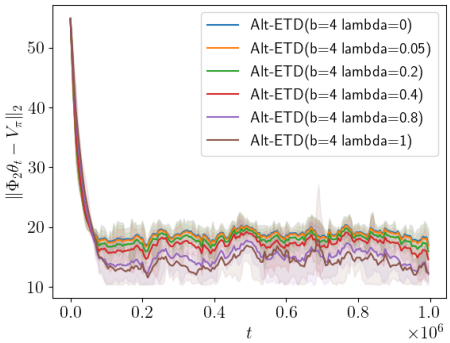
For PER-ETD(0), one-dimensional features $\Phi_1 = (0.35, 0.35, 0.35, 0.35, 0.35, 0.35, 0.37)^\top$ are adopted. This is a highly reduced feature space and the ground truth value function $V_\pi = (10, 10, 10, 10, 10, 10, 10)^\top$ does not lie inside the linear function class. In fig. 1a, we plot the distances between the ground truth and the learned approximate value functions of TD, vanilla ETD(0) and PER-ETD(0) with $b = 1, 2, 4, 6, 8, 10, 16$ and 20. In the figure, the TD diverges at a fast speed and out of the range at the very beginning; Vanilla ETD diverges due to very large variance. For our PER-ETD(0), the algorithm diverges when $b = 1$ and 2 and it converges for $b \geq 4$. Increasing the periodic length b, we observe that the algorithm converges with smaller distance to the ground truth value and larger variance. When b= 4, PER-ETD(0) achieves both small approximation error and variance.

For PER-ETD(λ), we randomly generate two dimensional features for each states,

$$\Phi_2 = ((0.3425, 0.0171)^\top, (0.1902, 0.4248)^\top, (0.1354, 0.76)^\top, (0.1357, 0.7973)^\top,$$

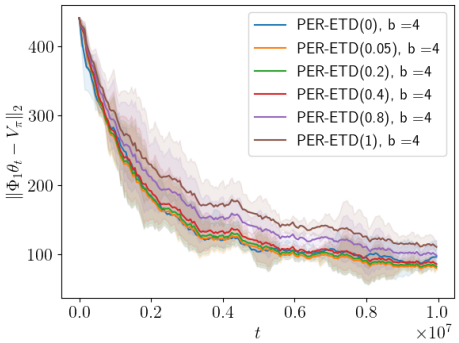


(a) Distance from the ground truth of TD, Vanilla ETD(0) and PER-ETD(0) with different b's

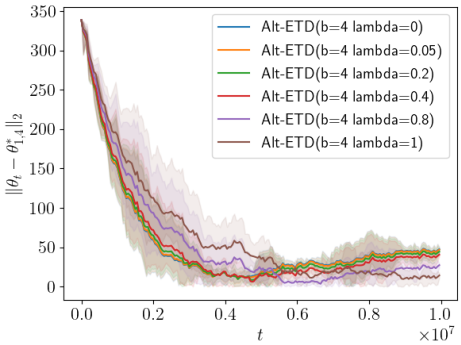


(b) Distance from the ground truth of PER-ETD(λ), fixed $b = 4$, with $\lambda = 0, 0.05, 0.2, 0.4, 0.8$ and 1

Figure 1: Comparison of PER-ETD and other algorithms



(a) Distance to the ground truth



(b) Distance to the fixed point of PER-ETD(1), with fixed $b = 4$

Figure 2: Comparison of PER-ETD(λ) with different λ

$$(0.8674, 0.8774)^\top, (0.5166, 0.9493)^\top, (0.3094, 0.8535)^\top)^\top.$$

We fixed the periodic length $b = 4$ because it converges and has a large bias to observe the effects of λ. As shown in fig. 1b, increasing λ brings a shorter distance from the ground truth and slightly alleviated variance. As we have discussed at the end of Section 4.3, the pattern may appear in a reversed way. More details can be found at .....

When used the one dimensional feature space, $\Phi_1 = (0.35, 0.35, 0.35, 0.35, 0.35, 0.35, 0.37)^\top$, the behavior of increasing λ shows an inverse phenomenon in fig. 2a, where increasing the λ will incur deteriorated performance of the algorithm. As we have explained at the end of Section 4.3, this phenomenon arises from the bias of the PER-ETD(λ). A more accurate understanding could be given though the fixed points of PER-ETD(λ) (for fixed periodic length b, define the fixed point as $\theta^*_{\lambda,b}$). Using the iterative equation we introduced in the next subsection, we could compute the $\theta^*_{\lambda,b}$. Based on the computation, $\theta^*_{0,b}$ may be closer to the orthogonal projection of ground truth than $\theta^*_{1,b}$. $\theta^*_{\lambda,b}$ will continuously transition from $\theta^*_{0,b}$ to $\theta^*_{1,b}$ when λ increases from 0 to 1. We plot the distance between $\theta_{1,4}$ and the learned parameters of different λ's in fig. 2b. The results are consistent to our expectation.

5.1 COMPUTATION OF THE FIXED POINT OF PER-ETD(λ)

From the theory of stochastic approximation, the PER-ETD(λ) algorithm will converge to the the point $A^{-1}c$ with adequate stepsize, where

$$A := \lim_{t\to\infty} \mathbb{E}\left[A_t \left(:= \left(\rho_t^b e_t^b (\phi_t^b - \gamma\phi_t^{b+1})\right)^\top\right)\right],$$

and

$$c := \lim_{t\to\infty} \mathbb{E}\left[c_t \left(:= \rho_t^b r_t^b e_t^b\right)\right],$$

where $\beta_b(s) = \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t \middle| s_t^b = s\right]$.

As shown in eqs. (40) and (44), we have

$$\begin{aligned}\mathbb{E}\left[A_t|\mathcal{F}_{t-1}\right] &= \sum_{s\in\mathcal{S}} \beta_b(s) \cdot \left((\Phi)_{(s,\cdot)} - \gamma(P_\pi\Phi)_{(s,\cdot)}\right) \\ &= \beta_b(I - \gamma P_\pi)\Phi.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[c_t|\mathcal{F}_{t-1}\right] &= \sum_{s\in\mathcal{S}} r_\pi(s)\beta_b(s) \\ &= \beta_b r_\pi.\end{aligned}$$

Further deductions in the proof of Proposition 3 yield eq. (42)

$$\beta_b = \lambda\Phi^\top D_{\mu,b} + (1-\lambda)\Phi^\top F_b + \gamma\lambda\beta_{b-1}P_\pi,$$

where $\beta_0 = \phi(s_t^0)$, $F_b = \mathrm{diag}(f_b)$ and f_b is determined iteratively by eq. (21) as follows

$$f_b = d_{\mu,b} + \gamma P_\pi^\top f_{b-1},$$

with $f_0 = d_{\mu,0}$.

Taking expectation with respect to $\mathcal{F}_{t-1}$ and letting $t \to \infty$, the above two equalities induce

$$\bar{\beta}_b = \lambda\Phi^\top D_\mu + (1-\lambda)\Phi^\top \bar{F}_b + \gamma\lambda\bar{\beta}_{b-1}P_\pi, \tag{10}$$

where $\bar{\beta}_b := \lim_{t\to\infty} \mathbb{E}\left[\beta_b\right]$, $\bar{F}_b = \mathrm{diag}(\bar{f}_b)$ and $\bar{\beta}_0 = \Phi^\top D_\mu$. And

$$\bar{f}_b = d_\mu + \gamma P_\pi^\top \bar{f}_{b-1}, \tag{11}$$

where $\bar{f}_b := \lim_{t\to\infty} \mathbb{E}[f_b]$, $\bar{f}_0 = d_\mu$.

$\bar{\beta}_b$ can be derived through applying eqs. (10) and (11) iteratively. And the fixed point of PER-ETD(λ) is obtained immediately through $\bar{\beta}_b$.

REFERENCES

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning*, pp. 30–37. Elsevier, 1995.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Annual Conference on Learning Theory (COLT)*, pp. 1691–1692. PMLR, 2018.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018a.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Proc. Annual Conference on Learning Theory (COLT)*, pp. 1199–1233. PMLR, 2018b.

Peter Dayan. The convergence of TD (λ) for general λ. *Machine learning*, 8(3-4):341–362, 1992.

Peter Dayan and Terrence J Sejnowski. TD (λ) converges with probability 1. *Machine Learning*, 14 (3):295–301, 1994.

Sina Ghiassian, Banafsheh Rafiee, and Richard S Sutton. A first empirical study of emphatic temporal difference learning. *arXiv preprint arXiv:1705.04185*, 2017.

Harsh Gupta, R Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 32:4704–4713, 2019.

Assaf Hallak, Aviv Tamar, Rémi Munos, and Shie Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. In *aaai*, 2016.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

Ray Jiang, Shangtong Zhang, Veronica Chelu, Adam White, and Hado van Hasselt. Learning expected emphatic traces for deep RL. *arXiv preprint arXiv:2107.05405*, 2021.

J Kolter. The fixed points of off-policy TD. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 24:2169–2177, 2011.

Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. Simple and optimal methods for stochastic variational inequalities, ii: Markovian noise and policy evaluation in reinforcement learning. *arXiv preprint arXiv:2011.08434*, 2020.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1347–1355. PMLR, 2018.

Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

A Rupam Mahmood, Huizhen Yu, Martha White, and Richard S Sutton. Emphatic temporal-difference learning. *arXiv preprint arXiv:1507.01569*, 2015.

Jingjiao Ni. Toward emphatic reinforcement learning. Master's thesis, University of Alberta, 2021.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 21(21):1609–1616, 2008.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research (JMLR)*, 17(1): 2603–2631, 2016.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

Yue Wang, Wei Chen, Yuting Liu, Zhi-Ming Ma, and Tie-Yan Liu. Finite sample analysis of the gtd policy evaluation algorithms in markov setting. *arXiv preprint arXiv:1809.08926*, 2018.

Huizhen Yu. Convergence of least squares temporal difference methods under general conditions. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1207–1214, 2010.

Huizhen Yu. On convergence of emphatic temporal-difference learning. In *Proc. Annual Conference on Learning Theory (COLT)*, pp. 1724–1751. PMLR, 2015.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *Proc. International Conference on Learning Representations (ICLR)*, 2020a.

Shangtong Zhang and Shimon Whiteson. Truncated emphatic temporal difference methods for prediction and control. *arXiv preprint arXiv:2108.05338*, 2021.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 11204–11213. PMLR, 2020b.

APPENDIX

A SUPPORTING LEMMAS

Lemma 2. *Consider a transition matrix $P \in \mathbb{R}^{N\times N}$, where $\sum_j P_{(i,j)} = 1$ for all i and $0 \leq P_{(i,j)} \leq 1$ for all j. We have for any $n \in \mathbb{N}$*

$$\|P^n\|_\infty = 1,$$

and,

$$\|(P^\top)^n\|_1 = 1.$$

Proof of lemma 2. This Lemma is well-known and we give the proof for completeness.

Using the fact that the ℓ_∞ norm of a matrix equals to the maximum of its row sums of a matrix, we have

$$\begin{aligned}\|P^n\|_\infty &= \max_{1\leq i\leq N} \sum_{1\leq j\leq N} (P^n)_{(i,j)} \\ &\overset{(i)}{=} \max_{1\leq i\leq N} 1 \\ &= 1,\end{aligned}$$

where (i) follows from the fact that P^n is still a transition kernel and its sum of row always equals to 1.

Using the fact that the ℓ_1 norm of a matrix equals to the maximum of its column sums, we can show

$$\|(P^\top)^n\|_1 = 1.$$

□

Lemma 3. *Consider $0 < p, q < 1$, with $p \neq q$. We have*

$$\sum_{k=0}^{n-1} p^k q^{n-k} \leq \frac{1}{|p-q|}\xi^n,$$

where $\xi = \max\{p, q\}$.

Proof of Lemma 3. If $p > q$, we have

$$\begin{aligned}\sum_{k=0}^{n-1} p^k q^{n-k} &\leq \sum_{k=0}^{n-1} p^{k+1} q^{n-1-k} \\ &\leq \sum_{m=0}^{n-1} p^{n-m} q^m.\end{aligned}$$

Without loss of generality, we assume $p < q$ and $\xi = q$. We have

$$\begin{aligned}\sum_{k=0}^{n-1} p^k q^{n-k} &= q^n \cdot \sum_{k=0}^{n-1} (\tfrac{p}{q})^k \\ &= q^n \cdot \frac{1-(p/q)^n}{1-p/q} \\ &\leq \frac{1}{q-p} \cdot q^n \\ &= \frac{1}{|p-q|} \cdot \xi^n.\end{aligned}$$

□

Lemma 4. *Consider a matrix $P \in \mathbb{R}^{N\times N}$ where $\sum_j P_{(i,j)} \le C$ for all i and $0 \le P_{(i,j)} \le 1$ for all j, and a positive vector x where $x \in \mathbb{R}^N$ and $x_i \ge 0$ for all i. We have*

$$\mathbf{1}^\top P x \le C \mathbf{1}^\top x.$$

Proof of Lemma 4. We have

$$\begin{aligned}
\mathbf{1}^\top P x &= \sum_{1\le i,j\le N} P_{i,j} x_i \\
&= \sum_{1\le i\le N} x_i \left(\sum_{1\le j\le N} P_{i,j} \right) \\
&\le C \sum_{1\le i\le N} x_i \\
&= C \mathbf{1}^\top x.
\end{aligned}$$

□

Lemma 5. *Consider a diagonal matrix $D \in \mathbb{R}^{N\times N}$, where $D = \mathrm{diag}(d(1), d(2), \ldots, d(N))$. a transition matrix $P \in \mathbb{R}^{N\times N}$ where $\sum_j P_{i,j} = 1$ for all $i = 1, 2, \ldots, N$ and $P_{i,j} \ge 0$ for all j, and an arbitrary vector $x \in \mathbb{R}^N$. We have*

$$\|\Phi^\top D P x\|_2 \le B_\phi \|d\|_1 \|x\|_\infty.$$

Proof of Lemma 5. We have

$$\Phi^\top D = \left(d(1)\phi(1), d(2)\phi(2), \ldots, d(N)\phi(N) \right),$$

and

$$\left(\Phi^\top D P\right)_{(\cdot,s)} = \sum_{\tilde{s}} P_{\tilde{s},s} d(\tilde{s}) \phi(\tilde{s}),$$

which implies

$$\begin{aligned}
\Phi^\top D P x &= \sum_s x(s) \left(\sum_{\tilde{s}} P_{\tilde{s},s} d(\tilde{s}) \phi(\tilde{s}) \right) \\
&= \sum_{\tilde{s}} \left(\sum_s x(s) P_{\tilde{s},s} \right) d(\tilde{s}) \phi(\tilde{s}).
\end{aligned}$$

Taking ℓ_2 norm on both sides of the above equality yields

$$\begin{aligned}
\left\|\Phi^\top D P x\right\|_2 &= \left\| \sum_{\tilde{s}} \left(\sum_s x(s) P_{\tilde{s},s} \right) d(\tilde{s}) \phi(\tilde{s}) \right\|_2 \\
&\le \sum_{\tilde{s}} \left| \sum_s x(s) P_{\tilde{s},s} \right| \cdot |d(\tilde{s})| \cdot \|\phi(\tilde{s})\|_2 \\
&= B_\phi \|d\|_1 \cdot \max_{\tilde{s}} \left| \sum_s x(s) P_{\tilde{s},s} \right| \\
&\le B_\phi \|d\|_1 \|x\|_\infty \max_{\tilde{s}} \left| \sum_s P_{\tilde{s},s} \right| \\
&= B_\phi \|d\|_1 \|x\|_\infty.
\end{aligned}$$

□

Lemma 6. *Consider a diagonal matrix $D \in \mathbb{R}^{N\times N}$ where $D = \mathrm{diag}(d(1), d(2), \ldots, d(N))$, a transition matrix $P \in \mathbb{R}^{N\times N}$ where $\sum_j P_{i,j} = 1$ for all $i = 1, 2, \ldots, N$ and $P_{i,j} \geq 0$ for all j, a matrix $Q \in \mathbb{R}^{N\times N}$ that satisfies $0 \leq Q_{i,j} \leq CP_{i,j}$, where $C > 1$ is a constant, and an arbitrary vector $x \in \mathbb{R}^N$. We have*

$$\mathrm{trace}\left(Q^m P^n \Phi\Phi^\top D\right) \leq C^m B_\phi^2 \|d\|_1,$$

and,

$$\mathrm{trace}\left(P^n Q^m \Phi\Phi^\top D\right) \leq C^m B_\phi^2 \|d\|_1,$$

for any m and $n \in \mathbb{N}_{\geq 0}$.

