Recurrent Natural Policy Gradient for POMDPs

Semih Cayci [1] **Atilla Eryilmaz** [2]

Abstract

In this paper, we study a natural policy gradient method based on recurrent neural networks (RNNs) for partially-observable Markov decision processes, whereby RNNs are used for policy parameterization and policy evaluation to address curse of dimensionality in non-Markovian reinforcement learning. We present finite-time and finite-width analyses for both the critic (recurrent temporal difference learning), and correspondingly-operated recurrent natural policy gradient method in the near-initialization regime. Our analysis demonstrates the efficiency of RNNs for problems with short-term memory with explicit bounds on the required network widths and sample complexity, and points out the challenges in the case of long-term dependencies.

1. Introduction

Reinforcement learning for partially-observable Markov decision processes (POMDPs) has been a particularly challenging problem due to the absence of an optimal stationary policy, which leads to a curse of dimensionality as the space of non-stationary policies grows exponentially over time (Krishnamurthy, 2016; Murphy, 2000). There has been a growing interest in finite-memory policies to address the curse of dimensionality in reinforcement learning for POMDPs (Yu & Bertsekas, 2008; Yu, 2012; Kara & Yüksel, 2023; Cayci et al., 2022). Among these, recurrent neural networks (RNNs) have been shown to achieve impressive *empirical* success in solving POMDPs (Whitehead & Lin, 1995; Wierstra et al., 2010; Mnih et al., 2014). However, theoretical understanding of RNN-based RL methods for POMDPs is still in a nascent stage.

In this paper, we aim to remedy this by studying a model-free policy optimization method based on a recurrent natural actor-critic (Rec-NAC) framework (Section 5), which

- utilizes an RNN-based policy parameterization for efficient history representation in non-stationary policies,
- incorporates an RNN-based temporal difference learning (Rec-TD) algorithm as the critic (Section 6), and
- performs policy updates by using RNN-based natural policy gradient (Rec-NPG) as the actor (Section 7),

for large POMDPs. We establish non-asymptotic (finite-time, finite-width) analyses of Rec-TD (in Theorem 6.3) and Rec-NPG (Theorem 7.3 and Propositions 7.6-7.8), and prove their near-optimality in the large-network limit for problems that require short-term memory. We identify pathological cases that cause exponentially growing iteration complexity and network size (Remarks 6.5-7.4). Our analysis reveals an interesting connection between (i) the memory (i.e., long-term dependencies) in the POMDP, (ii) continuity and smoothness of the parameters of the RNN, and (iii) global near-optimality of the Rec-NPG in terms of the required network size and iterations.

1.1. Previous work

Natural policy gradient method, proposed in (Kakade, 2001), has been extensively investigated for MDPs (Agarwal et al., 2020; Cen et al., 2020; Khodadadian et al., 2021), and analyses of NPG with feedforward neural networks (FNNs) have been established in (Wang et al., 2019; Liu et al., 2019; Cayci et al., 2024). As these works consider MDPs, the policies are stationary. In our case, the analysis of RNNs and POMDPs constitute a very significant challenge.

In (Yu, 2012; Singh et al., 1994; Kara & Yüksel, 2023; Cayci et al., 2022), finite-memory policies based on sliding-window approximations of the history were investigated. Alternatively, value- and policy-based model-free approaches based on RNNs have been widely considered in the literature to solve POMDPs (Lin & Mitchell, 1993; Whitehead & Lin, 1995; Wierstra et al., 2010; Mnih et al., 2014). However, these works are predominantly experimental, thus there is no theoretical analysis of RNN-based RL methods for POMDPs to the best of our knowledge. In this work, we also present theoretical guarantees for RNN-based NPG for POMDPs. For structural results on the hardness of RL for

[1] Department of Mathematics, RWTH Aachen University, Aachen, Germany [2] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH. Correspondence to: Semih Cayci <cayci@mathc.rwth-aachen.de>.

Workshop on Foundations of Reinforcement Learning and Control at the 41^{st} International Conference on Machine Learning, Vienna, Austria. Copyright 2024 by the author(s).

POMDPs, refer to (Liu et al., 2022; Singh et al., 1994).

1.2. Notation

For a vector $\Theta = (\Theta_1^\top, \dots, \Theta_m^\top)^\top \in \mathbb{R}^{m\cdot(d+1)}$, $m, d \in \mathbb{Z}_+$ with $\Theta_i = (V_i, U_i^\top)^\top \in \mathbb{R}^{d+1}$ for $V_i \in \mathbb{R}, U_i \in \mathbb{R}^d$ and $\rho = (\rho_1, \rho_2) \in \mathbb{R}^2_{\geq 0}$, we define $\mathcal{B}^{(m)}_{2,\infty}(\Theta, \rho) := \bigotimes_{i=1}^m \left(\mathcal{B}_1^{(1)}\left(V_i, \frac{\rho_1}{\sqrt{m}}\right), \mathcal{B}_2^{(d)}\left(U_i, \frac{\rho_2}{\sqrt{m}}\right)\right)$, where $\bigotimes$ is the Cartesian product, and $\mathcal{B}_p^{(d)}(x, \rho_0) := \{z \in \mathbb{R}^d : \|z - x\|_p \leq \rho_0\}$ for any $p \geq 1, x \in \mathbb{R}^d, \rho_0 \geq 0$. $\mathbb{M}_m$ denotes the set of all $m \times m$ diagonal matrices. $[m] := \{1, 2, \dots, m\}$ for any $m \in \mathbb{Z}_+$. $\Delta(\mathbb{Y})$ is the space of probability distributions on a set $\mathbb{Y}$. $\mathsf{Rad}(\alpha) = \mathsf{Unif}\{-\alpha, \alpha\}$ for $\alpha \in \mathbb{R}_{\geq 0}$.

2. Preliminaries on Partially-Observable Markov Decision Processes

In this paper, we consider a discrete-time infinite-horizon partially-observable Markov decision process (POMDP) with the (nonlinear) dynamics

$$\mathbb{P}(S_{t+1} \in B | \sigma(S_k, A_k, k \leq t)) =: \mathcal{P}((S_t, A_t), B),$$
$$\mathbb{P}(C | \sigma(S_t)) =: \phi(S_t, C),$$

for any $B \in \mathscr{B}(\mathbb{S})$ and $C \in \mathscr{B}(\mathbb{Y})$, where S_t is an $\mathbb{S}$-valued *state*, Y_t is a $\mathbb{Y}$-valued *observation*, and A_t is an $\mathbb{A}$-valued *control* process with the stochastic kernels $\mathcal{P} : \mathbb{S} \times \mathbb{A} \times \mathscr{B}(\mathbb{S}) \to [0, 1]$ and $\phi : \mathbb{S} \times \mathscr{B}(\mathbb{Y}) \to [0, 1]$. We consider finite but arbitrarily large $\mathbb{A} \subset \mathbb{R}^{d_1}, \mathbb{Y} \subset \mathbb{R}^{d_2}$ with $\mathbb{Y} \times \mathbb{A} \subset \mathcal{B}_2^{(d_1+d_2)}(0, 1)$ and $\mathbb{S}$. In this setting, the state process $(S_t)_{t \in \mathbb{N}}$ is not observable by the controller. Let

$$Z_t = \begin{cases} Y_0, & \text{if } t = 0, \\ (Z_{t-1}, A_{t-1}, Y_t), & \text{if } t > 0, \end{cases} \tag{1}$$

be the history process, which is available to the controller at time $t \in \mathbb{N}$, and

$$\bar{Z}_t := (Z_t, A_t) = (Y_0, A_0, \dots, Y_t, A_t),$$

be the history-action process.

Definition 2.1 (Admissible policy)**.** An admissible control policy $\pi = (\pi_t)_{t \in \mathbb{N}}$ is a sequence of measurable mappings $\pi_t : (\mathbb{Y} \times \mathbb{A})^t \times \mathbb{Y} \to \Delta(\mathbb{A})$, and the control at time t is chosen under π_t randomly as

$$\mathbb{P}(A_t = a | Z_t = z_t) = \pi_t(a | z_t),$$

for any $z_t \in (\mathbb{Y} \times \mathbb{A})^t \times \mathbb{Y}$. We denote the class of all admissible policies by Π_{NM}.

If an action a is taken at state a, then a reward $r(s, a)$ is obtained. For simplicity, we assume that the reward is deterministic, and $\max\limits_{s,a} |r(s, a)| \leq r_\infty < \infty$.

Definition 2.2 (Value function, $\mathcal{Q}$-function, advantage function)**.** Let π be an admissible policy, and $\mu \in \Delta(\mathbb{Y})$ be an initial observation distribution. Then, the value function under π with discount factor $\gamma \in (0, 1]$ is defined as

$$\mathcal{V}_t^\pi(z_t) := \mathbb{E}^\pi \Big[\sum_{k=t}^\infty \gamma^{k-t} r(S_k, A_k) \Big| Z_t = z_t \Big], \tag{2}$$

for any $z_t \in (\mathbb{Y} \times \mathbb{A})^t \times \mathbb{Y}$. Similarly, the state-action value function (also known as $\mathcal{Q}$-function) and the advantage function under π are defined as

$$\begin{aligned} \mathcal{Q}_t^\pi(\bar{z}_t) &:= \mathbb{E}^\pi \Big[\sum_{k=t}^\infty \gamma^{k-t} r(S_k, A_k) \Big| \bar{Z}_t = \bar{z}_t \Big], \\ \mathcal{A}_t^\pi(z_t, a) &:= \mathcal{Q}_t^\pi(z_t, a) - \mathcal{V}_t^\pi(z_t), \end{aligned} \tag{3}$$

for any $\bar{z}_t \in (\mathbb{Y} \times \mathbb{A})^{t+1}$, respectively.

Given an initial observation distribution $\mu \in \Delta(\mathbb{Y})$, the optimization problem is

$$\underset{\pi \in \Pi_{\mathsf{NM}}}{\text{maximize}} \int_{\mathbb{Y}} \mathcal{V}_0^\pi(z_0) \mu(dz_0) =: \mathcal{V}^\pi(\mu). \tag{4}$$

We denote $\pi^\star \in \underset{\pi \in \Pi_{\mathsf{NM}}}{\arg\max}\, \mathcal{V}^\pi(\mu)$ as an optimal policy.

Remark 2.3 (Curse of history in RL for POMDPs). Note that the problem in equation 4 is significantly more challenging than its subcase of (fully-observable) MDPs since there may not exist an optimal policy which is (i) stationary, or even Markovian, and (ii) deterministic (Krishnamurthy, 2016; Singh et al., 1994). As such, the policy search is over *non-Markovian* randomized policies of type $\pi = (\pi_0, \pi_1, \dots)$ where $\pi_t : (\mathbb{Y} \times \mathbb{A})^t \times \mathbb{Y} \to \Delta(\mathbb{A})$ depends on the history of observations $Z_t = (Y_0, A_0, Y_1, \dots, A_{t-1}, Y_t)$ for $t \in \mathbb{N}$. In this case, direct extensions of the existing reinforcement learning methods for MDPs become intractable, even for finite $\mathbb{Y}, \mathbb{A}$: the instantaneous memory complexity of a probabilistic admissible policy $\pi \in \Pi_{\mathsf{NM}}$ at epoch $t \in \mathbb{N}$ is $\mathscr{O}(|\mathbb{Y} \times \mathbb{A}|^{t+1})$, growing exponentially over t.

Recurrent neural networks (RNNs), which involve a parametric recurrent structure to efficiently represent the process history by using finite memory, are universal approximators for sequence-to-sequence mappings (Schäfer & Zimmermann, 2007; Grigoryeva & Ortega, 2018). As such, we consider using them in an actor-critic framework for approximation in (i) value space (for the critic), and (ii) policy space (for the actor). In the following section, we formally introduce the RNN architecture that we study in this paper.

3. Elman-Type Recurrent Neural Networks

We consider an Elman-type recurrent neural network (RNN) of width $m \in \mathbb{N}$ with $\mathbf{W} \in \mathbb{R}^{m \times m}$ and $\mathbf{U} \in \mathbb{R}^{m \times d}$, where

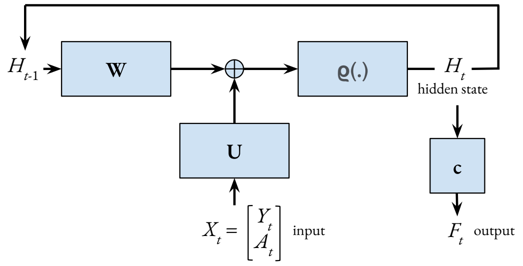


Figure 1. An Elman-type RNN in the reinforcement learning framework.

$d = d_1 + d_2$, and the rows of $\mathbf{U}$ are denoted as $U_i^\top$ for $i = 1, 2, \dots, m$. Given a smooth activation function $\varrho : \mathscr{C}^2(\mathbb{R}, \mathbb{R})$ with $\|\varrho\|_\infty \leq \varrho_0, \|\varrho'\|_\infty \leq \varrho_1, \|\varrho''\|_\infty \leq \varrho_2$, we denote $\vec{\varrho} : \mathbb{R}^m \to \mathbb{R}^m : \boldsymbol{z} \mapsto \begin{pmatrix} \varrho(z_1)) \\ \vdots \\ \varrho(z_m) \end{pmatrix}$. Let $X_t = \begin{pmatrix} Y_t \\ A_t \end{pmatrix}$, which is an $\mathbb{R}^d$-valued random variable with $d = d_1 + d_2$. The central structure in an RNN is the sequence of hidden states $H_t \in \mathbb{R}^m$, which evolves according to

$$H_t(\bar{Z}_t; \mathbf{W}, \mathbf{U}) = \vec{\varrho}\Big(\mathbf{W} H_{t-1}(\bar{Z}_{t-1}; \mathbf{W}, \mathbf{U}) + \mathbf{U} X_t\Big), \quad (5)$$

with $H_0(\bar{Z}_0; \mathbf{W}, \mathbf{U}) = \vec{\varrho}(\mathbf{U} X_0)$ and $\bar{Z}_t = (X_0, \dots, X_t)$ denoting the history. We denote the i^{th} element of H_t as $H_t^{(i)}$ for $i \in [m]$. We consider a linear readout layer with weights $c \in \mathbb{R}^m$, which leads to the output

$$F_t(\bar{Z}_t; \mathbf{W}, \mathbf{U}, c) = \frac{1}{\sqrt{m}} \sum_{i=1}^m c_i H_t^{(i)}(\bar{Z}_t; \mathbf{W}, \mathbf{U}). \quad (6)$$

The characteristic property of RNNs is ***weight-sharing***: throughout all time-steps $t \in \mathbb{N}$, the same weights are utilized, which enables the hidden state $(H_t)_{t>0}$ to summarize the entire history $\bar{Z}_t$ compactly with a fixed memory.

We consider diagonal $\mathbf{W}$ and general $\mathbf{U}$ in the paper, which simplifies the analysis, yet preserves the essential properties of RNNs. This diagonal structure for $\mathbf{W}$ is common in the study of deep linear networks for the aforementioned reason (Gunasekar et al., 2018; Nacson et al., 2022; Even et al., 2023; Woodworth et al., 2020), while our work also encompasses nonlinear activation functions and weight-sharing. The operation of an Elman-type recurrent neural network is illustrated in Figure 1. Following the neural tangent kernel literature, we omit the straightforward task of training the linear output layer $c \in \mathbb{R}^m$ for simplicity, and study the training dynamics of $(\mathbf{W}, \mathbf{U})$, which is the main challenge (Du et al., 2018; Oymak & Soltanolkotabi, 2020; Cai et al., 2019; Wang et al., 2019). Consequently, we denote the learnable parameters of a hidden unit $i \in [m]$ compactly as $\Theta_i = \begin{pmatrix} W_{ii} \\ U_i \end{pmatrix}$, and denote the learnable parameters of an RNN by $\Theta = \left[W_{11}, U_1^\top, W_{22}, U_2^\top, \dots, W_{mm}, U_m^\top\right]^\top \in \mathbb{R}^{m(d+1)}$. Given learnable parameters $(\mathbf{W}, \mathbf{U})$, we denote the sequence of recurrent neural network outputs as $\boldsymbol{F}(\cdot; \mathbf{W}, \mathbf{U}) = (F_t(\cdot; \mathbf{W}, \mathbf{U}))_{t \in \mathbb{N}}$, and use Θ and $(\mathbf{W}, \mathbf{U})$ interchangeably throughout the paper.

4. Infinite-Width Limit of Diagonal Recurrent Neural Networks

In this paper, we consider a class of systems that can be efficiently approximated and learned by the class of large recurrent neural networks in the near-initialization regime following (Cayci & Eryilmaz, 2024). To that end, we provide the following characterization of the infinite-width limit of RNNs in order to give our results in later sections. Let $w_0 \sim \mathsf{Rad}(\alpha)$ and $u_0 \sim \mathcal{N}(0, I_d)$ be independent random variables, and $\theta := \begin{pmatrix} w_0 \\ u_0 \end{pmatrix}$. Given a history-action realization $\bar{\boldsymbol{z}} = (x_0, x_1, \dots) \in (\mathbb{Y} \times \mathbb{A})^{\mathbb{Z}_+}$, define

$$h_t(\bar{z}_t; \theta_0) := \varrho(w_0 h_{t-1}(\bar{z}_{t-1}; \theta_0) + \langle u_0, x_t \rangle), \; t > 0,$$

with $h_{-1} := 0$ (thus $h_0(\bar{z}_0; \theta_0) = \varrho(\langle u_0, x_0 \rangle))$, and $\mathcal{I}_t(\bar{z}_t; \theta_0) := \varrho'(w_0 h_{t-1}(\bar{z}_{t-1}; \theta_0) + \langle u_0, x_t \rangle)$. Then, the neural tangent random feature (NTRF) mapping[1] at time t is defined as (with $\bar{\mathcal{I}}_{t,k}(\bar{z}_t; \theta_0) := \prod_{j=0}^k \mathcal{I}_{t-j}(\bar{z}_{t-j}; \theta_0)$):

$$\psi_t(\bar{z}_t; \theta_0) := \sum_{k=0}^t w_0^k \begin{pmatrix} h_{t-k-1}(\bar{z}_{t-k-1}; \theta_0) \\ x_{t-k} \end{pmatrix} \bar{\mathcal{I}}_{t,k}(\bar{z}_t; \theta_0),$$

We also define the NTRF matrix as follows:

$$\Psi_T(\bar{\boldsymbol{z}}; \theta_0) := \begin{pmatrix} \psi_0^\top(\bar{z}_0; \theta_0) \\ \psi_1^\top(\bar{z}_1; \theta_0) \\ \vdots \\ \psi_{T-1}^\top(\bar{z}_{T-1}; \theta_0) \end{pmatrix}, \; T \in \mathbb{N}, \quad (7)$$

with $\Psi(\bar{z}; \theta_0) := \Psi_\infty(\bar{z}; \theta_0)$.

Definition 4.1 (Transportation mapping)**.** Let $\mathscr{H}$ be the set of mappings $\boldsymbol{v} : \mathbb{R}^{1+d} \to \mathbb{R}^{1+d} : \theta_0 \mapsto \begin{pmatrix} v_w(w_0) \\ v_u(u_0) \end{pmatrix}$ with

$$\mathbb{E}[|v_w(w_0)|^2] = \frac{1}{2}\Big(|v_w(\alpha)|^2 + |v_w(-\alpha)|^2\Big) < \infty,$$
$$\mathbb{E}[\|v_u(u_0)\|_2^2] = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \|v_u(u)\|_2^2 e^{-\frac{1}{2}\|u\|_2^2} du < \infty.$$

We call each $\boldsymbol{v} \in \mathscr{H}$ a transportation mapping, following (Ji & Telgarsky, 2019; Ji et al., 2019).

[1] The feature uses a complicated weighted-sum of all past inputs $x_k, k \leq t$, leading to a discounted memory to tackle non-stationarity. x_{t-k} is scaled with $w_0^k \sim \mathsf{Rad}(\alpha)$, thus it yields a fading memory approximation of the history if $\alpha < 1$.

Definition 4.2 (Infinite-width limit)**.** We define the infinite-width limit of Elman-type RNNs as follows:

$$\mathscr{F} := \left\{(\mathbb{Y}\times\mathbb{A})^{\mathbb{Z}_+} \ni \bar{\boldsymbol{z}} \mapsto \mathbb{E}\left[\Psi(\bar{\boldsymbol{z}};\theta_0)\boldsymbol{v}(\theta_0)\right] : \boldsymbol{v}\in\mathscr{H}\right\}.$$

$\mathscr{F}$ consists of $f_t^\star(\bar{z}_t;\boldsymbol{v}) = \mathbb{E}[\langle\boldsymbol{v}(\theta_0),\psi_t(\bar{z}_t;\theta_0)\rangle]$ for any $\bar{\boldsymbol{z}}\in(\mathbb{Y}\times\mathbb{A})^{\mathbb{Z}_+}$. The same transportation mapping $\boldsymbol{v}$ is used to define the mapping $f_t^\star$ at each time t, which is a characteristic feature of weight-sharing in recurrent neural networks. Also, the input $\bar{\boldsymbol{z}}$ grows over time in a concatenated nature, which implies that $\boldsymbol{f}^\star\in\mathscr{F}$ is a representational assumption on the dynamical structure of the problem.

