MODEL-BASED REPARAMETERIZATION POLICY GRADIENT METHODS

Shenao Zhang, Yan Li & Tuo Zhao
Georgia Tech
Atlanta, GA 30332, USA
shenao@gatech.edu

Boyi Liu & Zhaoran Wang
Northwestern University
Evanston, IL 60208, USA

1 INTRODUCTION

Reinforcement learning (RL) has enjoyed great success in various sequential decision-making applications, including strategy games (Silver et al., 2017; Vinyals et al., 2019; Schrittwieser et al., 2020) and robotics (Duan et al., 2016; Wang et al., 2019b; Ibarz et al., 2021) by finding actions that maximize the accumulated long-term reward. As one of the mainstream algorithms, policy gradient method (PGM) (Sutton et al., 1999; Konda & Tsitsiklis, 1999; Kakade, 2001; Silver et al., 2014) seeks to directly search the optimal policy within a prespecified policy class (e.g. neural networks), by iteratively computing and following the stochastic gradient direction w.r.t policy parameters. Thus, the computation and quality of the gradient naturally play a crucial rule in the performance of PGM.

Most of the current stochastic gradient estimation schemes fall into two categories: the likelihood ratio (LR) estimator (Williams, 1992; Konda & Tsitsiklis, 1999; Kakade, 2001; Degris et al., 2012), and the reparameterization (RP) gradient estimator (Heess et al., 2015; Amos et al., 2021; Clavera et al., 2020; Mora et al., 2021; Suh et al., 2022; Xu et al., 2022) (see Section 2.2). In a nutshell, LR estimator performs zeroth-order estimation as it only requires sampling of function values, while RP estimator exploits the differentiability of the estimated function (e.g. model or critic), which offers a unique advantage when one can obtain accurate function approximation. Yet, the theoretical understandings of LR-based PGMs have to date overwhelmingly dominated their RP-based counterparts. Specifically, global convergence of LR-based PGMs has been studied in various settings (Agarwal et al., 2021; Wang et al., 2019a; Bhandari & Russo, 2019), while rare analysis has been established for general RP-based PGMs. The clear understanding of different RP gradient estimate schemes and the overall convergence of RP-based PGMs is still lacking. More importantly, previous work overlooked the contributing factors of the quality of model-based RP gradient estimation (e.g. variance and bias), and its connections to the properties of the model, such as accuracy, smoothness, and expansion step.

The lack of fundamental analysis for RP-based PGMs also has far-reaching effect beyond theoretical interest. As model-based RP gradients are experimentally observed to suffer from large variance and non-smooth loss landscapes (Parmas et al., 2018; Metz et al., 2021; Xu et al., 2022), identifying the key properties and determining constituents of RP estimator that affects its convergence could significantly improve the current algorithmic design of model-based RP PGMs.

Our contributions mainly reside in the following aspects:

- We propose a unified framework that can be instantiated to multiple RP policy gradient algorithms, including MAAC (Clavera et al., 2020; Amos et al., 2021), SVG (Heess et al., 2015), BPTT (Bastani, 2020; Xu et al., 2022), and DPG (Silver et al., 2014; Lillicrap et al., 2015).
- We prove the non-asymptotic global convergence of this general RP-based PGM framework. Moreover, we establish its explicit dependency on the variance and bias of the gradient estimator.
- We characterize how the gradient variance and bias are controlled by several key factors of the algorithm. Specifically, we show that the gradient variance and bias scale exponentially with the Lipschitz continuity of the estimated model, while the bias is also controlled by the gradient accuracy of the function approximations.
- Our observation also suggests potential algorithmic designs. In particular, for complex contact-rich systems, one can significantly reduce the variance and bias of RP gradients by learning a smooth transition kernel and policy. The tradeoff between RP gradient variance and bias further identifies the optimal model expansion step that depends on the gradient error of model and critic.

- We experimentally validate the explosion of gradient variance caused by the non-smooth transition, and demonstrate the benefits of spectral normalized models, policies which allow longer model unrolls and more informative training signals. Besides, we test different sampling schemes suggested by our analysis as well as different model learning methods.

The rest of the paper is organized as follows. Section 2 introduces the setup of RL and the stochastic gradient estimators. Section 3 describes the RP-based policy gradient methods. We propose the unified framework of model-based RP gradients in Section 4. Section 5 presents our main theoretical results, while the numerical studies are provided in Section 6.

2 Background

2.1 Reinforcement Learning

Consider learning to optimize an infinite-horizon γ-discounted Markov Decision Process (MDP) over repeated episodes of interaction. Denote the state space and action space as $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$, respectively. When taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, the agent receives reward $r(s,a)$ whose absolute value is bounded by $|r(s,a)| \leq r_{\max}$. Meanwhile, the environment transits to a new state according to probability $s' \sim f(\cdot|s,a)$.

We are interested in controlling the system by finding a policy π_θ that maximizes the expected cumulative reward. Denote the state and state-action value function associated with π by $V^\pi : \mathcal{S} \to \mathbb{R}$ and $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, respectively, which are defined as

$$V^\pi(s) = (1-\gamma) \cdot \mathbb{E}\Big[\sum_{i=0}^{\infty} \gamma^i \cdot r(s_i, a_i) \,\Big|\, \pi, f, s_0 = s\Big], \quad \forall s \in \mathcal{S}, \tag{2.1}$$

$$Q^\pi(s,a) = (1-\gamma) \cdot \mathbb{E}\Big[\sum_{i=0}^{\infty} \gamma^i \cdot r(s_i, a_i) \,\Big|\, \pi, f, s_0 = s, a_0 = a\Big], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{2.2}$$

Denote the initial state distribution as ζ. Under policy π, the state visitation measure $\nu_\pi(s)$ over $\mathcal{S}$ and the state-action visitation measure $\sigma_\pi(s,a)$ over $\mathcal{S} \times \mathcal{A}$ are defined as

$$\nu_\pi(s) = (1-\gamma) \cdot \sum_{i=0}^{\infty} \gamma^i \cdot \mathbb{P}(s_i = s), \quad \sigma_\pi(s,a) = (1-\gamma) \cdot \sum_{i=0}^{\infty} \gamma^i \cdot \mathbb{P}(s_i = s, a_i = a), \tag{2.3}$$

where $s_0 \sim \zeta(\cdot)$, $a_i \sim \pi(\cdot \,|\, s_i)$, and $s_{i+1} \sim f(\cdot \,|\, s_i, a_i)$. The objective is then

$$J(\pi) = \mathbb{E}_{s_0 \sim \zeta}[V^\pi(s_0)] = \mathbb{E}_{(s,a) \sim \sigma_\pi}[r(s,a)]. \tag{2.4}$$

2.2 Stochastic Gradient Estimation

Consider the general problem form underlying policy gradient, i.e., computing the gradient of a probabilistic objective w.r.t. the parameters of sampling distribution: $\nabla_\theta \mathbb{E}_{p(x;\theta)}[y(x)]$. In RL, $p(x;\theta)$ is the trajectory distribution conditioned on policy parameter θ, and $y(x)$ is the cumulative reward.

