

---

# Global Convergence of Pathwise Policy Gradient Through Model Backpropagation

---

**Shenao Zhang**

Georgia Institute of Technology

**Zhaoran Wang**

Northwestern University

**Tuo Zhao**

Georgia Institute of Technology

## 1 Introduction

For policy gradient RL algorithms with likelihood ratio estimation, recent works show the global convergence rates by leveraging the gradient domination property [3, 1]. Different from the likelihood ratio gradient estimator, the pathwise estimator does not sum over terms for which the parameter has no effect, allowing the estimator variance to be much lower [11].

The pathwise estimator is also attracting more attention in reinforcement learning. For example, in model-based RL, the gradient can be directly calculated by backpropagation through paths [8]. Instead of using entire trajectories, one can also leverage a learned value function and compute policy gradients from subsequences of trajectories [5], or directly backpropagate analytic action derivatives from a Q-function in a model-free manner. However, previous analysis for likelihood ratio policy gradient based on the score function estimator does not apply to pathwise policy gradient algorithms, which depend on the pathwise derivative and are among the most effective methods in model-based RL.

In this work, we derived the global convergence rates for several pathwise model-based policy gradient algorithms, including *Back-Propagation Through Time* (BPTT), *Stochastic Value Gradient* (SVG) [8], and *Model-Augmented Actor-Critic* (MAAC) [5]. We find that with the gradient calculated by model backpropagation, the convergence rates depend exponentially on the Jacobian of the learned dynamics. This indicates one source of sample inefficiency: the curse of chaos, *i.e.*, small changes in initial conditions result in diverging states. The analysis at the same time suggests several principled modifications to improve the efficiency of model-based analytic gradient algorithms, such as learning in a non-chaotic version of physics with spectral normalization. These modifications also share similarities with several recent works [10, 12, 4], which provide evidences from experimental observations.

Our main contributions include:

- We provide the global convergence rates for pathwise model-based policy gradient algorithms, with or without truncation (followed by critic function).
- We derive the optimal truncated timestep: smaller truncated time indicates smaller gradient variance; with small model error, larger truncated time indicates smaller gradient bias.
- We give a theoretical justification of the self-consistent update [6] in model-based PG.
- We also point out the distribution mismatch problem, which can be solved by training on the fixed point distribution.

## 2 Convergence Rates in Finite MDPs for Stochastic Policy with Direct Parameterization

PG with likelihood ratio estimator:

$$\nabla_{\theta} V_{\pi}(s) = \nabla_{\theta} \mathbb{E}_{\pi} [Q_{\pi}(s, a)] = \mathbb{E}_{\pi} [Q_{\pi}(s, a) \underbrace{\nabla_{\theta} \log \pi_{\theta}(a|s)}_{\text{score function}}]$$

PG with pathwise estimator:

$$\nabla_{\theta} V_{\pi}(s) = \mathbb{E}_{\pi} [\nabla_a Q(s, a) \nabla_{\theta} a]$$

First consider discounted finite MDPs where projected gradient ascent on the direct policy parametrization is performed.

$$\text{direct policy parametrization: } \pi_{\theta}(a|s) = \theta_{s,a}$$

$$\text{projected gradient ascent: } \pi_{t+1} = P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_{\theta} \hat{J}(\pi_{\theta_t}))$$

Here,  $J(\pi) = \mathbb{E}_{s_0 \sim \zeta(\cdot)} [V_{\pi}(s_0)]$  and  $\hat{J}(\pi) = \mathbb{E}_{s_0 \sim \zeta(\cdot)} [\hat{V}_{\pi}(s_0)]$ .

**Assumption 1.**  $\nabla_{\theta} V_{\pi_{\theta}}$  is  $L$ -Lipschitz in  $\theta$ .

This assumption holds when the reward  $r$  and transition function  $f$  are both Lipschitz continuous and smooth (i.e., twice continuously differentiable with Lipschitz continuous first derivative) [2, 14, 13].

**Fact 1.** (Bastani, 2021, Lemma D.2). Denote  $L_h$  as the Lipschitz constant for function  $h$  and  $\bar{L}_h = \max\{L_h, L_{\nabla h, 1}\}$ . Then  $\nabla V_{\pi_{\theta}}$  is  $L$ -Lipschitz, where  $L = 44H^5 \bar{L}_r \bar{L}_f^{4H}$ . Particularly,  $\nabla_{\theta} V_{\pi_{\theta}}$  is Lipschitz continuous in  $\theta$  with Lipschitz constant  $24H^5 \bar{L}_r \bar{L}_f^{4H}$ .

$$\text{gradient mapping: } \rho_t = \frac{1}{\eta} [P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_{\theta} J(\pi_{\theta_t}))]$$

$$\text{gradient bias: } b_t = \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2$$

$$\text{upper bound of squared variance: } v_t = \mathbb{E} \left[ \left\| \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right]$$

With a similar proof in [14], we have the following bound.

**Lemma 1.** For  $\eta \leq \frac{1}{L}$ ,

$$\min_{t \in [T]} \mathbb{E} [\|\rho_t\|_2^2] = \frac{4}{T} (\eta - L\eta^2)^{-1} \left( \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^T (\eta b_t + \frac{\eta}{2} v_t) \right) + \frac{4}{T} \sum_{t=1}^T (b_t^2 + v_t). \quad (1)$$

**Lemma 2.** (Agarwal, 2021, Lemma 4.1). Value function  $V$  satisfies gradient domination property. For all state distribution  $\mu_1, \mu_2 \in \Delta(\mathcal{S})$ , we have

$$\mathbb{E}_{s_0 \sim \mu_1} [V_{\pi^*}(s_0)] - \mathbb{E}_{s_0 \sim \mu_1} [V_{\pi}(s_0)] \leq \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_2}^{\pi}} \right\|_{\infty} \max_{\bar{\pi}} (\bar{\pi} - \pi)^{\top} \nabla_{\pi} \mathbb{E}_{s_0 \sim \mu_2} [V_{\pi}(s_0)].$$

**Lemma 3.** (Agarwal, 2021, Proposition B.1). Define  $G^{\eta}(\pi) = \frac{1}{\eta} [P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_{\pi} \mathbb{E}_{s_0} [V_{\pi}(s_0)])]$ . If  $\|G^{\eta}(\pi)\|_2 \leq \epsilon$ , then

$$\max_{\pi + e \in \Delta(\mathcal{A})^{|S|}, \|e\|_2 \leq 1} e^{\top} \nabla_{\pi} \mathbb{E}_{s_0} [V_{\pi}(s_0)] \leq \epsilon(\eta L + 1).$$

**Theorem 1.** For  $\eta \leq \frac{1}{L}$ ,

$$\min_{t \in [T]} J(\pi^*) - J(\pi_{\theta_t}) \leq \left\| \frac{d_{\zeta}^{\pi^*}}{d_{\zeta}^{\pi_{\theta_t}}} \right\|_{\infty} \frac{3}{\sqrt{T}} \left( (\eta - L\eta^2)^{-1} \left( \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^T (\eta b_t + \frac{\eta}{2} v_t) \right) + \sum_{t=1}^T (b_t^2 + v_t) \right)^{\frac{1}{2}}. \quad (2)$$

### 3 Convergence Rates Beyond Finite MDPs

Next, we consider discounted MDPs beyond finite settings, which enables function approximations in the next sections.

**Theorem 2.** For  $\eta \leq \frac{1}{L}$ ,

$$\min_{t \in [T]} J(\pi^*) - J(\pi_{\theta_t}) \leq \left\| \frac{d_{\zeta}^{\pi^*}}{d_{\zeta}^{\pi_{\theta_t}}} \right\|_{\infty} \min_{t \in [T]} \max_{s,a} \frac{1}{\|\nabla_{\theta_t} \pi(a|s)\|} \frac{4}{\sqrt{T}} \left( (\eta - L\eta^2)^{-1} \left( \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^T (\eta b_t + \frac{\eta}{2} v_t) \right) + \sum_{t=1}^T (b_t^2 + v_t) \right)^{\frac{1}{2}}. \quad (3)$$

The RHS of the inequality scales with  $\min_{t \in [T]} \max_{s,a} \frac{1}{\|\nabla_{\theta_t} \pi(a|s)\|}$ . We need to ensure this term not to be so large to avoid a vacuous upper bound. For this reason, we initialize the policy parameter  $\theta_0$  such that  $\max_{s,a} \frac{1}{\|\nabla_{\theta_0} \pi(a|s)\|}$  is upper bounded by a constant, i.e.,  $\inf_{s,a} \|\nabla_{\theta_0} \pi(a|s)\| > 0$ .