For any fixed time $t\in\mathbb{N}$, the completion of $\{\bar{z}_t\mapsto f_t^\star(\bar{z}_t;\boldsymbol{v}) : \boldsymbol{v}\in\mathscr{H}\}$ is exactly the reproducing kernel Hilbert space (RKHS) $\mathscr{G}_{\kappa_t}$ associated with the "recurrent" neural tangent kernel (NTK) κ_t (Rahimi et al., 2007; Ji et al., 2019). For any $t\in\mathbb{N}$, the inner product of two functions in $\mathscr{G}_{\kappa_t}$ associated with the transportation mappings $\boldsymbol{v},\boldsymbol{v}'$ is

$$\langle f_t^\star(\cdot;\boldsymbol{v}), f_t^\star(\cdot;\boldsymbol{v}')\rangle_{\mathscr{H}_{\kappa_t}} = \mathbb{E}\Big[\langle\boldsymbol{v}(\theta_0),\boldsymbol{v}'(\theta_0)\rangle\Big].$$

As such, the RKHS norm of any $f\in\mathscr{G}_{\kappa_t}$ is $\|f\|_{\mathscr{G}_{\kappa_t}} = \sqrt{\mathbb{E}\|\boldsymbol{v}(\theta_0)\|_2^2} = \sqrt{\mathbb{E}\|v_u(u_0)\|_2^2+\mathbb{E}|v_w(w_0)|^2}$.

Remark 4.3 (Reduction to FNNs). Consider $T=1$:

$$\mathscr{F}_1 := \left\{\bar{z}_0\mapsto\mathbb{E}\left[\psi_0^\top(\bar{z}_0;\theta_0)\boldsymbol{v}(\theta_0)\right] : \boldsymbol{v}\in\mathscr{H}\right\}.$$

In this case, we exactly recover the NTK (and the associated RKHS) for single-layer FFNs (Jacot et al., 2018; Wang et al., 2019; Liu et al., 2019). Furthermore, since the kernel κ_0 is universal, the associated RKHS $\mathscr{G}_{\kappa_0}$ is dense in the space of continuous functions on a compact set (Ji et al., 2019).

5. Rec-NAC Algorithm for POMDPs

In this section, we present a high-level description of our Recurrent Natural Actor-Critic (Rec-NAC) Algorithm with two inner loops, critic (called Rec-TD) and actor (called Rec-NPG), for policy optimization with RNNs. The details of the inner loops of the algorithm will be given in the succeeding sections. We use an admissible policy $\pi=(\pi_t)_{t\in\mathbb{N}}$ that is parameterized by a recurrent neural network $(F_t^{\mathsf{a}}(\cdot;\Phi))_{t\in\mathbb{N}}$ of the form given in equation 6 with a network width $m\in\mathbb{Z}_+$. To that end, for any $t\in\mathbb{N}$, let

$$\pi_t^\Phi(a|z_t) := \frac{\exp\left(F_t^{\mathsf{a}}((z_t,a);\Phi)\right)}{\sum_{a'\in\mathbb{A}}\exp\left(F_t^{\mathsf{a}}((z_t,a');\Phi)\right)}, \tag{8}$$

for any $z_t\in(\mathbb{Y}\times\mathbb{A})^t\times\mathbb{Y}$ and $a\in\mathbb{A}$ with the parameter $\Phi\in\mathbb{R}^{m(d+1)}$.

Rec-NAC operates as follows:

- **Initialization.** The actor RNN F^{a} is randomly initialized with parameter $\Phi(0)\sim\zeta_{\mathsf{init}}$ (see Def. A.1).
- **Natural policy gradient.** For $0\leq n<N$,
  - **Critic.** Estimate $\hat{\mathcal{Q}}_t^{(n)}(\cdot) := F_t^{\mathsf{c}}(\cdot;\bar{\Theta}^{(n)})$ $t<T$ of $\mathcal{Q}_t^{\pi^{\Phi(n)}}(\cdot),\ t<T$ via Rec-TD learning in Sec. 6. F^{c} is initialized independently for each n as Definition A.1.
  - **Actor.** By projected stochastic gradient descent (SGD), obtain a solution ω_n for the compatible function approximation problem

    $$\min_\omega \mathbb{E}\sum_{t=0}^{T-1}\gamma^t|\nabla\ln\pi_t^n(A_t|Z_t)\omega-\hat{\mathcal{A}}_t^{\pi^{\Phi(n)}}(\bar{Z}_t)|^2,$$
    such that $\omega\in\mathcal{B}_{2,\infty}(0,\rho)$,

    where for any $t\in\mathbb{N}$,

    $$\hat{\mathcal{A}}_t^{(n)}(z_t,a) := \hat{\mathcal{Q}}_t^{(n)}(z_t,a) - \mathop{\mathbb{E}}_{A'\sim\pi_t^\Phi(\cdot|z_t)}\hat{\mathcal{Q}}_t^{(n)}(z_t,A').$$

For information regarding the algorithmic tools, i.e., random initialization and max-norm regularization for RNNs, we refer to Section A.

6. Critic: Recurrent Temporal Difference Learning (Rec-TD)

In this section, we study a value prediction algorithm for policy evaluation in POMDPs, which will serve as the critic.

Policy evaluation problem. Consider the policy evaluation problem for POMDPs under a given non-Markovian policy $\pi\in\Pi_{\mathsf{NM}}$. Given an initial observation distribution $\mu\in\Delta(\mathbb{Y})$, policy evaluation aims to solve

$$\min_\Theta \mathcal{R}_T^\pi(\Theta) := \mathbb{E}_\mu^\pi\sum_{t=0}^{T-1}\gamma^t\big(F_t(\bar{Z}_t;\Theta)-\mathcal{Q}_t^\pi(\bar{Z}_t)\big)^2, \tag{9}$$
$$\text{such that } \Theta\in\Omega_{\rho,m} := \mathcal{B}_{2,\infty}^{(m)}(0,\rho),$$

where $T\in\mathbb{N}$ is the truncation level, and $\{F_t : t\in\mathbb{N}\}$ is an Elman-type recurrent neural network given in equation 6 – we drop the superscript a for simplicity throughout the discussion. The expectation in $\mathcal{R}_T^\pi(\Theta)$ is with respect to the joint probability law $P_T^{\pi,\mu}$ of the stochastic process $\{(S_t,A_t,Y_t) : t\in[0,T]\}$ where $Z_0\sim\mu$.

6.1. Recurrent TD Learning Algorithm

Given a sample trajectory $\bar{z}_T\in(\mathbb{Y}\times\mathbb{A})^{T+1}$, let

$$\delta_t(\bar{z}_{t+1};\Theta) := r_t+\gamma F_{t+1}(\bar{z}_{t+1};\Theta)-F_t(\bar{z}_t;\Theta), \tag{10}$$

be the temporal difference, and let

$$\check{\nabla}\mathcal{R}_T(\bar{z}_T;\Theta) = \sum_{t=0}^{T}\gamma^t\delta_t(\bar{z}_{t+1};\Theta)\nabla_\Theta F_t(\bar{z}_t;\Theta), \tag{11}$$

be the stochastic semi-gradient. Note that, despite the exponential growth in the dimension of $\bar{z}_t \in (\mathbb{Y} \times \mathbb{A})^{t+1}$ over $t \in \mathbb{N}$, the memory complexity for computing $\check{\nabla}\mathcal{R}_T(\bar{z}_T;\Theta)$ is only $\mathscr{O}(m^2 + md)$ thanks to the use of RNNs.

Assumption 6.1 (Sampling oracle)**.** Given an initial state distribution μ, we assume that the system can be independently started from $S_0 \sim \mu$, i.e., independent trajectories $\{(S_t, Y_t, A_t) : t \in [T]\} \sim P_T^{\pi,\mu}$ can be obtained.

Under Assumption 6.1, for $k \in \mathbb{N}$, let $\{(S_t^k, Y_t^k, A_t^k) : t \in [T]\} \sim P_T^{\pi,\mu}$ be an independent trajectory (for each $k \in \mathbb{N}$, i.e., a trajectory with an independent initial sample $S_0^k \sim \mu$), and $\{Z_t^k : t \in [T]\}$ and $\{\bar{Z}_t^k : t \in [T]\}$ be the resulting (truncated) history and history-action processes. Starting from a random initialization $(\mathbf{W}(0), \mathbf{U}(0), c)$, let

$$\check{\Theta}(k+1) = \Theta(k) + \eta \cdot \check{\nabla}\mathcal{R}_T(\bar{Z}_T^k;\Theta(k)), \qquad (12)$$

for $k \in \mathbb{N}$. For **Rec-TD**, one uses $\Theta(k+1) = \check{\Theta}(k+1)$. For **Rec-TD with max-norm regularization**, one uses

$$\Theta(k+1) = \mathbf{Proj}_{\Omega_{\rho,m}}[\check{\Theta}(k+1)],$$

for parameter $\rho = (\rho_w, \rho_u) \in \mathbb{R}^2_{>0}$.

Remark 6.2 (Intuition behind Rec-TD). In a stochastic optimization setting, the loss-minimization for $\mathcal{R}_T(\Theta)$ would be solved by using gradient descent, where the gradient is $\mathbb{E}_\mu^\pi \sum_{t=0}^{T-1} \gamma^t \left(F_t(\bar{Z}_t;\Theta) - \mathcal{Q}_t^\pi(\bar{Z}_t)\right) \nabla F_t(\bar{Z}_t;\Theta)$. On the other hand, the target function $\mathcal{Q}_t^\pi$ is unknown and to be learned. Following the bootstrapping idea for MDPs in (Sutton, 1988), we exploit an extended non-Markovian Bellman equation in Proposition B.3, and use $r_t + \gamma F_{t+1}(\bar{Z}_{t+1};\Theta)$ as a bootstrap estimate for the unknown $\mathcal{Q}_t^\pi(\bar{Z}_t)$. Note that, in the realizable case with $F_t(\cdot;\Theta^\star) = \mathcal{Q}_t^\pi(\cdot)$, $t \in \mathbb{Z}_+$ for some $\Theta^\star \in \mathbb{M}_m \times \mathbb{R}^{m\times d}$, we have $\mathbb{E}_\mu^\pi[\check{\nabla}\mathcal{R}_T(\bar{Z}_T;\Theta^\star)] = 0$, which implies that the stochastic approximation approach for MDPs can be used for the non-Markovian setting.

6.2. Theoretical Analysis of Rec-TD: Finite-Time Bounds and Global Near-Optimality

In the following, we prove that Rec-TD with max-norm regularization achieves global optimality in expectation. To characterize the impact of long-term dependencies on the performance of Rec-TD, let $p_t(x) = \sum_{k=0}^{t-1} |x|^k$, and $q_t(x) = \sum_{k=0}^{t-1}(k+1)|x|^k$, $x \in \mathbb{R}, t \in \mathbb{N}$.

Theorem 6.3 (Finite-time bounds for Rec-TD)**.** *Assume that $\{\mathcal{Q}_t^\pi : t \in \mathbb{N}\} \in \mathscr{F}$ with a transportation mapping $\boldsymbol{v} = (v_{\mathsf{w}}, v_{\mathsf{u}}) \in \mathscr{H}$ such that $\sup_{u\in\mathbb{R}^d} \|v_{\mathsf{u}}(u)\|_2 \leq \nu_{\mathsf{u}}$ and $\sup_{w\in\mathbb{R}} |v_{\mathsf{w}}(w)| \leq \nu_{\mathsf{w}}$. Then, for any projection radius $\rho \succeq \nu = (\nu_{\mathsf{w}}, \nu_{\mathsf{u}})$ and step-size $\eta > 0$, Rec-TD with max-norm regularization achieves the following error bound:*

$$\mathbb{E}\Big[\frac{1}{K}\sum_{k=0}^{K-1}\mathcal{R}_T^\pi(\Theta(k))\Big] \leq \frac{1}{\sqrt{K}}\left(\frac{\|\nu\|_2^2}{(1-\gamma)} + \frac{C_T^{(1)}}{(1-\gamma)^3}\right) + \frac{C_T^{(2)}}{(1-\gamma)^2\sqrt{m}} + \underbrace{\frac{\gamma^T}{(1-\gamma)K}\sum_{k=0}^{K-1}\omega_{T,k}^2}_{(\heartsuit)}. \qquad (13)$$

for any $K \in \mathbb{N}$, where

$$C_T^{(1)}, C_T^{(2)} = \mathsf{poly}\left(p_T\left(\varrho_1(\alpha + \frac{\rho_w}{\sqrt{m}})\right), \|\rho\|_2, \|\nu\|_2\right),$$

are instance-dependent constants that do not depend on K, and $\omega_{t,k} := \sqrt{\mathbb{E}[(F_t(\bar{Z}_t;\Theta(k)) - \mathcal{Q}_t^\pi(\bar{Z}_t^k))^2]}$ for $t, k \in \mathbb{N}$. For the average-iterate Rec-TD with $\bar{\Theta}_K := \frac{1}{K}\sum_{k=0}^{K-1}\Theta(k)$, we have

$$\mathbb{E}\left[\mathcal{R}_T^\pi\left(\bar{\Theta}_K\right)\right] \leq \frac{10}{(1-\gamma)\sqrt{K}}\left(\|\nu\|_2^2 + \frac{C_T^{(1)}}{(1-\gamma)^2}\right) + \frac{10C_T^{(2)}}{(1-\gamma)^2\sqrt{m}} + \frac{10\gamma^T}{(1-\gamma)K}\sum_{k=0}^{K-1}\omega_{T,k}^2.$$

The proof of Theorem 6.3 can be found in Section B.

Remark 6.4 (Truncation level T and the impact of long-term dependencies). From Proposition 7.2, we observe that the exact natural policy gradient update would require a large T. In the following, we discuss the impact of T on the performance of policy evaluation, depending on the inherent memory (i.e., long-term dependencies) in the system. As noted in (Goodfellow et al., 2016), the spectral radius of $\{\mathbf{W}(k) : k \in \mathbb{N}\}$ determines the degree of long-term dependencies in the problem as it scales H_t. Consistent with this observation, our bounds have a strong dependency on

$$\alpha_m := \alpha + \frac{\rho_{\mathsf{w}}}{\sqrt{m}} \geq \lambda_{\mathsf{max}}(\mathbf{W}(k)) = \|\mathbf{W}(k)\|_{\infty,\infty},$$

for any $k \in \mathbb{N}$.

Remark 6.5 (When is Rec-TD efficient?). Note that both constants $C_T^{(1)}, C_T^{(2)}$ polynomially depend on $p_T(\varrho_1\alpha_m)$. Let $\varepsilon > 0$ be any given target error.

- **Short-term memory.** If $\alpha_m < \frac{1}{\varrho_1}$, then it is easy to see that $p_T(\varrho_1\alpha_m) \leq \frac{1}{1-\varrho_1\alpha_m}$. Thus, the extra term $(\heartsuit)$ in equation 13 vanishes at a geometric rate as $T \to \infty$, yet m (network-width) and K (iteration-complexity) are still $\tilde{\mathscr{O}}(1/\varepsilon^2)$. Rec-TD is very efficient in that case.
- **Long-term memory.** If $\alpha_m > \frac{1}{\varrho_1}$, as $T \to \infty$, both m and K grow at a rate $\mathscr{O}\left((\varrho_1\alpha_m)^T/\varepsilon^2\right)$ while the extra term $(\heartsuit)$ in equation 13 vanishes at a geometric

rate. As such, the required network size and iterations grow at a geometric rate with T in systems with long-term memory, constituting the pathological case for Rec-TD.

The performance of Rec-TD is studied numerically in Random-POMDP instances in Section C.

Finally, note that the additional term ($\heartsuit$) in Theorem 6.3 is unique to Rec-TD learning, and stems from the use of bootstrapping in reinforcement learning.

7. Actor: Recurrent Natural Policy Gradient (Rec-NPG) for POMDPs

The goal is to solve the following problem:

$$\underset{\Theta\in\mathbb{R}^{m(d+1)}}{\text{maximize}}\ \mathcal{V}^{\pi^\Phi}(\mu) \text{ such that } \Phi\in\Omega_{\rho,m}, \tag{PO}$$

for a given initial distribution $\mu\in\Delta(\mathbb{Y})$ and $\rho\in\mathbb{R}^2_{>0}$. $\pi^\star$ denotes an optimal policy.

Remark 7.1 (Why RNNs for policy parameterization?). Recall that an admissible policy $\pi\in\Pi_{\mathsf{NM}}$ is a sequence of elements $\pi_t:\mathbb{Y}^{t+1}\times\mathbb{A}^t\rightarrow\Delta(\mathbb{A})$ for $t\geq 0$. Thus, the memory complexity of a *complete* policy at time t is $\mathscr{O}(|\mathbb{Y}\times\mathbb{A}|^{t+1})$, which is extremely infeasible to perform policy optimization. Restricted hypothesis classes are deployed to address this in the POMDP setting. The key concept is the use of internal states to summarize $(\bar{Z}_t)_{t\in\mathbb{N}}$ with finite memory. For that purpose, we use RNNs to parameterize policies with a parameter $\Phi\in\mathbb{R}^{m(d+1)}$ and use the hidden states $(H_t)_{t\in\mathbb{N}}$ to efficiently represent the process history. π^Φ and its gradient can be computed recursively, leading to a finite-memory policy with memory complexity $\mathscr{O}((d+1)m)$. As we will see in Prop. 7.6, the representation power of the RNN in the policy parameterization (as measured by an approximation error) to represent $\mathcal{Q}$-functions will play a vital role in achieving optimality.

7.1. Recurrent Natural Policy Gradient for POMDPs

In this section, we describe the recurrent natural policy gradient (Rec-NPG) algorithm for non-Markovian reinforcement learning. As proved in Prop. D.2, the policy gradient under partial observability is

$$\nabla_\Phi\mathcal{V}^{\pi^\Phi}(\mu):=\mathbb{E}_\mu^{\pi^\Phi}\sum_{t=0}^\infty\gamma^t\mathcal{Q}_t^{\pi^\Phi}(Z_t,A_t)\nabla_\Phi\ln\pi_t^\Phi(A_t|Z_t).$$

Fisher information matrix under a policy π^Φ is defined as

$$G_\mu(\Phi):=\mathbb{E}_\mu^{\pi^\Phi}\sum_{t=0}^\infty\gamma^t\nabla\ln\pi_t^\Phi(A_t|Z_t)\nabla^\top\ln\pi_t^\Phi(A_t|Z_t),$$

for an initial observation distribution $\mu\in\Delta(\mathbb{Y})$. Rec-NPG updates the policy parameters by

$$\Phi(n+1)=\Phi(n)+\eta\cdot G_\mu^+(\Phi(n))\nabla_\Phi\mathcal{V}^{\pi^{\Phi(n)}}(\mu), \tag{14}$$

for an initial parameter $\Phi(0)$ and step-size $\eta>0$, where G^+ denotes the Moore-Penrose inverse of a matrix G. This update rule is in the same spirit as the NPG introduced in (Kakade, 2001), however, due to the non-Markovian nature of the partially-observable MDP, it has significant complications that we will address.

In order to avoid computationally-expensive policy updates in equation 14, we utilize the following extension of the compatible function approximation in (Kakade, 2001) to the case of non-Markovian policies for POMDPs.