Likelihood Ratio (LR) Gradient Estimator. By leveraging the *score function*, LR gradient estimators only require samples of the function values. Specifically, with score function expanded as $\nabla_\theta \log p(x;\theta) = \nabla_\theta p(x;\theta)/p(x;\theta)$, the LR gradient is given by

$$\nabla_\theta \mathbb{E}_{p(x;\theta)}[y(x)] = \int f(x) \nabla_\theta p(x;\theta) dx = \mathbb{E}_{p(x;\theta)}[y(x) \nabla_\theta \log p(x;\theta)]. \tag{2.5}$$

Based on this zeroth-order estimator, several policy gradient algorithms are proposed, e.g., REINFORCE (Williams, 1992) that is defined with a stochastic policy $\pi(a|s) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

Reparameterization (RP) Gradient Estimator. It is also referred to as first-order estimator to characterize its usage of pathwise derivatives. RP gradients benefit from the structural characteristics of the objective, i.e., how the overall objective is affected by the operations applied to the sources of randomness as they pass through the measure and into the cost function (Mohamed et al., 2020).

From the simulation property of continuous distribution, we have the following equivalence between direct and an indirect ways of drawing samples: [advantages of RP over LR seems not that clear]

$$\widehat{x} \sim p(x;\theta) \;\equiv\; \widehat{x} = g(\epsilon;\theta), \epsilon \sim p(\epsilon) \tag{2.6}$$

Derived from the *law of the unconscious statistician* (LOTUS) (Grimmett & Stirzaker, 2020), i.e., $\mathbb{E}_{p(x;\theta)}[y(x)] = \mathbb{E}_{p(\epsilon)}[y(g(\epsilon;\theta))]$, the RP gradient can be expressed as

$$\nabla_\theta \mathbb{E}_{p(x;\theta)}\big[y(x)\big] = \nabla_\theta \int p(\epsilon) y\big(g(\epsilon;\theta)\big) d\epsilon = \mathbb{E}_{p(\epsilon)}\Big[\nabla_\theta y\big(g(\epsilon;\theta)\big)\Big]. \tag{2.7}$$

3 Reparameterization Gradient in RL

In this section, we first present a general form of reparameterization gradient in RL, and then provide a dynamical view for the gradient representation.

3.1 Reparameterization Policy-Value Gradient

We consider the stochastic policy taking the form $a = \pi(s, \varsigma; \theta)$ with noise variable $\varsigma \sim p(\varsigma)$. Notably, the policy class that RP gradient considers differs from the one in zeroth-order gradient estimator (e.g. REINFORCE), which maps the state-action pair to a probability distribution.

Assumption 3.1. The MDP and the policy satisfy that $f(s' \mid s, a)$, $\pi(s, \varsigma; \theta)$, $r(s, a)$, $\nabla_a r(s, a)$ are continuous in all parameters and variables s, a, s'.

The above assumption allows the first-order differentiation through the path of value computation. Recent work (Suh et al., 2022) assumes access to differentiable simulators and concerns the discontinuities caused by contact and geometrical constraints, which is orthogonal to our analysis.

The general form of RP gradient is as follows:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\zeta(\cdot),\varsigma\sim p(\cdot)}\big[\nabla_\theta Q^{\pi_\theta}(s, \pi_\theta(s, \varsigma))\big]. \tag{3.1}$$

By performing sequential decision making, any immediate action could lead to changes of all future states and rewards. Therefore, the value gradient $\nabla_\theta Q^{\pi_\theta}$ possesses a recursive formula. Adapted from the deterministic policy gradient theorem (Silver et al., 2014; Lillicrap et al., 2015) by taking stochasticity into consideration, we can rewrite (3.1) as

Model-Free: $$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\nu_\pi(\cdot),\varsigma\sim p(\varsigma)}\Big[\nabla_\theta \pi_\theta(s, \varsigma) \cdot \nabla_a Q^\pi(s, a)\big|_{a=\pi_\theta(s,\varsigma)}\Big],$$

where $\nabla_a Q^\pi$ can be estimated by the derivative of a learned critic, which yields model-free frameworks such as SVG(0) Heess et al. (2015); Amos et al. (2021). Notably, as a result of the recursive structure of $\nabla_\theta Q^{\pi_\theta}$, the expectation is taken over the state visitation ν_π instead of the initial distribution ζ.

Despite the model-free expression, the RP gradient can also be expanded in a dynamical way through transition paths, which we turn our attention to in the following sections.

3.2 RP Gradient Through Transition Path

Due to the simulation property of continuous distribution in (2.6), we interchangeably write $a \sim \pi(\cdot|s)$ and $a = \pi(s, \varsigma)$, $s' \sim f(\cdot|s, a)$ and $s' = f(s, a, \xi^*)$, with ξ^* sampled from unknown distribution $p(\xi^*)$. From the Bellman equation $V^\pi(s) = \mathbb{E}_\varsigma\big[(1-\gamma)\cdot r(s, \pi(s, \varsigma)) + \gamma \cdot \mathbb{E}_{\xi^*}[V^\pi(f(s, \pi(s, \varsigma), \xi^*))]\big]$, we obtain the backward recursions of gradient:

$$\nabla_\theta V^\pi(s) = \mathbb{E}_\varsigma\Big[\nabla_a r \nabla_\theta \pi + \gamma \mathbb{E}_{\xi^*}\big[\nabla_{s'} V^\pi(s') \nabla_a f \nabla_\theta \pi + \nabla_\theta V^\pi(s')\big]\Big], \tag{3.2}$$

$$\nabla_s V^\pi(s) = \mathbb{E}_\varsigma\Big[\nabla_s r + \nabla_a r \nabla_s \pi + \gamma \mathbb{E}_{\xi^*}\big[\nabla_{s'} V^\pi(s')(\nabla_s f + \nabla_a f \nabla_s \pi)\big]\Big]. \tag{3.3}$$

This gives us the model-based RP gradient calculated by backpropagation through transition paths:

Model-Based: $$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\zeta(\cdot)}\big[\nabla_\theta V^\pi(s)\big].$$

There remain problems that need to be solved for gradient estimation. Firstly, the above formulas require the derivatives of the transition function, i.e. $\nabla_a f$ and $\nabla_s f$. We assume that the transition function and its derivatives are not known and need to be learned.[1] It is thus natural to ask how the model properties (e.g., prediction accuracy and the model smoothness) affect the gradient estimation and the convergence of the RP gradient algorithms, which we investigate in this work.

Besides, even if we have access to an accurate model, unrolling a model over full sequences faces practical difficulties: the memory cost scales linearly with the unroll length as the intermediate outputs need to be stored for backpropagation; long unrolls can also lead to exploding gradients and chaotic non-smooth loss landscapes (Pascanu et al., 2013; Maclaurin et al., 2015; Vicol et al., 2021; Metz et al., 2019), which demands some form of truncation.

4 Model-Based RP Policy Gradient Methods

In this section, we first introduce a model-based value expansion technique to perform truncation while not bringing additional gradient bias. Built upon it, two model-based frameworks can be instantiated with difference lies in the model usage when estimating the RP gradients.

