Global Convergence of Analytic 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 analytic estimator does not sum over terms for which the parameter has no effect, allowing the estimator variance to be much lower [9].

The analytic 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 [6]. 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 analytic 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 analytic policy gradient algorithms, including *Back-Propagation Through Time* (BPTT), *Stochastic Value Gradient* (SVG) [6], and *Model-Augmented Actor-Critic* (MAAC) [5]. We find that with the gradient calculated by model backpropgation, the convergence rates depend exponentially on the Jacobian of the learned dynamics and the bias of the model gradient. This indicates two sources of sample inefficiency: the curse of chaos, *i.e.*, small changes in initial conditions result in diverging states, and the objective mismatch, *i.e.*, we want the gradient through the model, not the state prediction, to be accurate (c.f. Section 4).

Besides, the analysis at the same time suggests several principled modifications to improve the efficiency of model-based analytic gradient algorithms, such as learning a non-chaotic version of physics with spectral normalization and leveraging a gradient-aware objective for model learning. These modifications also share similarities with several recent works [8, 10, 4] based on their experimental observations.

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

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

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

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

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

Work in progress. To be submitted to ICML 2022.

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

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, 12, 11].

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 ∇V_{π_θ} is L-Lipschitz, where $L = 44H^5\bar{L}_r\bar{L}_f^{4H}$. Particularly, $\nabla_\theta V_{\pi_\theta}$ is Lipschitz continuous in θ with Lipschitz constant $24H^5\bar{L}_r\bar{L}_f^{4H}$.*

$$\begin{aligned}
\text{gradient mapping:} \quad & \rho_t = \frac{1}{\eta}\left[P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta\nabla_\theta J(\pi_{\theta_t}))\right] \\
\text{gradient bias:} \quad & b_t = \left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2 \\
\text{upper bound of squared variance:} \quad & v_t = \mathbb{E}\left[\left\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2^2\right]
\end{aligned}$$

With a similar proof in [12], we have the following bound.

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

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

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*

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

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

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

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

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

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 $\eta \le \frac{1}{L}$,*

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

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 θ_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 $V_\theta^{(h)}(s)$. Consider the transition $s_{h+1} = f(s_h, a_h) + \xi_h$, with $\xi_h \sim p(\xi)$ and assume $p(\xi)$ is σ_ξ-subgaussian. Denote $\hat{R}_\theta(s) = \mathbb{E}_{a\sim\pi}[r(s,a)]$ and $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_H)$. Write $\hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h; a_h \sim \pi, \xi_h \sim \boldsymbol{\xi})$ as $\hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h)$ for clarity. Then we have

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

For finite-horizon γ-discounted MDPs, we have the following results.

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

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

Proposition 2. *Define the "error" at iteration* t *as*

$$\delta_t = \sum_{h=1}^{H}\left(\gamma^{h-1}\left\|\nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(s)\right\|_2 + \gamma^h L \left\|\mathop{\mathbb{E}}_a\left[\nabla f(s,a) - \nabla \hat{f}(s,a)\right]\right\|_2\right).$$

Then we have

$$b_t = \left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2 \leq \left(HL_{\hat{f}}^H + 1\right)\delta_t. \tag{5}$$

We can see that both v_t and b_t contains terms that are exponential in $L_{\hat{f}}$. Some works [8, 10] 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).

5 SVG(0)

SVG(0) [6] is a model-free algorithm, and is the stochastic analogue of DPG.

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

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 [9].