Proposition 7.2 (Compatible function approximation for non-Markovian policies)**.** *For any $\Phi\in\mathbb{R}^{m(d+1)}$ and initial observation distribution μ, let*

$$\mathcal{L}_\mu(w;\Phi)=\mathbb{E}_\mu^{\pi^\Phi}\sum_{t=0}^\infty\gamma^t\big(\nabla^\top\ln\pi_t^\Phi(A_t|Z_t)\omega-\mathcal{A}_t^{\pi^\Phi}(\bar{Z}_t)\big)^2, \tag{15}$$

for $\omega\in\mathbb{R}^{m(d+1)}$. Then, we have

$$G_\mu^+(\Phi)\nabla_\Phi\mathcal{V}^{\pi^\Phi}(\mu)\in\underset{\omega\in\mathbb{R}^{m(d+1)}}{\arg\min}\,\mathcal{L}_\mu(\omega;\Phi). \tag{16}$$

Path-based compatible function approximation with truncation. For MDPs, the compatible function approximation error $\mathcal{L}_\mu(w;\Phi)$ can be expressed by using the discounted state-action occupancy measure, from which one can obtain unbiased samples (Agarwal et al., 2020; Konda & Tsitsiklis, 2003). Thus, the infinite-horizon can be handled without any loss. On the other hand, for general (non-Markovian) problems as in equation 15, this simplification is impossible due to the non-stationarity. As such, we use a path-based method under truncation for a given $T\in\mathbb{N}$ with

$$\ell_T(\omega;\Phi,\mathcal{Q}):=\sum_{t=0}^{T-1}\gamma^t(\nabla\ln\pi_t^\Phi(A_t|Z_t)\omega-\mathcal{A}_t(Z_t,A_t))^2,$$

where $\mathcal{A}_t(z_t,a_t)=\mathcal{Q}_t(z_t,a_t)-\sum_{a\in\mathbb{A}}\pi_t^\Phi(a|z_t)\mathcal{Q}_t(z_t,a)$. Given a policy with parameter $\Phi(n)$ and the corresponding output of the critic (Rec-TD with the average-iterate $\bar{\Theta}^{(n)}:=\frac{1}{K_{\text{td}}}\sum_{k<K_{\text{td}}}\Theta^{(n)}(k)$):

$$\hat{\mathcal{Q}}^{(n)}(\cdot):=F_t(\cdot;\bar{\Theta}^{(n)}),$$

the actor aims to solve the following problem:

$$\begin{aligned}&\min_\omega\mathbb{E}\Big[\ell_T\left(\omega;\Phi(n),\hat{\mathcal{Q}}^{(n)}\right)\Big|\bar{\Theta}^{(n)},\Phi(n),\ldots,\Phi(0)\Big]\\&\text{such that }\omega\in\mathcal{B}_{2,\infty}^{(m)}(0,\rho).\end{aligned}$$

To that end, we utilize stochastic gradient descent (SGD) to solve the above problem. Let $\bar{Z}_T^{n,k} \sim P_T^{\pi^{\Phi(n)},\mu}$ be an independent random sequence for $k = 0, 1, \ldots$, and let

$$\tilde{\omega}_n(k+1) = \hat{\omega}_n(k) - \eta_{\mathsf{sgd}} \nabla_\omega \ell_T\big(\hat{\omega}_n(k); \Phi(n), \hat{\mathcal{Q}}^{(n)}\big),$$
$$\hat{\omega}_n(k+1) = \mathbf{Proj}_{\mathcal{B}_{2,\infty}^{(m)}(0,\rho)}[\tilde{\omega}_n(k+1)],$$

with $\hat{\omega}_n(0) = 0$. Then, a biased stochastic approximation of the natural policy gradient $G_\mu^+(\Phi(n))\nabla_\Phi \mathcal{V}^{\pi^{\Phi(n)}}(\mu)$ is obtained as $\omega_n := \frac{1}{K_{\mathsf{sgd}}} \sum_{k<K_{\mathsf{sgd}}} \hat{\omega}_n(k)$, and the policy update is performed as

$$\Phi(n+1) = \Phi(n) + \eta_{\mathsf{npg}} \cdot \omega_n.$$

In the following, we present a non-asymptotic analysis of the above approach.

7.2. Theoretical Analysis of Rec-NAC for POMDPs

We establish an error bound on the best-iterate for the Rec-NPG. The significance of the following result is two-fold: (i) it will explicitly connect the optimality gap to the compatible function approximation error, and (ii) it will explicitly show the impact of truncation on the performance of path-based policy optimization for the non-stationary case.

Theorem 7.3. *Assume that $P_T^{\pi^\star,\mu} \ll P_T^{\pi^{\Phi(n)},\mu}$, $n < N$, and let*

$$\kappa := \max_{0 \le n < N} \left\| \frac{P_T^{\pi^\star,\mu}}{P_T^{\pi^{\Phi(n)},\mu}} \right\|_\infty.$$

We have the following result under Rec-NPG after $N \in \mathbb{Z}_+$ steps with step-size $\eta_{\mathsf{npg}} = 1/\sqrt{N}$ with projection radius $\rho \in \mathbb{R}^2_{>0}$:

$$\min_{0 \le n < N} \mathbb{E}_0[\mathcal{V}^{\pi^\star}(\mu) - \mathcal{V}^{\pi^{\Phi(n)}}(\mu)] \le \frac{\ln|\mathbb{A}|}{(1-\gamma)\sqrt{N}}$$
$$+ \sqrt{p_T(\gamma)}\mathbb{E}_0\left[\frac{1}{N}\sum_{n=0}^{N-1}\left(\kappa \varepsilon_{\mathsf{cfa}}^T(\Phi(n), \omega_n)\right)^{\frac{1}{2}}\right] + \frac{2\gamma^T r_\infty}{(1-\gamma)^2}$$
$$+ \|\rho\|_2^2 \sum_{t<T} \gamma^t \frac{2\beta_t + 12(\Lambda_t^2 \varrho_2 + \chi_t \varrho_1)\sqrt{N}}{\sqrt{m}}$$
$$+ \|\rho\|_2^2 \sum_{t<T} \gamma^t \frac{12 L_t \sqrt{\Lambda_T^2 \varrho_2 + \chi_t \varrho_1}}{m^{1/4}} + \frac{\|\rho\|_2^2}{2\sqrt{N}} \sum_{t<T} \gamma^t L_t^2,$$

where $\varepsilon_{\mathsf{cfa}}^T(\Phi, \omega) := \mathbb{E}_\mu^{\pi^{\Phi(n)}} \sum_{t<T} \gamma^t |\nabla^\top \ln \pi_t^\Phi(A_t|Z_t)\omega - \mathcal{A}_t^{\pi^\Phi}(Z_t, A_t)|^2$, and the sequence $(L_t, \beta_t, \Lambda_t, \chi_t)_t$ is defined in Lemma B.1.

Remark 7.4. We have the following remarks.

- The effectiveness of Rec-NPG is proportional to the approximation power of the RNN used for policy parameterization, as reflected in $\varepsilon_{\mathsf{cfa}}^T$ in Theorem 7.3. We further characterize this error term in Prop. 7.6-7.8.
- The terms $L_t, \beta_t, \Lambda_t, \chi_t$ grow at a rate $p_t(\varrho_1 \alpha_m)$. Thus, if $\alpha_m > \varrho_1^{-1}$, then m and N should grow at a rate $(\alpha_m \varrho_1)^T$, implying the curse of dimensionality (more generally, it is known as the exploding gradient problem (Goodfellow et al., 2016)). On the other hand, if $\alpha_m < \varrho_1^{-1}$, then $L_t, \beta_t, \Lambda_t, \chi_t$ are all $\mathscr{O}(1)$ for all t, implying efficient learning of POMDPs. This establishes a very interesting connection between the memory in the system, the continuity and smoothness of the RNN with respect to its parameters, and the optimality gap under Rec-NPG.
- The term $\frac{2\gamma^T r_\infty}{(1-\gamma)^2}$ is due to truncating the trajectory at T, and vanishes with large T.

Remark 7.5. The quantity κ in Proposition 7.8 is the so-called concentrability coefficient in policy gradient methods (Agarwal et al., 2020; Bhandari & Russo, 2019; Wang et al., 2019), and determines the complexity of exploration. Note that it is defined in terms of path probabilities $P_T^{\pi,\mu}$ in the non-stationary setting.

In the following, we decompose the compatible function approximation error $\varepsilon_{\mathsf{cfa}}^T$ into the approximation error for the RNN and the statistical errors. To that end, let

$$\varepsilon_{\mathsf{app},n} = \inf\Big\{\mathbb{E}\sum_{t=0}^{T-1} \gamma^t \big|\nabla^\top F_t(\bar{Z}_t; \Phi(0))\omega - \mathcal{Q}_t^{\pi^{\Phi(n)}}(\bar{Z}_t)\big|^2$$
$$: \omega \in \mathcal{B}_{2,\infty}^{(m)}(0,\rho)\Big\},$$

be the approximation error where the expectation is with respect to $P_T^{\pi^{\Phi(n)},\mu}$,

$$\varepsilon_{\mathsf{td},n} = \mathbb{E}[\mathcal{R}_T^{\pi^{\Phi(n)}}(\bar{\Theta}^{(n)})|\Phi(k), k \le n],$$

be the error in the critic (see equation 9), and finally let

$$\varepsilon_{\mathsf{sgd},n} = \mathbb{E}[\ell_T(\omega_n; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\bar{\Theta}^{(n)}, \Phi(k), k \le n]$$
$$- \inf_w \mathbb{E}[\ell_T(\omega; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\bar{\Theta}^{(n)}, \Phi(k), k \le n],$$

be the error in the policy update via compatible function approximation.

Proposition 7.6 (Error decomposition for $\varepsilon_{\mathsf{cfa}}^T$)**.** *We have*

$$\mathbb{E}\Big[\mathbb{E}_\mu^{\pi^{\Phi(n)}}\Big[\ell_T(\omega_n; \Phi(n), \mathcal{Q}^{(n)})\Big]\Big|\Phi(k), k \le n\Big]$$
$$\le \frac{8\|\rho\|_2^2}{m}\sum_{t=0}^{T-1}\gamma^t \beta_t^2 + 8\varepsilon_{\mathsf{app},n} + 6\varepsilon_{\mathsf{td},n} + 2\varepsilon_{\mathsf{sgd},n}.$$

for any $n \in \mathbb{Z}_+$.

From Theorem 6.3, we have, for $\eta_{\mathsf{td}} = \mathscr{O}(1/\sqrt{K_{\mathsf{td}}})$,

$$\varepsilon_{\mathsf{td},n} \le \mathbf{poly}(p_T(\varrho_1 \alpha_m))\mathscr{O}\left(\frac{1}{\sqrt{K_{\mathsf{td}}}} + \frac{1}{\sqrt{m_{\mathsf{critic}}}} + \gamma^T\right),$$

and by Theorem 14.8 in (Shalev-Shwartz & Ben-David, 2014), we have, for $\eta_{\mathsf{td}} = \mathscr{O}(1/\sqrt{K_{\mathsf{td}}})$,

$$\varepsilon_{\mathsf{sgd},n} \leq \mathbf{poly}(p_T(\varrho_1 \alpha_m), \|\rho\|_2)\mathscr{O}(1/\sqrt{K_{\mathsf{sgd}}}).$$

As such, the statistical errors in the critic and the policy update (i.e., $\varepsilon_{\mathsf{td},n}, \varepsilon_{\mathsf{sgd},n}$) can be made arbitrarily small by using larger $K_{\mathsf{td}}, K_{\mathsf{sgd}}$ and larger m_{critic}. The remaining quantity to characterize is the approximation error, which is of critical importance for a small optimality gap as shown in Theorem 7.3 and Proposition 7.6. In the following, we will provide a finer characterization of $\varepsilon_{\mathsf{app},n}$ and identify a class of POMDPs that can be efficiently solved using Rec-NPG.

Assumption 7.7. For an index set J and $\nu \in \mathbb{R}^2_{>0}$, we consider a class $\mathscr{H}_{J,\nu}$ of transportation mappings

$$\left\{ \boldsymbol{v}^{(j)} \in \mathscr{H} : j \in J, \begin{pmatrix} \sup\limits_{w \in \mathbb{R}, j \in J} |v_{\mathsf{w}}^{(j)}(w)| \\ \sup\limits_{u \in \mathbb{R}^d, j \in J} \|v_{\mathsf{u}}^{(j)}(u)\|_2 \end{pmatrix} \leq \begin{pmatrix} \nu_{\mathsf{w}} \\ \nu_{\mathsf{u}} \end{pmatrix} \right\},$$

and also the corresponding infinite-width limit

$$\mathscr{F}_{J,\nu} := \{\bar{z} \mapsto \mathbb{E}[\Psi(\bar{z}; \theta_0)\boldsymbol{v}(\theta_0)] : \boldsymbol{v} \in \mathbf{Conv}(\mathscr{H}_{J,\nu})\},$$

where $\Psi(\cdot; \theta_0)$ is the NTRF matrix, defined in equation 7.

We assume that there exists an index set J and $\nu \in \mathbb{R}^2_{>0}$ such that $\mathscr{Q}^{\pi^{\Phi(n)}} \in \mathscr{F}_{J,\nu}$ for all $n \in \mathbb{N}$.

This representational assumption implies that the $\mathscr{Q}$-functions under all iterate policies $\pi^{\Phi(n)}$ throughout the Rec-NPG iterations $n = 0, 1, \ldots$ can be represented by convex combinations of a *fixed* set of mappings in the NTK function class $\mathscr{F}$ indexed by J. As we will see, the richness of J as measured by a relevant Rademacher complexity will play an important role in bounding the approximation error. To that end, for $\bar{z}_t = (z_t, a_t) \in (\mathbb{Y} \times \mathbb{A})^{t+1}$, let

$$G_t^{\bar{z}_t} := \{\phi \mapsto \nabla_\phi^\top H_t^{(1)}(\bar{z}_t; \phi)\boldsymbol{v}(\phi) : \boldsymbol{v} \in \mathscr{H}_{J,\nu}\},$$

and

$$\mathfrak{Rad}_m(G_t^{\bar{z}_t}) := \mathop{\mathbb{E}}_{\substack{\epsilon \sim \mathsf{Rad}^m(1) \\ \Phi(0) \sim \zeta_{\mathsf{init}}}} \sup_{g \in G_t^{\bar{z}_t}} \frac{1}{m} \sum_{i=1}^m \epsilon_i g(\Phi_i(0)).$$

Note that $\boldsymbol{v} \in \mathscr{H}_{J,\nu}$ above can be replaced more with $\boldsymbol{v} \in \mathbf{Conv}(\mathscr{H}_{J,\nu})$ without any loss. In that case, since the mapping $\boldsymbol{v}^{(j)} \mapsto f_t^\star(\bar{z}_t; \boldsymbol{v}^{(j)}) \in G_t^{\bar{z}_t}$ is linear, $G_t^{\bar{z}_t}$ is replaced with $\mathbf{Conv}(G_t^{\bar{z}_t})$ without changing the Rademacher complexity (Mohri et al., 2018).

The following proposition provides a finer characterization of the function approximation error.

Proposition 7.8. *Under Assumption 7.7, if $\rho \succeq \nu$, then*

$$\epsilon_{\mathsf{app},n} \leq \frac{1}{1-\gamma}\Bigg(2 \max_{0 \leq t < T} \max_{\bar{z}_t \in (\mathbb{Y} \times \mathbb{A})^{t+1}} \mathfrak{Rad}_m(G_t^{\bar{z}_t}) \\ + L_T \|\rho\|_2 \sqrt{\frac{\ln\left(2T|\mathbb{Y} \times \mathbb{A}|^T/\delta\right)}{m}}\Bigg)^2,$$

for all n simultaneously with probability at least $1-\delta$ over the random initialization for any $\delta \in (0,1)$.

Remark 7.9. Two interesting cases that lead to a vanishing approximation error (as $m \to \infty$), thus global near-optimality, are as follows.

- **Finite J.** If $|J| < \infty$, then Proposition 7.8 reduces to (Cayci et al., 2024) (with $T = 1$ for FNNs) with the complexity term $\mathscr{O}\left(\sqrt{\frac{\ln(|J|/\delta)}{m}}\right)$ by the finite-class lemma (Mohri et al., 2018). In this case, the $\mathscr{Q}$-functions throughout $n = 0, 1, \ldots$ lie in the convex hull of $|J|$ basis functions in $\mathscr{F}$ generated by $\{\boldsymbol{v}^{(j)} \in \mathscr{H} : j \in J\}$.
- **Linear transportation mappings.** For a fixed map $\varpi : \mathbb{R}^{d+1} \to \mathbb{R}^{(d+1)\times(d+1)}$, let $\boldsymbol{v}^{(b)}(\theta) = \langle \varpi(\theta), b \rangle$, $b \in J$ where $J \subset \mathbb{R}^{d+1}$ is a compact set. Then, the approximation error vanishes at a rate $\mathscr{O}(1/\sqrt{m})$.

Remark 7.10. In a *static* problem (e.g., the regression problem in supervised learning or the policy evaluation problem in Section 6) with a target function $f \in \mathscr{F}$, the approximation error is easy to characterize:

$$\left|\nabla^\top F_t(\bar{z}_t; \Phi(0))\omega^\star - f_t(\bar{z}_t)\right| = \mathscr{O}\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right), \quad (17)$$

by Hoeffding inequality with $\omega^\star := \left[\frac{1}{\sqrt{m}} c_i \boldsymbol{v}(\Phi_i(0))\right]_{i \in [m]}$.

In the *dynamical* policy optimization problem, the representational assumption $\mathscr{Q}^{\pi^{\Phi(n)}} \in \mathscr{F}$ does not imply an arbitrarily small approximation error as $m \to \infty$ since the target function $\mathscr{Q}^{\pi^{\Phi(n)}}$ also depends on $\Phi(0)$. Thus, an approximation

$$\nabla^\top F_t(\bar{z}_t; \Phi(0))\omega_n^\star = \sum_{i=1}^m \frac{\nabla^\top H_t^{(i)}(\bar{z}_t; \Phi(0))\boldsymbol{v}^{\Phi(n)}(\Phi_i(0))}{m},$$

with $\omega_n^\star := [\frac{1}{\sqrt{m}} c_i \boldsymbol{v}^{\Phi(n)}(\Phi_i(0))]_{i \in [m]}$ for the transportation mapping $\boldsymbol{v}^{\Phi(n)} \in \mathscr{H}$ may not converge to the target function $\mathscr{Q}^{\pi^{\Phi(n)}}$ because of the correlated $\nabla^\top H_t^{(i)}(\bar{z}_t; \Phi(0))\boldsymbol{v}^{\Phi(n)}(\Phi_i(0))$ across $i \in [m]$ as argued in (Cayci et al., 2024). To address this, we characterize the uniform approximation error as in Proposition 7.8 for the random features of the actor RNN in approximating all $\mathscr{Q}^{\pi^{\Phi(n)}}$ for all n based on Rademacher complexity.

8. Conclusion

In this work, we have studied RNN-based policy evaluation and policy optimization methods with finite-time analyses. An important limitation of Rec-NPG is that it does not provide an effective solution in POMDPs that require long-term memory as we point out in Remarks 6.5-7.4. As an extension of this work, theoretical analyses of more complicated LSTM- (Hochreiter & Schmidhuber, 1997) and GRU-based (Chung et al., 2014) natural policy gradient algorithms can be considered as a future work. Alternatively, the study of hard- and soft-attention mechanisms to address the limitations of the RNNs (Murphy, 2022) in policy optimization is also a very interesting future direction.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

Cayci, S. and Eryilmaz, A. Convergence of gradient descent for recurrent neural networks: A nonasymptotic analysis. *arXiv preprint arXiv:2402.12241*, 2024.

Cayci, S., He, N., and Srikant, R. Finite-time analysis of natural actor-critic for pomdps. *arXiv preprint arXiv:2202.09753*, 2022.

Cayci, S., He, N., and Srikant, R. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=BkEqk7pS1I.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Cover, T. M. and Thomas, J. A. Elements of information theory (wiley series in telecommunications and signal processing), 2006.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Even, M., Pesme, S., Gunasekar, S., and Flammarion, N. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. Maxout networks. In *International conference on machine learning*, pp. 1319–1327. PMLR, 2013.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Grigoryeva, L. and Ortega, J.-P. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Ji, Z., Telgarsky, M., and Xian, R. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2019.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kara, A. D. and Yüksel, S. Convergence of finite memory q learning for pomdps and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 48(4):2066–2093, 2023.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

Konda, V. R. and Tsitsiklis, J. N. Onactor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Krishnamurthy, V. *Partially observed Markov decision processes*. Cambridge university press, 2016.

Lin, L.-J. and Mitchell, T. M. *Reinforcement Learning With Hidden States*, pp. 269–278. The MIT Press, April 1993. ISBN 9780262287159. doi: 10.7551/mitpress/3116.003.0038. URL http://dx.doi.org/10.7551/mitpress/3116.003.0038.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022.

Mnih, V., Heess, N., Graves, A., et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Murphy, K. P. A survey of pomdp solution techniques. *environment*, 2(10), 2000.

Murphy, K. P. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pp. 16270–16295. PMLR, 2022.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Rahimi, A., Recht, B., et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, pp. 5. Citeseer, 2007.

Schäfer, A. M. and Zimmermann, H.-G. Recurrent neural networks are universal approximators. *International journal of neural systems*, 17(04):253–263, 2007.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Singh, S. P., Jaakkola, T., and Jordan, M. I. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pp. 284–292. Elsevier, 1994.