What remains is to bound  $b_t$  and  $v_t$  in the convergence rate. We show that enforcing the smoothness of the state-action function in model-free algorithms or the smoothness of the learned model in model-based algorithms is key to obtaining accurate gradient estimation, in terms of small gradient variance and small bias. We study BPTT and SVG(0) below.

### 4 BPTT

BPTT algorithm first learns a model and then updates policy by backpropagation through the model. With slight abuse of notation, we write the state value at step  $h$  as  $V_{\theta}^{(h)}(s)$ . In the following sections, we assume the initial state  $s_0$  is deterministic and known for simplicity. This implies  $J(\pi) = V_{\pi}(s_0)$ .

Consider the transition  $s_{h+1} = f(s_h, a_h) + \xi_h$ , where  $a_h \sim \pi(s_h)$ ,  $\xi_h \sim p(\xi)$  and assume  $p(\xi)$  is  $\sigma_{\xi}$ -subgaussian. Denote  $\xi = (\xi_1, \dots, \xi_H)$ ,  $R_{\theta}(s) = \mathbb{E}_{a \sim \pi}[r(s, a)]$ , and  $\hat{R}_{\theta}(s) = r(s, a)$ , where  $a \sim \pi(\cdot|s)$ .

Denote the estimated value by unrolling model  $\hat{f}$  as  $V^{\hat{f}}$ , or  $\hat{V}$ . Write  $\hat{V}_{\theta}^{(h+1)}(\hat{f}(s_h, a_h) + \xi_h; a_h \sim \pi, \xi_h \sim \xi)$  as  $\hat{V}_{\theta}^{(h+1)}(\hat{f}(s_h, a_h) + \xi_h)$  for clarity. For  $\epsilon = \max \nabla_{\theta} \pi(a|s)$ , we have

$$\hat{V}_{\theta}^{(h)}(s_h) = \hat{R}_{\theta}(s_h) + \gamma \hat{V}_{\theta}^{(h+1)}(\hat{f}(s_h, a_h) + \xi_h).$$

For finite-horizon  $\gamma$ -discounted MDPs, we have the following results.

**Proposition 1.** Denote  $L_{\hat{f}} = \max \left( 1, \left\| \nabla \hat{f}(s, a) \right\|_2 \right)_{\infty}$ , then

$$\begin{aligned} v_t &= \mathbb{E} \left[ \left\| \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] \\ &\leq \left( H^2 L_{\hat{f}}^H L + H L (1 + L_{\hat{f}}) + \frac{L_{\hat{f}} \epsilon}{L_{\hat{f}}} + L_{\hat{f}} \right)^2 \left( 3H^2 \sigma_{\xi}^2 d_s + 4H^2 |\mathcal{A}|^2 L_{\hat{f}}^2 + 8H^2 \sigma_{\xi} \sqrt{d_s} |\mathcal{A}| L_{\hat{f}} + H \sigma_{\xi} \sqrt{d_s} \right) \end{aligned} \quad (4)$$

*Proof.* See Appendix 7.4. □

**Proposition 2.** Define the “error” at iteration  $t$  as

$$\begin{aligned} \delta_{t,s} &= \sum_{h=1}^H \left( \gamma^{h-1} \left\| \nabla_s R_{\theta}(s) - \nabla_s \hat{R}_{\theta}(\hat{s}) \right\|_2 + \gamma^h L \left\| \mathbb{E}_a \left[ \nabla_s f(s, a) - \nabla_s \hat{f}(\hat{s}, a) \right] \right\|_2 \right) \\ \delta_{t,a} &= \sum_{h=1}^H \left( \gamma^{h-1} \left\| \nabla_{\theta} R_{\theta}(s) - \nabla_{\theta} \hat{R}_{\theta}(\hat{s}) \right\|_2 + \gamma^h L \left\| \mathbb{E}_a \left[ \nabla_a f(s, a) \nabla_{\theta} \pi(a|s) - \nabla_a \hat{f}(\hat{s}, a) \nabla_{\theta} \pi(a|\hat{s}) \right] \right\|_2 \right) \end{aligned}$$

Then we have

$$b_t = \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2 \leq \epsilon H L_{\hat{f}}^H \delta_{t,s} + \delta_{t,a}. \quad (5)$$

*Proof.* See Appendix 7.5.  $\square$

We can see that both  $v_t$  and  $b_t$  contains terms that are exponential in  $L_{\hat{f}}$ . Some works [10, 12] also have similar observations of the curse of chaos, *i.e.*, small changes in initial conditions result in diverging states. Proposition 1 and 2 suggests a way to mitigate it: constraining the Jacobian of the learned dynamical system. In other words, we can learn a more well-behaved non-chaotic version of physics instead of the rigid body dynamics (*e.g.*, ball bouncing off of a wall causes sharp changes in object velocity).

In ball bouncing environment, should not use BPTT? In nearly deterministic tasks, use BPTT.

## 5 SVG(0)

SVG(0) [8] is a model-free algorithm, and is the stochastic analogue of DPG.

$$\begin{aligned} \text{stochastic policy: } a &= \mu_{\theta}(s, \epsilon), \epsilon \sim p(\epsilon) \\ \text{exact policy gradient: } \nabla_{\theta} J(\pi_{\theta}) &= \mathbb{E}_{\epsilon} [\nabla_a Q \nabla_{\theta} \mu_{\theta}(s, \epsilon)] \\ \text{with Q function approximation: } \nabla_{\theta} \hat{J}(\pi_{\theta}) &= \nabla_a \hat{Q} \nabla_{\theta} \mu_{\theta}(s, \epsilon)|_{\epsilon} \end{aligned}$$

There is a close connection between the direct policy parameterization  $\pi_{\theta}(a|s) = \theta_{s,a}$  and the stochastic policy  $a = \mu_{\theta}(s, \epsilon)$  where  $\epsilon \sim p(\epsilon)$ . We will always be able to find an equivalent base distribution and sampling path [11].

$$a \sim \pi_{\theta}(a|s) \equiv a = \mu_{\theta}(s, \epsilon), \epsilon \sim p(\epsilon)$$

One simple way is to set  $p(\epsilon)$  as the uniform distribution and  $\mu$  as inverse cumulative distribution function (CDF). Denote  $F(a; \theta)$  as the CDF of the categorical distribution  $\pi_{\theta}(a|s)$ . Then

$$\begin{aligned} a &= \mu_{\theta}(s, \epsilon) = F^{-1}(\epsilon; \theta), \text{ where } \epsilon \sim \mathcal{U}[0, 1] \\ \nabla_{\theta} a &= \nabla_{\theta} F^{-1}(\epsilon; \theta) \\ \text{or } \nabla_{\theta} a &= -\frac{\nabla_{\theta} F(a; \theta)}{\nabla_a F(a; \theta)} = -\frac{\nabla_{\theta} F(a; \theta)}{p(a; \theta)} \end{aligned}$$

To obtain an accurate gradient estimation, we need  $\nabla_a \hat{Q}(s, a) \approx \nabla_a Q(s, a)$ . In practice, it is common to leverage smooth function approximators so that learning  $\hat{Q}(s, a) \approx Q(s, a)$  can give good results.