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

One simple way is to set $p(\epsilon)$ as the uniform distribution and μ 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 $\delta_t = \left\|\mathbb{E}[\nabla_a Q - \nabla_a \hat{Q}]\right\|_2$, then*

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

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

The key point here is that we need $\hat{Q}(s,a)$ to be Lipschitz coninuous 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 Proofs

6.1 Proof of Lemma 1

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

$$\begin{aligned} J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) &\ge \nabla_\theta J(\pi_{\theta_t})^\top(\theta_{t+1}-\theta_t) - \frac{L}{2}\|\theta_{t+1}-\theta_t\|_2^2 \\ &= \eta \nabla_\theta J(\pi_{\theta_t})^\top \beta_t - \frac{L\eta^2}{2}\|\beta_t\|_2^2. \end{aligned} \tag{8}$$

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. (8) by bounding the resulting three terms.

$$\left|\left(\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right)^\top \beta_t\right| \le \|\beta_t\|_2 \|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2 = \|\beta_t\|_2 b_t \tag{9}$$

$$\left(\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right)^\top \beta_t \le \frac{\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2} \tag{10}$$

$$\nabla_\theta \hat{J}(\pi_{\theta_t})^\top \beta_t \ge \|\beta_t\|_2^2, \tag{11}$$

where Eq. (11) holds due to the fact that $\left(\theta_{t+1} - (\theta_t + \eta \nabla_\theta \hat{J}(\theta_t))\right)^\top (\theta_{t+1} - \theta_t) \le 0$.

Thus, we can bound Eq. (8) by

$$J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) \ge \eta \left(-\|\beta_t\|_2 b_t - \frac{\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2}\right) - \frac{L\eta^2}{2}\|\beta_t\|_2^2. \tag{12}$$

By taking expectation in Eq. (12), we have

$$\begin{aligned}\left(\frac{\eta}{2}-\frac{L\eta^2}{2}\right)\mathbb{E}[\|\beta_t\|_2^2] &\leq \mathbb{E}[J(\pi_{\theta_{t+1}})-J(\pi_{\theta_t})]+\eta\,\mathbb{E}[\|\beta_t\|_2]b_t+\frac{\eta}{2}v_t \\ &\leq \mathbb{E}[J(\pi_{\theta_{t+1}})-J(\pi_{\theta_t})]+\eta b_t+\frac{\eta}{2}v_t\end{aligned} \tag{13}$$

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

$$\begin{aligned}\mathbb{E}[\|\rho_t-\beta_t\|_2^2] &\leq \mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t})-\nabla_\theta\hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t})-\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]+\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]-\nabla_\theta\hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &\leq 2\,\mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t})-\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]\right\|_2^2\right]+2\,\mathbb{E}\left[\left\|\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]-\nabla_\theta\hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &= 2\left\|\nabla_\theta J(\pi_{\theta_t})-\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]\right\|_2^2+2\,\mathbb{E}\left[\left\|\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]-\nabla_\theta\hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &\leq 2b_t^2+2v_t.\end{aligned} \tag{14}$$

For $\eta\leq\frac{1}{L}$, $\frac{\eta}{2}-\frac{L\eta^2}{2}>0$. Then with Eq. (13) and Eq. (14), we can bound $\min_{t\in[T]}\mathbb{E}\left[\|\rho_t\|_2^2\right]$ by

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

□

6.2 Proof of Theorem 1

Proof. From Lemma 2, we know that for initial distribution ζ,

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

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

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

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

6.3 Proof of Theorem 2

Proof. From the performance difference lemma [7],

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

Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{s_0\sim\mu_1}[V_{\pi^*}(s_0)] - \mathop{\mathbb{E}}_{s_0\sim\mu_1}[V_\pi(s_0)] &= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi^*}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\pi^*}[A^\pi(s,a)] \\
&\le \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi^*}_{\mu_1}} \left[\max_{\bar a} A^\pi(s,\bar a)\right] \\
&= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \left[\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}} \max_{\bar a} A^\pi(s,\bar a)\right] \\
&\le \frac{1}{1-\gamma} \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \left[\max_{\bar a} A^\pi(s,\bar a)\right] \\
&= \frac{1}{1-\gamma} \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \max_{\bar\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\bar\pi}[\bar\pi(a|s)A^\pi(s,a)] \\
&= \frac{1}{1-\gamma} \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \max_{\bar\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\bar\pi}[(\bar\pi(a|s)-\pi(a|s))A^\pi(s,a)] \\
&= \frac{1}{1-\gamma} \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \max_{\bar\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\bar\pi}[(\bar\pi(a|s)-\pi(a|s))Q^\pi(s,a)].
\end{aligned} \tag{15}$$