4.1 h-Step Model Value Expansion

As a common technique used to alleviate the challenges brought by full unrolls, algorithms with direct truncation split the full rollouts and backpropagate through the shorter sub-sequences, e.g., Truncated Backpropagation Through Time (TBPTT) (Werbos, 1990). However, such naive truncation has biased gradients and favors short-term dependencies.

In the model-based RL regime, a possible modification involves the combination with h-step Model Value Expansion (MVE) (Feinberg et al., 2018; Clavera et al., 2020; Amos et al., 2021), which decomposes the value estimate $\widehat{V}^\pi(s)$ into the rewards associated with the learned model and the tail estimated by the critic. Formally,

$$\widehat{V}^\pi(s) = (1-\gamma)\cdot\left(\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_i,\widehat{a}_i)+\gamma^h\cdot\widehat{Q}_\omega(\widehat{s}_h,\widehat{a}_h)\right), \tag{4.1}$$

where $\widehat{s}_0 = s$, $\widehat{a}_i = \pi(\widehat{s}_i,\varsigma;\theta)$, $\widehat{s}_{i+1} = \widehat{f}(\widehat{s}_i,\widehat{a}_i,\xi;\psi)$ with critic $\widehat{Q}_\omega$. Here, ς and ξ can either be sampled from $p(\varsigma)$, $p(\xi)$ or inferred from real samples, which we will discuss in more details.

4.2 Model-Based Gradient Estimation

By taking the pathwise gradient w.r.t. the policy parameter θ in the MVE formula (4.1), we obtain the following frameworks with difference lies in whether the samples used for gradient estimation comes from model unrolls or from real trajectories.

Model Derivatives and Predictions. One intuitive way to estimate the first-order RP gradient is to link together the reward, model, policy, critic and backpropagate through them. Specifically, the differentiation is taken through the imagined trajectories with the model used for both derivative calculation and state prediction. The estimator of gradient $\nabla_\theta J(\pi_\theta)$ takes the form of

$$\widehat{\nabla}^{\text{DP}}_\theta J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^{N}\nabla_\theta\left(\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})+\gamma^h\cdot\widehat{Q}_\omega(\widehat{s}_{h,n},\widehat{a}_{h,n})\right), \tag{4.2}$$

where $\widehat{s}_{0,n}\sim\zeta(\cdot)$, $\widehat{a}_{i,n} = \pi(\widehat{s}_{i,n},\varsigma_n;\theta)$ and $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n;\psi)$ with $\varsigma_n\sim p(\varsigma)$, $\xi_n\sim p(\xi)$.

Various algorithms can be instantiated with different choices of h. When $h=0$, the framework reduces to the model-free version we discussed in Section (3.1) as no backpropagation through state is required. For example, MAAC(0) (Amos et al., 2021) and its deterministic counterpart DPG (Silver et al., 2014; Lillicrap et al., 2015). When $h=\infty$, the resulting algorithm is BPTT (Grzeszczuk

[1] It simplifies the problem if we use the differentiable simulators (Huang et al., 2021; Mora et al., 2021; Xu et al., 2022): all the error and complexity brought by model learning will be removed in the analysis below.

et al., 1998; Mozer, 1995; Bastani, 2020; Degrave et al., 2019) where only the model is learned. However, long chains of nonlinear mappings are harmful, leading to large gradient variance, chaotic and non-smooth loss landscapes, and exploding or vanishing gradients (Parmas et al., 2018; Metz et al., 2021). Thus, a proper h prevents such phenomenon and are adopted by recent work, e.g., MAAC (Clavera et al., 2020; Amos et al., 2021) and its variants (Parmas et al., 2018; Mora et al., 2021; Xu et al., 2022; Li et al., 2021).

Model Derivatives on Real Samples. An alternative RP gradient estimator replaces the ∇f term in (3.2) and (3.3) with $\nabla \widehat{f}$. In other words, the learned differentiable model is used for derivative calculation only and Monte-Carlo estimates are computed on *real* samples. Formally,

$$\begin{aligned}\widehat{\nabla}_\theta V^\pi(\widehat{s}_{i,n}) &= \nabla_a r(\widehat{s}_{i,n},\widehat{a}_{i,n})\nabla_\theta \pi(\widehat{s}_{i,n},\varsigma_n) \\ &\quad + \gamma \widehat{\nabla}_s V^\pi(\widehat{s}_{i+1,n})\nabla_a \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n)\nabla_\theta \pi(\widehat{s}_{i,n},\varsigma_n) + \gamma \widehat{\nabla}_\theta V^\pi(\widehat{s}_{i+1,n}),\end{aligned} \tag{4.3}$$

$$\begin{aligned}\widehat{\nabla}_s V^\pi(\widehat{s}_{i,n}) &= \nabla_s r(\widehat{s}_{i,n},\widehat{a}_{i,n}) + \nabla_a r(\widehat{s}_{i,n},\widehat{a}_{i,n})\nabla_s \pi(\widehat{s}_{i,n},\varsigma_n) \\ &\quad + \gamma \widehat{\nabla}_s V^\pi(\widehat{s}_{i+1,n})\big(\nabla_s \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n) + \nabla_a \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n)\nabla_s \pi(\widehat{s}_{i,n},\varsigma_n)\big),\end{aligned} \tag{4.4}$$

where $\widehat{s}_{0,n} \sim \zeta(\cdot)$, $\widehat{a}_{i,n} = \pi(\widehat{s}_{i,n},\varsigma_n;\theta)$, $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n;\psi)$, and ς_n, ξ_n are inferred from the tuple $(s_{i,n},a_{i,n},s_{i+1,n})$ such that $\widehat{a}_{i,n} = a_{i,n}$ and $\widehat{s}_{i+1,n} = s_{i+1,n}$. The h-th timestep terminal $\widehat{\nabla} V^\pi(\widehat{s}_{h,n})$ is zero if $h = \infty$, and is $\nabla \widehat{V}_\omega(\widehat{s}_{h,n})$ if $h < \infty$. The corresponding gradient estimator is

$$\widehat{\nabla}^{\text{DR}}_\theta J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^{N} \widehat{\nabla}_\theta V^\pi(\widehat{s}_{0,n}) = \frac{1}{N}\sum_{n=1}^{N} \nabla_\theta \left(\sum_{i=0}^{h-1} \gamma^i \cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n}) + \gamma^h \cdot \widehat{Q}_\omega(\widehat{s}_{h,n},\widehat{a}_{h,n})\right). \tag{4.5}$$

Example algorithms of this framework include SVG (Heess et al., 2015) and its deterministic counterparts (Abbeel et al., 2006; Atkeson, 2012; Kumpati et al., 1990).

4.3 Algorithmic Framework

For model-based RP gradient algorithms, three update procedures are performed iteratively. Namely, policy, model, and critic are updated in every iteration $t \in [T]$, which give us sequences of $\{\pi_{\theta_t}\}_{t\in[T+1]}$, $\{\widehat{f}_{\psi_t}\}_{t\in[T]}$, and $\{\widehat{Q}_{\omega_t}\}_{t\in[T]}$, respectively.