Proof of Lemma 6. For any given $\tau \geq m \geq 0$ and $n \geq 0$,

$$\begin{aligned}(Q^m P^n \Phi\Phi^\top)_{(i,j)} &= \left\langle (Q^m)_{(i,\cdot)}, (P^n \Phi\Phi^\top)_{(\cdot,j)} \right\rangle \\ &\overset{(i)}{\leq} \|(Q^m)_{(i,\cdot)}\|_1 \|(P^n \Phi\Phi^\top)_{(\cdot,j)}\|_\infty, \end{aligned} \tag{12}$$

where (i) follows from the Hölder's inequality.

Furthermore, for the term of $(P^n \Phi\Phi^\top)_{(\cdot,j)}$, we have,

$$\|(P^n \Phi\Phi^\top)_{(\cdot,j)}\|_\infty \leq \|P^n\|_\infty \|(\Phi\Phi^\top)_{(\cdot,j)}\|_\infty \overset{(i)}{=} \|(\Phi\Phi^\top)_{(\cdot,j)}\|_\infty,$$

where (i) follows from Lemma 2.

Moreover, for the ith entry of $(\Phi\Phi^\top)_{(\cdot,j)}$, we have

$$(\Phi\Phi^\top)_{(i,j)} = \phi(i)^\top \phi(j) \leq \|\phi(i)\|_2 \|\phi(j)\|_2 \leq B_\phi^2$$

The above uniform bounds over all i imply that $\|(\Phi\Phi^\top)_{(\cdot,j)}\|_\infty \leq B_\phi^2$. Hence,

$$\|(P^n \Phi\Phi^\top)_{(\cdot,j)}\|_\infty \leq B_\phi^2.$$

Substituting the above inequality back into eq. (12), we obtain

$$\begin{aligned}(Q^m P^n \Phi\Phi^\top)_{(i,j)} &\leq B_\phi^2 \|(Q^m)_{(i,\cdot)}\|_1 \\ &\overset{(i)}{\leq} C^m \|(P^{\tau-m})_{(i,\cdot)}\|_1 B_\phi^2 \\ &\overset{(ii)}{\leq} C^m B_\phi^2 \sum_j P^{\tau-m}(j|i) \\ &= C^m B_\phi^2, \end{aligned} \tag{13}$$

where (i) follows by the condition of Q, $Q_{(i,j)} \leq CP_{(i,j)}$ for all i, j.

Finally, we have

$$\begin{aligned}\mathrm{trace}\left(Q^m P^n \Phi\Phi^\top D\right) &= \sum_i d(i)(Q^m P^n \Phi\Phi^\top)_{(i,i)} \\ &\leq \sum_i |d(i)| \left|(Q^m P^n \Phi\Phi^\top)_{(i,i)}\right| \\ &\leq \|d\|_1 \max_i (Q^m P^n \Phi\Phi^\top)_{(i,i)} \\ &\overset{(i)}{\leq} C^m B_\phi^2 \|d\|_1, \end{aligned} \tag{14}$$

where (i) follows from eq. (13).

Following steps similar to those in eqs. (12) to (14), we can obtain

$$\mathrm{trace}\left(P^n Q^m \Phi\Phi^\top D\right) \leq C^m B_\phi^2 \|d\|_1.$$

□

Lemma 7. *The operators $\mathcal{T}(\theta)$ and $\mathcal{T}^{\lambda}(\theta)$ satisfy the generalized monotone variational inequality. There exist $\mu_0, \mu_\lambda > 0$, such that,*

$$\langle \mathcal{T}(\theta), \theta - \theta^* \rangle \geq \mu_0 \|\theta - \theta^*\|_2^2,$$

and

$$\left\langle \mathcal{T}^{\lambda}(\theta), \theta - \theta^* \right\rangle \geq \mu_\lambda \|\theta - \theta^*\|_2^2.$$

Proof of Lemma 7. We have

$$\begin{aligned}
\langle \mathcal{T}(\theta), \theta - \theta^* \rangle &\overset{(i)}{=} \left\langle \left(\Phi^\top F(I - \gamma P_\pi)\Phi\right)(\theta - \theta^*), \theta - \theta^* \right\rangle \\
&= (\theta - \theta^*)^\top \left(\Phi^\top F(I - \gamma P_\pi)\Phi\right)(\theta - \theta^*) \\
&\geq \lambda_{min}\left(\Phi^\top F(I - \gamma P_\pi)\Phi\right)\|\theta - \theta^*\|_2^2,
\end{aligned} \tag{15}$$

where (i) follows from the definition of the $\mathcal{T}$ and θ^*.

Recall that $F(I - \gamma P_\pi)$ is positive definite(Sutton et al., 2016; Mahmood et al., 2015) and Φ has linearly independent columns. For any $x \in \mathbb{R}^d$ with $x \neq 0$, we have $\Phi x \neq 0$ and

$$x^\top \Phi^\top F(I - \gamma P_\pi)\Phi x = (\Phi x)^\top F(I - \gamma P_\pi)(\Phi x) > 0.$$

The above inequality shows that $\Phi^\top F(I - \gamma P_\pi)\Phi$ is positive definite and thus, there exists $\mu_0 > 0$ such that $\mu_0 = \lambda_{min}\left(\Phi^\top F(I - \gamma P_\pi)\Phi\right)$.

Following steps similar to those in eq. (15) and applying the positive definiteness of $M(I - \gamma\lambda P_\pi)^{-1}(I - \gamma P_\pi)$ (Sutton et al., 2016; Mahmood et al., 2015) yield

$$\left\langle \mathcal{T}^{\lambda}(\theta), \theta - \theta^* \right\rangle \geq \mu_\lambda \|\theta - \theta^*\|_2^2.$$

□

Lemma 8. *The operators $\mathcal{T}(\theta)$ and $\mathcal{T}^{\lambda}(\theta)$ satisfy the Lipschitz condition. There exist $L_0, L_\lambda > 0$, such that,*

$$\|\mathcal{T}(\theta_1) - \mathcal{T}(\theta_2)\|_2 \leq L_0 \|\theta_1 - \theta_2\|_2,$$

and

$$\left\|\mathcal{T}^{\lambda}(\theta_1) - \mathcal{T}^{\lambda}(\theta_2)\right\|_2 \leq L_\lambda \|\theta_1 - \theta_2\|_2.$$

Proof of lemma 8. We have

$$\begin{aligned}
\|\mathcal{T}(\theta_1) - \mathcal{T}(\theta_2)\|_2 &= \left\|\left(\Phi^\top F(I - \gamma P_\pi)\Phi\right)(\theta_1 - \theta_2)\right\|_2 \\
&\leq \left\|\Phi^\top F(I - \gamma P_\pi)\Phi\right\|_2 \|\theta_1 - \theta_2\|_2.
\end{aligned} \tag{16}$$

Let $L_0 \coloneqq \left\|\Phi^\top F(I - \gamma P_\pi)\Phi\right\|_2 \|$, eq. (16) completes the proof of the first inequality in the Lemma. Let $L_\lambda \coloneqq \|\Phi^\top M(I - \gamma\lambda P_\pi)^{-1}(I - \gamma P_\pi)\Phi\|_2$, the steps similar to those in eq. (16) finalizes the proof of the second inequality in the Lemma. □

Lemma 9 (Three point lemma)**.** *Suppose Θ is a closed and bounded subset of $\mathbb{R}^d$, and θ^* is the solution of the following maximization problem,*

$$\max_{\theta \in \Theta} \eta \langle G, \theta \rangle + \frac{1}{2}\|\theta - \theta_0\|_2^2,$$

where $G \in \mathbb{R}^d$ is a vector. Then, we have, for any $\theta \in \Theta$,

$$\eta \langle G, \theta^* - \theta \rangle + \frac{1}{2}\|\theta_0 - \theta^*\|_2^2 \leq \frac{1}{2}\|\theta_0 - \theta\|_2^2 - \frac{1}{2}\|\theta^* - \theta\|_2^2.$$

Proof. The proof can be found in Lan (2020). □

B PROOFS OF PROPOSITIONS

B.1 PROOF OF PROPOSITION 1

First, by the definition of $\widehat{\mathcal{T}}_t(\theta_t)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t)\middle|\mathcal{F}_{t-1}\right] \\
&\overset{(i)}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \\
&\quad \cdot \mathbb{P}\left(s_t^b = s, a_t^b = a, s_t^{b+1} = s'\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t)\middle|\mathcal{F}_{t-1}, s_t^b = s, a_t^b = a, s_t^{b+1} = s'\right] \\
&\overset{(ii)}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mu(a|s)\mathsf{P}(s'|s,a) \\
&\quad \cdot \mathbb{E}\left[\rho_t^b F_t^b \phi(s)[\phi^\top(s)\theta_t - r(s,a) - \gamma\phi^\top(s')\theta_t]\middle|\mathcal{F}_{t-1}, s_t^b = s, a_t^b = a, s_t^{b+1} = s'\right] \\
&\overset{(iii)}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\pi(a|s)\mathsf{P}(s'|s,a)\phi(s)[\phi^\top(s)\theta_t - r(s,a) - \gamma\phi^\top(s')\theta_t] \\
&\quad \cdot \mathbb{E}\left[F_t^b\middle|\mathcal{F}_{t-1}, s_t^b = s\right] \\
&= \sum_{s\in\mathcal{S}} \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^b\middle|\mathcal{F}_{t-1}, s_t^b = s\right]\phi(s)\left[\phi^\top(s)\theta_t - \gamma[P_\pi\Phi]_{(s,\cdot)}\theta_t - r_\pi(s)\right] \qquad (17)
\end{aligned}
$$

where (i) follows from the law of total probability, (ii) follows from rewriting $\mathbb{P}\left(s_t^b = s, a_t^b = a, s_t^{b+1} = s'\middle|\mathcal{F}_{t-1}\right)$, and (iii) follows from the facts that F_t^b only depends on $(s_0, a_0, s_1, a_1, \ldots, s_{tb+b-1}, a_{tb+b-1})$ and the chain is Markov.

Recall the definition of $\mathcal{T}(\theta_t)$, we have

$$
\begin{aligned}
\mathcal{T}(\theta_t) &= \Phi^\top F\left[(I-\gamma P_\pi)\Phi\theta_t - r_\pi\right] \\
&= \sum_{s\in\mathcal{S}} f(s)\left(\phi^\top(s)\theta_t - \gamma[P_\pi\Phi]_{(s,\cdot)}\theta_t - r_\pi(s)\right)\phi(s). \qquad (18)
\end{aligned}
$$

Equations (17) and (18) together imply the following,

$$
\begin{aligned}
\mathcal{T}(\theta_t) - \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t)\middle|\mathcal{F}_{t-1}\right] &= \sum_{s\in\mathcal{S}}\left(f(s) - \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^b\middle|\mathcal{F}_{t-1}, s_t^b = s\right]\right) \\
&\qquad \cdot \left(\phi^\top(s)\theta_t - \gamma\left[P_\pi\Phi\right]_{(s,\cdot)}\theta_t - r_\pi(s)\right)\phi(s) \\
&\overset{(i)}{=} \sum_{s\in\mathcal{S}}\left(f(s) - f_b(s)\right)\left(\phi^\top(s)\theta_t - \gamma\left[P_\pi\Phi\right]_{(s,\cdot)}\theta_t - r_\pi(s)\right)\phi(s) \\
&= \sum_{s\in\mathcal{S}}\left(f(s) - f_b(s)\right)\left(\left(I-\gamma P_\pi\right)\Phi\theta_t - r_\pi\right)_s\phi(s),
\end{aligned}
$$

where in (i) We define $f_b(s) \coloneqq \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^b\middle|\mathcal{F}_{t-1}, s_t^b = s\right]$. Taking ℓ_2 norm on both sides of the above equality yields

$$
\begin{aligned}
\left\|\mathcal{T}(\theta_t) - \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t)\middle|\mathcal{F}_{t-1}\right]\right\|_2 &\le \left\|\sum_{s\in\mathcal{S}}\left(f(s) - f_b(s)\right)\left(\left(I-\gamma P_\pi\right)\Phi\theta_t - r_\pi\right)_s\phi(s)\right\|_2 \\
&\le \sum_{s\in\mathcal{S}}|f(s) - f_b(s)|\cdot\left|\left(\left(I-\gamma P_\pi\right)\Phi\theta_t - r_\pi\right)_s\right|\cdot\|\phi(s)\|_2 \\
&\le \max_{s\in\mathcal{S}}\left\{\|\phi(s)\|_2\right\}\max_{s\in\mathcal{S}}\left\{\left|\left(\left(I-\gamma P_\pi\right)\Phi\theta_t - r_\pi\right)_s\right|\right\}\sum_{s\in\mathcal{S}}|f(s) - f_b(s)| \\
&= B_\phi\|(I-\gamma P_\pi)\Phi\theta_t - r_\pi\|_\infty\|f - f_b\|_1. \qquad (19)
\end{aligned}
$$