Srebro, N., Rennie, J., and Jaakkola, T. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

Telgarsky, M. Deep learning theory lecture notes. https://mjt.cs.illinois.edu/dlt/, 2021. Version: 2021-10-27 v0.0-e7150f2d (alpha).

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Whitehead, S. D. and Lin, L.-J. Reinforcement learning of non-markov decision processes. *Artificial intelligence*, 73(1-2):271–306, 1995.

Wierstra, D., Förster, A., Peters, J., and Schmidhuber, J. Recurrent policy gradients. *Logic Journal of IGPL*, 18 (5):620–634, 2010.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Yu, H. A function approximation approach to estimation of policy gradient for pomdp with structured policies. *arXiv preprint arXiv:1207.1421*, 2012.

Yu, H. and Bertsekas, D. P. On near optimality of the set of finite-state controllers for average cost pomdp. *Mathematics of Operations Research*, 33(1):1–11, 2008.

A. Algorithmic Tools for Recurrent Neural Networks

A.1. Random Initialization for Recurrent Neural Networks

One key concept is random initialization, which is widely used in practice (Goodfellow et al., 2016) and yields the basis of the kernel analysis (Jacot et al., 2018; Chizat et al., 2019). In this work, we assume that m is even, and use the following symmetric initialization (Chizat et al., 2019).

Definition A.1 (Symmetric random initialization)**.** Let $c_i \sim \mathsf{Rad}(1), V_i \sim \mathsf{Rad}(\alpha), U_i(0) \sim \mathcal{N}(0, I_d)$ independently for all $i \in \{1, 2, \ldots, m/2\}$ and independently from each other, and $c_i = -c_{i-m/2}, V_i = V_{i-m/2}$ and $U_i(0) = U_{i-m/2}(0)$ for $i \in \{m/2+1, \ldots, m\}$. Then, $(\mathbf{W}(0), \mathbf{U}(0), c)$ is called a symmetric random initialization where $\mathbf{W}(0) = \mathsf{diag}_m(V)$ and $U_i^\top(0)$ is the i^{th}-row of $\mathbf{U}(0)$.

The symmetrization ensures that $F_t(\bar{z}_t; \mathbf{W}(0), \mathbf{U}(0), c) = 0$ for any $t \geq 0$ and input $\bar{z}_t$.

A.2. Max-Norm Regularization for Recurrent Neural Networks

Max-norm regularization, proposed by (Srebro et al., 2004), has been shown to be very effective across a broad spectrum of deep learning problems (Srivastava et al., 2014; Goodfellow et al., 2013). In this work, we incorporate max-norm regularization (around the random initialization) into the recurrent natural policy gradient for sharp convergence guarantees. To that end, given a random initialization $(\mathbf{W}(0), \mathbf{U}(0), c)$ as in Definition A.1 and a vector $\rho = (\rho_{\mathsf{w}}, \rho_{\mathsf{u}})^\top \in \mathbb{R}^2_{>0}$ of projection radii, we define the compactly-supported set of weights $\Omega_{\rho,m} \subset \mathbb{R}^{m(d+1)}$ as

$$\Omega_{\rho,m} = \mathcal{B}^{(m)}_{2,\infty}(\Theta(0), \rho). \tag{18}$$

Given any symmetric random initialization $(\mathbf{W}(0), \mathbf{U}(0), c)$ and $\rho \in \mathbb{R}^2_{>0}$, the set $\Omega_{\rho,m}$ is a compact and convex subset of $\mathbb{R}^{m(d+1)}$, and for any $\Theta \in \Omega_{\rho,m}$, we have

$$\max_{1\leq i\leq m} |W_{ii} - W_{ii}(0)| \leq \frac{\rho_{\mathsf{w}}}{\sqrt{m}},$$
$$\max_{1\leq i\leq m} \|U_i - U_i(0)\| \leq \frac{\rho_{\mathsf{u}}}{\sqrt{m}}.$$

Let

$$\mathbf{Proj}_{\Omega_{\rho,m}}[\Theta] = \left[\underset{w\in\mathcal{B}_2\left(W_{ii}(0), \frac{\rho_{\mathsf{w}}}{\sqrt{m}}\right)}{\arg\min} |W_{ii} - w_i|, \underset{u_i\in\mathcal{B}_2\left(U_i(0), \frac{\rho_{\mathsf{u}}}{\sqrt{m}}\right)}{\arg\min} \|\mathbf{U}_i - u_i\|_2\right]_{i\in[m]} \tag{19}$$

As such, the projection operator $\mathbf{Proj}_{\Omega_{\rho,m}}[\cdot]$ onto $\Omega_{\rho,m}$ is called the max-norm projection (or regularization).

Note that we have $\|\mathbf{W} - \mathbf{W}(0)\|_2 \leq \rho_{\mathsf{w}}$, $\|\mathbf{U} - \mathbf{U}(0)\|_2 \leq \rho_{\mathsf{u}}$ and $\|\Theta - \Theta(0)\|_2 \leq \|\rho\|_2$ in the ℓ_2 geometry for any $\Theta \in \Omega_{\rho,m}$. Therefore, although the max-norm parameter class $\Omega_{\rho,m} \subset \{\Theta \in \mathbb{R}^{m(d+1)} : \|\Theta - \Theta(0)\|_2 \leq \|\rho\|_2\}$, the ℓ_2-projected (Cai et al., 2019; Wang et al., 2019; Liu et al., 2019) and max-norm projected (Cayci et al., 2024) optimization algorithms recover exactly the same function class (i.e., RKHS associated with the neural tangent kernel studied in (Ji et al., 2019; Telgarsky, 2021), see Section 4).

B. Proofs for Section 6

An important quantity in the analysis of recurrent neural networks is the following:

$$\Gamma_t^{(i)}(\bar{z}_t; \Theta) := W_{ii} H_t^{(i)}(\bar{z}_t; \Theta),$$

for any hidden unit $i \in [m]$ and $\Theta \in \mathbb{R}^{m(d+1)}$. The following Lipschitzness and smoothness results for $\Theta_i \mapsto H_t^{(i)}(\bar{z}_t; \Theta)$ and $\Theta_i \mapsto \Gamma_t^{(i)}(\bar{z}_t; \Theta)$.

Lemma B.1 (Local continuity of hidden states; Lemma 1-2 in (Cayci & Eryilmaz, 2024))**.** *Given $\rho \in \mathbb{R}^2_{>0}$ and $\alpha \geq 0$, let $\alpha_m = \alpha + \frac{\rho_w}{\sqrt{m}}$. Then, for any $\bar{\boldsymbol{z}} \in (\mathbb{Y} \times \mathbb{A})^{\bar{Z}_+}$ with $\sup_{t\in\mathbb{N}} \left\|\begin{pmatrix} y_t \\ a_t \end{pmatrix}\right\|_2 \leq 1$, $t \in \mathbb{N}$ and $i \in [m]$,*

- $\Theta_i \mapsto H_t^{(i)}(\bar{z}_t;\Theta)$ *is* L_t*-Lipschitz continuous with* $L_t = (\varrho_0^2+1)\varrho_0^2 \cdot p_t^2(\alpha_m\varrho_1)$,
- $\Theta_i \mapsto H_t^{(i)}(\bar{z}_t;\Theta)$ *is* β_t*-smooth with* $\beta_t = \mathscr{O}\left(d \cdot p(\alpha_m\varrho_1)\cdot q(\alpha_m\varrho_1)\right)$,
- $\Theta_i \mapsto \Gamma_t^{(i)}(\bar{z}_t;\Theta)$ *is* Λ_t*-Lipschitz with* $\Lambda_t = \sqrt{2}(\varrho_0 + 1 + \alpha_m L_t)$,
- $\Theta_i \mapsto \Gamma_t^{(i)}(\bar{z}_t;\Theta)$ *is* χ_t*-smooth with* $\chi_t = \sqrt{2}(L_t + \alpha_m\beta_t)$,

in $\Omega_{\rho,m}$*. Consequently, for any* $\Theta\in\Omega_{\rho,m}$*,*

$$\sup_{\bar{\boldsymbol{z}}\in\bar{\mathbb{H}}_\infty}\max_{0\le t\le T}|F_t(\bar{z}_t;\Theta)| \le L_T\cdot\|\rho\|_2,\ T\in\mathbb{N}, \tag{20}$$

$$\sup_{\bar{\boldsymbol{z}}\in\bar{\mathbb{H}}_\infty}|F_t^{\mathsf{Lin}}(\bar{z}_t;\Theta) - F_t(\bar{z}_t;\Theta)| \le \frac{2}{\sqrt{m}}(\varrho_2\Lambda_t^2 + \varrho_1\chi_t)\|\Theta-\Theta(0)\|_2^2,\ t\in\mathbb{N}, \tag{21}$$

$$\sup_{\bar{\boldsymbol{z}}\in\bar{\mathbb{H}}_\infty}\left\langle\nabla F_t(\bar{z}_t;\Theta) - \nabla F_t(\bar{z}_t;\Theta(0)), \Theta-\bar{\Theta}\right\rangle \le \frac{2\beta_t^2\|\rho\|_2^2}{\sqrt{m}}, \tag{22}$$

with probability 1 over the symmetric random initialization $(\mathbf{W}(0),\mathbf{U}(0),c)$*.*

Lemma B.2 (Approximation error between RNN-NTRF and RNN-NTK)**.** *Let* $f^\star\in\mathscr{F}$ *with the transportation mapping* $\boldsymbol{v}\in\mathscr{H}$*, and let*

$$\bar{\Theta}_i = \Theta_i(0) + \frac{1}{\sqrt{m}}c_i\boldsymbol{v}(\Theta_i(0)), i\in[m]. \tag{23}$$

for any symmetric random initialization $(\mathbf{W}(0),\mathbf{U}(0),c)$ *in Def. A.1. Let*

$$F_t^{\mathsf{Lin}}(\cdot;\Theta) = \nabla_\Theta F_t(\cdot;\Theta(0))\cdot(\Theta-\Theta(0)).$$

If $P_T^{\pi,\mu}$ *induces a compactly-supported marginal distribution for* $X_t, t\in\mathbb{N}$ *such that* $\|X_t\|_2\le 1$ *a.s. and* $\{\bar{Z}_t : t\in\mathbb{N}\}$ *is independent from the random initialization* $(\mathbf{W}(0),\mathbf{U}(0),c)$*, then we have*

$$\mathbb{E}\left[\mathbb{E}_\mu^\pi\left[\left(f_t^\star(\bar{Z}_t) - F_t^{\mathsf{Lin}}(\bar{Z}_t;\bar{\Theta})\right)^2\right]\right] \le \frac{2\|\nu\|_2^2(1+\varrho_0^2)p_t^2(\alpha\varrho_1)}{m}, \tag{24}$$

where the outer expectation is with respect to the random initialization $(\mathbf{W}(0),\mathbf{U}(0),c)$*.*

Proof. For any hidden unit $i\in[m]$, let

$$\zeta_i = \left\langle\boldsymbol{v}(\Theta_i(0)), \sum_{k=0}^t {W_{ii}}^k(0)\begin{pmatrix}H_{t-k-1}^{(i)}(\bar{Z}_{t-k-1},\Theta_i(0))\\ X_{t-k}\end{pmatrix}\prod_{j=0}^k\mathcal{I}_{t-j}(\bar{Z}_{t-j};\Theta_i(0))\right\rangle.$$

Then, it is straightforward to see that

$$F_t^{\mathsf{Lin}}(\bar{Z}_t;\bar{\Theta}) = \frac{1}{m}\sum_{i=1}^m\zeta_i, \tag{25}$$

and $\mathbb{E}[\zeta_i|\bar{Z}_t] = \mathbb{E}[f_t^\star(\bar{Z}_t)|\bar{Z}_t]$ almost surely. Note that $\{\zeta_i : i\in[m/2]\}$ is independent and identically distributed and $\zeta_i = \zeta_{i+m/2}$ for any $i\in[m/2]$. Also, with probability 1 we have

$$\begin{aligned}
|\zeta_i| &\overset{(\spadesuit)}{\le} \|\boldsymbol{v}(\Theta_i(0))\|_2\cdot\left\|\sum_{k=0}^t {W_{ii}}^k(0)\begin{pmatrix}H_{t-k-1}^{(i)}(\bar{Z}_{t-k-1},\Theta_i(0))\\ X_{t-k}\end{pmatrix}\prod_{j=0}^k\mathcal{I}_{t-j}(\bar{Z}_{t-j};\Theta_i(0))\right\|_2, \\
&\overset{(\clubsuit)}{\le} \|\boldsymbol{v}(\Theta_i(0))\|_2\sum_{k=0}^{t-1}\alpha^k\varrho_1^{k+1}\sqrt{1+\varrho_0^2}, \\
&\overset{(\diamondsuit)}{\le} \|\nu\|_2\cdot\varrho_1\cdot\sqrt{1+\varrho_0^2}\cdot p_t(\alpha\varrho_1),
\end{aligned}$$

where (♠) follows from Cauchy-Schwarz inequality, (♣) follows from the uniform bound $\sup_{z\in\mathbb{R}}|\varrho(z)| \leq \varrho_1$ and almost-sure bounds $\|X_k\|_2 \leq 1$ and $|W_{ii}(0)| \leq \alpha$, and (♣) follows from $\boldsymbol{v} \in \mathscr{H}_\nu$. From these bounds,

$$\mathrm{Var}(\zeta_i) \leq \mathbb{E}[\mathbb{E}_\mu^\pi[|\zeta_i|^2]] \leq \|\nu\|_2^2\varrho_1^2(1+\varrho_0)^2p_t^2(\alpha\varrho_1),\ i\in[m]. \tag{26}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_\mu^\pi\left[\left(f_t^\star(\bar{Z}_t) - F_t^{\mathsf{Lin}}(\bar{Z}_t;\bar{\Theta})\right)^2\right]\right] &= \mathbb{E}_\mu^\pi\left[\mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^m\left(\zeta_i - \mathbb{E}[\zeta_i|\bar{Z}_t]\right)\right|^2\right]\right],\\
&= \mathbb{E}_\mu^\pi\left[\mathbb{E}\left[\left|\frac{2}{m}\sum_{i=1}^{m/2}\left(\zeta_i - \mathbb{E}[\zeta_i|\bar{Z}_t]\right)\right|^2\right]\right],\\
&= \frac{4}{m^2}\mathbb{E}_\mu^\pi\sum_{i=1}^{m/2}\sum_{j=1}^{m/2}\mathbb{E}\left[\left(\zeta_i - \mathbb{E}[\zeta_i|\bar{Z}_t]\right)\left(\zeta_j - \mathbb{E}[\zeta_j|\bar{Z}_t]\right)\right],\\
&= \frac{4}{m^2}\mathbb{E}_\mu^\pi\sum_{i=1}^{m/2}\mathrm{Var}(\zeta_i) \leq \frac{2}{m}\|\nu\|_2^2\varrho_1^2(1+\varrho_0)^2p_t^2(\alpha\varrho_1),
\end{aligned}$$

where the first identity is from Fubini's theorem, the second identity is from the symmetricity of the random initialization, the fourth identity is due to the independent initialization for $i \leq m/2$, and the inequality is from the bound in equation 26.

□

Proposition B.3 (Non-stationary Bellman equation)**.** *For* $\pi \in \Pi_{\mathsf{NM}}$*, we have*

$$\mathcal{Q}_t^\pi(\bar{z}_t) = \mathbb{E}^\pi\left[r(S_t,A_t) + \gamma\mathcal{Q}_{t+1}^\pi(\bar{Z}_{t+1})\Big|\bar{Z}_t = \bar{z}_t\right] = \mathbb{E}^\pi\left[r(S_t,A_t) + \gamma\mathcal{V}_{t+1}^\pi(Z_{t+1})\Big|\bar{Z}_t = \bar{z}_t\right],$$

for any $t \in \mathbb{Z}_+$*.*

Proof of Theorem 6.3. Since $\{\mathcal{Q}_t^\pi : t \in \mathbb{N}\} \in \mathscr{F}$, let the point of attraction $\bar{\Theta}$ be defined as in equation 23, and the potential function be defined as

$$\Psi(\Theta) = \|\Theta - \bar{\Theta}\|_2^2. \tag{27}$$

Then, from the non-expansivity of the projection operator onto the convex set $\Omega_{\rho,m}$, we have the following inequality:

$$\Psi(\Theta(k+1)) \leq \Psi(\Theta(k)) + 2\eta\sum_{t=0}^{T-1}\gamma^t\delta_t(\bar{Z}_{t+1}^k;\Theta(k))\left\langle\nabla F_t(\bar{Z}_t^k;\Theta(k)),\Theta(k)-\bar{\Theta}\right\rangle + 2\eta^2\|\check{\mathcal{R}}_T(\bar{Z}_T^k;\Theta(k))\|_2^2. \tag{28}$$

Let $\check{\mathbb{E}}_t^k[\cdot] := \mathbb{E}[\cdot|\Theta(k),\ldots,\Theta(0),\bar{Z}_t^k]$. Then, we obtain

$$\begin{aligned}
\mathbb{E}[\Psi(\Theta(k+1) - \Psi(\Theta(k))] \leq 2\eta\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma^t\underbrace{\check{\mathbb{E}}_t^k[\delta_t(\bar{Z}_{t+1}^k;\Theta(k))]\left\langle\nabla F_t(\bar{Z}_t^k;\Theta(k)),\Theta(k)-\bar{\Theta}\right\rangle}_{(♠)_t}\Big]&\\
+\eta^2\mathbb{E}\underbrace{\|\check{\nabla}\mathcal{R}_T(\bar{Z}_T^k;\Theta(k))\|_2^2}_{(♣)}&.
\end{aligned} \tag{29}$$

Bounding $\mathbb{E}(♠)_t$. By using the Bellman equation in the non-Markovian setting (cf. Proposition B.3), notice that

$$\begin{aligned}
\check{\mathbb{E}}_t^k\delta_t(\bar{Z}_{t+1}^k;\Theta(k)) &= \check{\mathbb{E}}_t^k[r_t^k + \gamma F_{t+1}(\bar{Z}_{t+1}^k;\Theta(k)] - F_t(\bar{Z}_t^k;\Theta(k)),\\
&= \gamma\check{\mathbb{E}}_t^k\left[F_{t+1}(\bar{Z}_{t+1}^k;\Theta(k)) - \mathcal{Q}_{t+1}^\pi(\bar{Z}_{t+1}^k)\right] + \mathcal{Q}_t^\pi(\bar{Z}_t) - F_t(\bar{Z}_t^k;\Theta(k)).
\end{aligned}$$

Secondly, we perform a change-of-feature as follows:

$$\left\langle \nabla F_t(\bar{Z}_t^k;\Theta(k)), \Theta(k)-\bar{\Theta}\right\rangle = \left\langle \nabla F_t(\bar{Z}_t^k;\Theta(0)), \Theta(k)-\bar{\Theta}\right\rangle + \mathsf{err}_{t,k}^{(1)}, \tag{30}$$

where

$$\mathsf{err}_{t,k}^{(1)} := \left\langle \nabla F_t(\bar{Z}_t^k;\Theta(k)) - \nabla F_t(\bar{Z}_t^k;\Theta(0)), \Theta(k)-\bar{\Theta}\right\rangle, \text{ and } |\mathsf{err}_{t,k}^{(1)}| \leq \frac{2\beta_t^2\|\rho\|_2^2}{\sqrt{m}} \leq \frac{2\beta_T^2\|\rho\|_2^2}{\sqrt{m}},$$

by Lemma B.1. Furthermore,

$$\begin{aligned} \left\langle \nabla F_t(\bar{Z}_t^k;\Theta(0)), \Theta(k)-\bar{\Theta}\right\rangle &= F_t^{\mathsf{Lin}}(\bar{Z}_t^k;\Theta(k)) - F_t^{\mathsf{Lin}}(\bar{Z}_t^k;\bar{\Theta}), & (31)\\ &= F_t(\bar{Z}_t^k;\Theta(k)) - \mathcal{Q}_t^\pi(\bar{Z}_t^k) + \mathsf{err}_{t,k}^{(2)} + \mathsf{err}_{t,k}^{(3)} & (32)\end{aligned}$$

where

$$\begin{aligned}\mathsf{err}_{t,k}^{(2)} &:= F_t^{\mathsf{Lin}}(\bar{Z}_t^k;\Theta(k)) - F_t(\bar{Z}_t^k;\Theta(k)),\\ \mathsf{err}_{t,k}^{(3)} &:= -F_t^{\mathsf{Lin}}(\bar{Z}_t^k;\bar{\Theta}) + \mathcal{Q}_t^\pi(\bar{Z}_t^k).\end{aligned}$$