Policy Update. The update rule for policy parameter θ with learning rate η is as follows:

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \widehat{\nabla}_\theta J(\pi_{\theta_t}), \tag{4.6}$$

where $\widehat{\nabla}_\theta J(\pi_{\theta_t})$ can be specified as either $\widehat{\nabla}^{\text{DP}}_\theta J(\pi_\theta)$ or $\widehat{\nabla}^{\text{DR}}_\theta J(\pi_\theta)$.

Model Update. To approximate the (stochastic) environment dynamics, traditional model-based RL learns a forward model that predicts how the system will evolve when applying action a at state s. A simple way is to learn a deterministic function as the mean of transition, by minimizing the mean squared error (MSE) between the predictions and ground-truth. An alternative approach is to fit a probabilistic function by performing maximum likelihood estimation (MLE) on the sampled data.

When applying RP gradient estimators, accurate state predictions, however, do not imply accurate gradient estimation. We adopt the notation $\epsilon_f(t)$ to represent the model gradient error at iteration t:

$$\epsilon_f(t) := \max_{i\in[h]} \mathbb{E}_{\mathbb{P}(s_i,a_i),\mathbb{P}(\widehat{s}_i,\widehat{a}_i)}\left[\left\|\frac{\partial s_i}{\partial s_{i-1}} - \frac{\partial \widehat{s}_i}{\partial \widehat{s}_{i-1}}\right\|_2 + \left\|\frac{\partial s_i}{\partial a_{i-1}} - \frac{\partial \widehat{s}_i}{\partial \widehat{a}_{i-1}}\right\|_2\right]. \tag{4.7}$$

Here, $\mathbb{P}(s_i,a_i)$ is the state-action distribution where $s_0 \sim \zeta$, $a_j \sim \pi_t(\cdot \mid s_j)$, and $s_{j+1} \sim f(\cdot \mid s_j,a_j)$. $\mathbb{P}(\widehat{s}_i,\widehat{a}_i)$ is the distribution that the gradient is estimated with samples from it, i.e., $\widehat{s}_0 \sim \zeta$, $\widehat{a}_j \sim \pi_t(\cdot \mid \widehat{s}_j)$, and $\widehat{s}_{j+1} \sim \widehat{f}(\cdot \mid \widehat{s}_j,\widehat{a}_j)$ when $\widehat{\nabla}^{\text{DP}}_\theta J(\pi_\theta)$; and $\mathbb{P}(\widehat{s}_i,\widehat{a}_i) = \mathbb{P}(s_i,a_i)$ when $\widehat{\nabla}^{\text{DR}}_\theta J(\pi_\theta)$.

Since the objective mismatch problem exists between improving state prediction accuracy and minimizing gradient error ϵ_f, recent work (Li et al., 2021) proposed to directly induce a constraint on the model such that its directional derivative is consistent with the given data. However, if traditional state-predictive models learned on visited regions can extrapolate, the model derivative error can

be bounded with finite difference approximation. In this case, ϵ_f can be represented by introducing additional measure of model complexity, with which the generalizability of the model class is captured.

Critic Update. For any policy π, its value function satisfies the Bellman equation, and is also the unique solution of the Bellman equation, i.e., if $Q = \mathcal{T}^\pi Q$ then $Q = Q^\pi$. The Bellman operator $\mathcal{T}^\pi$ is defined by

$$\mathcal{T}^\pi Q(s,a) = \mathbb{E}\big[(1-\gamma)\cdot r(s,a) + \gamma \cdot Q(s',a') \,\big|\, \pi, f\big], \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

We aim to approximate the state-action value Q with a critic $\widehat{Q}_\omega$. Due to the uniqueness of the Bellman equation solution, it can be achieved by minimizing the mean-squared Bellman error $\mathbb{E}\big[\big(\widehat{Q}_\omega(s,a) - \mathcal{T}^{\pi_t}\widehat{Q}_\omega(s,a)\big)^2\big]$, which can be done by Temporal Difference (TD) learning (Sutton, 1988; Cai et al., 2019). We define the critic gradient error ϵ_v as

$$\epsilon_v(t) := \mathbb{E}_{\mathbb{P}(s_h,a_h),\mathbb{P}(\widehat{s}_h,\widehat{a}_h)}\left[\left\|\frac{\partial Q^{\pi_t}}{\partial s} - \frac{\partial \widehat{Q}_t}{\partial \widehat{s}}\right\|_2 + \left\|\frac{\partial Q^{\pi_t}}{\partial a} - \frac{\partial \widehat{Q}_t}{\partial \widehat{a}}\right\|_2\right], \tag{4.8}$$

where $\mathbb{P}(s_h,a_h)$ and $\mathbb{P}(\widehat{s}_h,\widehat{a}_h)$ are distributions at timestep h with the same definition as in (4.7).

The pseudocode of model-based RP gradient methods is as follows.

Algorithm 1 Model-Based Reparameterization Policy Gradient Methods

Input: Number of iterations T, learning rate η, batch size N.
1: **for** iteration $t \in [T]$ **do**
2: Update the model parameter ψ_t by performing MSE or MLE.
3: Update the critic parameter ω_t by performing TD learning.
4: Estimate $\widehat{\nabla}_\theta J(\pi_{\theta_t})$ using (4.2) or (4.5).
5: Update the policy parameter θ_t by $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \widehat{\nabla}_\theta J(\pi_{\theta_t})$ and execute $\pi_{\theta_{t+1}}$.
6: **end for**
7: **Output:** $\{\pi_{\theta_t}\}_{t\in[T]}$.

5 Main Results

We provide our main results in this section. Specifically, we study the relationship between the critic error, model error, model smoothness and the gradient bias, variance, which we then use to find the optimal truncation step and establish the convergence of model-based RP policy gradient methods.

To begin with, we impose a regularity condition on the smoothness of the expected total reward $J(\pi_\theta)$. Assumption 5.1 holds under certain regularity conditions of the MDP, e.g. when the transition and rewards are Lipschitz continuous (Bastani, 2020; Pirotta et al., 2015; Wang et al., 2019a).

Assumption 5.1 (Lipschitz Continuous Gradient)**.** Assume $J(\pi_\theta)$ is L-smooth in θ, such that $\big\|\nabla_\theta J(\pi_{\theta_1}) - \nabla_\theta J(\pi_{\theta_2})\big\|_2 \le L\cdot\big\|\theta_1-\theta_2\big\|_2$.

Now we character the convergence of RP gradient algorithms with the following proposition.