We next proceed to bound $\|f - f_b\|_1$. Consider $f_b(s)$, we have

$$
f_b(s) = \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^b\middle|s_t^b = s, \mathcal{F}_{t-1}\right]
$$

$$
\begin{aligned}
&\overset{(i)}{=} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \\
&\quad \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a} \middle| s_t^b = s, \mathcal{F}_{t-1}\right) \\
&\qquad \cdot \mathbb{E}\left[\gamma \rho_t^{b-1} F_t^{b-1} + 1 \middle| \mathcal{F}_{t-1}, s_t^b = s, s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\right] \\
&\overset{(ii)}{=} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \\
&\cdot \left(1 + \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \frac{\mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mu(\tilde{a}|\tilde{s}) \mathsf{P}(s|\tilde{s},\tilde{a})}{\mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right)} \mathbb{E}\left[\gamma \frac{\pi(\tilde{a}|\tilde{s})}{\mu(\tilde{a}|\tilde{s})} F_t^{b-1} \middle| \mathcal{F}_{t-1}, s_t^{b-1} = \tilde{s}\right]\right) \\
&= \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) + \gamma \sum_{\tilde{s}\in\mathcal{S}} \mathbb{P}(s_t^{b-1} = \tilde{s}|\mathcal{F}_{t-1}) P_\pi(s|\tilde{s}) \mathbb{E}\left[F_t^{b-1} \middle| \mathcal{F}_{t-1}, s_t^{b-1} = s\right],
\end{aligned}
$$

where (i) follows from the law of total probability and (ii) follows from the Bayes rule and the facts that F_t^{b-1} only depends on the chain elements $(s_t^0, a_t^0, s_t^1, \ldots, s_t^{b-1}, a_t^{b-1})$ and the Markov property. Define $d_{\mu,b}(s) = \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right)$, the above equality can be rewritten as

$$
f_b(s) = d_{\mu,b}(s) + \gamma \sum_{\tilde{s}} P_\pi(s|\tilde{s}) f_{b-1}(\tilde{s}). \tag{20}
$$

Since eq. (20) holds for all $s \in \mathcal{S}$, we have

$$
f_b = d_{\mu,b} + \gamma P_\pi^\top f_{b-1}. \tag{21}
$$

Note that for f we have the following holds (Sutton et al., 2016; Zhang et al., 2020a)

$$
f = d_\mu + \gamma P_\pi^\top f.
$$

Equations (21) and (22) imply

$$
f - f_b = d_\mu - d_{\mu,b} + \gamma P_\pi^\top (f - f_{b-1}).
$$

Applying the above equality recursively yields

$$
f - f_b = \sum_{\tau=0}^{b-1} (\gamma P_\pi^\top)^\tau (d_\mu - d_{\mu,b-\tau}) + \gamma^b (P_\pi^\top)^b (f - f_0).
$$

Take ℓ_1 norm on both sides of the above equality, we have

$$
\begin{aligned}
\|f - f_b\|_1 &= \left\| \sum_{\tau=0}^{b-1} \gamma^\tau (P_\pi^\top)^\tau (d_\mu - d_{\mu,b-\tau}) + \gamma^b (P_\pi^\top)^b (f - f_0) \right\| \\
&\leq \sum_{\tau=0}^{b-1} \gamma^\tau \left\| (P_\pi^\top)^\tau (d_\mu - d_{\mu,b-\tau}) \right\|_1 + \gamma^b \left\| (P_\pi^\top)^b (f - f_0) \right\|_1 \\
&\leq \sum_{\tau=0}^{b-1} \gamma^\tau \left\| (P_\pi^\top)^\tau \right\|_1 \left\| d_\mu - d_{\mu,b-\tau} \right\|_1 + \gamma^b \left\| (P_\pi^\top)^b \right\| \left\| f - f_0 \right\|_1 \\
&\overset{(i)}{\leq} \sum_{\tau=0}^{b-1} \gamma^\tau \left\| d_\mu - d_{\mu,b-\tau} \right\|_1 + \gamma^b \left\| f - f_0 \right\|_1 \qquad (22) \\
&\overset{(ii)}{\leq} \sum_{\tau=0}^{b-1} C_M \gamma^\tau \chi^{b-\tau} + \gamma^b \|f - f_0\|_1 \\
&\overset{(ii)}{\leq} \frac{1}{|\chi - \gamma|} \cdot C_M \xi^b + \gamma^b (1 + \|f\|_1), \qquad (23)
\end{aligned}
$$

where (i) follows from Lemma 2, (ii) follows from Lemma 1 , and (iii) follows from Lemma 3 and defining $\xi := \max\{\chi, \gamma\}$.

To bound the term $\|(I-\gamma P_\pi)\Phi\theta_t - r_\pi\|_\infty$, we proceed as following

$$\begin{aligned}\|(I-\gamma P_\pi)\Phi\theta_t - r_\pi\|_\infty &\overset{(i)}{=} \|(I-\gamma P_\pi)(\Phi\theta_t - V_\pi)\|_\infty \\ &= \|(I-\gamma P_\pi)(\Phi\theta_t - \Phi\theta^* + \Phi\theta^* - V_\pi)\|_\infty \\ &\leq \|I-\gamma P_\pi\|_\infty(\|\Phi\theta_t - \Phi\theta^*\|_\infty + \|\Phi\theta^* - V_\pi\|_\infty) \\ &\overset{(ii)}{\leq} (1-\gamma)B_\phi\|\theta_t - \theta^*\|_2 + (1-\gamma)\epsilon_{approx}, \end{aligned} \tag{24}$$

where (i) follows from the fact $V_\pi = (I-\gamma P_\pi)^{-1}r_\pi$ and (ii) follows from the facts that $\|I - \gamma P_\pi\|_\infty = \max_i\{1-\gamma\sum_j(P_\pi)_{(i,j)}\} = 1-\gamma$, $\epsilon_{approx} := \|\Phi\theta^* - V_\pi\|_\infty$, and

$$\|\Phi\theta_t - \Phi\theta^*\|_\infty = \max_{s\in\mathcal{S}} \phi^\top(s)(\theta_t - \theta^*) \leq B_\phi\|\theta_t - \theta^*\|_2.$$

Substituting eqs. (23) and (24) into eq. (19) yields

$$\begin{aligned}\left\|\mathcal{T}(\theta_t) - \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t)\middle|\mathcal{F}_{t-1}\right]\right\|_2 &\leq \left(B_\phi^2\|\theta_t - \theta^*\|_2 + B_\phi\epsilon_{approx}\right)(1-\gamma)\left(\tfrac{C_M}{|\chi-\gamma|}\xi^b + \gamma^b(1+\|f\|_1)\right) \\ &\leq C_b\left(B_\phi\|\theta_t - \theta^*\|_2 + \epsilon_{approx}\right)\xi^b,\end{aligned}$$

where $\xi = \max\{\chi,\gamma\}$ and $C_b = B_\phi(1-\gamma)\left(\frac{C_M}{|\chi-\gamma|} + (1+\|f\|_1)\right)$.

B.2 Proof of Proposition 2

According to the definition of $\widehat{\mathcal{T}}_t(\theta_t)$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|^2\middle|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\left(\rho_t^b F_t^b\right)^2\left(\theta_t^\top\phi_t^b - r_t^b - \gamma\theta_t^\top\phi_t^{b+1}\right)^2\|\phi_t^b\|_2^2\middle|\mathcal{F}_{t-1}\right] \\ &\overset{(i)}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\mathbb{P}\left(s_t^b = s, a_t^b = a, s_t^{b+1} = s'\middle|\mathcal{F}_{t-1}\right)\cdot\left(\theta_t^\top\phi(s) - r(s,a) - \gamma\theta_t^\top\phi(s')\right)^2\|\phi(s)\|_2^2 \\ &\quad\cdot\mathbb{E}\left[(\rho_t^b F_t^b)^2\middle|\mathcal{F}_{t-1}, s_t^b = s, a_t^b = a, s_t^{b+1} = s'\right] \\ &\overset{(ii)}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mu(a|s)\mathsf{P}(s'|s,a)\cdot\left(\theta_t^\top\phi(s) - r(s,a) - \gamma\theta_t^\top\phi(s')\right)^2\|\phi(s)\|_2^2 \\ &\quad\frac{\pi^2(a|s)}{\mu^2(a|s)}\mathbb{E}\left[(F_t^b)^2\middle|\mathcal{F}_{t-1}, s_t^b = s, a_t^b = a, s_t^{b+1} = s'\right] \\ &= \sum_{s\in\mathcal{S}}\mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(F_t^b)^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right]\|\phi(s)\|_2^2 \\ &\quad\cdot\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\frac{\pi^2(a|s)}{\mu(a|s)}\mathsf{P}(s'|s,a)(\phi^\top(s)\theta_t - r(s,a) - \gamma\phi^\top(s')\theta_t)^2,\end{aligned} \tag{25}$$

where (i) follows from the law of total probability and (ii) follows from the fact that F_t^b is independent from previous states and actions given s_t^b.

Note that $\|\phi(s)\|_2 \leq B_\phi$ for all $s\in\mathcal{S}$, $r(s,a)\leq r_{max}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, and $\|\theta_t\|_2 \leq B_\theta$ for all t due to projection. We have

$$\begin{aligned}(\phi^\top(s)\theta_t - r(s,a) - \gamma\phi^\top(s')\theta_t)^2 &\leq 2[(\phi(s) - \gamma\phi(s'))^\top\theta_t]^2 + 2r^2(s,a) \\ &\leq 2\|\phi(s) - \gamma\phi(s')\|_2^2\|\theta_t\|_2^2 + 2r_{max}^2 \\ &\leq 4(\|\phi(s)\|_2^2 + \gamma^2\|\phi(s')\|_2^2)\|\theta_t\|_2^2 + 2r_{max}^2 \\ &\leq 4(1+\gamma^2)B_\phi^2B_\theta^2 + 2r_{max}^2.\end{aligned} \tag{26}$$

We also have

$$\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\frac{\pi^2(a|s)}{\mu(a|s)}\mathsf{P}(s'|s,a) \leq \rho_{max}\cdot\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a|s)\mathsf{P}(s'|s,a) = \rho_{max}.$$

Substituting the above two inequalities into eq. (25) yields

$$\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|^2\middle|\mathcal{F}_{t-1}\right] \leq \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \sum_{s\in\mathcal{S}} \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(F_t^b)^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right]. \quad (27)$$

Define

$$r_b(s) \coloneqq \mathbb{P}\left(s_t^b = s\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(F_t^b)^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right] = d_{\mu,b}(s)\mathbb{E}\left[(F_t^b)^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right].$$

We have the following equations hold for $r_b(s)$:

$$\begin{aligned} r_b(s) &= \mathbb{P}(s_t^b = s|\mathcal{F}_{t-1})\mathbb{E}\left[\left(\gamma\rho_t^{b-1}F_t^{b-1}+1\right)^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right] \\ &= d_{\mu,b}(s)\mathbb{E}\left[1 + 2\gamma\rho_t^{b-1}F_t^{b-1} + \gamma^2(\rho_t^{b-1})^2(F_t^{b-1})^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right] \\ &= d_{\mu,b}(s) + 2\gamma d_{\mu,b}(s)\mathbb{E}\left[\rho_t^{b-1}F_t^{b-1}\middle|\mathcal{F}_{t-1}, s_t^b = s\right] \\ &\quad + \gamma^2 d_{\mu,b}(s)\mathbb{E}\left[(\rho_t^{b-1})^2(F_t^{b-1})^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right]. \end{aligned} \quad (28)$$

For the second term in the RHS of eq. (28), we have

$$\begin{aligned} &d_{\mu,b}(s)\mathbb{E}\left[\rho_t^{b-1}F_t^{b-1}\middle|\mathcal{F}_{t-1}, s_t^b = s\right] \\ &\overset{(i)}{=} d_{\mu,b}(s)\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\middle|s_t^b = s, \mathcal{F}_{t-1}\right) \\ &\qquad\qquad \cdot \mathbb{E}\left[\rho_t^{b-1}F_t^{b-1}\middle|\mathcal{F}_{t-1}, s_t^b = s, s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\right] \\ &\overset{(ii)}{=} d_{\mu,b}(s)\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \frac{d_{\mu,b-1}(\tilde{s})\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a})}{d_{\mu,b}(s)} \cdot \mathbb{E}\left[\rho_t^{b-1}F_t^{b-1}\middle|\mathcal{F}_{t-1}, s_t^b = s, s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\right] \\ &\overset{(iii)}{=} \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} d_{\mu,b-1}(\tilde{s})\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a}) \cdot \frac{\pi(\tilde{a}|\tilde{s})}{\mu(\tilde{a}|\tilde{s})}\mathbb{E}\left[F_t^{b-1}\middle|\mathcal{F}_{t-1}, s_t^{b-1} = \tilde{s}\right] \\ &= \sum_{\tilde{s}\in\mathcal{S}} P_\pi(s|\tilde{s}) \cdot d_{\mu,b-1}(\tilde{s})\mathbb{E}\left[F_t^{b-1}\middle|\mathcal{F}_{t-1}, s_t^{b-1} = \tilde{s}\right] \\ &\overset{(iv)}{=} (P_\pi^\top f_{b-1})_s, \end{aligned} \quad (29)$$

where (i) follows from the law of total probability, (ii) follows from the Bayes rule, (iii) follows from Markov property and (iv) follow from the definition of f_b which is given above eq. (19).

For the third term on the RHS of eq. (28), we have

$$\begin{aligned} &d_{\mu,b}(s)\mathbb{E}\left[(\rho_t^{b-1}F_b^{b-1})^2\middle|\mathcal{F}_{t-1}, s_t^b = s\right] \\ &\overset{(i)}{=} d_{\mu,b}(s)\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\middle|s_t^b = s, \mathcal{F}_{t-1}\right) \\ &\qquad\qquad \cdot \mathbb{E}\left[(\rho_t^{b-1}F_b^{b-1})^2\middle|\mathcal{F}_{t-1}, s_t^b = s, s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\right] \\ &\overset{(ii)}{=} d_{\mu,b}(s)\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \frac{d_{\mu,b-1}(\tilde{s})\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a})}{d_{\mu,b}(s)} \cdot \mathbb{E}\left[(\rho_t^{b-1}F_t^{b-1})^2\middle|\mathcal{F}_{t-1}, s_t^b = s, s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\right] \\ &= \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} d_{\mu,b-1}(\tilde{s})\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a}) \cdot \frac{\pi^2(\tilde{a}|\tilde{s})}{\mu^2(\tilde{a}|\tilde{s})} \cdot \mathbb{E}\left[(F_t^{b-1})^2\middle|\mathcal{F}_{t-1}, s_t^{b-1} = \tilde{s}\right] \\ &\overset{(iii)}{=} \sum_{\tilde{s}\in\mathcal{S}} P_{\mu,\pi}(s|\tilde{s})r_{b-1}(\tilde{s}) \\ &= (P_{\mu,\pi}^\top r_{b-1})_s, \end{aligned} \quad (30)$$

where (i) follows from the law of total probability, (ii) follows from the Bayes' rule, and in (iii) we define $P_{\mu,\pi} \in \mathbb{R}^{N\times N}$ where $(P_{\mu,\pi})_{s,\tilde{s}} = \sum_{\tilde{a}\in\mathcal{A}} \frac{\pi^2(\tilde{a}|\tilde{s})}{\mu(\tilde{a}|\tilde{s})}\mathsf{P}(s|\tilde{s},\tilde{a})$ for each $(s,\tilde{s}) \in \mathcal{S}\times\mathcal{S}$.