**Proposition 3.** If  $\hat{Q}(s, a)$  is  $L_{\hat{Q}}$ -Lipschitz in  $a$ , and  $\delta_t = \left\| \mathbb{E}_a [\nabla_a Q(s_0, a) - \nabla_a \hat{Q}(s_0, a)] \right\|_2$ , then

$$v_t \leq L_{\hat{Q}}^2 \max_{\epsilon} \mathbb{E} \left[ \left\| \nabla_{\theta} F^{-1}(\epsilon; \theta) \right\|_2^2 \right] \quad (6)$$

$$b_t \leq \left\| \mathbb{E}_{\epsilon} [\nabla_{\theta} F^{-1}(\epsilon; \theta)] \right\|_2 \delta_t \quad (7)$$

*Proof.* See Appendix 7.6.  $\square$

The key point here is that we need  $\hat{Q}(s, a)$  to be Lipschitz continuous in  $a$ . Since SVG(0) and DPG relies on the *derivative* of the Q function, learning  $\hat{Q}(s, a) \approx Q(s, a)$  will not suffice.

## 6 MAAC

MAAC [5] can be viewed as differentiating through the  $H$ -step model value expansion, *i.e.*, BPTT unrolling  $H$  steps followed by a parameterized critic function, *e.g.*,  $\hat{Q}_\phi$  for state-action value function and  $\hat{V}_\phi$  for state value function. Denote the  $H$ -step value expansion as  $\hat{V}_{\pi,H}$ .

$$\hat{V}_{\pi,H}(s_0) = \hat{r}_0 + \gamma \hat{r}_1 + \dots + \gamma^H \hat{V}_\phi(\hat{s}_H) = V^{\hat{f}}(s_0) + \gamma^H \hat{V}_\phi(\hat{s}_H) = V^{\hat{f}}(s_0) + \gamma^H \mathbb{E}_a[\hat{Q}_\phi(\hat{s}_H, a_H)].$$

If  $\hat{Q}_\phi(s, a)$  is  $L_{\hat{Q}}$ -Lipschitz in  $a$ , we have the following proposition by simply combining Proposition 1, Proposition 2, and Proposition 3.

**Proposition 4.** Denote  $L_{\hat{f}} = \max\left(1, \left\|\nabla \hat{f}(s, a)\right\|_2\right)_\infty$ , then

$$\begin{aligned} v_t &= \mathbb{E} \left[ \left\| \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] \\ &\leq \left( H^2 L_{\hat{f}}^H L + H L (1 + L_{\hat{f}}) \right)^2 \left( 3H^2 \sigma_{\xi}^2 d_s + 4H^2 |\mathcal{A}|^2 L_{\hat{f}}^2 + 8H^2 \sigma_{\xi} \sqrt{d_s} |\mathcal{A}| L_{\hat{f}} + H \sigma_{\xi} \sqrt{d_s} \right) \\ &\quad + \gamma^H L_{\hat{Q}}^2 \max_{\epsilon} \mathbb{E} \left[ \left\| \nabla_{\theta} F^{-1}(\epsilon; \theta) \right\|_2^2 \right] \end{aligned} \quad (8)$$

**Connection with [6]:** Learn the value function  $\hat{Q}_\phi$  first, then update policy, since the gradient is not compounding as in BPTT and the gradient variance is thus smaller:

$$v_t = \mathbb{E} \left[ \left\| \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] = \mathbb{E} \left[ \left\| \nabla_a \hat{Q}_\phi \nabla_{\theta} a_{\epsilon} - \mathbb{E}_a[\nabla_a \hat{Q}_\phi \nabla_{\theta} a] \right\|_2^2 \right]. \quad (9)$$

If using self-consistent  $Q$  functions,  $\hat{Q}_\phi$  is trained to align the value unrolled with  $\hat{f}$ . However, the gradient bias is larger (compared to BPTT) since only  $\nabla \hat{Q}_\phi = \nabla Q$  (or  $\hat{Q}_\phi = \sum_{\hat{f}} \hat{r} = Q$  everywhere) can guarantee equivalent gradient bias as BPTT.

MVE didn't consider the true value under real model! Since the MSE in Eq. (3) needs to be minimized. Besides, if we choose  $v$  as the fixed point, minimizing the MSE in Eq. (3) is not exactly what we desire!

**Proposition 5.** The gradient bias is bounded by

$$\begin{aligned} b_t &\leq B_2^{(0)}(\pi) + \gamma^H \left( \left\| \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_\phi(\hat{s}_H)] \right\|_2 + L_{\nabla V^{\hat{f}}} \mathbb{E}[\|s_H - \hat{s}_H\|_2] + B_2^{(0)}(\pi) \right) \\ &\quad + \gamma^H \left\| \mathbb{E}[\nabla_{\theta} s_H] \right\|_2 \left( \left\| \mathbb{E}_{\hat{s}_H}[\nabla_s V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_s \hat{V}_\phi(\hat{s}_H)] \right\|_2 + \|L_{\nabla \hat{V}}\|_2 \mathbb{E}[\|s_H - \hat{s}_H\|_2] + B_3^{(0)}(\pi) \right) \\ &\quad + \gamma^H \left\| \nabla_s \hat{V}_\phi(\hat{s}_H) \right\|_2 \left\| \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} s_H - \nabla_{\theta} \hat{s}_H] \right\|_2 \end{aligned} \quad (10)$$

*Proof.* See Appendix 7.7.  $\square$

**Remark 1.** We introduce model expended value  $V^{\hat{f}}$  to optimize the critic  $\hat{V}_\phi$ , instead of directly learning  $\hat{V}_\phi$  from the real environment data as in previous actor-critic frameworks. To understand this, an alternative bound of Equation (34) is given by

$$\left\| \mathbb{E}_{s_H}[\nabla_{\theta} V(s_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_\phi(\hat{s}_H)] \right\|_2 \leq \left\| \mathbb{E}_{s_H}[\nabla_{\theta} V(s_H)] - \mathbb{E}_{s_H}[\nabla_{\theta} \hat{V}_\phi(s_H)] \right\|_2 + \left\| \mathbb{E}_{s_H}[\nabla_{\theta} \hat{V}_\phi(s_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_\phi(\hat{s}_H)] \right\|_2$$

From the first term, it seems the critic  $\hat{V}_\phi$  can also be learned with  $V(s_H)$  as target. However,  $V(s_H)$  can only be obtained by running in the real environment an additional  $H$  steps.

Proposition 5 implies a self-consistent way to update the value towards the model expended value, *i.e.*, minimize  $\left\| \mathbb{E}_\zeta[V^{\hat{f}}] - \mathbb{E}_\zeta[\hat{V}_\phi] \right\|_2$  while keeping the model accurate. However, another issue arises — distribution mismatch. Training the critic on the initial state distribution  $\zeta = s_0$  could still lead to a large  $\left\| \mathbb{E}_{\hat{s}_H}[V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H}[\hat{V}_\phi(\hat{s}_H)] \right\|_2$ , especially at the beginning period. With a similar insight of [7], we can set  $\zeta$  to a fixed point of  $\hat{f}$ , *i.e.*,  $\zeta = (\hat{f}^\pi)^H \zeta$ . To achieve this, sample  $T \sim \mathcal{U}[0, H - 1]$  and set the initial state as  $\hat{s}_T$ . Then unrolling the model  $H - T$  steps rather than  $H$  steps. This fixes distribution mismatch.