Proposition 5.2 (Convergence to Stationary Points)**.** Suppose $\eta \le 1/L$ and the policy parameter space satisfies $\mathbb{E}\big[\big\|\theta_T-\theta_0\big\|_2\big] \le \delta$. Denote $c := \big(\eta - L\eta^2\big)^{-1}$. It then holds that

$$\min_{t\in[T]}\mathbb{E}\Big[\big\|\nabla_\theta J(\pi_{\theta_t})\big\|_2^2\Big] \le \frac{4}{T}\cdot\left(\sum_{t=0}^{T-1} c\cdot(\delta\cdot b_t + \frac{\eta}{2}\cdot v_t) + b_t^2 + v_t\right) + \frac{4c}{T}\cdot\mathbb{E}\big[J(\pi_{\theta_T}) - J(\pi_{\theta_1})\big],$$

where the gradient bias b_t and variance v_t is defined as

$$b_t := \Big\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2, \quad v_t := \mathbb{E}\left[\Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2^2\right].$$

We now upper bound b_t and v_t by first introducing the following Lipschitz assumption, which is adopted in various previous work (Pirotta et al., 2015; Clavera et al., 2020; Li et al., 2021).

Assumption 5.3 (Lipschitz Continuous Functions)**.** We assume that $r(s,a)$, $\widehat{f}_\psi(s,a,\xi)$, $\pi_\theta(s,\varsigma)$ $\widehat{Q}_\omega(s,a)$ are Lipschitz continuous, such that

$$\begin{aligned}
\big|r(s_1,a_1)-r(s_2,a_2)\big| &\le L_r\cdot\big\|(s_1-s_2,a_1-a_2)\big\|_2,\\
\big\|\widehat{f}(s_1,a_1,\xi_1)-\widehat{f}(s_2,a_2,\xi_2)\big\|_2 &\le L_{\widehat{f}}\cdot\big\|(s_1-s_2,a_1-a_2,\xi_1-\xi_2)\big\|_2,\\
\big\|\pi(s_1,\varsigma_1)-\pi(s_2,\varsigma_2)\big\| &\le L_\pi\cdot\big\|(s_1-s_2,\varsigma_1-\varsigma_2)\big\|_2,\\
\big|\widehat{Q}(s_1,a_1)-\widehat{Q}(s_2,a_2)\big| &\le L_{\widehat{Q}}\cdot\big\|(s_1-s_2,a_1-a_2)\big\|_2.
\end{aligned}$$

Assume the policy $\pi_\theta(s,\varsigma)$ is Lipschitz continuous also in parameter space such that $\|\nabla_\theta\pi\|_2\le L_\theta$.

Denote $\widetilde{L}_{\widehat{f}}:=\max\{L_{\widehat{f}},1\}$ and $\widetilde{L}_\pi:=\max\{L_\pi,1\}$. We have the following results of the gradient variance.

Proposition 5.4 (Gradient Variance)**.** Under Assumption 5.3, for any $t\in[T]$, the gradient variance when the estimator $\widehat{\nabla}_\theta J(\pi_\theta)$ is specified as either $\widehat{\nabla}^{\mathrm{DP}}_\theta J(\pi_\theta)$ or $\widehat{\nabla}^{\mathrm{DR}}_\theta J(\pi_\theta)$ can be bounded by

$$v_t\le O\Big(h^8\widetilde{L}^{6h}_{\widehat{f}}\widetilde{L}^{4h}_\pi/N+\gamma^{2h}h^6\widetilde{L}^{6h}_{\widehat{f}}\widetilde{L}^{4h}_\pi/N\Big). \tag{5.1}$$

We observe that when $L_{\widehat{f}}>1$ and $L_\pi>1$, a large truncation step h will lead to exponentially increasing gradient variance. The intuition behind is that when the model is non-smooth, the dynamics can diverge and could even be chaotic (Bollt, 2000). As a result, the gradient has large variance since small randomness in training can lead to diverging trajectories and update directions.

Therefore, when the dynamics of the underlying MDP is complex and contact-rich (Suh et al., 2022; Xu et al., 2022), RP gradient algorithms can benefit training by regularizing the smoothness of the learned model to significantly reduce the gradient variance. To restrain the bias brought by doing so, we assume the system is controllable with bounded value gradient.

Assumption 5.5 (Lipschitz Q-Value)**.** Assume the state-action value is L_Q Lipschitz continuous.

Besides, since policy actions can lead to changes of future states and rewards, unless we know the exact state-action value function which gives us accurate $\nabla_\theta Q^\pi$, it cannot be simply represented by quantities in any finite timescale. In other words, we need to consider the recursive structure of the value function to measure the gradient bias brought by the critic. For this reason, we define the following state-action visitation measure over $\mathcal{S}$:

$$\overline{\sigma}_1(s,a)=\sum_{i=h}^{\infty}\gamma^{i-h}\cdot\mathbb{P}(s_i=s,a_i=a),\quad \overline{\sigma}_2(s,a)=\mathbb{P}(s_h=s,a_h=a), \tag{5.2}$$

where $s_0\sim\zeta$, $a_i\sim\pi(\cdot\,|\,s_i)$, and $s_{i+1}\sim f(\cdot\,|\,s_i,a_i)$.

In what follows, we impose a regularity condition on the discrepancy between $\overline{\sigma}_1(s,a)$ and $\overline{\sigma}_2(s,a)$.

Assumption 5.6 (Regularity Condition)**.** We assume that there exists $\kappa>0$ such that

$$\left\{\mathbb{E}_{\overline{\sigma}_1}\left[\Big(\frac{d\overline{\sigma}_1}{d\overline{\sigma}_2}(s,a)\Big)^2\right]\right\}^{1/2}\le\kappa, \tag{5.3}$$

where $d\overline{\sigma}_1/d\overline{\sigma}_2$ is the Radon-Nikodym derivative of $\overline{\sigma}_1$ with respect to $\overline{\sigma}_2$.

Notably, when $h\to\infty$, $\overline{\sigma}_1=\overline{\sigma}_2$ and $\kappa=1$. When $h=0$, κ measures the discrepancy between the initial state distribution ζ and state visitation ν. In this case, we are also able to remove Assumption 5.6 by sampling $\widehat{s}_{0,n}$ from ν instead of ζ in the gradient estimator. This suggests two different data sampling schemes when updating policy: from replay buffer when $h=0$ (e.g. SVG(0) (Heess et al., 2015; Amos et al., 2021) and DDPG (Silver et al., 2014; Lillicrap et al., 2015)); from the initial distribution when h is large (e.g. BPTT-style algorithms (Kurutach et al., 2018; Curi et al., 2020; Xu et al., 2022)).

Proposition 5.7 (Gradient Bias)**.** Denote $\kappa'=\max\{\kappa,1/\kappa\}$. Under Assumption 5.3, 5.5, and 5.6, for any $t\in[T]$, the gradient bias is bounded by

$$b_t\le O\Big(h^4L^{3h}_{\widehat{f}}L^h_\pi(\epsilon_f+\widetilde{L}^h_\pi)+\kappa' h^2\gamma^h L^{2h}_{\widehat{f}}L^h_\pi\epsilon_v\Big). \tag{5.4}$$

Proposition 5.4 indicates that with non-smooth model and policy, the loss landscapes are highly non-smooth, which together with large gradient bias results in slow convergence and even failure of training even in simple toy examples (Parmas et al., 2018; Metz et al., 2021; Suh et al., 2022). Our results suggest that we can add smoothness regularization (e.g. Spectral Normalization (Miyato et al., 2018; Bjorck et al., 2021)) on the model and policy to avoid exponentially increasing gradient variance and bias. Besides, the trade-off between the gradient bias and variance will result in an optimal truncation step $h^* \in [0, \infty)$ that achieves the best convergence rate. We represent h^* with the notation of model error ϵ_f and critic error ϵ_v as follows.