Substituting eqs. (29) and (30) into eq. (28) yields

$$r_b = d_{\mu,b} + 2\gamma P_\pi^\top f_{b-1} + \gamma^2 P_{\mu,\pi}^\top r_{b-1}.$$

We also have the following inequality holds

$$\begin{aligned}
\mathbf{1}^\top r_b &= \mathbf{1}^\top d_{\mu,b} + 2\gamma \mathbf{1}^\top P_\pi^\top f_{b-1} + \gamma^2 \mathbf{1}^\top P_{\mu,\pi}^\top r_{b-1} \\
&\overset{(i)}{=} 1 + 2\gamma \mathbf{1}^\top f_{b-1} + \gamma^2 \mathbf{1}^\top P_{\mu,\pi}^\top r_{b-1} \\
&\overset{(ii)}{\leq} 1 + 2\gamma \mathbf{1}^\top f_{b-1} + \gamma^2 \rho_{max} \mathbf{1}^\top r_{b-1},
\end{aligned} \tag{31}$$

where (i) follows from $\mathbf{1}^\top P_\pi^\top = (P_\pi \mathbf{1})^\top = \mathbf{1}^\top$, and (ii) follows from the facts that $r_{b-1} \succeq \mathbf{0}$ and

$$\begin{aligned}
\mathbf{1}^\top P_{\mu,\pi}^\top &= (P_{\mu,\pi}\mathbf{1})^\top \\
&= \mathrm{vec}\left(\sum_{s\in\mathcal{S}}\sum_{\tilde{a}\in\mathcal{A}} \frac{\pi^2(\tilde{a}|\tilde{s})}{\mu(\tilde{a}|\tilde{s})} \mathsf{P}(s|\tilde{s},\tilde{a})\right) \\
&= \mathrm{vec}\left(\sum_{\tilde{a}\in\mathcal{A}} \frac{\pi^2(\tilde{a}|\tilde{s})}{\mu(\tilde{a}|\tilde{s})}\right) \\
&\preceq \rho_{max}\mathbf{1}^\top.
\end{aligned}$$

Recursively applying eq. (31) yields

$$\mathbf{1}^\top r_b \leq \sum_{\tau=0}^{b-1} (\gamma^2\rho_{max})^\tau (1 + 2\gamma \mathbf{1}^\top f_{b-\tau-1}) + (\gamma^2\rho_{max})^b \mathbf{1}^\top r_0 \tag{32}$$

$$\overset{(i)}{=} \sum_{\tau=0}^{b-1} (\gamma^2\rho_{max})^\tau (1 + 2\gamma \mathbf{1}^\top f_{b-\tau-1}) + (\gamma^2\rho_{max})^b, \tag{33}$$

where (i) follows from the fact that $\mathbf{1}^\top r_0 = 1$.

Recall that $f_b = d_{\mu,b} + \gamma P_\pi^\top f_{b-1}$ and $f = d_\mu + \gamma P_\pi^\top f$. We have

$$\begin{aligned}
\mathbf{1}^\top (f_\tau - f) &= \mathbf{1}^\top (d_{\mu,\tau} - d_\mu) + \gamma \mathbf{1}^\top P_\pi^\top (f_{\tau-1} - f) \\
&\overset{(i)}{=} \gamma \mathbf{1}^\top (f_{\tau-1} - f) \\
&\overset{(ii)}{=} \gamma^\tau \mathbf{1}^\top (f_0 - f),
\end{aligned}$$

where (i) follows from the facts that $d_{\mu,\tau}$ and d_μ are both probability distributions and $\mathbf{1}^\top d_{\mu,\tau} = \mathbf{1}^\top d_\mu = 1$ and $\mathbf{1}^\top P_\pi^\top = \mathbf{1}^\top$ and (ii) follows from recursively applying (i).

Thus, we have

$$\begin{aligned}
|\mathbf{1}^\top f_\tau| &= |(1-\gamma^\tau)\mathbf{1}^\top f + \gamma^\tau \mathbf{1}^\top f_0| \\
&\leq |(1-\gamma^\tau)\mathbf{1}^\top f| + \gamma^\tau \\
&\leq \|f\|_1 + 1.
\end{aligned} \tag{34}$$

Substituting eq. (34) into eq. (33) yields

$$\mathbf{1}^\top r_b \leq \sum_{\tau=0}^{b-1} (\gamma^2\rho_{max})^\tau (3 + 2\gamma\|f\|_1) + (\gamma^2\rho_{max})^b$$

Under different conditions of ρ_{max}, the term $\mathbf{1}^\top r_b$ is upper bounded differently as following:

(a) $\gamma^2\rho_{max} > 1$

$$\mathbf{1}^\top r_b \leq \left(\frac{3 + 2\gamma\|f\|_1}{\gamma^2\rho_{max} - 1} + 1\right)\gamma^{2b}\rho_{max}^b. \tag{35}$$

(b) $\gamma^2\rho_{max} = 1$

$$\mathbf{1}^\top r_b \leq (3 + 2\gamma\|f\|_1)\, b + 1. \tag{36}$$

(c) $\gamma^2\rho_{max} < 1$

$$\mathbf{1}^\top r_b \leq \frac{3 + 2\gamma\|f\|_1}{1 - \gamma^2\rho_{max}} + 1. \tag{37}$$

Substituting the above inequalities into eq. (27), we can upper-bound the term $\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right]$ under different conditions accordingly:

(a) $\gamma^2\rho_{max} > 1$

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right] &\leq \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \left(\frac{3 + 2\gamma\|f\|_1}{\gamma^2\rho_{max} - 1} + 1\right)\gamma^{2b}\rho_{max}^b \\ &= C_{\sigma,1}\gamma^{2b}\rho_{max}^b,\end{aligned}$$

where we specify $C_{\sigma,1} = \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \left(\frac{3+2\gamma\|f\|_1}{\gamma^2\rho_{max}-1} + 1\right)$.

(b) $\gamma^2\rho_{max} = 1$

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right] &\leq \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \left((3 + 2\gamma\|f\|_1)\, b + 1\right) \\ &= C_{\sigma,2} b,\end{aligned}$$

where we specify $C_{\sigma,2} = \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 (4 + 2\gamma\|f\|_1)$.

(c) $\gamma^2\rho_{max} < 1$

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right] &\leq \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \left(\frac{3 + 2\gamma\|f\|_1}{1 - \gamma^2\rho_{max}} + 1\right) \\ &\coloneqq C_{\sigma,3}.\end{aligned}$$

where we specify $C_{\sigma,3} = \rho_{max}\left(4(1+\gamma^2)B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2\rho_{max}} + 1\right)$.

To summarize, the variance term $\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2$ can be bounded as following

$$\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right] \leq \sigma^2,$$

where

$$\sigma^2 = \begin{cases} \mathcal{O}(1), & \textit{if}\ \gamma^2\rho_{max} < 1. \\ \mathcal{O}(b), & \textit{if}\ \gamma^2\rho_{max} = 1. \\ \mathcal{O}((\gamma^2\rho_{max})^b), & \textit{if}\ \gamma^2\rho_{max} > 1. \end{cases}$$

B.3 Proof of Proposition 3

Define the matrix $A_t \coloneqq \rho_t^b e_t^b(\phi_t^b - \gamma\phi_t^{b+1})$ and $c_t \coloneqq r_t^b\rho_t^b e_t^b$. We have

$$\widehat{\mathcal{T}}_t^\lambda(\theta_t) = A_t\theta_t - c_t. \tag{38}$$

Recall that θ_t is $\mathcal{F}_{t-1}$-measurable. We have

$$\mathbb{E}\left[\widehat{\mathcal{T}}_t^\lambda(\theta_t) \middle| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[A_t|\mathcal{F}_{t-1}\right]\theta_t - \mathbb{E}\left[c_t|\mathcal{F}_{t-1}\right].$$

To bound the bias error term $\left\|\mathbb{E}\left[\widehat{\mathcal{T}}_t^\lambda(\theta_t) \middle| \mathcal{F}_{t-1}\right] - \mathcal{T}^\lambda(\theta_t)\right\|_2$, we first take conditional expectations on A_t and c_t, respectively, as following

$\mathbb{E}\left[A_t|\mathcal{F}_{t-1}\right]$

$$
\begin{aligned}
&= \mathbb{E}\left[\rho_t^b e_t^b (\phi_t^b - \gamma\phi_t^{b+1})^\top \middle| \mathcal{F}_{t-1}\right] \\
&\overset{(i)}{=} \sum_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}} \mathbb{P}\left(s_t^b = s, a_t^b = a, s_t^{b+1} = s' \middle| \mathcal{F}_{t-1}\right) \qquad (39) \\
&\qquad \cdot \mathbb{E}\left[\rho_t^b e_t^b (\phi_t^b - \gamma\phi_t^{b+1})^\top \middle| \mathcal{F}_{t-1}, s_t^b = s, a_t^b = a, s_t^{b+1} = s'\right] \\
&\overset{(ii)}{=} \sum_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mu(a|s) \mathsf{P}(s'|s,a) \cdot \frac{\pi(a|s)}{\mu(a|s)} \mathbb{E}\left[e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right] (\phi(s) - \gamma\phi(s'))^\top \\
&= \sum_{s\in\mathcal{S}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t^b \middle| \mathcal{F}_{t-1}, s_t^b = s\right] \sum_{a\in\mathcal{A}, s'\in\mathcal{S}} \pi(a|s) \mathsf{P}(s'|s,a) \left(\phi(s) - \gamma\phi(s')\right)^\top \\
&= \sum_{s\in\mathcal{S}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t^b \middle| \mathcal{F}_{t-1}, s_t^b = s\right] \cdot \left((\Phi)_{(s,\cdot)} - \gamma (P_\pi \Phi)_{(s,\cdot)}\right), \qquad (40)
\end{aligned}
$$

where (i) follows from the law of total probability and (ii) follows from the Markov property and the fact that e_t^b only depends on $(s_t^0, a_t^0, s_t^1, \ldots, s_t^b)$.

Define $\beta_\tau(s) = \mathbb{P}\left(s_t^\tau = s | \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t^b \middle| \mathcal{F}_{t-1}, s_t^\tau = s\right]$. We have

$$
\begin{aligned}
\beta_b(s) &= \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right] \\
&\overset{(i)}{=} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \\
&\qquad \cdot \sum_{\tilde{s}\in\mathcal{S}, \tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a} \middle| s_t^b = s, \mathcal{F}_{t-1}\right) \\
&\qquad\qquad \cdot \mathbb{E}\left[\gamma\lambda\rho_t^{b-1} e_t^{b-1} + (\lambda + (1-\lambda)(1 + \rho_t^{b-1}\gamma F_t^{b-1})\phi_t^b) \middle| s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}, s_t^b = s, \mathcal{F}_{t-1}\right] \\
&\overset{(ii)}{=} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \phi(s) \\
&\qquad + \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \sum_{\tilde{s}\in\mathcal{S}, \tilde{a}\in\mathcal{A}} \frac{\mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mu(\tilde{a}|\tilde{s}) \mathsf{P}(s|\tilde{s},\tilde{a})}{\mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right)} \\
&\qquad\qquad \cdot \frac{\pi(\tilde{a}|\tilde{s})}{\mu(\tilde{a}|\tilde{s})} \cdot \mathbb{E}\left[\gamma\lambda e_t^{b-1} + (1-\lambda)\gamma F_t^{b-1}\phi(s) \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&= \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \phi(s) + \sum_{\tilde{s}\in\mathcal{S}} \mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) P_\pi(s|\tilde{s}) \\
&\qquad\qquad \cdot \mathbb{E}\left[\gamma\lambda e_t^{b-1} + (1-\lambda)\gamma F_t^{b-1}\phi(s) \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\overset{(iii)}{=} (\lambda d_{\mu,b}(s) + (1-\lambda) f_b(s)) \cdot \phi(s) + \gamma\lambda (P_\pi^\top \beta_{b-1})_s, \qquad (41)
\end{aligned}
$$

where (i) follows from the law of total probability, (ii) follows from the Bayes rule and the Markov property, and (iii) follows from the following definitions: $d_{\mu,b}(s) = \mathbb{P}\left(s_t^b = s | \mathcal{F}_{t-1}\right)$, $f_b(s) = d_{\mu,b}(s) \mathbb{E}\left[F_t^b = s | s_t^b = s, \mathcal{F}_{t-1}\right]$, $f_b = d_{\mu,b} + \gamma P_\pi^\top f_{b-1}$, and $\beta_\tau(s) = \mathbb{P}\left(s_t^\tau = s | \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t^b \middle| \mathcal{F}_{t-1}, s_t^\tau = s\right]$.

Define the matrix $\beta_\tau \in \mathbb{R}^{d\times N}$, where $\beta_\tau = (\beta_\tau(1), \beta_\tau(2), \ldots, \beta_\tau(N))$. Then, eq. (41) implies that

$$
\beta_b = \lambda \Phi^\top D_{\mu,b} + (1-\lambda)\Phi^\top F_b + \gamma\lambda \beta_{b-1} P_\pi, \qquad (42)
$$

where $D_{\mu,b} \coloneqq \mathrm{diag}(d_{\mu,b}(1), d_{\mu,b}(2), \ldots, d_{\mu,b}(N))$ and $F_b = \mathrm{diag}(f_b)$.

Recursively applying the above equality yields

$$
\beta_b = (\gamma\lambda)^b \beta_0 P_\pi^b + \lambda \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top D_{\mu,b-\tau} P_\pi^\tau + (1-\lambda) \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top F_{b-\tau} P_\pi^\tau. \qquad (43)
$$

Taking expectation of c_t conditioned on $\mathcal{F}_{t-1}$, we have

$$
\begin{aligned}
\mathbb{E}\left[c_t | \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\rho_t^b e_t^b r_t^b \middle| \mathcal{F}_{t-1}\right] \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mathbb{P}\left(s_t^b = s, a_t^b = a \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[\rho_t^b e_t^b r_t^b \middle| s_t^b = s, a_t^b = a, \mathcal{F}_{t-1}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mu(a|s) \cdot \frac{\pi(a|s)}{\mu(a|s)} r(s,a) \mathbb{E}\left[e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right] \\
&= \sum_{s\in\mathcal{S}} r_\pi(s) \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right] \\
&= \sum_{s\in\mathcal{S}} r_\pi(s) \beta_b(s). \qquad (44)
\end{aligned}
$$

Substituting eqs. (40) and (44) into eq. (38) yields

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{T}}_t^\lambda(\theta_t) \middle| \mathcal{F}_{t-1}\right] &= \sum_{s\in\mathcal{S}} \beta_b(s) \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right)_s \\
&= \beta_b \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right).
\end{aligned}
$$

Recall the definition of $\mathcal{T}^\lambda(\theta)$. We have

$$
\mathcal{T}^\lambda(\theta_t) - \mathbb{E}\left[\widehat{\mathcal{T}}_t^\lambda(\theta_t) \middle| \mathcal{F}_{t-1}\right] = \left(\Phi^\top M (I - \gamma\lambda P_\pi)^{-1} - \beta_b\right) \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right). \qquad (45)
$$