## 7 Proofs

### 7.1 Proof of Lemma 1

*Proof.* Denote  $\beta_t = \frac{\theta_{t+1} - \theta_t}{\eta}$ . By Assumption 1, we have

$$\begin{aligned} J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) &\geq \nabla_{\theta} J(\pi_{\theta_t})^{\top} (\theta_{t+1} - \theta_t) - \frac{L}{2} \|\theta_{t+1} - \theta_t\|_2^2 \\ &= \eta \nabla_{\theta} J(\pi_{\theta_t})^{\top} \beta_t - \frac{L\eta^2}{2} \|\beta_t\|_2^2. \end{aligned} \quad (11)$$

Rewrite the exact gradient  $\nabla_{\theta} J(\pi_{\theta_t})$  as

$$\nabla_{\theta} J(\pi_{\theta_t}) = \left( \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right) - \left( \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right) + \nabla_{\theta} \hat{J}(\pi_{\theta_t}).$$

Then we bound  $\nabla_{\theta} J(\pi_{\theta_t})^{\top} \beta_t$  in Eq. (11) by bounding the resulting three terms.

$$\left| \left( \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right)^{\top} \beta_t \right| \leq \|\beta_t\|_2 \|\nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})]\|_2 = \|\beta_t\|_2 b_t \quad (12)$$

$$\left( \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right)^{\top} \beta_t \leq \frac{\|\nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2} \quad (13)$$

$$\nabla_{\theta} \hat{J}(\pi_{\theta_t})^{\top} \beta_t \geq \|\beta_t\|_2^2, \quad (14)$$

where Eq. (14) holds due to the fact that  $\left( \theta_{t+1} - (\theta_t + \eta \nabla_{\theta} \hat{J}(\theta_t)) \right)^{\top} (\theta_{t+1} - \theta_t) \leq 0$ .

Thus, we can bound Eq. (11) by

$$J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) \geq \eta \left( -\|\beta_t\|_2 b_t - \frac{\|\nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2} \right) - \frac{L\eta^2}{2} \|\beta_t\|_2^2. \quad (15)$$

By taking expectation in Eq. (15), we have

$$\begin{aligned} \left( \frac{\eta}{2} - \frac{L\eta^2}{2} \right) \mathbb{E}[\|\beta_t\|_2^2] &\leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta \mathbb{E}[\|\beta_t\|_2] b_t + \frac{\eta}{2} v_t \\ &\leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta b_t + \frac{\eta}{2} v_t \end{aligned} \quad (16)$$

Besides,

$$\begin{aligned} \|\rho_t - \beta_t\|_2 &= \frac{1}{\eta} \left\| P_{\Delta(\mathcal{A})|S|}(\theta_t + \eta \nabla_{\theta} J(\pi_{\theta_t})) - P_{\Delta(\mathcal{A})|S|}(\theta_t + \eta \nabla_{\theta} \hat{J}(\pi_{\theta_t})) \right\|_2 \\ &\leq \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \nabla_{\theta} \hat{J}(\pi_{\theta_t}) \right\|_2. \end{aligned}$$

Then due to the fact that  $\|x + y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$ , we have

$$\begin{aligned} \mathbb{E}[\|\rho_t - \beta_t\|_2^2] &\leq \mathbb{E} \left[ \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \nabla_{\theta} \hat{J}(\pi_{\theta_t}) \right\|_2^2 \right] \\ &= \mathbb{E} \left[ \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] + \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] - \nabla_{\theta} \hat{J}(\pi_{\theta_t}) \right\|_2^2 \right] \\ &\leq 2 \mathbb{E} \left[ \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] + 2 \mathbb{E} \left[ \left\| \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] - \nabla_{\theta} \hat{J}(\pi_{\theta_t}) \right\|_2^2 \right] \\ &= 2 \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 + 2 \mathbb{E} \left[ \left\| \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] - \nabla_{\theta} \hat{J}(\pi_{\theta_t}) \right\|_2^2 \right] \\ &\leq 2b_t^2 + 2v_t. \end{aligned} \quad (17)$$

For  $\eta \leq \frac{1}{L}$ ,  $\frac{\eta}{2} - \frac{L\eta^2}{2} > 0$ . Then with Eq. (16) and Eq. (17), we can bound  $\min_{t \in [T]} \mathbb{E} [\|\rho_t\|_2^2]$  by

$$\begin{aligned}
\min_{t \in [T]} \mathbb{E} [\|\rho_t\|_2^2] &\leq \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\rho_t\|_2^2] \\
&\leq \frac{2}{T} \sum_{t=1}^T (\mathbb{E} [\|\beta_t\|_2^2] + \mathbb{E} [\|\rho_t - \beta_t\|_2^2]) \\
&\leq \frac{4}{T} (\eta - L\eta^2)^{-1} \left( \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^T (\eta b_t + \frac{\eta}{2} v_t) \right) + \frac{4}{T} \sum_{t=1}^T (b_t^2 + v_t) \\
&= \frac{4}{T} \left( \sum_{t=1}^T (\eta - L\eta^2)^{-1} (\eta b_t + \frac{\eta}{2} v_t) + b_t^2 + v_t \right) + \frac{4}{T} (\eta - L\eta^2)^{-1} \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})]
\end{aligned}$$

□

## 7.2 Proof of Theorem 1

*Proof.* From Lemma 2, we know that for initial distribution  $\zeta$ ,

$$J(\pi^*) - J(\pi_{\theta_t}) \leq \left\| \frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}} \right\|_\infty \max_{\bar{\pi}} (\bar{\pi} - \pi_{\theta_t})^\top \nabla_{\pi} J(\pi_{\theta_t}).$$

Then by Lemma 3, Lemma 1 and  $\eta L + 1 \leq 2$ , we know

$$\begin{aligned}
\min_{t \in [T]} J(\pi^*) - J(\pi_{\theta_t}) &\leq \left\| \frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}} \right\|_\infty \min_{t \in [T]} \mathbb{E} [\|\rho_t\|_2] (\eta L + 1) \\
&\leq \left\| \frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}} \right\|_\infty \min_{t \in [T]} \sqrt{\mathbb{E} [\|\rho_t\|_2^2]} (\eta L + 1) \\
&\leq \left\| \frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}} \right\|_\infty \frac{4}{\sqrt{T}} \left( (\eta - L\eta^2)^{-1} \left( \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^T (\eta b_t + \frac{\eta}{2} v_t) \right) + \sum_{t=1}^T (b_t^2 + v_t) \right)^{\frac{1}{2}},
\end{aligned}$$

where the second inequality follows Jensen's inequality with the concave square root function. □

## 7.3 Proof of Theorem 2

*Proof.* From the performance difference lemma [9],

$$V_\pi(s_0) - V_{\pi'}(s_0) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{s_0}^\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} [A^{\pi'}(s, a)].$$



Then

$$\begin{aligned}
\mathbb{E}_{s_0 \sim \mu_1} [V_{\pi^*}(s_0)] - \mathbb{E}_{s_0 \sim \mu_1} [V_{\pi}(s_0)] &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu_1}^{\pi^*}} \mathbb{E}_{a \sim \pi^*} [A^{\pi}(s, a)] \\
&\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu_1}^{\pi^*}} \left[ \max_{\bar{a}} A^{\pi}(s, \bar{a}) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu_1}^{\pi^*}} \left[ \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \max_{\bar{a}} A^{\pi}(s, \bar{a}) \right] \\
&\leq \frac{1}{1-\gamma} \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \left[ \max_{\bar{a}} A^{\pi}(s, \bar{a}) \right] \\
&= \frac{1}{1-\gamma} \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \max_{\bar{\pi}} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \mathbb{E}_{a \sim \bar{\pi}} [\bar{\pi}(a|s) A^{\pi}(s, a)] \\
&= \frac{1}{1-\gamma} \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \max_{\bar{\pi}} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \mathbb{E}_{a \sim \bar{\pi}} [(\bar{\pi}(a|s) - \pi(a|s)) A^{\pi}(s, a)] \\
&= \frac{1}{1-\gamma} \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \max_{\bar{\pi}} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \mathbb{E}_{a \sim \bar{\pi}} [(\bar{\pi}(a|s) - \pi(a|s)) Q^{\pi}(s, a)].
\end{aligned} \tag{18}$$

It holds that

$$\nabla_{\theta} J(\theta) = \nabla_{\theta} \mathbb{E}_{s_0 \sim \mu_1} [V_{\pi_{\theta}}(s_0)] = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu_1}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}} [Q^{\pi_{\theta}}(s, a) \nabla_{\theta} \pi_{\theta}(a|s)].$$