Proposition 5.8 (Optimal Model Expansion Step)**.** Suppose $L_{\widehat{f}} < 1$ and $L_\pi < 1$, then the optimal model expansion step h^* is with the following form:

$$h^* := \operatorname*{argmin}_h c \cdot (\delta \cdot b_t + \frac{\eta}{2} \cdot v_t) + b_t^2 + v_t = O\Big(W\big(2(\log\gamma)^2\big)/N + W\big(\frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}\big)\Big),$$

where the Lambert W function is the inverse function of $x \cdot e^x$ such that $W(x \cdot e^x) = x$.

Notably, for $x \in (0, \infty)$, the Lambert W function $W(x)$ is positive and increases monotonically. Therefore, h^* is positive and increases with ϵ_v/ϵ_f. This result can guide the algorithms to perform more model expansion steps when the model error ϵ_f is small; while avoiding long model unrolls when the critic error ϵ_f is relatively smaller.

A stationary point $\widehat{\theta}$ of $J(\pi_\theta)$ is defined as $\nabla_\theta J(\pi_{\widehat{\theta}}) = 0$.

Proposition 5.9 (Global Optimality of Stationary Point)**.** The stationary point $\widehat{\theta}$ satisfies

$$(1-\gamma) \cdot \big(J(\pi^*) - J(\pi_{\widehat{\theta}})\big) \leq 2r_{\max} \cdot \inf_\theta \Big\| u_{\widehat{\theta}}(s,a) - \big(\nabla_\theta \log \pi_{\widehat{\theta}}(a|s)\big)^\top \theta \Big\|_{\sigma_{\pi_{\widehat{\theta}}}}, \tag{5.5}$$

where $u_{\widehat{\theta}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$u_{\widehat{\theta}}(s,a) := \frac{d\sigma_{\pi^*}}{d\sigma_{\pi_{\widehat{\theta}}}}(s,a) - \frac{d\nu_{\pi^*}}{d\nu_{\pi_{\widehat{\theta}}}}(s), \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Here, $d\sigma_{\pi^*}/d\sigma_{\pi_{\widehat{\theta}}}$ and $d\nu_{\pi^*}/d\nu_{\pi_{\widehat{\theta}}}$ are the Radon-Nikodym derivatives and $\|\cdot\|_{\sigma_{\pi_{\widehat{\theta}}}}$ is the $L_2(\sigma_{\pi_{\widehat{\theta}}})$-norm.

6 Experiments

6.1 Gradient Variance and Loss Landscape

6.2 Benefit of Model (and Policy) Regularization and Large h

6.3 State Prediction Error vs Model Gradient Error

6.4 Data Sampling Schemes

References

Pieter Abbeel, Morgan Quigley, and Andrew Y Ng. Using inaccurate models in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1–8, 2006.

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.

Brandon Amos, Samuel Stanton, Denis Yarats, and Andrew Gordon Wilson. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, pp. 6–20. PMLR, 2021.

Christopher G Atkeson. Efficient robust policy optimization. In *2012 American Control Conference (ACC)*, pp. 5220–5227. IEEE, 2012.

Osbert Bastani. Sample complexity of estimating the policy gradient for nearly deterministic dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 3858–3869. PMLR, 2020.

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

Johan Bjorck, Carla P Gomes, and Kilian Q Weinberger. Towards deeper deep reinforcement learning. *arXiv preprint arXiv:2106.01151*, 2021.

Erik M Bollt. Controlling chaos and the inverse frobenius–perron problem: global stabilization of arbitrary invariant measures. *International Journal of Bifurcation and Chaos*, 10(05):1033–1050, 2000.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

Ignasi Clavera, Violet Fu, and Pieter Abbeel. Model-augmented actor-critic: Backpropagating through paths. *arXiv preprint arXiv:2005.08068*, 2020.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170, 2020.

Jonas Degrave, Michiel Hermans, Joni Dambre, et al. A differentiable physics engine for deep learning in robotics. *Frontiers in neurorobotics*, pp. 6, 2019.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. Model-based value expansion for efficient model-free reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

Radek Grzeszczuk, Demetri Terzopoulos, and Geoffrey Hinton. Neuroanimator: Fast neural network emulation and control of physics-based models. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pp. 9–20, 1998.

Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. *Advances in neural information processing systems*, 28, 2015.

Zhiao Huang, Yuanming Hu, Tao Du, Siyuan Zhou, Hao Su, Joshua B Tenenbaum, and Chuang Gan. Plasticinelab: A soft-body manipulation benchmark with differentiable physics. *arXiv preprint arXiv:2104.03311*, 2021.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

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

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

S Narendra Kumpati, Parthasarathy Kannan, et al. Identification and control of dynamical systems using neural networks. *IEEE Transactions on neural networks*, 1(1):4–27, 1990.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592*, 2018.

Chongchong Li, Yue Wang, Wei Chen, Yuting Liu, Zhi-Ming Ma, and Tie-Yan Liu. Gradient information matters in policy optimization by back-propagating through model. In *International Conference on Learning Representations*, 2021.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*, pp. 4556–4565. PMLR, 2019.

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

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

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.

Miguel Angel Zamora Mora, Momchil P Peychev, Sehoon Ha, Martin Vechev, and Stelian Coros. Pods: Policy optimization via differentiable simulation. In *International Conference on Machine Learning*, pp. 7805–7817. PMLR, 2021.

Michael C Mozer. A focused backpropagation algorithm for temporal. *Backpropagation: Theory, architectures, and applications*, 137, 1995.

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*, pp. 4065–4074. PMLR, 2018.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013.

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

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

HJ Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do differentiable simulators give better policy gradients? *arXiv preprint arXiv:2202.00817*, 2022.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Paul Vicol, Luke Metz, and Jascha Sohl-Dickstein. Unbiased gradient estimation in unrolled computation graphs with persistent evolution strategies. In *International Conference on Machine Learning*, pp. 10553–10563. PMLR, 2021.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019a.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019b.

Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Jie Xu, Viktor Makoviychuk, Yashraj Narang, Fabio Ramos, Wojciech Matusik, Animesh Garg, and Miles Macklin. Accelerated policy learning with parallel differentiable simulation. *arXiv preprint arXiv:2204.07137*, 2022.