We then proceed to bound the term $\Phi^\top M (I - \gamma\lambda P_\pi)^{-1} - \beta_b$

$$
\begin{aligned}
&\Phi^\top M (I - \gamma\lambda P_\pi)^{-1} - \beta_b \\
&\quad \overset{(i)}{=} \Phi^\top M \left(\sum_{\tau=0}^{\infty} (\gamma\lambda)^\tau P_\pi^\tau\right) - \beta_b \\
&\quad \overset{(ii)}{=} \Phi^\top \left(\lambda D_\mu + (1-\lambda) F\right) \left(\sum_{\tau=0}^{\infty} (\gamma\lambda)^\tau P_\pi^\tau\right) \\
&\qquad - \left((\gamma\lambda)^b \beta_0 P_\pi^b + \lambda \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top D_{\mu,b-\tau} P_\pi^\tau + (1-\lambda) \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top F_{b-\tau} P_\pi^\tau\right) \\
&\quad = \lambda \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top \left(D_\mu - D_{\mu,b-\tau}\right) P_\pi^\tau + (1-\lambda) \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top \left(F - F_{b-\tau}\right) P_\pi^\tau \\
&\qquad + \sum_{\tau=b}^{\infty} (\gamma\lambda)^\tau \Phi^\top (\lambda D_\mu + (1-\lambda) F) P_\pi^\tau - \lambda (\gamma\lambda)^b \Phi^\top D_{\mu,0} P_\pi^b,
\end{aligned}
$$

where (i) follows from the fact that $(I - \gamma P_\pi)^{-1} = \sum_{\tau=0}^{\infty} \gamma^\tau P_\pi^\tau$, and (ii) follows from eq. (43). Substituting the above equality into eq. (45) and taking ℓ_2 norm on the both sides yield

$$
\begin{aligned}
&\left\| \mathcal{T}^\lambda(\theta_t) - \mathbb{E}\left[\widehat{\mathcal{T}}_t^\lambda(\theta_t) \middle| \mathcal{F}_{t-1}\right] \right\|_2 \\
&\quad = \left\| \left(\lambda \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top \left(D_\mu - D_{\mu,b-\tau}\right) P_\pi^\tau + (1-\lambda) \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \Phi^\top \left(F - F_{b-\tau}\right) P_\pi^\tau \right.\right. \\
&\qquad \left.\left. + \sum_{\tau=b}^{\infty} (\gamma\lambda)^\tau \Phi^\top (\lambda D_\mu + (1-\lambda) F) P_\pi^\tau - \lambda (\gamma\lambda)^b \Phi^\top D_{\mu,0} P_\pi^b \right) \left(\Phi\theta_t - \gamma P_\pi \theta_t - r_\pi\right) \right\|_2 \\
&\quad \le \lambda \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \left\| \Phi^\top \left(D_\mu - D_{\mu,b-\tau}\right) P_\pi^\tau \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right) \right\|_2 \\
&\qquad + (1-\lambda) \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \left\| \Phi^\top \left(F - F_{b-\tau}\right) P_\pi^\tau \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right) \right\|_2 \\
&\qquad + \sum_{\tau=b}^{\infty} (\gamma\lambda)^\tau \left\| \Phi^\top \left(\lambda D_\mu + (1-\lambda) F\right) P_\pi^\tau \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right) \right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \lambda(\gamma\lambda)^b \left\|\Phi^\top D_{\mu,0} P_\pi^b \left(\Phi\theta_t - \gamma P_\pi \Phi\theta_t - r_\pi\right)\right\|_2 \\
&\overset{(i)}{\leq} \lambda \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau B_\phi \|d_\mu - d_{\mu,b-\tau}\|_1 (1-\gamma) \left(B_\phi \|\theta_t - \theta_\lambda^*\|_2 + \epsilon_{approx}\right) \\
&\quad + (1-\lambda) \sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau B_\phi \left\|f - f_{b-\tau}\right\|_1 (1-\gamma) \left(B_\phi \|\theta_t - \theta_\lambda^*\|_2 + C\epsilon_{approx}\right) \\
&\quad + \sum_{\tau=b}^{\infty} (\gamma\lambda)^\tau B_\phi \left(\lambda\|d_\mu\|_1 + (1-\lambda)\|f\|_1\right) (1-\gamma) \left(B_\phi \|\theta_t - \theta_\lambda^*\|_2 + \epsilon_{approx}\right) \\
&\quad + \lambda(\gamma\lambda)^b B_\phi \|d_{\mu,0}\|_1 (1-\gamma) \left(B_\phi \|\theta_t - \theta_\lambda^*\|_2 + \epsilon_{approx}\right) \\
&\overset{(ii)}{\leq} \left(B_\phi^2 \|\theta_t - \theta_\lambda^*\|_2 + B_\phi \epsilon_{approx}\right) \\
&\quad \left(\sum_{\tau=0}^{b-1} (\gamma\lambda)^\tau \left(\lambda(1-\gamma) C_M \chi^{b-\tau} + (1-\lambda)(1-\gamma)\left(\tfrac{C_M}{|\gamma-\chi|}\xi^{b-\tau} + \gamma^{b-\tau}(1+\|f\|_1)\right)\right)\right) \\
&\quad + (1-\gamma)\left(B_\phi^2 \|\theta_t - \theta_\lambda^*\|_2 + B_\phi \epsilon_{approx}\right) \left(\left(\frac{\lambda + (1-\lambda)\|f\|_1}{1-\gamma\lambda} + \lambda\right)(\gamma\lambda)^b\right) \\
&\overset{(iii)}{\leq} \left(B_\phi^2 \|\theta_t - \theta_\lambda^*\|_2 + B_\phi \epsilon_{approx}\right) \\
&\qquad \cdot (1-\gamma)\left(\tfrac{\lambda}{|\chi-\gamma\lambda|} C_M \xi^b + \tfrac{C_M(1-\lambda)}{|\gamma-\chi|(\xi-\gamma\lambda)}\xi^b + (1+\|f\|_1)\gamma^b + \left(\tfrac{\lambda+(1-\lambda)\|f\|_1}{1-\gamma\lambda} + \lambda\right)(\gamma\lambda)^b\right) \\
&\overset{(iv)}{\leq} C_{b,\lambda} \left(B_\phi \|\theta_t - \theta_\lambda^*\|_2 + \epsilon_{approx}\right) \xi^b,
\end{aligned}
$$

where (i) follows from Lemma 5 and eq. (24), (ii) follows from Lemma 1 and eq. (23), in (iii) we define

$$
C_{b,\lambda} \coloneqq B_\phi (1-\gamma)\left(\tfrac{\lambda}{|\chi-\gamma\lambda|} C_M + \tfrac{C_M(1-\lambda)}{|\gamma-\chi|(\xi-\gamma\lambda)} + 1 + \|f\|_1 + \left(\tfrac{\lambda+(1-\lambda)\|f\|_1}{1-\gamma\lambda} + \lambda\right)\right),
$$

and (iv) follows from Lemma 3.

B.4 Proof of Proposition 4

According to the definition of $\widehat{\mathcal{T}}_t^\lambda$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[(\rho_t^b)^2 \left(r_t^b + \gamma\theta_t^\top \phi_t^{b+1} - \theta_t^\top \phi_t^b\right)^2 (e_t^b)^\top e_t^b \middle| \mathcal{F}_{t-1}\right] \\
&\overset{(i)}{=} \sum_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}} \mathbb{P}\left(s_t^b = s, a_t^b = a, s_t^{b+1} = s' \middle| \mathcal{F}_{t-1}\right) \\
&\qquad \cdot \mathbb{E}\left[(\rho_t^b)^2 \left(r_t^b + \gamma\theta_t^\top \phi_t^{b+1} - \theta_t^\top \phi_t^b\right)^2 (e_t^b)^\top e_t^b \middle| s_t^b = s, a_t^b = a, s_t^{b+1} = s', \mathcal{F}_{t-1}\right] \\
&\overset{(ii)}{=} \sum_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mu(a|s) \mathsf{P}(s'|s,a) \\
&\qquad \frac{\pi^2(a|s)}{\mu^2(a|s)} (r(s,a) + \gamma\theta_t^\top \phi(s') - \theta_t^\top \phi(s))^2 \mathbb{E}\left[(e_t^b)^\top e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right] \\
&\overset{(iii)}{\leq} \sum_{s\in\mathcal{S}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[(e_t^b)^\top e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right] \\
&\qquad \cdot \sum_{s'\in\mathcal{S}} P_{\mu,\pi}(s'|s) (r(s,a) + \gamma\phi(s')^\top \theta_t - \phi(s)^\top \theta_t)^2 \\
&\overset{(iv)}{\leq} \rho_{max}\left(4(1+\gamma^2) B_\phi^2 B_\theta^2 + 2r_{max}^2\right) B_\phi^2 \sum_{s\in\mathcal{S}} \mathbb{P}\left(s_t^b = s \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[(e_t^b)^\top e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right],
\end{aligned}
\tag{46}
$$

where (i) follows from the law of total probability, (ii) follows from the Markov property and the fact that e_t^b only depends on $(s_t^0, a_t^0, \ldots, s_t^b)$, and (iii) follows from eq. (26) and the fact

$$\sum_{s'} P_{\mu,\pi}(s'|s) \leq \rho_{max}.$$

Define $\Delta_b(s) = \mathbb{P}\left(s_t^b = s|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(e_t^b)^\top e_t^b \middle| s_t^b = s, \mathcal{F}_{t-1}\right]$. We then proceed to bound the term $\Delta_b(s)$. We have

$$\begin{aligned}
&\Delta_b(s) \\
&\overset{(i)}{=} \mathbb{P}\left(s_t^b = s|\mathcal{F}_{t-1}\right) \\
&\quad \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}\middle| s_t^b = s, \mathcal{F}_{t-1}\right)\mathbb{E}\left[(e_t^b)^\top e_t^b \middle| s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}, s_t^b = s, \mathcal{F}_{t-1}\right] \\
&\overset{(ii)}{=} \mathbb{P}\left(s_t^b = s|\mathcal{F}_{t-1}\right) \\
&\quad \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \frac{\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a})}{\mathbb{P}\left(s_t^b = s|\mathcal{F}_{t-1}\right)}\mathbb{E}\left[(e_t^b)^\top e_t^b \middle| s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}, s_t^b = s, \mathcal{F}_{t-1}\right] \\
&\overset{(iii)}{=} \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a}) \\
&\quad \cdot \mathbb{E}\Big[\left(\gamma\lambda\rho_t^{b-1}e_t^{b-1} + \left(\lambda + (1-\lambda)\left(\gamma\rho_t^{b-1}F_t^{b-1}+1\right)\right)\phi_t^b\right)^\top \\
&\qquad \cdot \left(\gamma\lambda\rho_t^{b-1}e_t^{b-1} + \left(\lambda + (1-\lambda)\left(\gamma\rho_t^{b-1}F_t^{b-1}+1\right)\right)\phi_t^b\right)\Big| s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}, s_t^b = s, \mathcal{F}_{t-1}\Big] \\
&= \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mu(\tilde{a}|\tilde{s})\mathsf{P}(s|\tilde{s},\tilde{a}) \\
&\quad \mathbb{E}\Big[(\gamma\lambda)^2(\rho_t^{b-1})^2(e_t^{b-1})^\top e_t^{b-1} + (1-\lambda)^2\gamma^2(\rho_t^{b-1}F_t^{b-1})^2\phi(s)^\top\phi(s) + \phi(s)^\top\phi(s) \\
&\qquad +2\gamma^2\lambda(1-\lambda)(\rho_t^{b-1})^2F_t^{b-1}\phi(s)^\top e_t^{b-1} + 2\gamma\lambda\rho_t^{b-1}\phi(s)^\top e_t^{b-1} \\
&\qquad\quad +2(1-\lambda)\gamma\rho_t^{b-1}F_t^{b-1}\phi(s)^\top\phi(s)\Big| s_t^{b-1} = \tilde{s}, a_t^{b-1} = \tilde{a}, s_t^b = s, \mathcal{F}_{t-1}\Big] \\
&= \mathbb{P}\left(s_t^b = s|\mathcal{F}_{t-1}\right)\|\phi(s)\|_2^2 \\
&\quad + \phi^\top(s)\sum_{\tilde{s}\in\mathcal{S}} P_\pi(s|\tilde{s})\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[2\gamma\lambda e_t^{b-1} + 2\gamma(1-\lambda)F_t^{b-1}\phi(s)\middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\qquad + \lambda^2\gamma^2\sum_{\tilde{s}\in\mathcal{S}} P_{\mu,\pi}(s|\tilde{s})\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(e_t^{b-1})^\top e_t^{b-1}\middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\qquad\quad + 2\gamma^2\lambda(1-\lambda)\sum_{\tilde{s}\in\mathcal{S}} P_{\mu,\pi}(s|\tilde{s})\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^{b-1}\phi(s)^\top e_t^{b-1}\middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\qquad\qquad + \|\phi(s)\|_2^2\gamma^2(1-\lambda)^2\sum_{\tilde{s}\in\mathcal{S}} P_{\mu,\pi}(s|\tilde{s})\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(F_t^{b-1})^2\middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\leq B_\phi^2 d_{\mu,b}(s) + 2\gamma\lambda\phi^\top(s)(\beta_{b-1}P_\pi)_{(\cdot,s)} + 2\gamma(1-\lambda)B_\phi^2(P_\pi^\top f_{b-1})_s \\
&\quad + \lambda^2\gamma^2(P_{\mu,\pi}^\top\Delta_{b-1})_s + \gamma^2(1-\lambda)^2B_\phi^2(P_{\mu,\pi}^\top r_{b-1})_s \\
&\qquad + 2\gamma^2\lambda(1-\lambda)\sum_{\tilde{s}\in\mathcal{S}} P_{\mu,\pi}(s|\tilde{s})\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^{b-1}\phi(s)^\top e_t^{b-1}\middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\overset{(iv)}{=} B_\phi^2 d_{\mu,b}(s) + \lambda^2\gamma^2(P_{\mu,\pi}^\top\Delta_{b-1})_s + 2\gamma(1-\lambda)B_\phi^2(P_\pi^\top f_{b-1})_s + 2\gamma\lambda\phi^\top(s)(\beta_{b-1}P_\pi)_{(\cdot,s)} \\
&\quad + \gamma^2(1-\lambda)^2B_\phi^2(P_{\mu,\pi}^\top r_{b-1})_s + 2\gamma^2\lambda(1-\lambda)\phi(s)^\top(\delta_{b-1}P_{\mu,\pi})_{(\cdot,s)}, \qquad (47)
\end{aligned}$$

where (i) follows from the law of total probability, (ii) follows from the Bayes rule, (iii) follows the update rule of e_t^b and in (iv) we define

$$\delta_\tau(s) = \mathbb{P}\left(s_t^\tau = s|\mathcal{F}_{t-1}\right)\mathbb{E}\left[F_t^\tau e_t^\tau | s_t^\tau = s, \mathcal{F}_{t-1}\right].$$

Summing eq. (47) over $\mathcal{S}$ yields

$$\mathbf{1}^\top\Delta_b$$

$$
\begin{aligned}
&= \sum_s \Delta_b(s) \\
&\leq B_\phi^2 \mathbf{1}^\top d_{\mu,b}(s) + \lambda^2\gamma^2 \mathbf{1}^\top P_{\mu,\pi}^\top \Delta_{b-1} + 2\gamma(1-\lambda) B_\phi^2 \mathbf{1}^\top P_\pi^\top f_{b-1} + 2\gamma\lambda \text{trace}\left(\Phi^\top \beta_{b-1} P_\pi\right) \\
&\quad + \gamma^2(1-\lambda)^2 B_\phi^2 \mathbf{1}^\top P_{\mu,\pi}^\top r_{b-1} + 2\gamma^2\lambda(1-\lambda)\text{trace}\left(\Phi^\top \delta_{b-1} P_{\mu,\pi}\right) \\
&\overset{(i)}{\leq} \lambda^2\gamma^2\rho_{max} \mathbf{1}^\top \Delta_{b-1} + B_\phi^2 + 2\gamma(1-\lambda) B_\phi^2 \mathbf{1}^\top f_{b-1} + 2\gamma\lambda \text{trace}\left(\Phi^\top \beta_{b-1} P_\pi\right) \\
&\quad + \gamma^2(1-\lambda)^2 B_\phi^2 \rho_{max} \mathbf{1}^\top r_{b-1} + 2\gamma^2\lambda(1-\lambda)\text{trace}\left(\Phi^\top \delta_{b-1} P_{\mu,\pi}\right)
\end{aligned}
$$

where (i) follows from Lemma 4 with $P = P_{\mu,\pi}$.