Then Eq. (18) can be further bounded by

$$\begin{aligned}
&\mathbb{E}_{s_0 \sim \mu_1} [V_{\pi^*}(s_0)] - \mathbb{E}_{s_0 \sim \mu_1} [V_{\pi}(s_0)] \\
&\leq \frac{1}{1-\gamma} \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \max_{\bar{\pi}} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \mathbb{E}_{a \sim \bar{\pi}} [(\bar{\pi}(a|s) - \pi(a|s)) Q^{\pi}(s, a)] \\
&\leq \frac{1}{1-\gamma} \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \max_{s, a} \frac{1}{\|\nabla_{\theta} \pi(a|s)\|} \max_{\bar{\pi}} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \mathbb{E}_{a \sim \bar{\pi}} [(\bar{\pi}(a|s) - \pi(a|s)) Q^{\pi}(s, a) \|\nabla_{\theta} \pi(a|s)\|] \\
&= \left\| \frac{d_{\mu_1}^{\pi^*}}{d_{\mu_1}^{\pi}} \right\|_{\infty} \max_{s, a} \frac{1}{\|\nabla_{\theta} \pi(a|s)\|} \max_{\bar{\pi}} \mathbb{E}_{s \sim d_{\mu_1}^{\pi}} \mathbb{E}_{a \sim \bar{\pi}} [(\bar{\pi}(a|s) - \pi(a|s)) \|\nabla_{\theta} J(\theta)\|]
\end{aligned} \tag{19}$$

Similar with the proof in Theorem 1, we have

$$\begin{aligned}
&\min_{t \in [T]} J(\pi^*) - J(\pi_{\theta_t}) \\
&\leq \left\| \frac{d_{\zeta}^{\pi^*}}{d_{\zeta}^{\pi_{\theta_t}}} \right\|_{\infty} \min_{t \in [T]} \max_{s, a} \frac{1}{\|\nabla_{\theta_t} \pi(a|s)\|} \mathbb{E}[\|\rho_t\|_2] (\eta L + 1) \\
&\leq \left\| \frac{d_{\zeta}^{\pi^*}}{d_{\zeta}^{\pi_{\theta_t}}} \right\|_{\infty} \min_{t \in [T]} \max_{s, a} \frac{1}{\|\nabla_{\theta_t} \pi(a|s)\|} \frac{4}{\sqrt{T}} \left( (\eta - L\eta^2)^{-1} \left( \mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^T (\eta b_t + \frac{\eta}{2} v_t) \right) + \sum_{t=1}^T (b_t^2 + v_t) \right)^{\frac{1}{2}}.
\end{aligned}$$

□

#### 7.4 Proof of Proposition 1

Before the proofs of Prop. 1 and 2, we note the following formulas in BPTT.

$$\begin{aligned}
\hat{V}_\theta^{(h)}(s) &= \hat{R}_\theta(s) + \gamma \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) \\
\nabla_\theta \hat{V}_\theta^{(h)}(s) &= \nabla_\theta \hat{R}_\theta(s) + \gamma \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) + \gamma \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) \nabla_\theta \hat{f}(s, a_h) \\
\nabla_s \hat{V}_\theta^{(h)}(s) &= \nabla_s \hat{R}_\theta(s) + \gamma \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) \nabla_s \hat{f}(s, a_h) \\
\nabla_\theta V_\theta^{(h)}(s) &= \nabla_\theta R_\theta(s) + \gamma \mathbb{E}_{\substack{a \sim \pi \\ \xi \sim \xi}} \left[ \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) + \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \nabla_\theta \hat{f}(s, a) \right] \\
\nabla_s V_\theta^{(h)}(s) &= \nabla_s R_\theta(s) + \gamma \mathbb{E}_{\substack{a \sim \pi \\ \xi \sim \xi}} \left[ \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \nabla_s \hat{f}(s, a) \right]
\end{aligned}$$

*Proof.* Suppose

$$\begin{aligned}
\left\| \nabla_\theta \hat{V}_\theta^{(h)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_\theta^{(h)}(s)] \right\|_2 &\leq B_0^{(h)}(\pi, \xi) \\
\left\| \nabla_s \hat{V}_\theta^{(h)}(s) - \mathbb{E}[\nabla_s \hat{V}_\theta^{(h)}(s)] \right\|_2 &\leq B_1^{(h)}(\pi, \xi)
\end{aligned}$$

By induction,

$$\begin{aligned}
&\left\| \nabla_\theta \hat{V}_\theta^{(h)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_\theta^{(h)}(s)] \right\|_2 \\
&\leq \left\| \nabla_\theta \hat{r}(s, a_h) - \mathbb{E}_a[\nabla_\theta \hat{r}(s, a)] \right\|_2 + \gamma \left\| \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \mathbb{E}_{a, \xi}[\nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi)] \right\|_2 \\
&+ \gamma \left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) \nabla_\theta \hat{f}(s, a_h) - \mathbb{E}_{a, \xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \nabla_\theta \hat{f}(s, a)] \right\|_2 \\
&\leq \left\| \nabla_a \hat{r}(s, a_h) \nabla_\theta a_h - \mathbb{E}_a[\nabla_a \hat{r}(s, a) \nabla_\theta a] \right\|_2 + \gamma \left\| \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \mathbb{E}_{a, \xi}[\nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h)] \right\|_2 \\
&+ \gamma \mathbb{E}_{a, \xi} \left[ \left\| \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \right\|_2 \right] \\
&+ \gamma \left\| \nabla_\theta \hat{f}(s, a_h) \right\|_2 \left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \mathbb{E}_{a, \xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h)] \right\|_2 \\
&+ \gamma \left\| \mathbb{E}_a[\nabla_\theta \hat{f}(s, a)] \right\|_2 \mathbb{E}_{a, \xi} \left[ \left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \right\|_2 \right] \\
&\leq 2|\mathcal{A}|L_{\hat{r}}\epsilon + \gamma B_0^{(h+1)}(\pi, \xi) + \gamma L \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) \\
&+ \gamma \left\| \nabla_\theta \hat{f}(s, a_h) \right\|_2 B_1^{(h+1)}(\pi, \xi) + \gamma L \left\| \mathbb{E}_a[\nabla_\theta \hat{f}(s, a)] \right\|_2 \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) \\
&\leq 2|\mathcal{A}|L_{\hat{r}}\epsilon + \gamma B_0^{(h+1)}(\pi, \xi) + \gamma \left\| \nabla_\theta \hat{f}(s, a_h) \right\|_2 B_1^{(h+1)}(\pi, \xi) \\
&+ \gamma L \left( 1 + \left\| \mathbb{E}_a[\nabla_\theta \hat{f}(s, a)] \right\|_2 \right) \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) \\
&= B_0^{(h)}(\pi, \xi).
\end{aligned} \tag{20}$$

Similarly, we also have

$$\begin{aligned}
& \left\| \nabla_s \hat{V}_\theta^{(h)}(s) - \mathbb{E}[\nabla_s \hat{V}_\theta^{(h)}(s)] \right\|_2 \\
&= \left\| \nabla_s \hat{r}(s, a_h) - \mathbb{E}_a[\nabla_s \hat{r}(s, a)] \right\|_2 + \gamma \left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) \nabla_s \hat{f}(s, a_h) - \mathbb{E}_{a, \xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \nabla_s \hat{f}(s, a)] \right\|_2 \\
&\leq 2|\mathcal{A}|L_{\hat{r}} + \gamma \left\| \nabla_s \hat{f}(s, a_h) \right\|_2 \left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \mathbb{E}_{a, \xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h)] \right\|_2 \\
&+ \gamma \left\| \mathbb{E}[\nabla_s \hat{f}(s, a_h)] \right\|_2 \mathbb{E}_{a, \xi} \left[ \left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a_h) + \xi_h) - \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s, a) + \xi) \right\|_2 \right] \\
&\leq 2|\mathcal{A}|L_{\hat{r}} + \gamma \left\| \nabla_s \hat{f}(s, a_h) \right\|_2 B_1^{(h+1)}(\pi, \xi) + \gamma L \left\| \mathbb{E}[\nabla_s \hat{f}(s, a_h)] \right\|_2 \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) \\
&= B_1^{(h)}(\pi, \xi). \tag{21}
\end{aligned}$$