Thus,

$$\begin{aligned}(\spadesuit)_t = -(\mathcal{Q}_t^\pi(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\Theta(k)))^2 + \gamma\check{\mathbb{E}}_t^k\left[F_{t+1}(\bar{Z}_{t+1}^k;\Theta(k)) - \mathcal{Q}_{t+1}^\pi(\bar{Z}_{t+1}^k)\right]\cdot(\mathcal{Q}_t^\pi(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\Theta(k)))\\ + \check{\mathbb{E}}_t^k\delta_t(\bar{Z}_{t+1}^k;\Theta(k))\sum_{j=1}^3\mathsf{err}_{t,k}^{(j)}.\end{aligned}$$

By equation 20, we have

$$\sup_{\bar{\boldsymbol{z}}\in\bar{\mathbb{H}}_\infty}|\delta_t(\bar{z}_{t+1};\Theta(k))| \leq r_\infty + 2L_T\|\rho\|_2 =: \delta_{\mathsf{max}}$$

Now, let $\omega_{t,k} := \left(\mathbb{E}[(\mathcal{Q}_t^\pi(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\Theta(k)))^2]\right)^{1/2}$, where the expectation is over the joint distribution of $\Theta(k)$ and $\bar{Z}_T^k$. Then,

$$\mathbb{E}[(\spadesuit)_t] \leq -\omega_{t,k}^2 + \gamma\omega_{t+1,k}\omega_{t,k} + \delta_{\mathsf{max}}\sum_{j=1}^3\mathbb{E}|\mathsf{err}_{t,k}^{(j)}|.$$

From equation 21, we have

$$\mathbb{E}|\mathsf{err}_{t,k}^{(2)}| \leq \frac{2}{\sqrt{m}}(\varrho_2\Lambda_T^2 + \varrho_1\chi_T)\|\rho\|_2^2.$$

From the approximation bound in Lemma B.2, we get

$$\mathbb{E}|\mathsf{err}_{t,k}^{(3)}| \leq \sqrt{\mathbb{E}|\mathsf{err}_{t,k}^{(3)}|^2} \leq \frac{2\|\nu\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\alpha\varrho_1)}{\sqrt{m}}.$$

Also, note that $\omega_{t+1,k}\omega_{t,k} \leq \frac{1}{2}(\omega_{t,k}^2 + \omega_{t+1,k}^2)$. Putting these together, we obtain the following bound for every $t \in \{0,1,\ldots,T-1\}$:

$$\mathbb{E}[(\spadesuit)_t] \leq -\omega_{t,k}^2 + \frac{\gamma}{2}(\omega_{t+1,k}^2 + \omega_{t,k}^2) + \delta_{\mathsf{max}}\cdot\frac{C_T}{\sqrt{m}},$$

where

$$C_T := 2\beta_T^2\|\rho\|_2^2 + 2(\varrho_2\Lambda_T^2 + \varrho_1\chi_T)\|\rho\|_2^2 + 2\|\nu\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\alpha\varrho_1).$$

Hence, we obtain the following upper bound:

$$\begin{aligned}\sum_{t=0}^{T-1}\gamma^t\mathbb{E}[(\spadesuit)_t] &\leq -(1-\gamma/2)\sum_{t<T}\gamma^t\omega_{t,k}^2 + \frac{\delta_{\mathsf{max}}\cdot C_T}{(1-\gamma)\sqrt{m}} + \underbrace{\frac{1}{2}\sum_{t<T}\gamma^{t+1}\omega_{t+1,k}^2}_{\leq\frac{1}{2}(\sum_{t<T}\gamma^t\omega_{t,k}^2 + \gamma^T\omega_{T,k}^2)}\\ &\leq -\frac{1-\gamma}{2}\sum_{t<T}\gamma^t\omega_{t,k}^2 + \frac{1}{2}\gamma^T\omega_{T,k}^2 + \frac{C_T\cdot\delta_{\mathsf{max}}}{(1-\gamma)\sqrt{m}}. \qquad (33)\end{aligned}$$

Bounding $\mathbb{E}[(\clubsuit)]$. Using the triangle inequality, we obtain:

$$\|\sum_{t<T}\gamma^t\delta_t(\bar{Z}_{t+1}^k;\Theta(k))\nabla F_t(\bar{Z}_t;\Theta(k))\|_2 \leq \sum_{t<T}\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\Theta(k))|\cdot\|\nabla F_t(\bar{Z}_t;\Theta(k))\|_2.$$

Since $\Theta(k)\in\Omega_{\rho,m}$ for every $k\in\mathbb{N}$ as a consequence of the max-norm regularization, we have

$$|\delta_t(\bar{Z}_{t+1}^k;\Theta(k))| \leq \delta_{\mathsf{max}} = r_\infty + 2L_T\|\rho\|_2,$$
$$\|\nabla F_t(\bar{Z}_t^k;\Theta(k))\|_2^2 = \frac{1}{m}\sum_{i=1}^m\|\nabla_{\Theta_i}H_t^{(i)}(\bar{Z}_t^k;\Theta(k))\|_2^2 \leq L_t^2 \leq L_T^2,$$

for every $t<T$ with probability 1 since $\Theta_i\mapsto H_t^{(i)}(\bar{z}_t;\Theta_i)$ is L_t-Lipschitz continuous by Lemma B.1. Hence, we obtain:

$$\|\check{\nabla}\mathcal{R}_T(\bar{Z}_T^k;\Theta(k))\|_2 \leq \frac{\delta_{\mathsf{max}}L_T}{1-\gamma}. \tag{34}$$

Final step. Now, taking expectation over $(\bar{Z}_t^k,\Theta(k))$ in equation 29, and substituting equation 33 and equation 34, we obtain:

$$\mathbb{E}[\Psi(\Theta(k+1))-\Psi(\Theta(k))] \leq -\eta(1-\gamma)\sum_{t=0}^{T-1}\gamma^t\omega_{t,k}^2 + \eta\gamma^T\omega_{T,k}^2 + \eta\frac{\delta_{\mathsf{max}}\cdot C_T}{(1-\gamma)\sqrt{m}} + \eta^2\frac{\delta_{\mathsf{max}}^2L_T^2}{(1-\gamma)^2},$$

for every $k\in\mathbb{N}$. Note that $\Psi(\Theta(0))\leq\|\nu\|_2^2$. Thus, telescoping sum over $k=0,1,\ldots,K-1$ yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathcal{R}_T(\Theta(k)) \leq \frac{\|\nu\|_2^2}{\eta(1-\gamma)K} + \frac{\eta\delta_{\mathsf{max}}^2L_T^2}{(1-\gamma)^3} + \frac{\delta_{\mathsf{max}}\cdot C_T}{(1-\gamma)^2\sqrt{m}} + \frac{\gamma^T}{(1-\gamma)K}\sum_{k=0}^{K-1}\omega_{T,k}^2. \tag{35}$$

The final inequality in the proof stems from the linearization result Lemma B.2, and directly follows from

$$\mathcal{R}_T\left(\frac{1}{K}\sum_{k<K}\Theta(k)\right) \leq \frac{4}{K}\sum_{k<K}\mathcal{R}_T(\Theta(k)) + \frac{6}{\sqrt{m}}\left(\varrho_2\Lambda_T^2+\varrho_1\chi_T\right)\|\rho\|_2^2,$$

which directly follows from (Cayci & Eryilmaz, 2024), Corollary 1. □

In the following, we study the error under mean-path Rec-TD learning algorithm.

Theorem B.4 (Finite-time bounds for mean-path Rec-TD). *For $K\in\mathbb{N}$, with the step-size choice $\eta=\frac{(1-\gamma)^2}{64L_T^2}$, mean-path Rec-TD learning achieves the following error bound:*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k<K}\mathcal{R}_T^\pi(\Theta(k))\right] \leq \frac{2\|\nu\|_2^2}{(1-\gamma)\eta K} + \frac{\gamma^T\omega_{T,k}}{1-\gamma} + \frac{C_T\delta_{\mathsf{max}}}{(1-\gamma)^2\sqrt{m}} + \eta\left(\frac{(C_T')^2}{m} + 16\gamma^{2T}L_T^4(\|\rho\|_2^2+\|\nu\|_2^2)\right),$$

where C_T' and L_T are terms that do not depend on K.

Theorem B.4 indicates that if a noiseless semi-gradient is used in Rec-TD, then the rate can be improved from $\mathscr{O}\left(\frac{1}{\sqrt{K}}\right)$ to $\mathscr{O}\left(\frac{1}{K}\right)$, indicating the potential limits of using variance-reduction schemes.

Proof of Theorem B.4. At any iteration $k\in\mathbb{N}$, let

$$\bar{\nabla}\mathcal{R}_T(\Theta(k)) := \mathbb{E}_\mu^\pi\Big[\check{\nabla}\mathcal{R}(\bar{Z}_t^k;\Theta(k))\Big], \tag{36}$$

be the ***mean-path semi-gradient***. First, note that

$$\|\bar{\nabla}\mathcal{R}_T(\Theta(k))\|_2^2 \leq 2\|\bar{\nabla}\mathcal{R}_T(\Theta(k))-\bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2 + 2\|\bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2. \tag{37}$$

Bounding $\|\bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2$. For any $k \in \mathbb{N}, t \leq T$, we have

$$\mathbb{E}\big[\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})|\bar{Z}_t^k,\Theta(0),c\big] = \gamma\mathbb{E}[F_{t+1}(\bar{Z}_{t+1}^k;\bar{\Theta}) - \mathcal{Q}_{t+1}^{\pi}(\bar{Z}_{t+1}^k)|\bar{Z}_t^k,\Theta(0),c] + \mathcal{Q}_t^{\pi}(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\bar{\Theta}).$$

Since $\|\nabla F_t(\bar{z}_t;\bar{\Theta})\|_2 \leq L_t$, the following inequality holds:

$$\begin{aligned}\big\|\mathbb{E}\big[\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})\nabla F_t(\bar{Z}_t^k;\bar{\Theta})\big]\big\|_2 &\leq \mathbb{E}\,\big\|\mathbb{E}\big[\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})|\bar{Z}_t^k,\Theta(0),c\big]\nabla F_t(\bar{Z}_t^k;\bar{\Theta})\big\|_2\,,\\ &\leq L_T\mathbb{E}\,\big|\mathbb{E}\big[\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})|\bar{Z}_t^k,\Theta(0),c\big]\big|\,,\\ &\leq L_T\left(\gamma\mathbb{E}\left|F_{t+1}(\bar{Z}_{t+1}^k;\bar{\Theta}) - \mathcal{Q}_{t+1}^{\pi}(\bar{Z}_{t+1}^k)\right| + \mathbb{E}\left|\mathcal{Q}_t^{\pi}(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\bar{\Theta})\right|\right), \end{aligned} \tag{38}$$

where we used Jensen's inequality, the law of iterated expectations, and triangle inequality. From the above inequality, we obtain

$$\begin{aligned}\|\bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2 &\overset{\textcircled{1}}{\leq} \sum_{t=0}^{T-1}\gamma^t\big\|\mathbb{E}\big[\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})\nabla F_t(\bar{Z}_t^k;\bar{\Theta})\big]\big\|_2,\\ &\overset{\textcircled{2}}{\leq} L_T\gamma\sum_{t<T}\gamma^t\mathbb{E}|F_{t+1}(\bar{Z}_{t+1}^k;\bar{\Theta}) - \mathcal{Q}_{t+1}^{\pi}(\bar{Z}_{t+1}^k)| + L_T\sum_{t<T}\gamma^t\mathbb{E}|\mathcal{Q}_t^{\pi}(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\bar{\Theta})|,\\ &\overset{\textcircled{3}}{\leq} \frac{L_T}{\sqrt{1-\gamma}}\left(\gamma\mathbb{E}\sqrt{\sum_{t<T}\gamma^t|F_{t+1}(\bar{Z}_{t+1}^k;\bar{\Theta}) - \mathcal{Q}_{t+1}^{\pi}(\bar{Z}_{t+1}^k)|^2} + \mathbb{E}\sqrt{\sum_{t<T}\gamma^t|F_t(\bar{Z}_t^k;\bar{\Theta}) - \mathcal{Q}_t^{\pi}(\bar{Z}_t^k)|^2}\right),\\ &\overset{\textcircled{4}}{\leq} \frac{L_T}{\sqrt{1-\gamma}}\left(\gamma\sqrt{\mathbb{E}\sum_{t<T}\gamma^t|F_{t+1}(\bar{Z}_{t+1}^k;\bar{\Theta}) - \mathcal{Q}_{t+1}^{\pi}(\bar{Z}_{t+1}^k)|^2} + \sqrt{\mathbb{E}\sum_{t<T}\gamma^t|F_t(\bar{Z}_t^k;\bar{\Theta}) - \mathcal{Q}_t^{\pi}(\bar{Z}_t^k)|^2}\right),\\ &\overset{\textcircled{5}}{\leq} \frac{\sqrt{2}(1+\gamma)L_T}{\sqrt{1-\gamma}}\frac{\|\nu\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\varrho_1\alpha)}{\sqrt{m}}.\end{aligned}$$

where ① follows from triangle inequality, ② follows from equation 38, ③ follows from Cauchy-Schwarz inequality and the monotonicity of the geometric series $T \mapsto \sum_{t<T}\gamma^t$, ④ follows from Jensen's inequality, and finally ⑤ follows from Lemma B.2. Hence, we obtain

$$\|\bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2 \leq \frac{8L_T^2\|\nu\|_2^2(1+\varrho_0^2)p_T^2(\varrho_1\alpha)}{(1-\gamma)m}. \tag{39}$$

Bounding $\|\bar{\nabla}\mathcal{R}_T(\Theta(k)) - \bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2$. First, note that

$$\|\bar{\nabla}\mathcal{R}_T(\Theta(k)) - \bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2 = \|\mathbb{E}\Big[\sum_{t<T}\gamma^t\left(\delta_t(\bar{Z}_{t+1}^k;\Theta(k))\nabla F_t(\bar{Z}_t^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})\nabla F_t(\bar{Z}_t^k;\bar{\Theta})\right)\Big]\|_2$$

We make the following decomposition for each $t < T$:

$$\begin{aligned}\delta_t(\bar{Z}_{t+1}^k;\Theta(k))\nabla F_t(\bar{Z}_t^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})\nabla F_t(\bar{Z}_t^k;\bar{\Theta}) &= \delta_t(\bar{Z}_{t+1}^k;\Theta(k))\left(\nabla F_t(\bar{Z}_t^k;\Theta(k)) - \nabla F_t(\bar{Z}_t^k;\bar{\Theta})\right)\\ &\quad + \nabla F_t(\bar{Z}_t^k;\Theta(k))\left(\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta}) - \delta_t(\bar{Z}_{t+1}^k;\Theta(k))\right)\end{aligned} \tag{40}$$

By Lemma B.1, we have $|\delta_t(\bar{Z}_{t+1}^k;\Theta)| \leq \delta_{\mathsf{max}}$ and $\|\nabla F_t(\bar{Z}_t^k;\Theta)\|_1 \leq L_t \leq L_T$ almost surely for any $\Theta \in \Omega_{\rho,m}$, which holds for $\Theta(k)$ (due to the max-norm projection) and $\bar{\Theta}$. As such, by triangle inequality,

$$\begin{aligned}\|\bar{\nabla}\mathcal{R}_T(\Theta(k)) - \bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2 &\leq \sum_{t<T}\gamma^t\left(\delta_{\mathsf{max}}\frac{\beta_t^2\mathbb{E}\|\Theta(k) - \bar{\Theta}\|_2^2}{m} + L_t\mathbb{E}|\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta}) - \delta_t(\bar{Z}_{t+1}^k;\Theta(k))|\right),\\ &\leq \underbrace{\frac{\delta_{\mathsf{max}}\beta_T^2(\|\rho\|_2^2 + \|\nu\|_2^2)}{m(1-\gamma)}}_{=:\frac{C_T^{(4)}}{m}} + L_T\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\bar{\Theta}) - \delta_t(\bar{Z}_{t+1}^k;\Theta(k))|\right]\end{aligned} \tag{41}$$

Note that

$$\begin{aligned}\sum_{t<T}\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})| &= \sum_{t<T}\gamma^t\Big(|F_{t+1}(\bar{Z}_{t+1}^k;\bar{\Theta}) - F_{t+1}(\bar{Z}_{t+1}^k;\Theta(k))| + |F_t(\bar{Z}_t^k;\bar{\Theta}) - F_t(\bar{Z}_t^k;\Theta(k))|\Big),\\ &\leq 2\sum_{t<T}\gamma^t\Big|F_t(\bar{Z}_t^k;\bar{\Theta}) - F_t(\bar{Z}_t^k;\Theta(k))\Big| + \gamma^T L_T\|\Theta(k)-\bar{\Theta}\|_2, \qquad (42)\end{aligned}$$

where the second line follows from the Lipschitz continuity of $\Theta \mapsto F_t(\cdot;\Theta)$. Then, adding and subtracting $\mathcal{Q}_t^\pi$ to each term, we obtain

$$\begin{aligned}\sum_{t<T}\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})| \leq 2\sum_{t<T}\gamma^t\left(|F_t(\bar{Z}_t^k;\bar{\Theta}) - \mathcal{Q}_t^\pi(\bar{Z}_t^k)| + |\mathcal{Q}_t^\pi(\bar{Z}_t^k) - F_t(\bar{Z}_t^k;\Theta(k))|\right)\\ + \gamma^T L_T\|\Theta(k)-\bar{\Theta}\|_2. \quad (43)\end{aligned}$$

Taking expectation, we obtain

$$\begin{aligned}\mathbb{E}\sum_{t<T}\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})| &\leq \frac{2}{\sqrt{1-\gamma}}\sqrt{\mathbb{E}\left[\sum_{t<T}\gamma^t|F_t(\bar{Z}_t^k;\Theta(k)) - \mathcal{Q}_t^\pi(\bar{Z}_t^k)|^2\right]}\\ &+ \frac{2}{\sqrt{1-\gamma}}\sqrt{\mathbb{E}\left[\sum_{t<T}\gamma^t|F_t(\bar{Z}_t^k;\bar{\Theta}) - \mathcal{Q}_t^\pi(\bar{Z}_t^k)|^2\right]} + \gamma^T L_T\|\Theta(k)-\bar{\Theta}\|_2.\end{aligned}$$

By Lemma B.2 and equation 21, we have

$$\mathbb{E}|F_t(\bar{Z}_t^k;\bar{\Theta}) - \mathcal{Q}_t^\pi(\bar{Z}_t^k)|^2 \leq \frac{4}{m}\|\nu\|_2^2\varrho_1^2(1+\varrho_0)^2p_t^2(\alpha\varrho_1) + \frac{4}{m}(\varrho_2\Lambda_T^2+\varrho_1\chi_T)^2\|\rho\|_2^4,$$

for any $t<T$. Thus,

$$\begin{aligned}\mathbb{E}\sum_{t<T}\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})| &\leq \frac{2}{\sqrt{1-\gamma}}\sqrt{\mathbb{E}\left[\sum_{t<T}\gamma^t|F_t(\bar{Z}_t^k;\Theta(k)) - \mathcal{Q}_t^\pi(\bar{Z}_t^k)|^2\right]}\\ &+ \frac{1}{\sqrt{m}}\underbrace{\frac{4}{\sqrt{(1-\gamma)^3}}\left(\|\nu\|_2\varrho_1(1+\varrho_0)p_T(\alpha\varrho_1) + (\varrho_2\Lambda_T^2+\varrho_1\chi_T)\|\rho\|_2^2\right)}_{=:C_T^{(3)}} + \gamma^T L_T\underbrace{\|\Theta(k)-\bar{\Theta}\|_2}_{\leq\|\rho\|_2+\|\nu\|_2}.\end{aligned}$$

This results in the following bound:

$$\mathbb{E}\sum_{t<T}\left[\gamma^t|\delta_t(\bar{Z}_{t+1}^k;\Theta(k)) - \delta_t(\bar{Z}_{t+1}^k;\bar{\Theta})|\right] \leq \frac{2}{\sqrt{1-\gamma}}\sqrt{\mathcal{R}_T(\Theta(k))} + \frac{C_T^{(3)}}{\sqrt{m}} + \gamma^T L_T(\|\rho\|_2+\|\nu\|_2). \qquad (44)$$

Substituting the local smoothness result in equation 44 into equation 41, we obtain

$$\|\bar{\nabla}\mathcal{R}_T(\Theta(k)) - \bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2 \leq L_T\left(\frac{2}{\sqrt{1-\gamma}}\sqrt{\mathcal{R}_T(\Theta(k))} + \frac{C_T^{(3)}}{\sqrt{m}} + \gamma^T L_T(\|\rho\|_2+\|\nu\|_2)\right) + \frac{C_T^{(4)}}{m}.$$