Recursively applying the above inequality, we have

$$
\begin{aligned}
&\mathbf{1}^\top \Delta_b \\
&\leq (\lambda^2\gamma^2)^b \rho_{max}^b \mathbf{1}^\top \Delta_0 + \sum_{\tau=1}^{b} (\lambda^2\gamma^2)^{b-\tau} (\rho_{max})^{b-\tau} \left(B_\phi^2 + 2\gamma(1-\lambda) B_\phi^2 \mathbf{1}^\top f_{\tau-1} \right. \\
&\quad \left. + 2\gamma\lambda \text{trace}\left(\Phi^\top \beta_{\tau-1} P_\pi\right) + \gamma^2(1-\lambda)^2 B_\phi^2 \rho_{max} \mathbf{1}^\top r_{\tau-1} + 2\gamma^2\lambda(1-\lambda)\text{trace}\left(\Phi^\top \delta_{\tau-1} P_{\mu,\pi}\right)\right).
\end{aligned} \tag{48}
$$

Substituting eq. (43) into $\text{trace}\left(\Phi\beta_\tau P_\pi\right)$ with b and τ replaced by τ and m respectively yields

$$
\begin{aligned}
&\text{trace}\left(\Phi\beta_\tau P_\pi\right) \\
&\overset{(i)}{=} (\gamma\lambda)^\tau \text{trace}\left(\Phi\beta_0 P_\pi^{\tau+1}\right) \\
&\quad + \lambda \sum_{m=0}^{\tau-1} (\gamma\lambda)^m \left(\text{trace}\left(\Phi\Phi^\top D_{\mu,\tau-m} P_\pi^{m+1}\right) + (1-\lambda)\text{trace}\left(\Phi\Phi^\top F_{\tau-m} P_\pi^{m+1}\right)\right) \\
&\overset{(ii)}{=} (\gamma\lambda)^\tau \text{trace}\left(P_\pi^{\tau+1}\Phi\Phi^\top D_{\mu,0}\right) \\
&\quad + \lambda \sum_{m=0}^{\tau-1} (\gamma\lambda)^m \left(\text{trace}\left(P_\pi^{m+1}\Phi\Phi^\top D_{\mu,\tau-m}\right) + (1-\lambda)\text{trace}\left(P_\pi^{m+1}\Phi\Phi^\top F_{\tau-m}\right)\right) \\
&\overset{(iii)}{\leq} (\gamma\lambda)^\tau B_\phi^2 \mathbf{1}^\top d_{\mu,0} + B_\phi^2 \sum_{m=0}^{\tau-1} (\gamma\lambda)^m \left(\lambda\|d_{\mu,\tau-m}\|_1 + (1-\lambda)\|f_{\tau-m}\|_1\right) \\
&\overset{(iv)}{\leq} (\gamma\lambda)^\tau B_\phi^2 + \frac{1-(\gamma\lambda)^\tau}{1-\gamma\lambda} \left(\lambda + (1-\lambda)(1+\|f\|_1)\right) B_\phi^2 \\
&\leq \frac{B_\phi^2}{1-\gamma\lambda} \left(1 + (1-\lambda)\|f\|_1\right),
\end{aligned} \tag{49}
$$

where (i) follows from eq. (43), (ii) follows from the facts that $\text{trace}(AB) = \text{trace}(BA)$ and $\delta_0 = \Phi^\top D_{\mu,0}$, (iii) follows from Lemma 6 with $P = P_\pi$, and (iv) follows from eq. (34).

Next, we proceed to bound the term $\text{trace}\left(\Phi^\top \delta_{\tau-1} P_{\mu,\pi}\right)$.

$$
\begin{aligned}
\delta_{b-1}(\tilde{s}) &\overset{(i)}{=} \mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[F_t^{b-1}\left(\gamma\lambda\rho_t^{b-2} e_t^{b-2} + (\lambda + (1-\lambda)F_t^{b-1})\phi(\tilde{s})\right) \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&\overset{(ii)}{=} (1-\lambda)\mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[(F_t^{b-1})^2 \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \phi(\tilde{s}) \\
&\quad + \lambda\mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[F_t^{b-1} \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \phi(\tilde{s}) \\
&\quad\quad + \gamma\lambda\mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[\rho_t^{b-2} e_t^{b-2} \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \phi(\tilde{s}) \\
&\quad\quad\quad + \gamma^2\lambda\mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \mathbb{E}\left[(\rho_t^{b-2})^2 F_t^{b-2} e_t^{b-2} \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \phi(\tilde{s}) \\
&\overset{(iii)}{=} (1-\lambda) r_{b-1}(\tilde{s})\phi(\tilde{s}) + \lambda f_{b-1}(\tilde{s})\phi(\tilde{s}) \\
&\quad + \mathbb{P}\left(s_t^{b-1} = \tilde{s} \middle| \mathcal{F}_{t-1}\right) \sum_{s'',a''} \mathbb{P}\left(s_t^{b-2} = s'', a_t^{b-2} = a'' \middle| s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \mathbb{E}\left[\gamma\lambda\rho_t^{b-2}e_t^{b-2} + \gamma^2\lambda(\rho_t^{b-2})^2F_t^{b-2}e_t^{b-2}\middle| s_t^{b-2} = s'', a_t^{b-2} = a'', s_t^{b-1} = \tilde{s}, \mathcal{F}_{t-1}\right] \\
&= (1-\lambda)r_{b-1}(\tilde{s})\phi(\tilde{s}) + \lambda f_{b-1}(\tilde{s})\phi(\tilde{s}) \\
&\quad + \mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)\sum_{s'',a''}\frac{\mathbb{P}\left(s_t^{b-2} = s''\middle|\mathcal{F}_{t-1}\right)\mu(a''|s'')\mathsf{P}(\tilde{s}|s'',a'')}{\mathbb{P}\left(s_t^{b-1} = \tilde{s}\middle|\mathcal{F}_{t-1}\right)} \\
&\quad \cdot \mathbb{E}\left[\gamma\lambda\frac{\pi(a''|s'')}{\mu(a''|s'')}e_t^{b-2} + \gamma^2\lambda\frac{\pi^2(a''|s'')}{\mu^2(a''|s'')}F_t^{b-2}e_t^{b-2}\middle| s_t^{b-2} = s'', \mathcal{F}_{t-1}\right] \\
&= (1-\lambda)r_{b-1}(\tilde{s})\phi(\tilde{s}) + \lambda f_{b-1}(\tilde{s})\phi(\tilde{s}) + \gamma\lambda(\beta_{b-2}P_\pi)_{(\cdot,s)} + \gamma^2\lambda(\delta_{b-2}P_{\mu,\pi})_{(\cdot,s)}.
\end{aligned}
$$

where (i) follows from the update of e_t^{b-1}, (ii) follows from the update rule of F_t^{b-1}, and (iii) follows from the law of total probability.

The above equality implies that

$$
\delta_{b-1} = (1-\lambda)\Phi^\top \mathrm{diag}(r_{b-1}) + \lambda\Phi^\top\mathrm{diag}(f_{b-1}) + \gamma\lambda\beta_{b-2}P_\pi + \gamma^2\lambda\delta_{b-2}P_{\pi,\mu}.
$$

Recursively applying the above equality yields

$$
\delta_{b-1} = \sum_{m=1}^{b-1}(\Phi^\top((1-\lambda)\mathrm{diag}(r_m) + \lambda\mathrm{diag}(f_m)) + \gamma\lambda\beta_{m-1}P_\pi)(P_{\pi,\mu})^{b-1-m} + (\gamma^2\lambda)^{b-1}\delta_0(P_{\pi,\mu})^{b-1}.
$$

Note that the above inequality holds for any fixed $b >= 2$. As a result, by changing of notation, for all $\tau \geq 1$, we have

$$
\delta_\tau = \sum_{m=1}^{\tau}(\Phi^\top((1-\lambda)\mathrm{diag}(r_m) + \lambda\mathrm{diag}(f_m)) + \gamma\lambda\beta_{m-1}P_\pi)(P_{\pi,\mu})^{\tau-m} + (\gamma^2\lambda)^\tau\delta_0(P_{\pi,\mu})^\tau. \tag{50}
$$

Substituting eq. (50) into $\mathrm{trace}\left(\Phi\delta_\tau P_{\mu,\pi}\right)$, we have

$$
\begin{aligned}
&\mathrm{trace}\left(\Phi\delta_\tau P_{\mu,\pi}\right) \\
&= \mathrm{trace}\Bigg(\Phi\Bigg(\sum_{m=1}^{\tau}(\Phi^\top((1-\lambda)\mathrm{diag}(r_m) + \lambda\mathrm{diag}(f_m)) + \gamma\lambda\beta_{m-1}P_\pi)(P_{\pi,\mu})^{\tau-m} \\
&\qquad\qquad + (\gamma^2\lambda)^\tau\delta_0(P_{\pi,\mu})^\tau\Bigg)P_{\mu,\pi}\Bigg) \qquad (51) \\
&\overset{(i)}{=} \sum_{m=1}^{\tau}\mathrm{trace}\left((P_{\pi,\mu})^{\tau-m+1}\left(\Phi\Phi^\top((1-\lambda)\mathrm{diag}(r_m) + \lambda\mathrm{diag}(f_m)) + \gamma\lambda\Phi\beta_{m-1}P_\pi\right)\right) \\
&\qquad + \mathrm{trace}\left((\gamma^2\lambda)^\tau(P_{\pi,\mu})^{\tau+1}\Phi\delta_0\right), \qquad (52)
\end{aligned}
$$

where (i) follows from the fact that $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ and $\mathrm{trace}(A+B) = \mathrm{trace}(A) + \mathrm{trace}(B)$.

Applying Lemma 6 with $Q = P_{\mu,\pi}$, $C = \rho_{max}$, $P = P_\pi$ and $D = (1-\lambda)\mathrm{diag}(r_m)$, we have

$$
\mathrm{trace}\left((P_{\pi,\mu})^{\tau-m+1}\Phi\Phi^\top(1-\lambda)\mathrm{diag}(r_m)\right) \leq (1-\lambda)\rho_{max}^{\tau-m+1}B_\phi^2\|r_m\|_1. \tag{53}
$$

Applying Lemma 6 with $Q = P_{\mu,\pi}$, $C = \rho_{max}$, $P = P_\pi$ and $D = \lambda\mathrm{diag}(f_m)$, we have

$$
\begin{aligned}
\mathrm{trace}\left((P_{\pi,\mu})^{\tau+1-m}\Phi\Phi^\top\lambda\mathrm{diag}(f_m)\right) &\leq \lambda\rho_{max}^{\tau+1-m}B_\phi^2\|f_m\|_1 \\
&\overset{(i)}{\leq} \lambda\rho_{max}^{\tau+1-m}B_\phi^2(1+\|f\|_1) \qquad (54)
\end{aligned}
$$

where (i) follows from eq. (34).

For the term $\mathrm{trace}\left((P_{\pi,\mu})^{\tau+1-m}\Phi\beta_{m-1}P_\pi\right)$, we have

$$
\mathrm{trace}\left((P_{\pi,\mu})^{\tau+1-m}\Phi\beta_{m-1}P_\pi\right)
$$

$$
\begin{aligned}
&\overset{(i)}{=} (\gamma\lambda)^{m-1}\text{trace}\left(P_\pi^m (P_{\pi,\mu})^{\tau+1-m}\Phi\Phi^\top D_{\mu,0}\right) \\
&\quad + \sum_{l=0}^{m-2}(\gamma\lambda)^l\left(\lambda\text{trace}\left(P_\pi^{l+1}(P_{\pi,\mu})^{\tau+1-m}\Phi\Phi^\top D_{\mu,m-1-l}\right)\right. \\
&\quad \left. +(1-\lambda)\text{trace}\left(P_\pi^{l+1}(P_{\pi,\mu})^{\tau-m+1}\Phi\Phi^\top F_{\pi,m-1-l}\right)\right) \\
&\overset{(ii)}{\leq} B_\phi^2\rho_{max}^{\tau+1-m}\left((\gamma\lambda)^{m-1}\mathbf{1}^\top d_{\mu,0} + \sum_{l=0}^{m-2}(\gamma\lambda)^l\left(\lambda\mathbf{1}^\top d_{\mu,m-1-l} + (1-\lambda)\mathbf{1}^\top f_{m-1-l}\right)\right) \\
&\overset{(iii)}{\leq} \frac{B_\phi^2\rho_{max}^{\tau+1-m}}{1-\gamma\lambda}\left(1+(1-\lambda)\|f\|_1\right),
\end{aligned} \tag{55}
$$

where (i) follows from eq. (43) and the facts that $\text{trace}(A+B) = \text{trace}(A) + \text{trace}(B)$ and $\text{trace}(AB) = \text{trace}(BA)$, (ii) follows from Lemma 6 with $Q = P_{\mu,\pi}$, $P = P_\pi$, and $D = F_{m-1-l}$ and $D_{\pi,m-1-l}$ respectively, and (iii) follow from the eq. (34).

Recall $\delta_0 = \Phi^\top D_{\mu,0}$. Applying Lemma 6 with $Q = P_{\mu,\pi}$ and $D = D_{\mu,0}$ yields

$$
\text{trace}\left((\gamma^2\lambda)^\tau (P_{\pi,\mu})^{\tau+1}\Phi\delta_0\right) \leq (\gamma^2\lambda)^\tau B_\phi^2\rho_{max}^{\tau+1}. \tag{56}
$$

Substituting eqs. (53) to (56) into eq. (52), we have

$$
\begin{aligned}
&\text{trace}\left(\Phi\delta_\tau P_{\mu,\pi}\right) \\
&\quad \leq (\gamma^2\lambda)^\tau B_\phi^2\rho_{max}^{\tau+1} \\
&\qquad + \sum_{m=1}^{\tau} B_\phi^2\rho_{max}^{\tau-m+1}\left(\lambda(\|f\|_1+1) + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right) + \sum_{m=1}^{\tau}(1-\lambda)\rho_{max}^{\tau-m+1}B_\phi^2\|r_m\|_1.
\end{aligned} \tag{57}
$$

Substituting eqs. (49) and (57) into eq. (48), we have

$$
\begin{aligned}
\mathbf{1}^\top\Delta_b \leq\ & \lambda^{2b}\gamma^{2b}\rho_{max}^b B_\phi^2 + \left(1+2\gamma(1-\lambda)(1+\|f\|_1) + 2\gamma\lambda C_\beta\right)B_\phi^2\sum_{\tau=1}^{b}\lambda^{2(b-\tau)}\gamma^{2(b-\tau)}\rho_{max}^{b-\tau} \\
&+ \gamma^2(1-\lambda)^2\rho_{max}^{b+1}B_\phi^2\sum_{\tau=1}^{b}\lambda^{2b-2\tau}\gamma^{2b-2\tau}\rho_{max}^{-\tau}\|r_{\tau-1}\|_1 + 2\lambda^b\gamma^{2b}(1-\lambda)\rho_{max}^b B_\phi^2\sum_{\tau=1}^{b}\lambda^{b-\tau} \\
&+ 2\gamma^2\lambda(1-\lambda)B_\phi^2\left(\lambda\|f\|_1 + 1 + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^b\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau-1}\rho_{max}^{-m} \\
&+ 2\gamma^2\lambda(1-\lambda)^2\rho_{max}^b B_\phi^2\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau-1}\rho_{max}^{-m}\|r_m\|_1,
\end{aligned} \tag{58}
$$

where we let $C_\beta \coloneqq \frac{B_\phi^2}{1-\gamma\lambda}\left(1+(1-\lambda)\|f\|_1\right)$.