To obtain the desired  $v_t$ , we need  $B_0^{(1)}(\pi, \xi)$ , which depends on  $B_1^{(h)}(\pi, \xi)$ . From Eq. (21), we have the recursive expression of  $B_1^{(h)}(\pi, \xi)$ . Denote  $L_{\hat{f}} = \max \left( 1, \left\| \nabla \hat{f}(s, a) \right\|_2 \right)$ . Then

$$\begin{aligned}
B_1^{(h)} &= \gamma \left\| \nabla_s \hat{f}(s, a_h) \right\|_2 B_1^{(h+1)} + \gamma L \left\| \mathbb{E}[\nabla_s \hat{f}(s, a_h)] \right\|_2 \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + 2|\mathcal{A}|L_{\hat{r}} \\
&\leq \gamma L_{\hat{f}} B_1^{(h+1)} + \gamma L_{\hat{f}} L \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + 2|\mathcal{A}|L_{\hat{r}} \\
&\leq \gamma^2 L_{\hat{f}}^2 B_1^{(h+2)} + \gamma^2 L_{\hat{f}}^2 L \left( \|\xi_{h+1}\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + \gamma L_{\hat{f}} L \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + 2(1 + \gamma)|\mathcal{A}|L_{\hat{r}} \\
&\leq \gamma^2 L_{\hat{f}}^2 B_1^{(h+2)} + \gamma^2 L_{\hat{f}}^2 L \left( \|\xi_{h+1}\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + \gamma L_{\hat{f}}^2 L \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + 2(1 + \gamma)|\mathcal{A}|L_{\hat{r}} \\
&\leq \dots \\
&\leq L_{\hat{f}}^H L \sum_{h'=h}^H \gamma^{h'-h+1} \left( \|\xi_{h'}\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + 2 \sum_{h'=0}^{H-h} \gamma^{h'} |\mathcal{A}|L_{\hat{r}}. \tag{22}
\end{aligned}$$

Thus,

$$\begin{aligned}
B_1^{(1)}(\pi, \xi) &\leq L_{\hat{f}}^H L \sum_{h=1}^H \gamma^h \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) + 2 \sum_{h=0}^{H-h} \gamma^h |\mathcal{A}|L_{\hat{r}} \\
&\leq HL_{\hat{f}}^H L \left( H\sigma_\xi \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \sum_{h=1}^H \|\xi_h\|_2 \right) + 2H|\mathcal{A}|L_{\hat{r}}. \tag{23}
\end{aligned}$$

Finally,

$$\begin{aligned}
& \left\| \nabla_\theta \hat{V}_\theta^{(1)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_\theta^{(1)}(s)] \right\|_2 \leq B_0^{(1)}(\pi, \xi) \\
&= 2|\mathcal{A}|L_{\hat{r}}\epsilon + \gamma B_0^{(2)}(\pi, \xi) + \gamma \left\| \nabla_\theta \hat{f}(s, a_1) \right\|_2 B_1^{(2)}(\pi, \xi) \\
&+ \gamma L \left( 1 + \left\| \mathbb{E}_a[\nabla_\theta \hat{f}(s, a)] \right\|_2 \right) \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) \\
&\leq \sum_{h=1}^H \gamma^{h-1} 2|\mathcal{A}|L_{\hat{r}}\epsilon + \gamma^h \left( L_{\hat{f}} B_1^{(h+1)}(\pi, \xi) + L(1 + L_{\hat{f}}) \left( \|\xi_h\|_2 + \sigma_\xi \sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}} \right) \right) \\
&\leq \left( H^2 L_{\hat{f}}^H L + HL(1 + L_{\hat{f}}) \right) \left( H\sigma_\xi \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \sum_{h=1}^H \|\xi_h\|_2 \right) + 2H|\mathcal{A}|L_{\hat{r}}(\epsilon + L_{\hat{f}}) \\
&\leq \left( H^2 L_{\hat{f}}^H L + HL(1 + L_{\hat{f}}) + \frac{L_{\hat{r}}\epsilon}{L_{\hat{f}}} + L_{\hat{r}} \right) \left( H\sigma_\xi \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \sum_{h=1}^H \|\xi_h\|_2 \right). \tag{24}
\end{aligned}$$

And

$$\begin{aligned}
v_t &= \mathbb{E} \left[ \left\| \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] \\
&\leq \left( H^2 L_{\hat{f}}^H L + HL(1 + L_{\hat{f}}) + \frac{L_{\hat{r}} \epsilon}{L_{\hat{f}}} + L_{\hat{r}} \right)^2 \\
&\quad \left( \left( H\sigma_{\xi} \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} \right)^2 + 2H\sigma_{\xi} \sqrt{d_s} \left( H\sigma_{\xi} \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} \right) + \mathbb{E} \left[ \left( \sum_{h=1}^H \|\xi_h\|_2 \right)^2 \right] \right) \\
&= \left( H^2 L_{\hat{f}}^H L + HL(1 + L_{\hat{f}}) + \frac{L_{\hat{r}} \epsilon}{L_{\hat{f}}} + L_{\hat{r}} \right)^2 \\
&\quad \left( \left( H\sigma_{\xi} \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} \right)^2 + 2H\sigma_{\xi} \sqrt{d_s} \left( H\sigma_{\xi} \sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \frac{1}{2} \right) \right) \\
&= \left( H^2 L_{\hat{f}}^H L + HL(1 + L_{\hat{f}}) + \frac{L_{\hat{r}} \epsilon}{L_{\hat{f}}} + L_{\hat{r}} \right)^2 \left( 3H^2 \sigma_{\xi}^2 d_s + 4H^2 |\mathcal{A}|^2 L_{\hat{f}}^2 + 8H^2 \sigma_{\xi} \sqrt{d_s} |\mathcal{A}| L_{\hat{f}} + H\sigma_{\xi} \sqrt{d_s} \right)
\end{aligned} \tag{25}$$

□

## 7.5 Proof of Proposition 2

*Proof.* Suppose

$$\begin{aligned}
\left\| \nabla_{\theta} V_{\theta}^{(h)}(s) - \mathbb{E}[\nabla_{\theta} \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 &= \left\| \mathbb{E}[\nabla_{\theta} V_{\theta}^{(h)}(s)] - \mathbb{E}[\nabla_{\theta} \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 \leq B_2^{(h)}(\pi) \\
\left\| \nabla_s V_{\theta}^{(h)}(s) - \mathbb{E}[\nabla_s \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 &= \left\| \mathbb{E}[\nabla_s V_{\theta}^{(h)}(s)] - \mathbb{E}[\nabla_s \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 \leq B_3^{(h)}(\pi)
\end{aligned}$$

By induction,

$$\begin{aligned}
&\left\| \nabla_{\theta} V_{\theta}^{(h)}(s) - \mathbb{E}[\nabla_{\theta} \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 \\
&\leq \left\| \nabla_{\theta} R_{\theta}(s) - \nabla_{\theta} \hat{R}_{\theta}(\hat{s}) \right\|_2 + \gamma \left\| \mathbb{E}_{a,\xi} \left[ \nabla_{\theta} V_{\theta}^{(h+1)}(f(s, a) + \xi) \right] - \mathbb{E}_{a,\xi} \left[ \nabla_{\theta} \hat{V}_{\theta}^{(h+1)}(\hat{f}(\hat{s}, a) + \xi) \right] \right\|_2 \\
&+ \gamma \left\| \mathbb{E}_{a,\xi} \left[ \nabla_s V_{\theta}^{(h+1)}(f(s, a) + \xi) \nabla_{\theta} f(s, a) \right] - \mathbb{E}_{a,\xi} \left[ \nabla_s \hat{V}_{\theta}^{(h+1)}(\hat{f}(\hat{s}, a) + \xi) \nabla_{\theta} \hat{f}(\hat{s}, a) \right] \right\|_2 \\
&\leq \left\| \nabla_{\theta} R_{\theta}(s) - \nabla_{\theta} \hat{R}_{\theta}(\hat{s}) \right\|_2 + \gamma B_2^{(h+1)}(\pi) + \gamma \left\| \mathbb{E}_a [\nabla_{\theta} \hat{f}(\hat{s}, a)] \right\|_2 B_3^{(h+1)}(\pi) + \gamma L \left\| \mathbb{E}_a [\nabla_{\theta} \hat{f}(\hat{s}, a) - \nabla_{\theta} f(s, a)] \right\|_2 \\
&= B_2^{(h)}(\pi),
\end{aligned} \tag{26}$$