It holds that

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

Then Eq. (15) can be further bounded by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{s_0\sim\mu_1}[V_{\pi^*}(s_0)] - \mathop{\mathbb{E}}_{s_0\sim\mu_1}[V_\pi(s_0)] \\
&\le \frac{1}{1-\gamma} \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \max_{\bar\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\bar\pi}[(\bar\pi(a|s)-\pi(a|s))Q^\pi(s,a)] \\
&\le \frac{1}{1-\gamma} \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \max_{s,a} \frac{1}{\|\nabla_\theta\pi(a|s)\|} \max_{\bar\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\bar\pi}[(\bar\pi(a|s)-\pi(a|s))Q^\pi(s,a)\,\|\nabla_\theta\pi(a|s)\|] \\
&= \left\|\frac{d^{\pi^*}_{\mu_1}}{d^{\pi}_{\mu_1}}\right\|_\infty \max_{s,a} \frac{1}{\|\nabla_\theta\pi(a|s)\|} \max_{\bar\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu_1}} \mathop{\mathbb{E}}_{a\sim\bar\pi}[(\bar\pi(a|s)-\pi(a|s))\,\|\nabla_\theta J(\theta)\|]
\end{aligned} \tag{16}$$

Similar with the proof in Theorem 1, we have

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

□

6.4 Proof of Proposition 1

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

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

Proof. Suppose

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

By induction,

$$
\begin{aligned}
&\left\| \nabla_\theta \hat{V}_\theta^{(h)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_\theta^{(h)}(s)] \right\|_2 \\
&\le \gamma \left\| \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h) - \mathop{\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) - \mathop{\mathbb{E}}_{a,\xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a) + \xi) \nabla_\theta \hat{f}(s,a)] \right\|_2 \\
&\le \gamma \left\| \nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h) - \mathop{\mathbb{E}}_{a,\xi}[\nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h)] \right\|_2 \\
&+ \gamma \mathop{\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) - \mathop{\mathbb{E}}_{a,\xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h)] \right\|_2 \\
&+ \gamma \left\| \mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a)] \right\|_2 \mathop{\mathbb{E}}_{a,\xi} \left[\left\| \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h) + \xi_h) - \nabla_\theta \hat{V}_s^{(h+1)}(\hat{f}(s,a) + \xi) \right\|_2 \right] \\
&\le \gamma B_0^{(h+1)}(\pi, \boldsymbol{\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, \boldsymbol{\xi}) + \gamma L \left\| \mathop{\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) \\
&\le \gamma B_0^{(h+1)}(\pi, \boldsymbol{\xi}) + \gamma \left\| \nabla_\theta \hat{f}(s,a_h) \right\|_2 B_1^{(h+1)}(\pi, \boldsymbol{\xi}) \\
&+ \gamma L \left(1 + \left\| \mathop{\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, \boldsymbol{\xi}).
\end{aligned}
\tag{17}
$$

Similarly, we also have

$$
\begin{aligned}
&\left\|\nabla_\theta \hat{V}_s^{(h)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_s^{(h)}(s)]\right\|_2 \\
&= \gamma \left\|\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h)+\xi_h)\nabla_s \hat{f}(s,a_h) - \mathop{\mathbb{E}}_{a,\xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a)+\xi)\nabla_s \hat{f}(s,a)]\right\|_2 \\
&\le \gamma \left\|\nabla_s \hat{f}(s,a_h)\right\|_2 \left\|\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h)+\xi_h) - \mathop{\mathbb{E}}_{a,\xi}[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h)+\xi_h)]\right\|_2 \\
&+ \gamma \left\|\mathbb{E}[\nabla_s \hat{f}(s,a_h)]\right\|_2 \mathop{\mathbb{E}}_{a,\xi}\left[\left\|\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a_h)+\xi_h) - \nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a)+\xi)\right\|_2\right] \\
&\le \gamma \left\|\nabla_s \hat{f}(s,a_h)\right\|_2 B_1^{(h+1)}(\pi,\boldsymbol{\xi}) + \gamma L \left\|\mathbb{E}[\nabla_s \hat{f}(s,a_h)]\right\|_2 \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&= B_1^{(h)}(\pi,\boldsymbol{\xi}).
\end{aligned}
\tag{18}
$$