Under different conditions of ρ_{max}, the term $\mathbf{1}^\top\Delta_b$ and $\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2\middle|\mathcal{F}_{t-1}\right]$ can be upper bounded differently as following:

(a) $\rho_{max} = 1$, $\gamma^2\rho_{max} = \gamma^2 < 1$. Proceed with eq. (58). We have

$$
\begin{aligned}
\mathbf{1}^\top\Delta_b \overset{(i)}{\leq}\ & \lambda^{2b}\gamma^{2b}B_\phi^2 + \left(1+2\gamma(1-\lambda)(1+\|f\|_1) + 2\gamma\lambda C_\beta\right)\frac{B_\phi^2}{1-\gamma^2\lambda^2} \\
&+ \frac{\gamma^2(1-\lambda)^2B_\phi^2}{1-\gamma^2\lambda^2}\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2} + 1\right) + 2\lambda^b\gamma^{2b}B_\phi^2 \\
&+ 2\gamma^2\lambda(1-\lambda)B_\phi^2\left(\lambda\|f\|_1 + 1 + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\tau
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\gamma^2\lambda(1-\lambda)^2\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right)B_\phi^2\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\tau \\
&\overset{(ii)}{\leq} \lambda^{2b}\gamma^{2b}B_\phi^2 + \left(1+2\gamma(1-\lambda)(1+\|f\|_1)+2\gamma\lambda C_\beta\right)\frac{B_\phi^2}{1-\gamma^2\lambda^2} \\
&\quad + \frac{\gamma^2(1-\lambda)^2B_\phi^2}{1-\gamma^2\lambda^2}\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right) + 2\lambda^b\gamma^{2b}B_\phi^2 \\
&\quad + 2\gamma^2\lambda(1-\lambda)B_\phi^2\left(\lambda\|f\|_1+1+\frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\frac{b}{1-\gamma^2\lambda^2} \\
&\quad + 2\gamma^2\lambda(1-\lambda)^2B_\phi^2\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right)\frac{b}{1-\gamma^2\lambda^2}. \qquad (59)
\end{aligned}
$$

where (i) follows from eq. (37) and (ii) follows from the following fact,

$$
\sum_{\tau=1}^{b}(\gamma\lambda)^{2(b-\tau)}\tau = \frac{b}{1-\gamma^2\lambda^2} - \frac{\gamma^2\lambda^2-(\gamma\lambda)^{2(b+1)}}{(1-\gamma\lambda)^2} \leq \frac{b}{1-\gamma^2\lambda^2}.
$$

Substituting eq. (59) into eq. (46) yields

$$
\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2\middle|\mathcal{F}_{t-1}\right] \leq C_{\sigma,\lambda,1}b, \qquad (60)
$$

where $C_{\sigma,\lambda,1} > 0$ is a constant and be determined by eq. (59).

(b) $\rho_{max} > 1$ and $\gamma^2\rho_{max} < 1$, substituting eq. (37) into eq. (58) yields

$$
\begin{aligned}
\mathbf{1}^\top\Delta_b &\leq \lambda^{2b}\gamma^{2b}\rho_{max}^bB_\phi^2 + \left(1+2\gamma(1-\lambda)(1+\|f\|_1)+2\gamma\lambda C_\beta\right)B_\phi^2\sum_{\tau=1}^{b}\lambda^{2(b-\tau)}\gamma^{2(b-\tau)}\rho_{max}^{b-\tau} \\
&\quad + \gamma^2(1-\lambda)^2\rho_{max}^{b+1}B_\phi^2\sum_{\tau=1}^{b}\lambda^{2b-2\tau}\gamma^{2b-2\tau}\rho_{max}^{-\tau}\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right) \\
&\quad + 2\lambda^b\gamma^{2b}(1-\lambda)\rho_{max}^bB_\phi^2\sum_{\tau=1}^{b}\lambda^{b-\tau} \\
&\quad + 2\gamma^2\lambda(1-\lambda)B_\phi^2\left(\lambda\|f\|_1+1+\frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^b \\
&\qquad \cdot\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau-1}\rho_{max}^{-m} \\
&\quad + 2\gamma^2\lambda(1-\lambda)^2\rho_{max}^bB_\phi^2\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau-1}\rho_{max}^{-m}\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right) \\
&\leq \lambda^{2b}\gamma^{2b}\rho_{max}^bB_\phi^2 + \frac{\left(1+2\gamma(1-\lambda)(1+\|f\|_1)+2\gamma\lambda C_\beta\right)B_\phi^2}{1-\gamma^2\rho_{max}\lambda} \\
&\quad + \frac{\gamma^2(1-\lambda)^2\rho_{max}B_\phi^2}{1-\gamma^2\rho_{max}\lambda}\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right) + 2\lambda^{2b}\gamma^b\rho_{max}^bB_\phi^2 \\
&\quad + \frac{2\gamma^2\lambda(1-\lambda)B_\phi^2}{(1-\gamma^2\lambda^2)(1-\rho_{max}^{-1})}\left(\lambda\|f\|_1+1+\frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^{b+1} \\
&\quad + \frac{2\gamma^2\lambda(1-\lambda)^2B_\phi^2}{(1-\gamma^2\lambda^2)(1-\rho_{max}^{-1})}\left(\frac{3+2\gamma\|f\|_1}{1-\gamma^2}+1\right)\rho_{max}^{b+1}, \qquad (61)
\end{aligned}
$$

where the last two terms of the above inequality are of the order $\mathcal{O}\left(\rho_{max}^b\right)$. Therefore, we have $\mathbf{1}^\top\Delta_b \leq C\rho_{max}^b$ for some $C > 0$. Substituting eq. (61) into eq. (46) yields

$$
\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2\middle|\mathcal{F}_{t-1}\right] \leq C_{\sigma,\lambda,2}\rho_{max}^b,
$$

where $C_{\sigma,\lambda,2} > 0$ is a constant and is determined by eq. (61).

(c) $\gamma^2\rho_{max} = 1$, substituting eq. (36) into eq. (58) yields

$$\begin{aligned}
\mathbf{1}^\top\Delta_b \leq\ & \lambda^{2b}B_\phi^2 + \left(1 + 2\gamma(1-\lambda)(1+\|f\|_1) + 2\gamma\lambda C_\beta\right)B_\phi^2\sum_{\tau=1}^{b}\lambda^{2(b-\tau)} \\
& + \gamma^2(1-\lambda)^2\rho_{max}B_\phi^2\sum_{\tau=1}^{b}\lambda^{2b-2\tau}\left(4+2\gamma\|f\|_1\right)\tau + 2\lambda^b(1-\lambda)B_\phi^2\sum_{\tau=1}^{b}\lambda^{b-\tau} \\
& + 2\gamma^2\lambda(1-\lambda)B_\phi^2\left(\lambda\|f\|_1 + 1 + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^b\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau}\rho_{max}^{-m} \\
& + 2\gamma^2\lambda(1-\lambda)^2\left(4+2\gamma\|f\|_1\right)B_\phi^2\sum_{\tau=1}^{b}\lambda^{2b-2\tau}\sum_{m=1}^{\tau}\rho_{max}^{\tau-m}m. \\
\leq\ & \lambda^{2b}B_\phi^2 + \frac{\left(1+2\gamma(1-\lambda)(1+\|f\|_1)+2\gamma\lambda C_\beta\right)B_\phi^2}{1-\lambda^2} + 2\lambda^b B_\phi^2 \\
& + \left(\frac{\gamma^2(1-\lambda)^2\rho_{max}B_\phi^2}{1-\lambda^2}\left(4+2\gamma\|f\|_1\right) + \frac{2\gamma^2\lambda(1-\lambda)^2\left(4+2\gamma\|f\|_1\right)B_\phi^2}{(1-\lambda^2)(1-\rho_{max}^{-1})}\right)b \\
& + \frac{2\gamma^2\lambda(1-\lambda)B_\phi^2}{(1-\gamma^2\lambda^2)(1-\rho_{max}^{-1})}\left(\lambda\|f\|_1 + 1 + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^{b+1}, \qquad (62)
\end{aligned}$$

where the last term of the above inequality is of the order $\mathcal{O}\left(\rho_{max}^b\right)$. Therefore, we have $\mathbf{1}^\top\Delta_b \leq C\rho_{max}^b$ for some $C > 0$. Substituting eq. (62) into eq. (46) yields

$$\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2\middle|\mathcal{F}_{t-1}\right] \leq C_{\sigma,\lambda,3}\rho_{max}^b,$$

where $C_{\sigma,\lambda,3} > 0$ is a constant and is determined by eq. (62).

(d) $\gamma^2\rho_{max} > 1$, substituting eq. (35) into eq. (58) yields

$$\begin{aligned}
\mathbf{1}^\top\Delta_b \leq\ & \lambda^{2b}\gamma^{2b}\rho_{max}^b B_\phi^2 + \left(1+2\gamma(1-\lambda)(1+\|f\|_1)+2\gamma\lambda C_\beta\right)B_\phi^2\sum_{\tau=1}^{b}\lambda^{2(b-\tau)}\gamma^{2(b-\tau)}\rho_{max}^{b-\tau} \\
& + \gamma^{2b+2}(1-\lambda)^2\rho_{max}^{b+1}B_\phi^2\left(\frac{3+2\gamma\|f\|_1}{\gamma^2\rho_{max}-1}+1\right)\sum_{\tau=1}^{b}\lambda^{2b-2\tau} + 2\lambda^b\gamma^{2b}(1-\lambda)\rho_{max}^b B_\phi^2\sum_{\tau=1}^{b}\lambda^{b-\tau} \\
& + 2\gamma^2\lambda(1-\lambda)B_\phi^2\left(\lambda\|f\|_1 + 1 + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^b\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau-1}\rho_{max}^{-m} \\
& + 2\gamma^2\lambda(1-\lambda)^2\rho_{max}^b B_\phi^2\left(\frac{3+2\gamma\|f\|_1}{\gamma^2\rho_{max}-1}+1\right)\sum_{\tau=1}^{b}\gamma^{2b-2\tau}\lambda^{2b-2\tau}\sum_{m=1}^{\tau-1}\gamma^{2m} \\
\leq\ & \lambda^{2b}\gamma^{2b}\rho_{max}^b B_\phi^2 + \frac{\left(1+2\gamma(1-\lambda)(1+\|f\|_1)+2\gamma\lambda C_\beta\right)B_\phi^2}{\gamma^2\rho_{max}-1}\cdot\gamma^{2b}\rho_{max}^b \\
& + \frac{\gamma^{2b+2}(1-\lambda)^2\rho_{max}^{b+1}B_\phi^2}{1-\lambda^2}\left(\frac{3+2\gamma\|f\|_1}{\gamma^2\rho_{max}-1}+1\right) + 2\lambda^b\gamma^{2b}\rho_{max}^b B_\phi^2 \\
& + \frac{2\gamma^2\lambda(1-\lambda)B_\phi^2}{(1-\gamma^2\lambda^2)(1-\rho_{max}^{-1})}\left(\lambda\|f\|_1 + 1 + \frac{1+(1-\lambda)\|f\|_1}{1-\gamma\lambda}\right)\rho_{max}^{b+1} \\
& + \frac{2\gamma^2\lambda(1-\lambda)^2}{(1-\gamma^2)(1-\gamma^2\lambda^2)}B_\phi^2\left(\frac{3+2\gamma\|f\|_1}{\gamma^2\rho_{max}-1}+1\right)\rho_{max}^b, \qquad (63)
\end{aligned}$$

where the last term of the above inequality is of the order $\mathcal{O}\left(\rho_{max}^b\right)$. Therefore, we have $\mathbf{1}^\top\Delta_b \leq C\rho_{max}^b$ for some $C > 0$. Substituting eq. (63) into eq. (46) yields

$$\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^\lambda(\theta_t)\right\|_2^2\middle|\mathcal{F}_{t-1}\right] \leq C_{\sigma,\lambda,4}\rho_{max}^b,$$

where $C_{\sigma,\lambda,4} > 0$ is a constant and is determined by eq. (63).

To summarize, the variance term $\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right]$ can be bounded by

$$\mathbb{E}\left[\left\|\widehat{\mathcal{T}}_t^{\lambda}(\theta_t)\right\|_2^2 \middle| \mathcal{F}_{t-1}\right] \leq \sigma_\lambda^2,$$

where

$$\sigma_\lambda^2 = \begin{cases} \mathcal{O}(b), & \textit{if } \rho_{max} = 1. \\ \mathcal{O}(\rho_{max}^b), & \textit{if } \rho_{max} > 1. \end{cases}$$

Tengyu: I have a question. When $\rho_{\max} = 1$, we would have $\pi = \pi_b$, right? If that is the case then it is simply on-policy setting, thus there is no point to discuss this case here.

C Proof of the Main Theorems

C.1 Proof of Theorem 1

Note the θ update specified in Algorithm 1 is the closed form solution of the following maximization problem.

$$\theta_{t+1} = \operatorname*{argmax}_{\theta\in\Theta} \eta_t \left\langle \widehat{\mathcal{T}}_t(\theta_t), \theta \right\rangle + \frac{1}{2}\|\theta - \theta_t\|_2^2.$$

Applying Lemma 9 with $\theta^* = \theta_{t+1}$, $\eta = \eta_t$, $G = \widehat{\mathcal{T}}_t(\theta_t)$, and $\theta_0 = \theta_t$ yields, for any $\theta \in \Theta$,

$$\eta_t \left\langle \widehat{\mathcal{T}}_t(\theta_t), \theta_{t+1} - \theta \right\rangle + \frac{1}{2}\|\theta_t - \theta_{t+1}\|_2^2 \leq \frac{1}{2}\|\theta_t - \theta\|_2^2 - \frac{1}{2}\|\theta_{t+1} - \theta\|_2^2. \tag{64}$$

Proceed with the first term in the above inequality as follows

$$\begin{aligned}
&\left\langle \widehat{\mathcal{T}}_t(\theta_t), \theta_{t+1} - \theta \right\rangle \\
&= \langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta\rangle + \langle \mathcal{T}(\theta_t) - \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta\rangle + \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_{t+1} - \theta \right\rangle \\
&\overset{(i)}{\geq} \langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta\rangle - L_0\|\theta_t - \theta_{t+1}\|_2\|\theta_{t+1} - \theta\|_2 + \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_{t+1} - \theta \right\rangle \\
&= \langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta\rangle - L_0\|\theta_t - \theta_{t+1}\|_2\|\theta_{t+1} - \theta\|_2 + \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_{t+1} - \theta_t \right\rangle \\
&\qquad + \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_t - \theta \right\rangle \\
&\geq \langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta\rangle - L_0\|\theta_t - \theta_{t+1}\|_2\|\theta_{t+1} - \theta\|_2 - \left\|\widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t)\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2 \\
&\qquad + \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_t - \theta \right\rangle,
\end{aligned}$$

where (i) follows from the Cauchy-Schwartz inequality and Lemma 8.