Thus, we obtain

$$\|\bar{\nabla}\mathcal{R}_T(\Theta(k)) - \bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2 \leq \frac{16L_T^2}{1-\gamma}\mathcal{R}_T(\Theta(k)) + \frac{4(C_T^{(3)})^2L_T^2 + 4(C_T^{(4)})^2}{m} + 8\gamma^{2T}L_T^4(\|\rho\|_2^2+\|\nu\|_2^2). \qquad (45)$$

Using equation 39 and equation 45 together, we obtain

$$\begin{aligned}\|\bar{\nabla}\mathcal{R}_T(\Theta(k))\|_2^2 &\leq 2\|\bar{\nabla}\mathcal{R}_T(\Theta(k)) - \bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2 + 2\|\bar{\nabla}\mathcal{R}_T(\bar{\Theta})\|_2^2,\\ &\leq \frac{32L_T^2\mathcal{R}_T(\Theta(k))}{1-\gamma} + \frac{(C_T')^2}{m} + 16\gamma^{2T}L_T^4(\|\rho\|_2^2+\|\nu\|_2^2). \qquad (46)\end{aligned}$$

In the final step, we use equation 29, equation 33 and equation 46 together:

$$\mathbb{E}\left[\Psi(\Theta(k+1)) - \Psi(\Theta(k))\right] \leq -\eta(1-\gamma)\mathbb{E}\mathcal{R}_T(\Theta(k)) + \eta\gamma^T\omega_{T,k} + \eta\frac{C_T\delta_{\mathsf{max}}}{(1-\gamma)\sqrt{m}} \\ + \eta^2\left(\frac{32L_T^2\mathbb{E}\mathcal{R}_T(\Theta(k))}{1-\gamma} + \frac{(C_T')^2}{m} + 16\gamma^{2T}L_T^4(\|\rho\|_2^2 + \|\nu\|_2^2)\right), \quad (47)$$

where the expectation is over the random initialization. Choosing $\eta = \frac{(1-\gamma)^2}{64L_T^2}$, we obtain

$$\mathbb{E}[\Psi(\Theta(k+1)) - \Psi(\Theta(k))] \leq -\frac{\eta(1-\gamma)}{2}\mathbb{E}\mathcal{R}_T(\Theta(k)) + \eta\gamma^T\omega_{T,k} + \eta\frac{C_T\delta_{\mathsf{max}}}{(1-\gamma)\sqrt{m}} \\ + \eta^2\left(\frac{(C_T')^2}{m} + 16\gamma^{2T}L_T^4(\|\rho\|_2^2 + \|\nu\|_2^2)\right). \quad (48)$$

Telescoping sum over $k = 0, 1, \ldots, K-1$, and re-arranging terms, we obtain:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k<K}\mathcal{R}_T(\Theta(k))\right] \leq \frac{2\|\nu\|_2^2}{(1-\gamma)\eta K} + \frac{\gamma^T\omega_{T,k}}{1-\gamma} + \frac{C_T\delta_{\mathsf{max}}}{(1-\gamma)^2\sqrt{m}} + \eta\left(\frac{(C_T')^2}{m} + 16\gamma^{2T}L_T^4(\|\rho\|_2^2 + \|\nu\|_2^2)\right). \quad (49)$$

□

C. Numerical Experiments for Rec-TD

In the following, we will demonstrate the numerical performance of Rec-TD for a given non-Markovian policy π^{greedy}.

POMDP setting. We consider a randomly-generated finite POMDP instance with $|\mathbb{S}| = |\mathbb{Y}| = 8$, $|\mathbb{A}| = 4$, $r(s,a) \sim \mathsf{Unif}[0,1]$ for all $(s,a) \in \mathbb{S}\times\mathbb{A}$. For a fixed ambient dimension $d = 8$, we use a random feature mapping $(y,a) \mapsto \varphi(y,a) \sim \mathcal{N}(0, I_d)$, $\forall (y,a) \sim \mathbb{Y}\times\mathbb{A}$.

Greedy policy. Let

$$j^\star(t) \in \arg\max_{0\leq j<t} r_j,$$

be the instance before t at which the maximum reward was obtained, and let

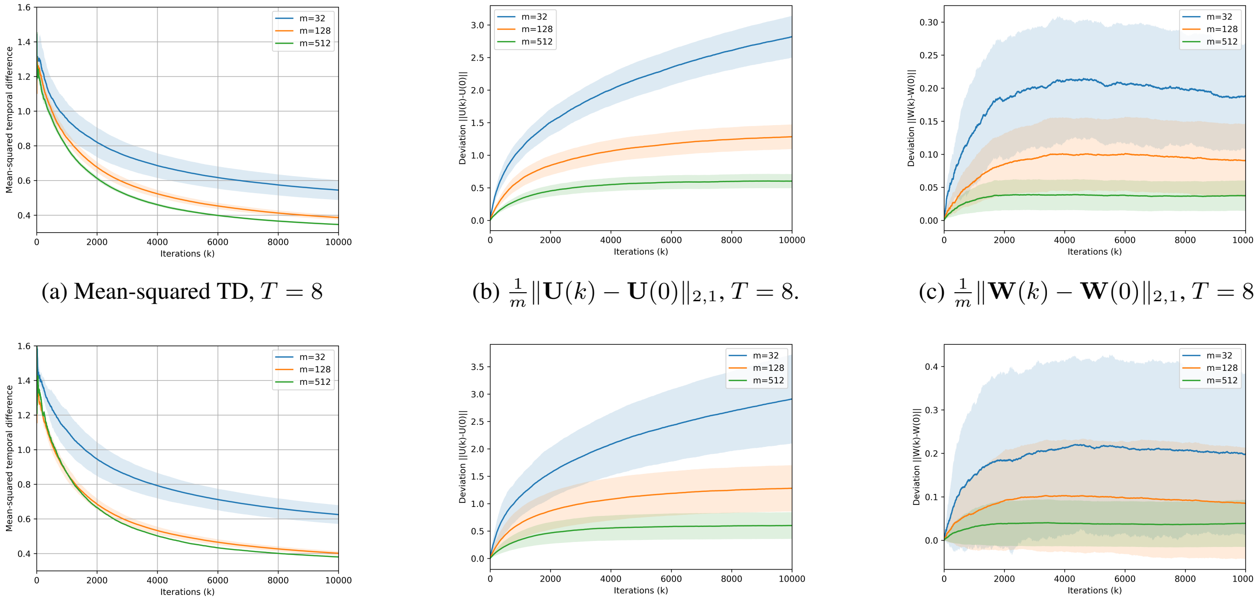
$$\pi_t^{\mathsf{greedy}}(a|Z_t) = \begin{cases} \frac{1}{|\mathbb{A}|}, & \text{w.p. } \min\{\frac{2+t}{10}, p_{\mathsf{exp}}\}, \\ \mathbb{1}_{a=A_{j^\star(t)}}, & \text{w.p. } 1-\min\{\frac{2+t}{10}, p_{\mathsf{exp}}\}, \end{cases} \quad (50)$$

be the greedy policy with a user-specified exploration probability $p_{\mathsf{exp}} \in (0,1)$. The long-term dependencies in this greedy policy is obviously controlled by p_{exp}: a small exploration probability will make the policy (thus, the corresponding $\mathcal{Q}$-functions) more history-dependent. Since the exact computation of $(\mathcal{Q}_t^\pi)_{t\in\mathbb{N}}$ is highly intractable for POMDPs, we use (empirical) mean-square temporal difference (MSTD) [2] as a surrogate loss.

Example 1 (Short-term memory). We first consider the performance of Rec-TD with learning rate $\eta = 0.05$, discount factor $\gamma = 0.7$ and RNNs with various choices of network width m. For $p_{\mathsf{exp}} = 0.8$, the performance of Rec-TD is demonstrated in Figure 2. Consistent with the theoretical results in Theorem 6.3, Rec-TD (1) achieves smaller error with larger network width m, (2) requires smaller deviation from the random initialization $\Theta(0)$, which is known as the *lazy training* phenomenon. Since $\|\mathbf{W}(k)\|_{2,\infty} \leq 1$ due to large enough p_{exp} that avoids long-term dependencies, the problem exhibits a weak memory behavior. This is observed in Figures 2d-2f without a visible increase in the MSTD performance despite a significant 3-fold increase in T, consistent with the theoretical findings in Theorem 6.3.

Example 2 (Long-term memory). In the second example, we consider the same POMDP with a discount factor $\gamma = 0.9$. The exploration probability is reduced to $p_{\mathsf{exp}} = 0.3$, which leads to longer dependency on the history. This impact can be observed in Figure 3b-3d, which implies a larger spectral radius compared to Example 1 (in comparison with Figures 2c-2f). As a consequence of the long-term dependencies, increasing T from 8 to 32 leads to a dramatic increase in the MSTD unlike the weak-memory system in Example 1. The impact of a larger network size (i.e., m) is very significant in this example: choosing $m = 512$ leads to a dramatic improvement in the performance.

[2] the empirical mean of independently sampled $\left\{\frac{1}{k}\sum_{s<k}\hat{\mathcal{R}}_T^{\mathsf{TD}}(\Theta(s)) : k \in \mathbb{N}\right\}$ where $\hat{\mathcal{R}}_T^{\mathsf{TD}}(\Theta(k)) = \sum_{t=0}^{T-1}\gamma^t\delta_t^2(\bar{Z}_t^k; \Theta(k))$.



(a) Mean-squared TD, $T = 8$

(b) $\frac{1}{m}\|\mathbf{U}(k) - \mathbf{U}(0)\|_{2,1}, T = 8.$

(c) $\frac{1}{m}\|\mathbf{W}(k) - \mathbf{W}(0)\|_{2,1}, T = 8.$

(d) Mean-squared TD, $T = 32$

(e) $\frac{1}{m}\|\mathbf{U}(k) - \mathbf{U}(0)\|_{2,1}, T = 32.$

(f) $\frac{1}{m}\|\mathbf{W}(k) - \mathbf{W}(0)\|_{2,1}, T = 32.$

Figure 2. Mean-square TD and parameter movement under Rec-TD for the case $p_{\mathsf{min}} = 0.8$ and $\gamma = 0.7$. The mean curve and confidence intervals (90%) in Figures 2a and 2d stem from 5 trials. The 90% confidence intervals in Figures 2b-2c and 2e-2f correspond to deviations (i.e., $\|U_i(k) - U_i(0)\|_2$ and $|W_{ii}(k) - W_{ii}(0)|$) across different units $i \in [m]$ in a single trial.

D. Policy Gradients under Partial Observability

In this section, we will provide basic results for policy gradients under POMDPs, which is critical to develop the natural policy gradient method for POMDPs.

Proposition D.1. *Let $\pi' \in \Pi_{\mathsf{NM}}$ be an admissible policy, and let $\bar{Z}_T \sim P_T^{\pi',\mu}$. Then, for any $t < T$, conditional distribution of S_t given $\bar{Z}_t$ is independent of π'. Furthermore, for any $\pi \in \Pi_{\mathsf{NM}}$, the conditional distribution of $r(S_t, A_t) + \gamma \mathcal{V}_{t+1}^{\pi}(Z_{t+1})$ given $\bar{Z}_t$ is independent of π'.*

Proof of Prop. D.1. Let the belief at time $t \in \mathbb{N}$ be defined as

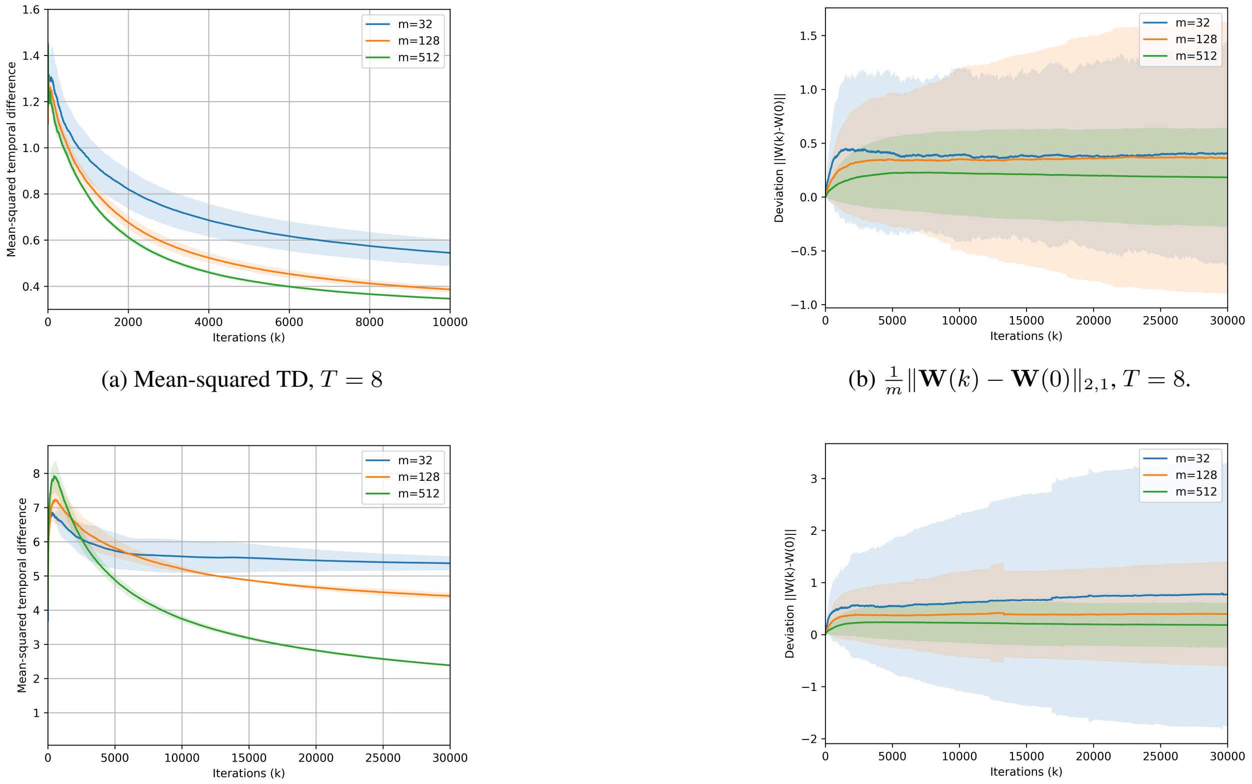
$$b_t(s) := \mathbb{P}(S_t = s|\bar{Z}_t). \tag{51}$$

For any non-stationary admissible policy π, the belief function is policy-independent. To see this, note that

$$\begin{aligned}
\mathbb{P}(S_t = s_t, \bar{Z}_t = \bar{z}_t) &= \sum_{(s_0,\ldots,s_{t-1})\in\mathbb{S}^t} \mathbb{P}(S_0 = s_0|Y_0 = y)\pi_0(a_0|z_0) \prod_{k=0}^{t-1} \mathcal{P}(s_{k+1}|s_k, a_k)\phi(y_{k+1}|s_{k+1})\pi_{k+1}(a_{k+1}|z_{k+1}), \\
&= \left(\prod_{k=0}^{t} \pi_k(a_k|z_k)\right) \sum_{(s_0,\ldots,s_{t-1})\in\mathbb{S}^t} \mathbb{P}(S_0 = s_0|Y_0 = y) \prod_{k=0}^{t-1} \mathcal{P}(s_{k+1}|s_k, a_k)\phi(y_{k+1}|s_{k+1}),
\end{aligned}$$

since $\prod_{k=0}^{t} \pi_k(a_k|z_k)$ does not depend on the summands $(s_0, \ldots, s_{t-1})$ – note that we use the notation $\mathcal{P}(s_{k+1}|s_k, a_k) := \mathcal{P}(s_k, a_k, \{S_{k+1} = s_{k+1}\})$ and $\phi(y_k|s_k) := \phi(s_k, \{Y_k = y_k\})$. Thus,

$$b_t(s_t) = \frac{\sum_{(s_0,\ldots,s_{t-1})\in\mathbb{S}^t} \mathbb{P}(S_0 = s_0|Y_0 = y) \prod_{k=0}^{t-1} \mathcal{P}(s_{k+1}|s_k, a_k)\phi(y_{k+1}|s_{k+1})}{\sum_{(s'_0,\ldots,s'_{t-1},s'_t)\in\mathbb{S}^{t+1}} \mathbb{P}(S_0 = s'_0|Y_0 = y) \prod_{k=0}^{t-1} \mathcal{P}(s'_{k+1}|s'_k, a_k)\phi(y_{k+1}|s'_{k+1})},$$



(a) Mean-squared TD, $T = 8$

(b) $\frac{1}{m}\|\mathbf{W}(k) - \mathbf{W}(0)\|_{2,1}$, $T = 8$.

(c) Mean-squared TD, $T = 32$

(d) $\frac{1}{m}\|\mathbf{W}(k) - \mathbf{W}(0)\|_{2,1}$, $T = 32$.

Figure 3. Mean-square TD and parameter deviation under Rec-TD for the case $p_{\text{min}} = 0.3$ and $\gamma = 0.9$. The mean curve and confidence intervals (90%) in Figures 3a and 3c stem from 5 trials. The 90% confidence intervals in Figures 3b and 3d correspond to deviations (i.e., $|W_{ii}(k) - W_{ii}(0)|$) across different units $i \in [m]$ in a single trial.

independent of π. As such, we have

$$\begin{aligned}
\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}^{\pi}(Z_{t+1})|\bar{Z}_t] &= \sum_{s\in\mathbb{S}} b_t(s)\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}^{\pi}_{t+1}(Z_{t+1})|\bar{Z}_t = \bar{z}_t, S_t = s],\\
&= \sum_{s_t,s_{t+1}\in\mathbb{S}}\sum_{y\in\mathbb{Y}} b_t(s_t)\left(r(s_t, A_t) + \gamma\mathcal{P}(s_{t+1}|s_t, A_t)\phi(y|s_{t+1})\mathcal{V}^{\pi}_{t+1}(Z_t, y_{t+1})\right),\\
&= \mathbb{E}[r_t + \gamma\mathcal{V}^{\pi}_{t+1}(Z_{t+1})|\bar{Z}_t = \bar{z}_t],
\end{aligned}$$

in other words, the conditional distribution of $r(S_t, A_t) + \gamma\mathcal{V}^{\pi}_{t+1}(Z_{t+1})$ given $\{\bar{Z}_t = \bar{z}_t\}$ is independent of π'. We also know from Prop. B.3 that

$$\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}^{\pi}_{t+1}(Z_{t+1})|\bar{Z}_t = \bar{z}_t] = \mathbb{E}[r_t + \gamma\mathcal{V}^{\pi}_{t+1}(Z_{t+1})|\bar{Z}_t = \bar{z}_t] = \mathcal{Q}^{\pi}_t(\bar{z}_t).$$

□

The next result generalizes the policy gradient theorem to POMDPs. We note that there is an extension of REINFORCE-type policy gradient for POMDPs in (Wierstra et al., 2010). The following result is a different and improved version as it ① provides a variance-reduced unbiased estimate of the policy gradient for POMDPs, and more importantly ② yields the compatible function approximation (Prop. 7.2) that yields natural policy gradient (NPG) for POMDPs.

Proposition D.2 (Policy gradient – POMDPs). *For any $\Phi \in \mathbb{R}^{m(d+1)}$, we have*

$$\nabla_{\Phi}\mathcal{V}^{\pi^{\Phi}}(\mu) = \mathbb{E}^{\pi^{\Phi}}_{\mu}\left[\sum_{t=0}^{\infty}\gamma^t \cdot \mathcal{Q}^{\pi^{\Phi}}_t(Z_t, A_t) \cdot \nabla_{\Phi}\ln\pi^{\Phi}_t(A_t|Z_t)\right], \tag{52}$$

for any $\mu \in \Delta(\mathbb{Y})$.