where the last inequality holds since

$$\begin{aligned}
&\mathbb{E}_{a,\xi} \left[ \nabla_s V_{\theta}^{(h+1)}(f(s, a) + \xi) \nabla_{\theta} f(s, a) \right] - \mathbb{E}_{a,\xi} \left[ \nabla_s \hat{V}_{\theta}^{(h+1)}(\hat{f}(\hat{s}, a) + \xi) \nabla_{\theta} \hat{f}(\hat{s}, a) \right] \\
&= \mathbb{E}_{a,\xi} \left[ \left( \nabla_s V_{\theta}^{(h+1)}(f(s, a) + \xi) - \nabla_s \hat{V}_{\theta}^{(h+1)}(\hat{f}(\hat{s}, a) + \xi) \right) \nabla_{\theta} \hat{f}(\hat{s}, a) \right] \\
&- \mathbb{E}_{a,\xi} \left[ \nabla_s V_{\theta}^{(h+1)}(f(s, a) + \xi) \left( \nabla_{\theta} \hat{f}(\hat{s}, a) - \nabla_{\theta} f(s, a) \right) \right].
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
& \left\| \nabla_s V_\theta^{(h)}(s) - \mathbb{E}[\nabla_s \hat{V}_\theta^{(h)}(\hat{s})] \right\|_2 \\
& \leq \left\| \nabla_s R_\theta(s) - \nabla_s \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma \left\| \mathbb{E}_{a,\xi} \left[ \nabla_s V_\theta^{(h+1)}(f(s,a) + \xi) \nabla_s f(s,a) \right] - \mathbb{E}_{a,\xi} \left[ \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(\hat{s},a) + \xi) \nabla_s \hat{f}(\hat{s},a) \right] \right\|_2 \\
& \leq \left\| \nabla_s R_\theta(s) - \nabla_s \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma \left\| \mathbb{E}_a [\nabla_s \hat{f}(\hat{s},a)] \right\|_2 B_3^{(h+1)}(\pi) + \gamma L \left\| \mathbb{E}_a [\nabla_s \hat{f}(\hat{s},a) - \nabla_s f(s,a)] \right\|_2 \\
& = B_3^{(h)}(\pi),
\end{aligned} \tag{27}$$

Thus, with the recursive structure, we can write  $B_3^{(h)}(\pi)$  as

$$B_3^{(h)}(\pi) = \sum_{h'=h}^H \gamma^{h'-h} \left\| \mathbb{E}_a [\nabla_s \hat{f}(\hat{s},a)] \right\|_2^{h'-h} \left( \left\| \nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma L \left\| \mathbb{E}_a [\nabla_s \hat{f}(\hat{s},a) - \nabla_s f(s,a)] \right\|_2 \right) \tag{28}$$

Recall that  $L_{\hat{f}} = \max \left( 1, \left\| \nabla \hat{f}(s,a) \right\|_2 \right)$ . Following the definition that

$$\begin{aligned}
\delta_{t,s} &= \sum_{h=1}^H \left( \gamma^{h-1} \left\| \nabla_s R_\theta(s) - \nabla_s \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma^h L \left\| \mathbb{E}_a [\nabla_s f(s,a) - \nabla_s \hat{f}(\hat{s},a)] \right\|_2 \right) \\
\delta_{t,a} &= \sum_{h=1}^H \left( \gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma^h L \left\| \mathbb{E}_a [\nabla_a f(s,a) \nabla_\theta \pi(a|s) - \nabla_a \hat{f}(\hat{s},a) \nabla_\theta \pi(a|\hat{s})] \right\|_2 \right)
\end{aligned}$$

Then

$$\begin{aligned}
B_3^{(1)}(\pi) &= \sum_{h=1}^H \gamma^{h-1} \left\| \mathbb{E}_a [\nabla_s \hat{f}(\hat{s},a)] \right\|_2^{h-1} \left( \left\| \nabla_s R_\theta(s) - \nabla_s \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma L \left\| \mathbb{E}_a [\nabla_s \hat{f}(\hat{s},a) - \nabla_s f(s,a)] \right\|_2 \right) \\
&\leq L_{\hat{f}}^H \delta_{t,s}.
\end{aligned} \tag{29}$$

Finally, recall the definition  $\epsilon = \max \nabla_\theta \pi(a|s)$ . The gradient bias is bounded by

$$\begin{aligned}
b_t &= \left\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2 \\
&= B_2^{(1)}(\pi) \\
&= \sum_{h=1}^H \gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma^h \left\| \mathbb{E}_a [\nabla_\theta \hat{f}(\hat{s},a)] \right\|_2 B_3^{(h+1)}(\pi) + \gamma^h L \left\| \mathbb{E}_a [\nabla_\theta \hat{f}(\hat{s},a) - \nabla_\theta f(s,a)] \right\|_2 \\
&\leq \epsilon H L_{\hat{f}}^H \delta_{t,s} + \sum_{h=1}^H \gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma^h L \left\| \mathbb{E}_a [\nabla_\theta \hat{f}(\hat{s},a) - \nabla_\theta f(s,a)] \right\|_2 \\
&= \epsilon H L_{\hat{f}}^H \delta_{t,s} + \sum_{h=1}^H \gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(\hat{s}) \right\|_2 + \gamma^h L \left\| \mathbb{E}_a [\nabla_a \hat{f}(\hat{s},a) \nabla_\theta \pi(a|\hat{s}) - \nabla_a f(s,a) \nabla_\theta \pi(a|s)] \right\|_2 \\
&\leq \epsilon H L_{\hat{f}}^H \delta_{t,s} + \delta_{t,a}
\end{aligned} \tag{30}$$

□

## 7.6 Proof of Proposition 3

*Proof.*

$$\begin{aligned}
v_t &= \mathbb{E} \left[ \left\| \nabla_{\theta} \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] \\
&= \mathbb{E} \left[ \left\| \nabla_a \hat{Q} \nabla_{\theta} a_{\epsilon} - \mathbb{E}_a[\nabla_a \hat{Q} \nabla_{\theta} a] \right\|_2^2 \right] \\
&\leq L_{\hat{Q}}^2 \max_{\epsilon} \mathbb{E} \left[ \left\| \nabla_{\theta} F^{-1}(\epsilon; \theta) \right\|_2^2 \right].
\end{aligned} \tag{31}$$

Besides,

$$\begin{aligned}
b_t &= \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2 \\
&= \left\| \nabla_{\theta} \mathbb{E}_a[Q(s_0, a)] - \mathbb{E}_a[\nabla_a \hat{Q} \nabla_{\theta} a] \right\|_2 \\
&= \left\| \mathbb{E}_a[\nabla_a Q \nabla_{\theta} a] - \mathbb{E}_a[\nabla_a \hat{Q} \nabla_{\theta} a] \right\|_2 \\
&\leq \left\| \mathbb{E}_{\epsilon}[\nabla_{\theta} F^{-1}(\epsilon; \theta)] \right\|_2 \delta_t
\end{aligned} \tag{32}$$

□

## 7.7 Proof of Proposition 5

*Proof.* For

$$\hat{V}_{\pi, H}(s_0) = V^{\hat{f}}(s_0) + \gamma^H \hat{V}_{\phi}(\hat{s}_H),$$

we have

$$\nabla_{\theta} \hat{V}_{\pi, H}(s_0) = \nabla_{\theta} V^{\hat{f}}(s_0) + \gamma^H \nabla_{\theta} \hat{V}_{\phi}(\hat{s}_H) + \gamma^H \nabla_s \hat{V}_{\phi}(\hat{s}_H) \nabla_{\theta} \hat{s}_H$$