Substituting the above inequality into eq. (64) yields

$$\begin{aligned}
&\eta_t \langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta\rangle - \eta_t L_0\|\theta_t - \theta_{t+1}\|_2\|\theta_{t+1} - \theta\|_2 - \eta_t \left\|\widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t)\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2 \\
&\quad + \eta_t \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_t - \theta \right\rangle + \frac{1}{2}\|\theta_t - \theta_{t+1}\|_2^2 \\
&\qquad \leq \frac{1}{2}\|\theta_t - \theta\|_2^2 - \frac{1}{2}\|\theta_{t+1} - \theta\|_2^2.
\end{aligned} \tag{65}$$

Applying Young's inequality to $\eta_t \left\|\widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t)\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2$ yields

$$\eta \left\|\widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t)\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2 \leq \frac{1}{4}\|\theta_{t+1} - \theta_t\|_2^2 + \eta_t^2 \left\|\widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t)\right\|_2^2,$$

and applying Young's inequality to $\eta_t L_0 \|\theta_t - \theta_{t+1}\|_2 \|\theta_{t+1} - \theta\|_2$ yields

$$\eta_t L_0 \|\theta_t - \theta_{t+1}\|_2 \|\theta_{t+1} - \theta\|_2 \leq \frac{1}{4}\|\theta_t - \theta_{t+1}\|_2^2 + \eta_t^2 L_0^2 \|\theta_{t+1} - \theta\|_2^2.$$

Substituting the above two inequalities into eq. (65) yields

$$\begin{aligned}\frac{1}{2}\|\theta_t - \theta\|_2^2 \geq \eta_t \left\langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta \right\rangle + \left(\frac{1}{2} - \eta_t^2 L_0^2\right) \|\theta_{t+1} - \theta\|_2^2 \\ + \eta_t \left\langle \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t), \theta_t - \theta \right\rangle - \eta_t^2 \left\| \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t) \right\|_2^2.\end{aligned}$$

Taking expectation conditioned on $\mathcal{F}_{t-1}$ on the both sides of the above inequality, we obtain

$$\begin{aligned}\frac{1}{2}\|\theta_t - \theta\|_2^2 \geq \eta_t \mathbb{E}\left[\langle \mathcal{T}(\theta_{t+1}), \theta_{t+1} - \theta \rangle | \mathcal{F}_{t-1}\right] + \left(\frac{1}{2} - \eta_t^2 L_0^2\right) \mathbb{E}\left[\|\theta_{t+1} - \theta\|_2^2 \middle| \mathcal{F}_{t-1}\right] \\ + \eta_t \left\langle \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t) \middle| \mathcal{F}_{t-1}\right], \theta_t - \theta \right\rangle - \eta_t^2 \mathbb{E}\left[\left\| \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t) \right\|_2^2 \middle| \mathcal{F}_{t-1}\right].\end{aligned} \tag{66}$$

Letting $\theta = \theta^*$ and applying Lemma 7 to eq. (66) yields

$$\begin{aligned}\frac{1}{2}\|\theta_t - \theta^*\|_2^2 &\geq \left(\frac{1}{2} + \mu_0 \eta_t - \eta_t^2 L_0^2\right) \mathbb{E}\left[\|\theta_{t+1} - \theta^*\|_2^2 \middle| \mathcal{F}_{t-1}\right] \\ &\quad - \eta_t C_b B_\phi \xi^b \|\theta_t - \theta^*\|_2^2 - \eta_t C_b \epsilon_{approx} \xi^b \|\theta_t - \theta^*\|_2 \\ &\quad - \eta_t^2 \mathbb{E}\left[\left\| \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t) \right\|_2^2 \middle| \mathcal{F}_{t-1}\right] \\ &\overset{(i)}{\geq} \left(\frac{1}{2} + \mu_0 \eta_t - \eta_t^2 L_0^2\right) \mathbb{E}\left[\|\theta_{t+1} - \theta^*\|_2^2 \middle| \mathcal{F}_{t-1}\right] \\ &\quad - \eta_t C_b B_\phi \xi^b \|\theta_t - \theta^*\|_2^2 - \eta_t C_b \epsilon_{approx} \xi^b \|\theta_t - \theta^*\|_2 - 4\eta_t^2 C_b^2 B_\phi^2 \xi^{2b} \|\theta_t - \theta^*\|_2^2 \\ &\quad - 4\eta_t^2 C_b^2 \xi^{2b} \epsilon_{approx}^2 - 2\sigma^2 \eta_t^2,\end{aligned}$$

where (i) follows from Propositions 1 and 2, and the facts that $(x+y)^2 \leq 2x^2 + 2y^2$ and

$$\begin{aligned}\left\| \widehat{\mathcal{T}}_t(\theta_t) - \mathcal{T}(\theta_t) \right\|_2^2 &\leq 2 \left\| \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t) \middle| \mathcal{F}_{t-1}\right] - \mathcal{T}(\theta_t) \right\|_2^2 + 2 \left\| \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t) \middle| \mathcal{F}_{t-1}\right] - \widehat{\mathcal{T}}_t(\theta_t) \right\|_2^2 \\ &\leq 2 \left\| \widehat{\mathcal{T}}(\theta_t) \right\|_2^2 + 2 \left\| \mathbb{E}\left[\widehat{\mathcal{T}}_t(\theta_t) \middle| \mathcal{F}_{t-1}\right] - \widehat{\mathcal{T}}_t(\theta_t) \right\|_2^2.\end{aligned}$$

Taking expectation on both sides of the above inequality yields

$$\begin{aligned}&\left(\frac{1}{2} + \mu_0 \eta_t - \eta_t^2 L_0^2\right) \mathbb{E}\left[\|\theta_{t+1} - \theta^*\|_2^2\right] \\ &\leq \left(\frac{1}{2} + C_b B_\phi \xi^b \eta_t + 4 C_b^2 B_\phi^2 \xi^{2b} \eta_t^2\right) \mathbb{E}\left[\|\theta_t - \theta^*\|_2\right] + \eta_t C_b B_\theta \epsilon_{approx} \xi^b \\ &\quad + 4\eta_t^2 C_b^2 \xi^{2b} \epsilon_{approx}^2 + 2\sigma^2 \eta_t^2.\end{aligned}$$

Recall that we set $t_0 = \frac{8L_0^2}{\mu_0^2}$. Let $\alpha_t = (t + t_0 + 1)(t + t_0 + 2)$. Multiplying $2\alpha_t$ on both sides of the above inequality and telescoping from $t = 0, 1, 2, \ldots, T-1$ yields

$$\begin{aligned}&\sum_{t=0}^{T-1} \alpha_t \left(1 + 2\mu_0 \eta_t - 2\eta_t^2 L_0^2\right) \mathbb{E}\left[\|\theta_{t+1} - \theta^*\|_2^2\right] \\ &\leq \sum_{t=0}^{T-1} \alpha_t \left(1 + 2C_b B_\phi \xi^b \eta_t + 8 C_b^2 B_\phi^2 \xi^{2b} \eta_t^2\right) \mathbb{E}\left[\|\theta_t - \theta^*\|_2\right] \\ &\quad + \left(4\sigma^2 + 8C_b^2 \xi^{2b} \epsilon_{approx}^2\right) \sum_{t=0}^{T-1} \alpha_t \eta_t^2 + 2C_b B_\theta \epsilon_{approx} \xi^b \sum_{t=0}^{T-1} \alpha_t \eta_t.\end{aligned} \tag{67}$$

Recall the setting of η_t, we have

$$\begin{aligned}
1+2\mu_0\eta_t-2\eta_t^2L_0^2 &= 1+\frac{3\mu_0\eta_t}{2}\left(\frac{4}{3}-\frac{4}{3\mu_0}\eta_tL_0^2\right)\\
&= 1+\frac{3\mu_0\eta_t}{2}\left(1+\frac{1}{3}\left(1-\frac{4}{\mu_0}\eta_tL_0^2\right)\right)\\
&\overset{(i)}{\geq} 1+\frac{3\mu_0\eta_t}{2},
\end{aligned}$$

where (i) follows from the fact that $1-\frac{4\eta_tL_0^2}{\mu_0}\geq 1-\frac{8L_0^2}{\mu_0^2t_0}\geq 0$. Multiplying α_t on both sides of the above inequality yields

$$\begin{aligned}
\alpha_t(1+2\mu_0\eta_t-2\eta_t^2L_0^2) &\geq (t+t_0+1)(t+t_0+2)\left(1+\frac{3\mu_0}{2}\frac{2}{\mu_0(t+t_0)}\right)\\
&= (t+t_0+1)(t+t_0+2)(t+t_0+3)/(t+t_0). \qquad (68)
\end{aligned}$$

Under appropriate value of b, we have $C_bB_\phi\xi^b\leq\frac{\mu_0}{5}$. Which implies that

$$\begin{aligned}
2C_bB_\phi\xi^b\eta_t+8C_b^2B_\phi^2\xi^{2b}\eta_t^2 &= \frac{\mu_0\eta_t}{2}+\frac{\mu_0\eta_t}{2}\left(\frac{4C_bB_\phi\xi^b}{\mu_0}+\frac{16C_b^2B_\phi^2\xi^{2b}\eta_t}{\mu_0}-1\right)\\
&\leq \frac{\mu_0\eta_t}{2}+\frac{\mu_0\eta_t}{2}\left(\frac{4}{5}+\frac{16\mu_0\eta_t}{25}-1\right)\\
&\leq \frac{\mu_0\eta_t}{2}+\frac{\mu_0\eta_t}{2}\left(\frac{4\mu_0^2}{25L_0^2}-\frac{1}{5}\right)\\
&\leq \frac{\mu_0\eta_t}{2}.
\end{aligned}$$

Multiplying α_{t+1} on both sides of the above inequality yields

$$\begin{aligned}
\alpha_{t+1}(1+2C_bB_\phi\xi^b\eta_t+8C_b^2B_\phi^2\xi^{2b}\eta_t^2) &\leq \alpha_{t+1}\left(1+\frac{\mu_0\eta_{t+1}}{2}\right)\\
&\leq \alpha_{t+1}\left(1+\frac{\mu_0\eta_{t+1}}{2}\right)\\
&= (t+t_0+2)(t+t_0+3)\left(1+\frac{\mu_0}{2}\frac{2}{\mu_0(t+t_0+1)}\right)\\
&= (t+t_0+2)^2(t+t_0+3)/(t+t_0+1). \qquad (69)
\end{aligned}$$

Equations (68) and (69) together imply that

$$\begin{aligned}
&\alpha_t(1+2\mu\eta_t-2\eta_t^2L_0^2)-\alpha_{t+1}(1+2C_bB_\phi\xi^b\eta_t+8C_b^2B_\phi^2\xi^{2b}\eta_t^2)\\
&\quad\geq \frac{(t+t_0+1)(t+t_0+2)(t+t_0+3)}{t+t_0}-\frac{(t+t_0+2)^2(t+t_0+3)}{t+t_0+1}\\
&\quad= \frac{(t+t_0+2)(t+t_0+3)}{(t+t_0)(t+t_0+1)}\left((t+t_0+1)^2-(t+t_0)(t+t_0+3)\right)\\
&\quad= \frac{4(t+t_0+2)(t+t_0+3)}{(t+t_0)(t+t_0+1)}\\
&\quad> 0.
\end{aligned}$$

The above inequality shows that the $\|\theta_t-\theta^*\|_2^2$, $t=1,\ldots,T-1$, terms on both sides of eq. (67) can be canceled, which indicates the following

$$\begin{aligned}
&(T+t_0)(T+t_0+1)\left(1+2\mu_0\eta_{T-1}-2\eta_{T-1}^2L_0^2\right)\mathbb{E}\left[\|\theta_T-\theta^*\|_2^2\right]\\
&\quad\leq (t_0+1)(t_0+2)\left(1+2C_bB_\phi\xi^b\eta_0+8C_b^2B_\phi^2\xi^{2b}\eta_0^2\right)\|\theta_0-\theta^*\|_2^2\\
&\qquad+\left(4\sigma^2+8C_b^2\xi^{2b}\epsilon_{approx}^2\right)\sum_{t=0}^{T-1}\alpha_t\eta_t^2+2C_bB_\theta\epsilon_{approx}\xi^b\sum_{t=0}^{T-1}\alpha_t\eta_t. \qquad (70)
\end{aligned}$$

Note that

$$\sum_{t=0}^{T-1} \alpha_t \eta_t^2 \leq \sum_{t=0}^{T-1} \frac{6}{\mu_0^2} \leq \frac{6T}{\mu_0^2},$$
$$1 + 2\mu_0 \eta_{T-1} - 2\eta_{T-1}^2 L_0^2 \geq 1,$$

and

$$\sum_{t=0}^{T-1} \alpha_t \eta_t \leq \frac{4}{\mu_0} \sum_{t=0}^{T-1} (t + t_0 + 2) \leq \frac{2}{\mu_0} (T + t_0 + 2)^2.$$

Dividing $(T + t_0)(T + t_0 + 1)\left(1 + 2\mu_0 \eta_{T-1} - 2\eta_{T-1}^2 L_0^2\right)$ on both sides of eq. (70) yields

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_T - \theta^*\|_2^2\right] \\
&\quad \leq \frac{(t_0+1)(t_0+2)}{(T+t_0)(T+t_0+1)} \left(1 + \frac{\mu_0 \eta_0}{2}\right) \|\theta_0 - \theta^*\|_2^2 \\
&\qquad + \frac{24\sigma^2 + 48 C_b^2 \xi^{2b} \epsilon_{approx}^2}{\mu_0^2} \frac{1}{T + t_0 + 1} + \frac{4 C_b B_\theta \epsilon_{approx} \xi^b}{\mu_0} \frac{(T + t_0 + 2)^2}{(T + t_0 + 1)(T + t_0)} \\
&\quad = \mathcal{O}\left(\frac{\|\theta_0 - \theta^2\|_2^2}{T^2}\right) + \mathcal{O}\left(\frac{\sigma^2}{T}\right) + \mathcal{O}\left(\frac{C_b^2 \xi^{2b}}{T}\right) + \mathcal{O}\left(C_b \xi^b\right).
\end{aligned}$$

C.2 Proof of Theorem 2

The proof follows the same steps as Theorem 1 by replaces the terms $\widehat{\mathcal{T}}_t$, $\mathcal{T}$, t_0, L_0, μ_0, C_b, σ^2 and θ^* by $\widehat{\mathcal{T}}_t^\lambda$, $\mathcal{T}^\lambda$, t_λ, L_λ, μ_λ, $C_{b,\lambda}$, σ_λ^2 and θ_λ^* respectively.