Proof of Prop. D.2. For any $t \in \mathbb{N}$, we have

$$\mathcal{V}_t^{\pi^\Phi}(z_t) = \sum_{a_t} \pi_t^\Phi(a_t|z_t)\mathcal{Q}_t^{\pi^\Phi}(z_t, a_t), \tag{53}$$

by Prop. B.3. Thus, we obtain

$$\begin{aligned}\nabla \mathcal{V}_t^{\pi^\Phi}(z_t) &= \sum_{a_t} \pi_t^\Phi(a_t|z_t)\nabla \ln \pi_t^\Phi(a_t|z_t)\mathcal{Q}_t^{\pi^\Phi}(z_t, a_t) + \sum_{a_t} \pi_t^\Phi(a_t|z_t)\nabla \mathcal{Q}_t^{\pi^\Phi}(z_t, a_t),\\ &= \mathbb{E}^{\pi^\Phi}[\nabla \ln \pi_t^\Phi(A_t|Z_t)\mathcal{Q}_t^{\pi^\Phi}(Z_t, A_t) + \nabla \mathcal{Q}_t^{\pi^\Phi}(Z_t, A_t)|Z_t = z_t]. \end{aligned}\tag{54}$$

Now, note that

$$\begin{aligned}\mathcal{Q}_t^{\pi^\Phi}(z_t, a_t) &= \mathbb{E}[r(S_t, A_t) + \gamma \mathcal{V}_{t+1}^{\pi^\Phi}(Z_{t+1})|\bar{Z}_t = (z_t, a_t)],\\ &= \sum_{s_t} b_t(s_t)\left(r(s_t, a_t) + \gamma \sum_{s_{t+1}} \mathcal{P}(s_{t+1}|s_t, a_t) \sum_{y_{t+1}} \phi(y_{t+1}|s_{t+1})\mathcal{V}_{t+1}^{\pi^\Phi}(z_{t+1})\right),\end{aligned}$$

where $z_{t+1} = (z_t, a_t, y_{t+1})$. As a consequence of Prop. D.1, we have $\nabla_\Phi \sum_{s_t} b_t(s_t)r(s_t, a_t) = 0$, and also

$$\begin{aligned}\nabla_\Phi \mathcal{Q}_t^{\pi^\Phi}(z_t, a_t) &= \gamma \sum_{s_t} b_t(s_t) \sum_{s_{t+1}} \mathcal{P}(s_{t+1}|s_t, a_t) \sum_{y_{t+1}} \phi(y_{t+1}|s_{t+1})\nabla_\Phi \mathcal{V}_{t+1}^{\pi^\Phi}(z_{t+1}),\\ &= \gamma \mathbb{E}[\nabla \ln \pi_{t+1}^\Phi(A_{t+1}|Z_{t+1})\mathcal{Q}_{t+1}^{\pi^\Phi}(Z_{t+1}, A_{t+1}) + \nabla_\Phi \mathcal{Q}_{t+1}^{\pi^\Phi}(Z_{t+1}, A_{t+1})|\bar{Z}_t = (z_t, a_t)],\\ &= \gamma \mathbb{E}^{\pi^\Phi}\Big[\sum_{k=t+1}^{\infty} \gamma^{k-t-1}\nabla_\Phi \ln \pi_k^\Phi(A_k|Z_k)\mathcal{Q}_k^{\pi^\Phi}(Z_k, A_k)\Big|\bar{Z}_t = (z_t, a_t)\Big].\end{aligned}$$

Using the above recursive formula for $\nabla_\Phi \mathcal{Q}_t^{\pi^\Phi}$ along with the law of iterated expectations in equation 54, we obtain

$$\nabla_\Phi \mathcal{V}_t^{\pi^\Phi}(z_t) = \mathbb{E}^{\pi^\Phi}\Big[\sum_{k=t}^{\infty} \gamma^{k-t}\nabla_\Phi \ln \pi_k^\Phi(A_k|Z_k)\mathcal{Q}_k^{\pi^\Phi}(Z_k, A_k)\Big|Z_t = z_t\Big]. \tag{55}$$

Since we have $\mathcal{V}^\pi := \mathcal{V}_0^\pi$, and also $\nabla_\Phi \mathcal{V}^{\pi^\Phi}(\mu) = \nabla_\Phi \sum_{z_0} \mu(z_0)\mathcal{V}^{\pi^\Phi}(z_0) = \sum_{z_0} \mu(z_0)\nabla_\Phi \mathcal{V}^{\pi^\Phi}(z_0)$ by the linearity of gradient, we conclude the proof.

Note on the baseline. Similar to the case of fully-observable MDPs, adding a baseline $q_t^{\pi^\Phi}(z_t)$ to the $\mathcal{Q}$-function does not change the policy gradients since $\sum_a \pi_t(a|z_t)\nabla \ln \pi_t^\Phi(a|z_t)q_t^{\pi^\Phi}(z_t) = q_t^{\pi^\Phi}(z_t)\sum_a \nabla \pi_t^\Phi(a|z_t) = q_t^{\pi^\Phi}(z_t)\nabla \sum_a \pi_t^\Phi(a|z_t) = 0$. Thus, we also have

$$\nabla_\Phi \mathcal{V}^{\pi^\Phi}(\mu) = \mathbb{E}_\mu^{\pi^\Phi}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{A}_t^{\pi^\Phi}(Z_t, A_t)\nabla_\Phi \ln \pi_t^\Phi(A_t|Z_t)\right], \tag{56}$$

which uses $q_t^{\pi^\Phi} = \mathcal{V}_t^{\pi^\Phi}$ as the baseline, akin to the fully-observable case. □

The following result extends the compatible function approximation theorem in (Kakade, 2001) to POMDPs.

Proof of Prop. 7.2. The proof is identical to (Kakade, 2001). By first-order condition for optimality, we have

$$2\mathbb{E}_\mu^{\pi^\Phi} \sum_{t=0}^{\infty} \gamma^t \nabla \ln \pi_t^\Phi(A_t|Z_t)\left(\nabla^\top \ln \pi_t^\Phi(A_t|Z_t)\omega^\star - \mathcal{A}_t^{\pi^\Phi}(\bar{Z}_t)\right) = 2\left(G_\mu(\Phi)\omega^\star - \nabla_\Phi \mathcal{V}^{\pi^\Phi}(\mu)\right) = 0,$$

which concludes the proof. □

E. Theoretical Analysis of Rec-NPG

First, we prove structural results for RNNs in the kernel regime, which will be key in the analysis later.

E.1. Log-Linearization of SOFTMAX Policies Parameterized by RNNs

The key idea behind the neural tangent kernel (NTK) analysis is linearization around the random initialization. To that end, let

$$F_t^{\mathsf{Lin}}(\bar{z}_t;\Theta) := \langle \nabla F_t(\bar{z}_t;\Theta(0)), \Theta - \Theta(0)\rangle, \tag{57}$$

for any $\Theta \in \mathbb{R}^{m(d+1)}$. We define the log-linearized policy as follows:

$$\tilde{\pi}_t^{\Phi}(a|z_t) := \frac{\exp(F_t^{\mathsf{Lin}}(z_t,a;\Phi))}{\sum_{a'\in\mathbb{A}}\exp(F_t^{\mathsf{Lin}}(z_t,a';\Phi))},\ t\in\mathbb{N}. \tag{58}$$

The first result bounds the Kullback-Leibler divergence between π_t^{Φ} and its log-linearized version $\tilde{\pi}_t^{\Phi}$. In the case of FNNs with ReLU activation functions, a similar result was presented in (Cayci et al., 2024). The following result extends this idea to (i) RNNs, and (ii) smooth activation functions.

Proposition E.1 (Log-linearization error)**.** *For any $t\in\mathbb{N}$ and $(z_t,a)\in(\mathbb{Y}\times\mathbb{A})^{t+1}$, we have*

$$\sup_{(z_t,a)\in(\mathbb{Y}\times\mathbb{A})^{t+1}}\left|\ln\frac{\tilde{\pi}_t^{\Phi}(a|z_t)}{\pi_t^{\Phi}(a|z_t)}\right| \leq \frac{6}{\sqrt{m}}\left(\Lambda_t^2\varrho_2+\chi_t\varrho_1\right)\|\Phi-\Phi(0)\|_2^2, \tag{59}$$

for any $t\in\mathbb{N}$. Consequently, we have $\pi_t(\cdot|z_t)\ll\tilde{\pi}_t(\cdot|z_t)$ and $\tilde{\pi}_t(\cdot|z_t)\ll\pi_t(\cdot|z_t)$, and

$$\max\left\{\mathscr{D}_{\mathsf{KL}}(\pi_t^{\Phi}(\cdot|z_t)\|\tilde{\pi}_t^{\Phi}(\cdot|z_t)),\mathcal{D}_{\mathsf{KL}}(\tilde{\pi}_t^{\Phi}(\cdot|z_t)\|\pi_t^{\Phi}(\cdot|z_t))\right\} \leq \frac{6}{\sqrt{m}}\left(\Lambda_t^2\varrho_2+\chi_t\varrho_1\right)\|\Phi-\Phi(0)\|_2^2, \tag{60}$$

for all $z_t\in(\mathbb{Y}\times\mathbb{A})^{t+1}$ and $t\in\mathbb{N}$.

Proof. Fix $(z_t,a)\in(\mathbb{Y}\times\mathbb{A})^{t+1}$. By the log-sum inequality (Cover & Thomas, 2006), we have

$$\ln\frac{\sum_a\exp(F_t^{\mathsf{Lin}}(z_t,a;\Phi))}{\sum_a\exp(F_t(z_t,a;\Phi))} \leq \sum_{a\in\mathbb{A}}\tilde{\pi}_t^{\Phi}(a|z_t)\left(F_t^{\mathsf{Lin}}(z_t,a;\Phi)-F_t(z_t,a;\Phi)\right).$$

Using the same argument, we obtain

$$\left|\ln\frac{\sum_a\exp(F_t^{\mathsf{Lin}}(z_t,a;\Phi))}{\sum_a\exp(F_t(z_t,a;\Phi))}\right| \leq \sum_{a\in\mathbb{A}}\left(\tilde{\pi}_t^{\Phi}(a|z_t)+\pi_t^{\Phi}(a|z_t)\right)\cdot\left|F_t^{\mathsf{Lin}}(z_t,a;\Phi)-F_t(z_t,a;\Phi)\right|. \tag{61}$$

Thus, we have

$$\left|\ln\frac{\tilde{\pi}_t^{\Phi}(a|z_t)}{\pi_t^{\Phi}(a|z_t)}\right| \leq (1+\tilde{\pi}_t^{\Phi}(a|z_t)+\pi_t^{\Phi}(a|z_t))\left|F_t^{\mathsf{Lin}}(z_t,a;\Phi)-F_t(z_t,a;\Phi)\right|.$$

By using Lemma B.1, we have $\sup_{\bar{z}_t\in(\mathbb{Y}\times\mathbb{A})^{t+1}}\left|F_t^{\mathsf{Lin}}(\bar{z}_t';\Phi)-F_t(\bar{z}_t';\Phi)\right| \leq \frac{2}{\sqrt{m}}(\Lambda_t^2\varrho_2+\chi_t\varrho_1)\|\Phi-\Phi(0)\|_2^2$. By using the last two inequalities together, and noting that $1+\tilde{\pi}_t^{\Phi}(a|z_t)+\pi_t^{\Phi}(a|z_t)\leq 3$, we conclude that

$$\left|\ln\frac{\tilde{\pi}_t^{\Phi}(a|z_t)}{\pi_t^{\Phi}(a|z_t)}\right| \leq \frac{6}{\sqrt{m}}(\Lambda_t^2\varrho_2+\chi_t\varrho_1)\|\Phi-\Phi(0)\|_2^2.$$

Since the righthand-side of the above inequality is independent of (z_t,a), we deduce that the result holds for all (z_t,a), thus concluding the proof. □

The following result will be important in establishing the Lyapunov drift analysis of Rec-NPG.

Proposition E.2 (Smoothness of $\ln \tilde{\pi}_t^{\Phi}(a|z_t)$)**.** *For any $t \in \mathbb{N}$, we have*

$$\sup_{(z_t,a)\in(\mathbb{Y}\times\mathbb{A})^{t+1}} \|\nabla \ln \tilde{\pi}_t^{\Phi}(a|z_t) - \nabla \ln \tilde{\pi}_t^{\Phi'}(a|z_t)\|_2 \leq L_t^2 \|\Phi - \Phi'\|_2,$$

for any $\Phi, \Phi' \in \mathbb{R}^{m(d+1)}$.

Proof. Consider a general log-linear parameterization

$$p_\theta(x) \propto \exp(\phi_x^\top \theta),\ x \in \mathbf{X}.$$

Then, if $\sup_{x\in\mathbf{X}} \|\phi_x\|_2 \leq B < \infty$, then $\theta \mapsto \ln p_\theta(x)$ has B^2-Lipschitz continuous gradients for each $x \in \mathbf{X}$ (Agarwal et al., 2020). The remaining part is to prove a uniform upper bound for $\|\nabla_\Phi F_t(\bar{z}_t; \Phi(0))\|_2$. To that end, notice that

$$\nabla_{\Phi_i} F_t(\bar{z}_t; \Phi(0)) = \frac{1}{\sqrt{m}} c_i \nabla H_t^{(i)}(\bar{z}_t; \Phi(0)),\ \bar{z}_t \in (\mathbb{Y}\times\mathbb{A})^{t+1}, i \in [m].$$

From the local Lipschitz continuity result in Lemma B.1, we have $\sup_{\bar{z}_t : \max_{j\leq t}\|(y_j,a_j)\|_2 \leq 1} \|\nabla_{\Phi_i} H_t^{(i)}(\bar{z}_t;\Phi(0))\|_2 \leq L_t$ for any $i \in [m]$. Thus, for any $\bar{z}_t$, we have

$$\|\nabla_\Phi F_t(\bar{z}_t; \Phi(0))\|_2^2 = \frac{1}{m}\sum_{i=1}^{m} \|\nabla_{\Phi_i} H_t^{(i)}(\bar{z}_t; \Phi(0))\|_2^2 \leq L_t^2. \tag{62}$$

□

E.2. Theoretical Analysis of Rec-NPG

For any $\pi \in \Pi_{\mathsf{NM}}$, we define the potential function as

$$\mathscr{L}(\pi) := \mathbb{E}_\mu^{\pi^\star}\left[\sum_{t=0}^{T-1} \gamma^t \mathscr{D}_{\mathsf{KL}}\left(\pi_t^\star(\cdot|Z_t)\|\pi_t(\cdot|Z_t)\right)\right]. \tag{63}$$

Then, we have the following drift inequality.

Proposition E.3 (Drift inequality)**.** *For any $n \in \mathbb{N}$, the drift can be bounded as follows:*

$$\begin{aligned}\mathscr{L}(\pi^{\Phi(n+1)}) - \mathscr{L}(\pi^{\Phi(n)}) \leq -\eta_{\mathsf{npg}}(\mathcal{V}^{\pi^\star}(\mu) - \mathcal{V}^{\pi^{\Phi(n)}}(\mu)) &\underbrace{-\eta_{\mathsf{npg}}\mathbb{E}_\mu^{\pi^\star}\left[\sum_{t=0}^{T-1}\gamma^t\left(\nabla^\top \ln \pi_t^{\Phi(n)}(A_t|Z_t)\omega_n - \mathcal{A}_t^{\pi^{\Phi(n)}}(\bar{Z}_t)\right)\right]}_{\textcircled{1}} \\ &\underbrace{+\,\eta_{\mathsf{npg}}\mathbb{E}_\mu^{\pi^\star}\sum_{t=T}^{\infty}\gamma^t\mathcal{A}_t^{\pi^{\Phi(n)}}(\bar{Z}_t)}_{\textcircled{2}}\underbrace{-\eta_{\mathsf{npg}}\mathbb{E}_\mu^{\pi^\star}\sum_{t=0}^{T-1}\gamma^t\left(\nabla\ln\tilde{\pi}_t^{\Phi(n)}(A_t|Z_t) - \nabla\ln\pi_t^{\Phi(n)}(A_t|Z_t)\right)^\top\omega_n}_{\textcircled{3}} \\ &+\frac{1}{2}\eta_{\mathsf{npg}}^2\|\rho\|_2^2\sum_{t=0}^{T-1}\gamma^t L_t^2 + \frac{12\|\rho\|_2^2}{\sqrt{m}}\sum_{t=0}^{T-1}\gamma^t(\Lambda_t^2\varrho_2 + \chi_t\varrho_1).\end{aligned}$$

Proof. First, note that the drift can be expressed as

$$\mathscr{L}(\pi^{\Phi(n+1)}) - \mathscr{L}(\pi^{\Phi(n)}) = \mathbb{E}_\mu^{\pi^\star}\sum_{t=0}^{T-1}\gamma^t\sum_{a\in\mathbb{A}}\pi_t^\star(A_t|Z_t)\ln\frac{\pi_t^{\Phi(n)}(A_t|Z_t)}{\pi_t^{\Phi(n+1)}(A_t|Z_t)}.$$

Then, with a log-linear transformation,

$$\mathscr{L}(\pi^{\Phi(n+1)}) - \mathscr{L}(\pi^{\Phi(n)}) = \mathbb{E}_\mu^{\pi^\star}\sum_{t=0}^{T-1}\gamma^t\sum_{a\in\mathbb{A}}\pi_t^\star(A_t|Z_t)\left(\ln\frac{\tilde{\pi}_t^{\Phi(n)}(A_t|Z_t)}{\tilde{\pi}_t^{\Phi(n+1)}(A_t|Z_t)} + \ln\frac{\pi_t^{\Phi(n)}(A_t|Z_t)}{\tilde{\pi}_t^{\Phi(n)}(A_t|Z_t)} + \ln\frac{\tilde{\pi}_t^{\Phi(n+1)}(A_t|Z_t)}{\pi_t^{\Phi(n+1)}(A_t|Z_t)}\right).$$

By using the log-linearization bound in Prop. E.1 twice in the above inequality, we obtain

$$\mathscr{L}(\pi^{\Phi(n+1)}) - \mathscr{L}(\pi^{\Phi(n)}) \leq \mathbb{E}_\mu^{\pi^\star} \sum_{t=0}^{T-1} \gamma^t \sum_{a\in\mathbb{A}} \pi_t^\star(A_t|Z_t) \ln \frac{\tilde{\pi}_t^{\Phi(n)}(A_t|Z_t)}{\tilde{\pi}_t^{\Phi(n+1)}(A_t|Z_t)} + \frac{12}{\sqrt{m}} \sum_{t=0}^{T-1} \gamma^t (\Lambda_t^2 \varrho_2 + \chi_t \varrho_1) \|\rho\|_2^2. \tag{64}$$

By the smoothness result in Prop. E.2, we have

$$|\ln \tilde{\pi}_t^{\Phi(n+1)}(a_t|z_t) - \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t)(\Phi(n+1) - \Phi(n))| \leq \frac{1}{2} L_t^4 \|\Phi(n+1) - \Phi(n)\|_2^2.$$

Thus, we obtain

$$-\eta_{\mathsf{npg}}^2 L_t^4 \|\rho\|_2^2 \leq -\eta_{\mathsf{npg}}^2 L_t^4 \|\omega_n\|_2^2 \leq -\ln \frac{\tilde{\pi}_t^{\Phi(n)}(a_t|z_t)}{\tilde{\pi}_t^{\Phi(n+1)}(a_t|z_t)} - \eta_{\mathsf{npg}} \nabla^\top \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) \omega_n,$$

because of the max-norm gradient clipping that yields $\|\omega_n\|_2 \leq \|\rho\|_2$ and $\Phi(n+1) = \Phi(n) + \eta_{\mathsf{npg}} \omega_n$ for any $n \in \mathbb{N}$. Using this in equation 64, we get

$$\mathscr{L}(\pi^{\Phi(n+1)}) - \mathscr{L}(\pi^{\Phi(n)}) \leq -\eta_{\mathsf{npg}} \mathbb{E}_\mu^{\pi^\star} \sum_{t=0}^{T-1} \gamma^t \nabla^\top \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) \omega_n + \frac{12}{\sqrt{m}} \sum_{t=0}^{T-1} \gamma^t (\Lambda_t^2 \varrho_2 + \chi_t \varrho_1) \|\rho\|_2^2 + \frac{1}{2} \eta_{\mathsf{npg}}^2 L_t^4 \|\rho\|_2^2. \tag{65}$$

An important technical result that will be useful in our analysis is the *pathwise* performance difference lemma, which was originally developed in (Kakade & Langford, 2002) for fully-observable MDPs.

Lemma E.4 (Pathwise Performance Difference Lemma)**.** *Let $\Phi, \Phi' \in \mathbb{R}^{m(d+1)}$ be two parameters. Then, we have*

$$\mathcal{V}^{\pi^{\Phi'}}(\mu) - \mathcal{V}^{\pi^{\Phi}}(\mu) = \mathbb{E}_\mu^{\pi^{\Phi'}} \sum_{t=0}^{\infty} \gamma^t \mathcal{A}_t^{\pi^\Phi}(Z_t, A_t).$$

The proof of Lemma E.4 is an extension of (Agarwal et al., 2020) to non-stationary policies, and can be found at the end of this subsection.