A Proofs

A.1 Proof of Proposition 5.2

Proof. From the policy update rule, we know that $\widehat{\nabla}_\theta J(\pi_{\theta_t}) = (\theta_{t+1} - \theta_t)/\eta$. By Assumption 5.3, we have

$$\begin{aligned} J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) &\geq \nabla_\theta J(\pi_{\theta_t})^\top (\theta_{t+1} - \theta_t) - \frac{L}{2}\left\|\theta_{t+1} - \theta_t\right\|_2^2 \\ &= \eta \nabla_\theta J(\pi_{\theta_t})^\top \widehat{\nabla}_\theta J(\pi_{\theta_t}) - \frac{L\eta^2}{2}\left\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2^2. \end{aligned} \tag{A.1}$$

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

$$\nabla_\theta J(\pi_{\theta_t}) = \Big(\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big) - \Big(\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big) + \widehat{\nabla}_\theta J(\pi_{\theta_t}).$$

Then we bound $\nabla_\theta J(\pi_{\theta_t})^\top \widehat{\nabla}_\theta J(\pi_{\theta_t})$ in (A.1) by bounding the following three terms.

$$\begin{aligned} \left|\Big(\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big)^\top \widehat{\nabla}_\theta J(\pi_{\theta_t})\right| &\leq \left\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2 \cdot \left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\right\|_2 \\ &= \left\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2 \cdot b_t, \\ \Big(\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}[\widehat{\nabla}_\theta J(\pi_{\theta_t})]\Big)^\top \widehat{\nabla}_\theta J(\pi_{\theta_t}) &\leq \frac{\left\|\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\right\|_2^2}{2} + \frac{\left\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2^2}{2}, \\ \widehat{\nabla}_\theta J(\pi_{\theta_t})^\top \widehat{\nabla}_\theta J(\pi_{\theta_t}) &\geq \left\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2^2. \end{aligned} \tag{A.2}$$

Thus, we can bound (A.1) by

$$\begin{aligned}
&J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) \\
&\quad \geq \frac{\eta}{2} \cdot \left(-\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2 \cdot 2b_t - \left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}[\widehat{\nabla}_\theta J(\pi_{\theta_t})] \right\|_2^2 + \left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2 \right) \\
&\qquad - \frac{L\eta^2}{2} \cdot \left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2. \qquad \text{(A.3)}
\end{aligned}$$

By taking expectation in (A.3), we have

$$\mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] \geq -\eta \cdot \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2 \right] \cdot b_t - \frac{\eta}{2} \cdot v_t + \frac{\eta - L\eta^2}{2} \cdot \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2 \right].$$

Rearranging terms gives us

$$\frac{\eta - L\eta^2}{2} \cdot \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2 \right] \leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta \mathbb{E}[\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \|_2] b_t + \frac{\eta}{2} v_t. \qquad \text{(A.4)}$$

We now turn our attention to characterize $\| \nabla_\theta J(\pi_{\theta_t}) - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \|_2$.

$$\begin{aligned}
\mathbb{E}\left[\left\| \nabla_\theta J(\pi_{\theta_t}) - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2 \right] &= \mathbb{E}\left[\left\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] + \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2 \right] \\
&\leq 2 \left\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \right\|_2^2 + 2\mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \right\|_2^2 \right] \\
&= 2b_t^2 + 2v_t, \qquad \text{(A.5)}
\end{aligned}$$

where the second inequality holds since

$$\|y + z\|_2^2 \leq \|y\|_2^2 + \|z\|_2^2 + 2\|y\|_2 \cdot \|z\|_2 \leq 2\|y\|_2^2 + 2\|z\|_2^2. \qquad \text{(A.6)}$$

Since $\eta \leq 1/L$ implies that $(\eta - L\eta^2)/2 > 0$, combining (A.4) and (A.5) we have

$$\begin{aligned}
&\min_{t \in [T]} \mathbb{E}\left[\left\| \nabla_\theta J(\pi_{\theta_t}) \right\|_2^2 \right] \\
&\quad \leq \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \nabla_\theta J(\pi_{\theta_t}) \right\|_2^2 \right] \\
&\quad \leq \frac{2}{T} \cdot \sum_{t=0}^{T-1} \left(\mathbb{E}\left[\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \|_2^2 \right] + \mathbb{E}\left[\left\| \nabla_\theta J(\pi_{\theta_t}) - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2^2 \right] \right) \\
&\quad \leq \frac{4c}{T} \cdot \left(\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=0}^{T-1} \left(\eta \cdot \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2 \right] \cdot b_t + \frac{\eta}{2} \cdot v_t \right) \right) + \frac{4}{T} \cdot \sum_{t=0}^{T-1} (b_t^2 + v_t) \\
&\quad = \frac{4}{T} \cdot \left(\sum_{t=0}^{T-1} c \cdot \left(\eta \cdot \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2 \right] \cdot b_t + \frac{\eta}{2} \cdot v_t \right) + b_t^2 + v_t \right) + \frac{4c}{T} \cdot \mathbb{E}\big[J(\pi_{\theta_T}) - J(\pi_{\theta_1}) \big],
\end{aligned}$$

where recall that $c := \left(\eta - L\eta^2 \right)^{-1}$.

We also have

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta \cdot \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2 \right] \cdot b_t &\leq \sum_{t=0}^{T-1} \eta \mathbb{E}\left[\left\| \widehat{\nabla}_\theta J(\pi_{\theta_t}) \right\|_2 \right] \cdot \sum_{t=0}^{T-1} b_t \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \theta_{t+1} - \theta_t \right\|_2 \right] \cdot \sum_{t=0}^{T-1} b_t \\
&= \mathbb{E}\left[\left\| \theta_T - \theta_0 \right\|_2 \right] \cdot \sum_{t=0}^{T-1} b_t.
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
&\min_{t\in[T]} \mathbb{E}\Big[\big\|\rho_t\big\|_2^2\Big] \\
&\quad\le \frac{4}{T}\cdot\left(\sum_{t=0}^{T-1} c\cdot\big(\eta\cdot\mathbb{E}\Big[\big\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\big\|_2\Big]\cdot b_t+\frac{\eta}{2}\cdot v_t\big)+b_t^2+v_t\right)+\frac{4c}{T}\cdot\mathbb{E}\big[J(\pi_{\theta_T})-J(\pi_{\theta_1})\big] \\
&\quad\le \frac{4}{T}\cdot\left(\sum_{t=0}^{T-1} c\cdot\big(\delta\cdot b_t+\frac{\eta}{2}\cdot v_t\big)+b_t^2+v_t\right)+\frac{4c}{T}\cdot\mathbb{E}\big[J(\pi_{\theta_T})-J(\pi_{\theta_1})\big]. \qquad \text{(A.7)}
\end{aligned}$$

□

A.2 Proof of Proposition 5.4

Proof. For any random variable y, the following holds due to the Cauchy–Schwarz inequality:

$$\big\|\mathbb{E}[y]\big\|_2=\Big\|\sum_{k=1}^{\infty} y_k\cdot p(y_k)\Big\|_2\le\sum_{k=1}^{\infty} p(y_k)\cdot\|y_k\|_2=\mathbb{E}\big[\|y\|_2\big]. \qquad \text{(A.8)}$$

Denote $g:=\widehat{\nabla}_\theta J(\pi_{\theta_t})$ as the estimated gradient at iteration t. We have from (A.8) that

$$\begin{aligned}
\Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t})-\mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2 &= \Big\|g-\mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2 \\
&= \Big\|\mathbb{E}\big[g-\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2 \\
&\le \mathbb{E}\Big[\big\|g-\widehat{\nabla}_\theta J(\pi_{\theta_t})\big\|_2\Big], \qquad \text{(A.9)}
\end{aligned}$$

where the expectation is taken over the randomness in the policy and model dynamics.