To obtain the desired v_t, we need $B_0^{(1)}(\pi,\boldsymbol{\xi})$, which depends on $B_1^{(h)}(\pi,\boldsymbol{\xi})$. From Eq. (18), we have the recursive expression of $B_1^{(h)}(\pi,\boldsymbol{\xi})$. Denote $L_{\hat{f}} = \max\left(1, \left\|\|\nabla \hat{f}(s,a)\|_2\right\|_\infty\right)$. Then

$$
\begin{aligned}
B_1^{(h)} &= \gamma \left\|\nabla_s \hat{f}(s,a_h)\right\|_2 B_1^{(h+1)} + \gamma L \left\|\mathbb{E}[\nabla_s \hat{f}(s,a_h)]\right\|_2 \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&\le \gamma L_{\hat{f}} B_1^{(h+1)} + \gamma L_{\hat{f}} L \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&\le \gamma^2 L_{\hat{f}}^2 B_1^{(h+2)} + \gamma^2 L_{\hat{f}}^2 L \left(\|\xi_{h+1}\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) + \gamma L_{\hat{f}} L \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&\le \gamma^2 L_{\hat{f}}^2 B_1^{(h+2)} + \gamma^2 L_{\hat{f}}^2 L \left(\|\xi_{h+1}\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) + \gamma L_{\hat{f}}^2 L \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&\le \cdots \\
&\le L_{\hat{f}}^H L \sum_{h'=h}^{H} \gamma^{h'-h+1} \left(\|\xi_{h'}\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right).
\end{aligned}
\tag{19}
$$

Thus,

$$
\begin{aligned}
B_1^{(1)}(\pi,\boldsymbol{\xi}) &\le L_{\hat{f}}^H L \sum_{h=1}^{H} \gamma^h \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&\le H L_{\hat{f}}^H L \left(H\sigma_\xi\sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \sum_{h=1}^{H}\|\xi_h\|_2\right).
\end{aligned}
\tag{20}
$$

Finally,

$$
\begin{aligned}
\left\|\nabla_\theta \hat{V}_\theta^{(1)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_\theta^{(1)}(s)]\right\|_2 &\le B_0^{(1)}(\pi,\boldsymbol{\xi}) \\
&= \gamma B_0^{(2)}(\pi,\boldsymbol{\xi}) + \gamma \left\|\nabla_\theta \hat{f}(s,a_1)\right\|_2 B_1^{(2)}(\pi,\boldsymbol{\xi}) \\
&+ \gamma L \left(1 + \left\|\mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a)]\right\|_2\right) \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right) \\
&\le \sum_{h=1}^{H} \gamma^h \left(L_{\hat{f}} B_1^{(h+1)}(\pi,\boldsymbol{\xi}) + L(1+L_{\hat{f}}) \left(\|\xi_h\|_2 + \sigma_\xi\sqrt{d_s} + 2|\mathcal{A}|L_{\hat{f}}\right)\right) \\
&\le \left(H^2 L_{\hat{f}}^H L + HL(1+L_{\hat{f}})\right) \left(H\sigma_\xi\sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \sum_{h=1}^{H}\|\xi_h\|_2\right).
\end{aligned}
\tag{21}
$$