Using Lemma E.4 in equation 65, we obtain

$$\begin{aligned}\mathscr{L}(\pi^{\Phi(n+1)}) - \mathscr{L}(\pi^{\Phi(n)}) \leq -\eta_{\mathsf{npg}}(\mathcal{V}^{\pi^\star}(\mu) - \mathcal{V}^{\pi^{\Phi(n)}}(\mu)) - \eta_{\mathsf{npg}} \mathbb{E}_\mu^{\pi^\star} \sum_{t=0}^{T-1} \gamma^t \left(\nabla^\top \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) \omega_n - \mathcal{A}_t^{\pi^{\Phi(n)}}(\bar{Z}_t) \right) \\ + \eta_{\mathsf{npg}} \mathbb{E}_\mu^{\pi^\star} \sum_{t=T}^{\infty} \mathcal{A}_t^{\pi^{\Phi(n)}}(\bar{Z}_t) + \frac{12}{\sqrt{m}} \sum_{t=0}^{T-1} \gamma^t (\Lambda_t^2 \varrho_2 + \chi_t \varrho_1) \|\rho\|_2^2 + \frac{1}{2} \eta_{\mathsf{npg}}^2 L_t^4 \|\rho\|_2^2.\end{aligned} \tag{66}$$

Finally, we replace the term $\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t)$ with $\nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)$ by including the corresponding error term, and conclude the proof by considering the telescoping sum, and noting that $\mathscr{L}(\pi^{\Phi(0)}) = \log |\mathbb{A}|$ since $F_t(\cdot; \Phi(0)) = 0$ by symmetric initialization. □

Proof of Theorem 7.3. We prove Theorem 7.3 by bounding the numbered terms in Prop. E.3.

Bounding ① in Prop. E.3. Recall that $p_T(\gamma) = \sum_{t<T} \gamma^t$. Then, by using Jensen's inequality,

$$\begin{aligned}\mathbb{E}_\mu^{\pi^\star} \sum_{t=0}^{T-1} \gamma^t \left(\nabla^\top \ln \pi_t^{\Phi(n)}(A_t|Z_t) \omega_n - \mathcal{A}_t^{\pi^{\Phi(n)}}(\bar{Z}_t) \right) &\leq \sqrt{p_T(\gamma) \mathbb{E}_\mu^{\pi^\star} \sum_{t=0}^{T-1} \gamma^t \left| \nabla^\top \ln \pi_t^{\Phi(n)}(A_t|Z_t) \omega_n - \mathcal{A}_t^{\pi^{\Phi(n)}}(\bar{Z}_t) \right|^2}, \\ &=: \sqrt{p_T(\gamma)} \sqrt{\kappa \varepsilon_{\mathsf{cfa}}^T(\Phi(n), \omega_n)},\end{aligned}$$

where κ yields a change-of-measure argment from $P_T^{\pi^\star,\mu}$ to $P_T^{\pi^{\Phi(n)},\mu}$.

Bounding ② in Prop. E.3. $\sup_{s,a}|r(s,a)| \leq r_\infty$, therefore $|\mathcal{A}_t^\pi(\bar{z}_t)| \leq \frac{2r_\infty}{1-\gamma}$ for any $t \in \mathbb{N}, \bar{z}_t \in (\mathbb{Y}\times\mathbb{A})^{t+1}$, and $\pi \in \Pi_{\mathsf{NM}}$.

Bounding ③ in Prop. E.3. For any $t \in \mathbb{N}$, Cauchy-Schwarz inequality implies

$$\left(\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\right)^\top \omega_n \leq \|\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\|_2\|\rho\|_2.$$

Recall that

$$\begin{aligned}\nabla \ln \tilde{\pi}_t^{\Phi}(a_t|z_t) &= \nabla F_t(z_t, a_t; \Phi(0)) - \sum_{a'} \tilde{\pi}_t^{\Phi}(a'|z_t)\nabla F_t(z_t, a'; \Phi(0)),\\ \nabla \ln \pi_t^{\Phi}(a_t|z_t) &= \nabla F_t(z_t, a_t; \Phi) - \sum_{a'} \pi_t^{\Phi}(a'|z_t)\nabla F_t(z_t, a'; \Phi).\end{aligned}$$

First, from local β_t-Lipschitzness of $\Phi_i \mapsto \nabla H_t^{(i)}(\bar{z}_t; \Phi_i)$ for $\Phi \in \Omega_{\rho,m}$ by Lemma B.1, we have

$$\begin{aligned}\|\nabla_{\Phi_i} F_t(\bar{z}_t; \Phi(n)) - \nabla_{\Phi_i} F_t(\bar{z}_t; \Phi(0))\|_2 &= \frac{1}{\sqrt{m}}\|\nabla_{\Phi_i} H_t^{(i)}(\bar{z}_t; \Phi_i(n)) - \nabla_{\Phi_i} H_t^{(i)}(\bar{z}_t; \Phi_i(0))\|_2,\\ &\leq \frac{\beta_t\|\rho\|_2}{m},\end{aligned}$$

for any $n \in \mathbb{N}$ since $\max_i \|\Phi_i(n) - \Phi_i(0)\|_2 \leq \frac{\|\rho\|_2}{\sqrt{m}}$ by max-norm projection. Thus,

$$\|\nabla_\Phi F_t(\bar{z}_t; \Phi(n)) - \nabla_\Phi F_t(\bar{z}_t; \Phi(0))\|_2 \leq \frac{\beta_t\|\rho\|_2}{\sqrt{m}},\ t \in \mathbb{N}. \tag{67}$$

Thus,

$$\begin{aligned}\|\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\|_2 \leq \frac{\beta_t\|\rho\|_2}{\sqrt{m}} &+ \sum_a |\pi_t^{\Phi(n)}(a|z_t) - \tilde{\pi}_t^{\Phi(n)}(a|z_t)|\|\nabla F_t(\bar{z}_t; \Phi(0))\|_2\\ &+ \sum_a \pi_t^{\Phi(n)}(a|z_t)\|\nabla F_t(z_t, a; \Phi(n)) - \nabla F_t(z_t, a; \Phi(0))\|_2.\end{aligned}$$

From equation 62, we have

$$\|\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\|_2 \leq \frac{2\beta_t\|\rho\|_2}{\sqrt{m}} + 2L_t\mathscr{D}_{\mathsf{TV}}\left(\pi_t^{\Phi(n)}(\cdot|z_t)\|\tilde{\pi}_t^{\Phi(n)}(\cdot|z_t)\right),$$

where $\mathscr{D}_{\mathsf{TV}}$ denotes the total-variation distance between two probability measures. By Pinsker's inequality (Cover & Thomas, 2006), we obtain

$$\|\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\|_2 \leq \frac{2\beta_t\|\rho\|_2}{\sqrt{m}} + \sqrt{2}L_t\sqrt{\mathscr{D}_{\mathsf{KL}}\left(\pi_t^{\Phi(n)}(\cdot|z_t)\|\tilde{\pi}_t^{\Phi(n)}(\cdot|z_t)\right)}. \tag{68}$$

By the log-linearization result in Prop. E.1, we have

$$\|\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\|_2 \leq \frac{2\beta_t\|\rho\|_2}{\sqrt{m}} + \sqrt{12}L_t\|\rho\|_2\sqrt{\frac{\Lambda_t^2\varrho_2 + \chi_t\varrho_1}{\sqrt{m}}}. \tag{69}$$

Thus, we have

$$\left(\nabla \ln \tilde{\pi}_t^{\Phi(n)}(a_t|z_t) - \nabla \ln \pi_t^{\Phi(n)}(a_t|z_t)\right)^\top \omega_n \leq \|\rho\|_2^2\left(\frac{2\beta_t}{\sqrt{m}} + \sqrt{12}L_t\frac{\sqrt{\Lambda_t\varrho_2 + \chi_t\varrho_1}}{m^{1/4}}\right).$$

□

Proof of Lemma E.4. For any $y_0 \in \mathbb{Y}$, we have:

$$\begin{aligned}
\mathcal{V}^{\pi'}(y_0) - \mathcal{V}^{\pi}(y_0) &= \mathbb{E}_\mu^{\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t r_t \Big| Z_0 = y_0\Big] - \mathcal{V}^{\pi}(y_0), \\
&= \mathbb{E}_\mu^{\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t \Big(r_t + \mathcal{V}_t^{\pi}(Z_t) - \mathcal{V}_t^{\pi}(Z_t)\Big)\Big| Z_0 = y_0\Big] - \mathcal{V}^{\pi}(y_0), \\
&= \mathbb{E}_\mu^{\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t (r_t + \gamma\mathcal{V}_{t+1}^{\pi}(Z_{t+1}) - \mathcal{V}_t^{\pi}(Z_t)\Big| Z_0 = y_0\Big],
\end{aligned}$$

where $r_t = r(S_t, A_t)$ and the last identity holds since

$$\sum_{t=0}^{\infty}\gamma^t\mathcal{V}_t^{\pi}(z_t) = \mathcal{V}_0^{\pi}(z_0) + \gamma\sum_{t=0}^{\infty}\gamma^t\mathcal{V}_{t+1}^{\pi}(z_{t+1}).$$

Then, letting $r_t = r(s_t, a_t)$ and by using law of iterated expectations,

$$\mathcal{V}^{\pi'}(y_0) - \mathcal{V}^{\pi}(y_0) = \mathbb{E}_\mu^{\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t\Big(\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}_{t+1}^{\pi}(Z_{t+1})|\bar{Z}_t, S_t] - \mathcal{V}_t^{\pi}(Z_t)\Big)\Big| Z_0 = y_0\Big], \tag{70}$$

which holds because

$$\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}^{\pi}(Z_{t+1})|\bar{Z}_t] = \mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}^{\pi}(Z_{t+1})|\bar{Z}_t, Z_0].$$

The conditional expectation of $r_t + \gamma\mathcal{V}_{t+1}^{\pi}$ given $\{\bar{Z}_t = \bar{z}_t\}$ is independent of π':

$$\begin{aligned}
\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}^{\pi}(Z_{t+1})|\bar{Z}_t] &= \sum_{s\in\mathbb{S}} b_t(s)\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}_{t+1}^{\pi}(Z_{t+1})|\bar{Z}_t = \bar{z}_t, S_t = s], \\
&= \sum_{s_t, s_{t+1}\in\mathbb{S}}\sum_{y\in\mathbb{Y}} b_t(s_t)\left(r(s_t, A_t) + \gamma\mathcal{P}(s_{t+1}|s_t, A_t)\phi(y|s_{t+1})\mathcal{V}_{t+1}^{\pi}(Z_t, y_{t+1})\right), \\
&= \mathbb{E}[r_t + \gamma\mathcal{V}_{t+1}^{\pi}(Z_{t+1})|\bar{Z}_t = \bar{z}_t],
\end{aligned}$$

based on Prop. D.1. We also know from Prop. B.3 that

$$\mathbb{E}^{\pi'}[r_t + \gamma\mathcal{V}_{t+1}^{\pi}(Z_{t+1})|\bar{Z}_t = \bar{z}_t] = \mathbb{E}[r_t + \gamma\mathcal{V}_{t+1}^{\pi}(Z_{t+1})|\bar{Z}_t = \bar{z}_t] = \mathcal{Q}_t^{\pi}(\bar{z}_t).$$

Using the above identity in equation 70, we obtain

$$\mathcal{V}^{\pi'}(y_0) - \mathcal{V}^{\pi}(y_0) = \mathbb{E}_\mu^{\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t\Big(\mathcal{Q}_t^{\pi}(\bar{Z}_t) - \mathcal{V}^{\pi}(Z_t)\Big)\Big| Z_0 = y_0\Big], \tag{71}$$

which concludes the proof. □

Proof of Prop. 7.6. For any ω, we have

$$\ell_T(\omega; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}}) \leq 2\ell_T(\omega; \Phi(n), \hat{\mathcal{Q}}^{(n)}) + 2\sum_{t=0}^{\infty}\gamma^t(\mathcal{A}_t^{\pi^{\Phi(n)}}(Z_t, A_t) - \hat{\mathcal{A}}_t^{(n)}(Z_t, A_t))^2. \tag{72}$$

Let $\mathcal{G}_n := \sigma(\Phi(k), k \leq n)$ and $\mathcal{H}_n := \sigma(\bar{\Theta}^{(n)}, \Phi(k), k \leq n)$. Then, since

$$\varepsilon_{\mathsf{sgd},n} = \mathbb{E}[\ell_T(\omega_n; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\mathcal{H}_n] - \inf_{\omega\in\mathcal{B}_{2,\infty}^{(m)}(0,\rho)}\mathbb{E}[\ell_T(\omega; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\mathcal{H}_n],$$

we obtain

$$\mathbb{E}[\ell_T(\omega_n; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n] \leq 2\mathbb{E}\Big[\inf_{\omega}\mathbb{E}[\ell_T(\omega; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\mathcal{H}_n]\Big|\mathcal{G}_n\Big] + 2(\varepsilon_{\mathsf{td},n} + \varepsilon_{\mathsf{sgd},n}), \tag{73}$$

which uses the fact that $Var(X|\mathcal{G}_n) \leq \mathbb{E}[|X|^2|\mathcal{G}_n]$ for any square-integrable X. We also have

$$\inf_{\omega} \mathbb{E}[\ell_T(\omega; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\mathcal{H}_n] \leq 2 \inf_{\omega} \mathbb{E}[\ell_T(\omega; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n] + 2\sum_{t=0}^{\infty} \gamma^t (\mathcal{A}_t^{\pi^{\Phi(n)}}(Z_t, A_t) - \hat{\mathcal{A}}_t^{(n)}(Z_t, A_t))^2, \tag{74}$$

which further implies that

$$\mathbb{E}[\inf_{\omega} \mathbb{E}[\ell_T(\omega; \Phi(n), \hat{\mathcal{Q}}^{(n)})|\mathcal{H}_n]|\mathcal{G}_n] \leq 2\mathbb{E}[\inf_{\omega} \mathbb{E}[\ell_T(\omega; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n]|\mathcal{G}_n] + 2\varepsilon_{\mathsf{td},n}.$$

Thus,

$$\mathbb{E}[\ell_T(\omega_n; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n] \leq 4\mathbb{E}\Big[\inf_{\omega} \mathbb{E}[\ell_T(\omega; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n]\Big|\mathcal{G}_n\Big] + 6\varepsilon_{\mathsf{td},n} + 2\varepsilon_{\mathsf{sgd},n}. \tag{75}$$

For any $\omega \in \mathcal{B}_{2,\infty}^{(m)}(0, \rho)$,

$$\begin{aligned}\mathbb{E}[\ell_T(\omega; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n] &\leq \mathbb{E}[\sum_{t<T} \gamma^t (\nabla_{\Phi}^{\top} F_t(\bar{Z}_t; \Phi(n))\omega - \mathcal{Q}_t^{\pi^{\Phi(n)}}(\bar{Z}_t))^2|\mathcal{H}_n], \\ &\leq 2\mathbb{E}[\sum_{t<T} \gamma^t (\nabla_{\Phi}^{\top} F_t(\bar{Z}_t; \Phi(0))\omega - \mathcal{Q}_t^{\pi^{\Phi(n)}}(\bar{Z}_t))^2 + (\nabla F_t(\bar{Z}_t; \Phi(n)) - \nabla F_t(\bar{Z}_t; \Phi(0))^{\top}\omega)^2|\mathcal{H}_n],\end{aligned}$$

which implies that

$$\begin{aligned}\inf_{\omega} \mathbb{E}[\ell_T(\omega; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n] &\leq 2\varepsilon_{\mathsf{app},n} + 2\|\rho\|_2^2 \mathbb{E}[\sum_{t<T} \gamma^t \|\nabla F_t(\bar{Z}_t; \Phi(n)) - \nabla F_t(\bar{Z}_t; \Phi(0)\|_2^2|\mathcal{H}_n], \\ &\leq 2\varepsilon_{\mathsf{app},n} + \frac{2\|\rho\|_2^4}{m} \sum_{t<T} \gamma^t \beta_t^2,\end{aligned}$$

using equation 67. Hence,

$$\mathbb{E}[\ell_T(\omega_n; \Phi(n), \mathcal{Q}^{\pi^{\Phi(n)}})|\mathcal{H}_n] \leq \frac{8\|\rho\|_2^4}{m} \sum_{t<T} \gamma^t \beta_t^2 + 8\varepsilon_{\mathsf{app},n} + 6\varepsilon_{\mathsf{td},n} + 2\varepsilon_{\mathsf{sgd},n},$$

concluding the proof. □

Proof of Prop. 7.8. Under Assumption 7.7, consider $f_t^{(j)}(\bar{z}_t) := \mathbb{E}[\psi_t^{\top}(\bar{z}_t; \phi_0)\boldsymbol{v}^{(j)}(\phi_0)]$ for $\boldsymbol{v}^{(j)} \in \mathscr{H}_{\mathcal{J},\nu}$. Let

$$\omega_i^{(j)} := \frac{1}{\sqrt{m}} c_i \boldsymbol{v}^{(j)}(\Phi_i(0)),\ i = 1, 2, \ldots, m, \tag{76}$$

for any $j \in \mathcal{J}$. Since $\|\omega^{(j)}\|_2 \leq \|\nu\|_2$ and $\rho \succeq \nu$, we have

$$\inf_{\omega \in \mathcal{B}_{2,\infty}^{(m)}(0,\rho)} \left|\nabla^{\top} F_t(\bar{z}_t; \Phi(0))\omega - f_t^{(j)}(\bar{z}_t)\right| \leq \left|\nabla^{\top} F_t(\bar{z}_t; \Phi(0))\omega^{(j)} - f_t^{(j)}(\bar{z}_t)\right|. \tag{77}$$

Thus, we aim to find a uniform upper bound for the second term over $j \in \mathcal{J}$. For each $\bar{z}_t$, we have

$$\nabla^{\top} F_t(\bar{z}_t; \Phi(0))\omega^{(j)} = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\Phi_i}^{\top} H_t^{(i)}(\bar{z}_t; \Phi_i(0))\boldsymbol{v}^{(j)}(\Phi_i(0)),$$

thus $\mathbb{E}[\nabla^{\top} F_t(\bar{z}_t; \Phi(0))\omega^{(j)}] = f_t^{(j)}(\bar{z}_t)$. Furthermore, from Lemma B.1, since $\Phi(0) \in \Omega_{\rho,m}$ obviously, we have

$$\max_{1 \leq i \leq m} \|\nabla_{\Phi_i}^{\top} H_t^{(i)}(\bar{z}_t; \Phi_i(0))\boldsymbol{v}^{(j)}(\Phi_i(0))\|_2 \leq L_t \|\nu\|_2 \leq L_t \|\rho\|_2, \text{ a.s..}$$

Thus, by McDiarmid's inequality (Mohri et al., 2018), we have with probability at least $1 - \delta$,

$$\sup_{j \in \mathcal{J}} \left|\nabla^{\top} F_t(\bar{z}_t; \Phi(0))\omega^{(j)} - f_t^{(j)}(\bar{z}_t)\right| \leq 2\mathfrak{Rad}_m(G_t^{\bar{z}_t}) + L_t \|\rho\|_2 \sqrt{\frac{\log(2/\delta)}{m}}, \tag{78}$$

for each $t < T$ and $\bar{z}_t$. By union bound,

$$\sup_{j\in\mathcal{J}} \max_{\bar{z}_t} \left|\nabla^\top F_t(\bar{z}_t;\Phi(0))\omega^{(j)} - f_t^{(j)}(\bar{z}_t)\right| \leq 2\max_{\bar{z}_t} \mathfrak{Rad}_m(G_t^{\bar{z}_t}) + L_t\|\rho\|_2\sqrt{\frac{\log(2T|\mathbb{Y}\times\mathbb{A}|^{t+1}/\delta)}{m}}, \tag{79}$$

$$\leq 2\max_{0\leq t<T}\max_{\bar{z}_t} \mathfrak{Rad}_m(G_t^{\bar{z}_t}) + L_T\|\rho\|_2\sqrt{\frac{\log(2T|\mathbb{Y}\times\mathbb{A}|^{T}/\delta)}{m}}, \tag{80}$$

simultaneously for all $t < T$ with probability $\geq 1-\delta$. Therefore,

$$\inf_\omega \mathbb{E}_\mu^{\pi^{\Phi(n)}} \sum_{t<T} \gamma^t |\nabla^\top F_t(\bar{Z}_t;\Phi(0))\omega - f_t^{(j)}|^2 \leq \mathbb{E}_\mu^{\pi^{\Phi(n)}} \sum_{t<T} \gamma^t \sup_{j\in\mathcal{J}} |\nabla^\top F_t(\bar{Z}_t;\Phi(0))\omega^{(j)} - f_t^{(j)}|^2,$$

$$\leq \frac{1}{1-\gamma}\left(2\max_{0\leq t<T}\max_{\bar{z}_t}\mathfrak{Rad}_m(G_t^{\bar{z}_t}) + L_T\|\rho\|_2\sqrt{\frac{\log(2T|\mathbb{Y}\times\mathbb{A}|^{T}/\delta)}{m}}\right)^2.$$

□