Thus, the gradient bias is bounded by

$$\begin{aligned}
b_t &= \left\| \nabla_{\theta} J(\pi_{\theta_t}) - \mathbb{E}[\nabla_{\theta} \hat{J}(\pi_{\theta_t})] \right\|_2 \\
&\leq \left\| \nabla_{\theta} V(s_0) - \mathbb{E}[\nabla_{\theta} V^{\hat{f}}(s_0)] \right\|_2 + \gamma^H \left\| \mathbb{E}_{s_H}[\nabla_{\theta} V(s_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_{\phi}(\hat{s}_H)] \right\|_2 \\
&\quad + \gamma^H \left\| \mathbb{E}_{s_H}[\nabla_s V(s_H) \nabla_{\theta} s_H] - \mathbb{E}_{\hat{s}_H}[\nabla_s \hat{V}_{\phi}(\hat{s}_H) \nabla_{\theta} \hat{s}_H] \right\|_2
\end{aligned} \tag{33}$$

The first term is bounded by Proposition 2. The second term is bounded by

$$\begin{aligned}
&\left\| \mathbb{E}_{s_H}[\nabla_{\theta} V(s_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_{\phi}(\hat{s}_H)] \right\|_2 \\
&\leq \left\| \mathbb{E}_{s_H}[\nabla_{\theta} V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_{\phi}(\hat{s}_H)] \right\|_2 + \left\| \mathbb{E}_{s_H}[\nabla_{\theta} V^{\hat{f}}(s_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} V^{\hat{f}}(\hat{s}_H)] \right\|_2 + \left\| \mathbb{E}_{s_H}[\nabla_{\theta} V(s_H)] - \mathbb{E}_{s_H}[\nabla_{\theta} V^{\hat{f}}(s_H)] \right\|_2 \\
&\leq \left\| \mathbb{E}_{s_H}[\nabla_{\theta} V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H}[\nabla_{\theta} \hat{V}_{\phi}(\hat{s}_H)] \right\|_2 + L_{\nabla V^{\hat{f}}} \mathbb{E}[\|s_H - \hat{s}_H\|_2] + B_2^{(0)}(\pi)
\end{aligned} \tag{34}$$

Recall the definitions in the proof of Proposition 2

$$\begin{aligned}
\left\| \nabla_{\theta} V_{\theta}^{(h)}(s) - \mathbb{E}[\nabla_{\theta} \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 &= \left\| \mathbb{E}[\nabla_{\theta} V_{\theta}^{(h)}(s)] - \mathbb{E}[\nabla_{\theta} \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 \leq B_2^{(h)}(\pi) \\
\left\| \nabla_s V_{\theta}^{(h)}(s) - \mathbb{E}[\nabla_s \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 &= \left\| \mathbb{E}[\nabla_s V_{\theta}^{(h)}(s)] - \mathbb{E}[\nabla_s \hat{V}_{\theta}^{(h)}(\hat{s})] \right\|_2 \leq B_3^{(h)}(\pi)
\end{aligned}$$

The Lipschitz constant  $L_{\nabla V^{\hat{f}}}$  can be further specified in a similar manner with Fact 1.

Similarly,

$$\begin{aligned}
& \left\| \mathbb{E}_{s_H} [\nabla_s V(s_H) \nabla_\theta s_H] - \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H) \nabla_\theta \hat{s}_H] \right\|_2 \\
&= \left\| \mathbb{E}_{s_H} [\nabla_s V(s_H) \nabla_\theta s_H] - \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H) \nabla_\theta s_H] + \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H) \nabla_\theta s_H] - \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H) \nabla_\theta \hat{s}_H] \right\|_2 \\
&\leq \|\mathbb{E}[\nabla_\theta s_H]\|_2 \left\| \mathbb{E}_{s_H} [\nabla_s V(s_H)] - \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H)] \right\|_2 + \left\| \nabla_s \hat{V}_\phi(\hat{s}_H) \right\|_2 \left\| \mathbb{E}_{\hat{s}_H} [\nabla_\theta s_H - \nabla_\theta \hat{s}_H] \right\|_2 \\
&\leq \|\mathbb{E}[\nabla_\theta s_H]\|_2 \left\| \mathbb{E}_{\hat{s}_H} [\nabla_s V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H)] \right\|_2 + \|\mathbb{E}[\nabla_\theta s_H]\|_2 L_{\nabla \hat{V}} \mathbb{E}[\|s_H - \hat{s}_H\|_2] + \|\mathbb{E}[\nabla_\theta s_H]\|_2 B_3^{(0)}(\pi) \\
&+ \left\| \nabla_s \hat{V}_\phi(\hat{s}_H) \right\|_2 \left\| \mathbb{E}_{\hat{s}_H} [\nabla_\theta s_H - \nabla_\theta \hat{s}_H] \right\|_2
\end{aligned} \tag{35}$$

Thus,

$$\begin{aligned}
b_t &\leq B_2^{(0)}(\pi) + \gamma^H \left( \left\| \mathbb{E}_{\hat{s}_H} [\nabla_\theta V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H} [\nabla_\theta \hat{V}_\phi(\hat{s}_H)] \right\|_2 + L_{\nabla V^{\hat{f}}} \mathbb{E}[\|s_H - \hat{s}_H\|_2] + B_2^{(0)}(\pi) \right) \\
&+ \gamma^H \|\mathbb{E}[\nabla_\theta s_H]\|_2 \left( \left\| \mathbb{E}_{\hat{s}_H} [\nabla_s V^{\hat{f}}(\hat{s}_H)] - \mathbb{E}_{\hat{s}_H} [\nabla_s \hat{V}_\phi(\hat{s}_H)] \right\|_2 + \|\nabla_s \hat{V}_\phi(\hat{s}_H)\|_2 L_{\nabla \hat{V}} \mathbb{E}[\|s_H - \hat{s}_H\|_2] + B_3^{(0)}(\pi) \right) \\
&+ \gamma^H \left\| \nabla_s \hat{V}_\phi(\hat{s}_H) \right\|_2 \left\| \mathbb{E}_{\hat{s}_H} [\nabla_\theta s_H - \nabla_\theta \hat{s}_H] \right\|_2
\end{aligned} \tag{36}$$

□

## References

- [1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- [2] Osbert Bastani. Sample complexity of estimating the policy gradient for nearly deterministic dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 3858–3869. PMLR, 2020.
- [3] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.
- [4] Johan Bjorck, Carla P Gomes, and Kilian Q Weinberger. Towards deeper deep reinforcement learning. *arXiv preprint arXiv:2106.01151*, 2021.
- [5] Ignasi Clavera, Violet Fu, and Pieter Abbeel. Model-augmented actor-critic: Backpropagating through paths. *arXiv preprint arXiv:2005.08068*, 2020.
- [6] Greg Farquhar, Kate Baumli, Zita Marinho, Angelos Filos, Matteo Hessel, Hado P van Hasselt, and David Silver. Self-consistent models and values. *Advances in Neural Information Processing Systems*, 34, 2021.
- [7] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018.
- [8] Nicolas Heess, Greg Wayne, David Silver, Timothy Lillicrap, Yuval Tassa, and Tom Erez. Learning continuous control policies by stochastic value gradients. *arXiv preprint arXiv:1510.09142*, 2015.
- [9] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.
- [10] Luke Metz, C Daniel Freeman, Samuel S Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021.
- [11] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.
- [12] Paavo Parmas, Carl Edward Rasmussen, Jan Peters, and Kenji Doya. Pippis: Flexible model-based policy search robust to the curse of chaos. In *International Conference on Machine Learning*, pages 4065–4074. PMLR, 2018.
- [13] Matteo Pirodda, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, 2015.
- [14] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.