And

$$
\begin{aligned}
v_t &= \mathbb{E}\left[\left\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2^2\right] \\
&\leq \left(H^2 L_{\hat{f}}^H L + HL(1+L_{\hat{f}})\right)^2 \\
&\left(\left(H\sigma_\xi\sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}}\right)^2 + 2H\sigma_\xi\sqrt{d_s}\left(H\sigma_\xi\sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}}\right) + \mathbb{E}\left[\left(\sum_{h=1}^{H}\|\xi_h\|_2\right)^2\right]\right) \\
&= \left(H^2 L_{\hat{f}}^H L + HL(1+L_{\hat{f}})\right)^2 \\
&\left(\left(H\sigma_\xi\sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}}\right)^2 + 2H\sigma_\xi\sqrt{d_s}\left(H\sigma_\xi\sqrt{d_s} + 2H|\mathcal{A}|L_{\hat{f}} + \frac{1}{2}\right)\right) \\
&= \left(H^2 L_{\hat{f}}^H L + HL(1+L_{\hat{f}})\right)^2 \left(3H^2\sigma_\xi^2 d_s + 2H^2|\mathcal{A}|L_{\hat{f}}^2 + 8H^2\sigma_\xi\sqrt{d_s}|\mathcal{A}|L_{\hat{f}} + H\sigma_\xi\sqrt{d_s}\right)
\end{aligned} \tag{22}
$$

□

6.5 Proof of Proposition 2

Proof. Suppose

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

By induction,

$$
\begin{aligned}
&\left\|\nabla_\theta V_\theta^{(h)}(s) - \mathbb{E}[\nabla_\theta \hat{V}_\theta^{(h)}(s)]\right\|_2 \\
&\leq \left\|\nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(s)\right\|_2 + \gamma\left\|\mathop{\mathbb{E}}_{a,\xi}\left[\nabla_\theta V_\theta^{(h+1)}(f(s,a)+\xi)\right] - \mathop{\mathbb{E}}_{a,\xi}\left[\nabla_\theta \hat{V}_\theta^{(h+1)}(\hat{f}(s,a)+\xi)\right]\right\|_2 \\
&+ \gamma\left\|\mathop{\mathbb{E}}_{a,\xi}\left[\nabla_s V_\theta^{(h+1)}(f(s,a)+\xi)\nabla_\theta f(s,a)\right] - \mathop{\mathbb{E}}_{a,\xi}\left[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a)+\xi)\nabla_\theta \hat{f}(s,a)\right]\right\|_2 \\
&\leq \left\|\nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(s)\right\|_2 + \gamma B_2^{(h+1)}(\pi, \boldsymbol{\xi}) + \gamma\left\|\mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a)]\right\|_2 B_3^{(h+1)}(\pi, \boldsymbol{\xi}) + \gamma L\left\|\mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a) - \nabla_\theta f(s,a)]\right\|_2 \\
&= B_2^{(h)}(\pi, \boldsymbol{\xi}),
\end{aligned} \tag{23}
$$

where the last inequality holds since

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

Similarly, we have

$$
\begin{aligned}
&\left\|\nabla_s V_\theta^{(h)}(s) - \mathbb{E}[\nabla_s \hat{V}_\theta^{(h)}(s)]\right\|_2 \\
&\leq \left\|\nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(s)\right\|_2 + \gamma\left\|\mathop{\mathbb{E}}_{a,\xi}\left[\nabla_s V_\theta^{(h+1)}(f(s,a)+\xi)\nabla_s f(s,a)\right] - \mathop{\mathbb{E}}_{a,\xi}\left[\nabla_s \hat{V}_\theta^{(h+1)}(\hat{f}(s,a)+\xi)\nabla_s \hat{f}(s,a)\right]\right\|_2 \\
&\leq \left\|\nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(s)\right\|_2 + \gamma\left\|\mathbb{E}_a[\nabla_s \hat{f}(s,a)]\right\|_2 B_3^{(h+1)}(\pi, \boldsymbol{\xi}) + \gamma L\left\|\mathbb{E}_a\left[\nabla_s \hat{f}(s,a) - \nabla_s f(s,a)\right]\right\|_2 \\
&= B_3^{(h)}(\pi, \boldsymbol{\xi}),
\end{aligned} \tag{24}
$$

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

$$B_3^{(h)}(\pi, \boldsymbol{\xi}) = \sum_{h'=h}^{H} \gamma^{h'-h} \left\| \mathop{\mathbb{E}}_{a}[\nabla_s \hat{f}(s,a)] \right\|_2^{h'-h} \left(\left\| \nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(s) \right\|_2 + \gamma L \left\| \mathop{\mathbb{E}}_{a} \left[\nabla_s \hat{f}(s,a) - \nabla_s f(s,a) \right] \right\|_2 \right) \tag{25}$$

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

$$\delta_t = \sum_{h=1}^{H} \left(\gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(s) \right\|_2 + \gamma^h L \left\| \mathop{\mathbb{E}}_{a} \left[\nabla f(s,a) - \nabla \hat{f}(s,a) \right] \right\|_2 \right)$$

Then

$$\begin{aligned} B_3^{(1)}(\pi, \boldsymbol{\xi}) &= \sum_{h=1}^{H} \gamma^{h-1} \left\| \mathop{\mathbb{E}}_{a}[\nabla_s \hat{f}(s,a)] \right\|_2^{h-1} \left(\left\| \nabla_\theta R_\theta(s) - \nabla_s \hat{R}_\theta(s) \right\|_2 + \gamma L \left\| \mathop{\mathbb{E}}_{a} \left[\nabla_s \hat{f}(s,a) - \nabla_s f(s,a) \right] \right\|_2 \right) \\ &\leq L_{\hat{f}}^H \delta_t. \end{aligned} \tag{26}$$

Finally, the gradient bound is bounded by

$$\begin{aligned} b_t &= \left\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2 \\ &= B_2^{(1)}(\pi, \boldsymbol{\xi}) \\ &= \sum_{h=1}^{H} \gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(s) \right\|_2 + \gamma^h \left\| \mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a)] \right\|_2 B_3^{(h+1)}(\pi, \boldsymbol{\xi}) + \gamma^h L \left\| \mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a) - \nabla_\theta f(s,a)] \right\|_2 \\ &\leq H L_{\hat{f}}^H \delta_t + \sum_{h=1}^{H} \gamma^{h-1} \left\| \nabla_\theta R_\theta(s) - \nabla_\theta \hat{R}_\theta(s) \right\|_2 + \gamma^h L \left\| \mathop{\mathbb{E}}_{a}[\nabla_\theta \hat{f}(s,a) - \nabla_\theta f(s,a)] \right\|_2 \\ &\leq \left(H L_{\hat{f}}^H + 1 \right) \delta_t \end{aligned} \tag{27}$$

□

6.6 Proof of Proposition 3

Proof.

$$\begin{aligned} v_t &= \mathbb{E}\left[\left\| \nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right] \\ &= \mathbb{E}\left[\left\| \nabla_a \hat{Q} \nabla_\theta a_\epsilon - \mathop{\mathbb{E}}_{a}[\nabla_a \hat{Q} \nabla_\theta a] \right\|_2^2 \right] \\ &\leq L_{\hat{Q}}^2 \max_{\epsilon} \mathbb{E}\left[\left\| \nabla_\theta F^{-1}(\epsilon; \theta) \right\|_2^2 \right]. \end{aligned} \tag{28}$$

Besides,

$$\begin{aligned} b_t &= \left\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2 \\ &= \left\| \mathbb{E}[\nabla_a Q \nabla_\theta a] - \mathbb{E}[\nabla_a \hat{Q} \nabla_\theta a] \right\|_2 \\ &\leq \left\| \mathop{\mathbb{E}}_{\epsilon}[\nabla_\theta F^{-1}(\epsilon; \theta)] \right\|_2 \delta_t \end{aligned} \tag{29}$$

□

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] 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.

[7] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[8] Luke Metz, C Daniel Freeman, Samuel S Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021.

[9] 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.

[10] Paavo Parmas, Carl Edward Rasmussen, Jan Peters, and Kenji Doya. Pipps: Flexible model-based policy search robust to the curse of chaos. In *International Conference on Machine Learning*, pages 4065–4074. PMLR, 2018.

[11] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, 2015.

[12] 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.