The proof is established by showing the result in the two gradient estimators, i.e., when $\widehat{\nabla}_\theta J(\pi_\theta)=\widehat{\nabla}_\theta^{\mathrm{DP}} J(\pi_\theta)$ and when $\widehat{\nabla}_\theta J(\pi_\theta)=\widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta)$, with the form of (4.2) and (4.5), respectively.

In both frameworks, it holds that $\widehat{s}_{i+1,n}=\widehat{f}(\widehat{s}_{i,n},\xi_n)$. The difference lies in that $\xi_n\sim p(\xi)$ in (4.2), while in (4.5), ξ_n is inferred from the real samples $(s_{i,n},a_{i,n},s_{i+1,n})$ such that $\widehat{f}(s_{i,n},\xi_n)=s_{i+1,n}$. For example, with Gaussian transition model $\mathcal{N}\big(\phi(s_{i,n},a_{i,n}),\sigma^2\big)$, noise ξ_n can be inferred as $\xi_n=\big(s_{i+1,n}-\phi(s_{i,n},a_{i,n})\big)/\sigma$.

Denote $\widehat{x}_{i,n}:=(\widehat{s}_{i,n},\widehat{a}_{i,n})$. For sample size $N=1$, we have

$$\begin{aligned}
\mathbb{E}\Big[\big\|g-\widehat{\nabla}_\theta J(\pi_{\theta_t})\big\|_2\Big] \le \sum_{i=0}^{h-1}\gamma^i\cdot\mathbb{E}_{\overline{x}_i}\Big[\big\|\nabla_\theta r(\widehat{x}_{i,n})-\nabla_\theta r(\overline{x}_i)\big\|_2\Big] \\
+\gamma^h\cdot\mathbb{E}_{\overline{x}_h}\Big[\big\|\nabla\widehat{Q}(\widehat{x}_{h,n})\nabla_\theta\widehat{x}_{h,n}-\nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\overline{x}_h\big\|_2\Big]. \qquad \text{(A.10)}
\end{aligned}$$

Here and hereafter, the expectation is taken over $\overline{x}_i\sim\widehat{\mathbb{P}}_i$ when $\widehat{\nabla}_\theta J(\pi_\theta)=\widehat{\nabla}_\theta^{\mathrm{DP}} J(\pi_\theta)$ and taken over $\overline{x}_i\sim\mathbb{P}_i$ when $\widehat{\nabla}_\theta J(\pi_\theta)=\widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta)$.

For $i\ge 1$, we have the following relationship:

$$\frac{d\widehat{a}_{i,n}}{d\theta}=\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}+\frac{\partial\widehat{a}_{i,n}}{\partial\theta}. \qquad \text{(A.11)}$$

Iterating above gives us

$$\begin{aligned}\left\|\frac{d\widehat{a}_{i,n}}{d\theta}\right\|_2 &\leq \left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}+\frac{\partial\widehat{a}_{i,n}}{\partial\theta}\right\|_2 \\ &\leq L_\pi L_{\widehat{f}}\cdot\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2+L_\theta \\ &\leq L_{\widehat{f}}^i L_\pi^i\cdot L_\theta+L_\theta\cdot\sum_{j=0}^{i-1}L_{\widehat{f}}^j L_\pi^j \\ &\leq L_\theta\cdot(i+1)\cdot\widetilde{L}_{\widehat{f}}^i\widetilde{L}_\pi^i. \end{aligned} \tag{A.12}$$

Here, we apply the fact that for a sequence $x_0,\cdots,x_i$ where $x_i=ax_{i-1}+b$, it holds that

$$x_i=ax_{i-1}+b=a\cdot(ax_{i-2}+b)+b=a^i\cdot x_0+b\cdot\sum_{j=0}^{i-1}a^j. \tag{A.13}$$

Similarly, we can bound $\left\|\nabla_\theta\widehat{s}_{i,n}\right\|_2$ by recursion as follows

$$\begin{aligned}\left\|\frac{d\widehat{s}_{i,n}}{d\theta}\right\|_2 &= \left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}+\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2 \\ &\leq L_{\widehat{f}}\cdot\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}\right\|_2+L_\theta\cdot i\cdot\widetilde{L}_{\widehat{f}}^i\widetilde{L}_\pi^i \\ &\leq L_\theta\cdot\widetilde{L}_{\widehat{f}}^i\cdot\widetilde{L}_\pi+L_\theta\cdot i\cdot\widetilde{L}_{\widehat{f}}^i\widetilde{L}_\pi^i\cdot\sum_{j=0}^{i-1}L_{\widehat{f}}^j \\ &\leq L_\theta\cdot(i^2+1)\cdot\widetilde{L}_{\widehat{f}}^{2i}\widetilde{L}_\pi^i. \end{aligned} \tag{A.14}$$

Combining (A.12) and (A.14) we obtain

$$\left\|\frac{d\widehat{x}_{i,n}}{d\theta}\right\|_2\leq\widehat{K}(i):=L_\theta\cdot(i+1)\cdot\widetilde{L}_{\widehat{f}}^i\widetilde{L}_\pi^i+L_\theta\cdot(i^2+1)\cdot\widetilde{L}_{\widehat{f}}^{2i}\widetilde{L}_\pi^i, \tag{A.15}$$

where $\widehat{K}(i)$ is introduced for notation simplicity.

Therefore, the second term in (A.10) can be bounded by

$$\begin{aligned}&\mathbb{E}_{\overline{x}_h}\left[\left\|\nabla\widehat{Q}(\widehat{x}_{h,n})\nabla_\theta\widehat{x}_{h,n}-\nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\overline{x}_h\right\|_2\right] \\ &\quad\leq\mathbb{E}_{\overline{x}_h}\left[\left\|\nabla\widehat{Q}(\widehat{x}_{h,n})\nabla_\theta\widehat{x}_{h,n}-\nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\widehat{x}_{h,n}\right\|_2\right]+\mathbb{E}_{\overline{x}_h}\left[\left\|\nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\widehat{x}_{h,n}-\nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\overline{x}_h\right\|_2\right] \\ &\quad\leq 2L_{\widehat{Q}}\cdot\widehat{K}(i)+L_{\widehat{Q}}\cdot\left(\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{d\widehat{s}_{i,n}}{d\theta}-\frac{d\overline{s}_i}{d\theta}\right\|_2\right]+\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta}-\frac{d\overline{a}_i}{d\theta}\right\|_2\right]\right), \end{aligned} \tag{A.16}$$

where the last step follows from the Lipschitz critic assumption.

By the chain rule, we also have

$$\begin{aligned}&\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla_\theta r(\widehat{x}_{i,n})-\nabla_\theta r(\overline{x}_i)\right\|_2\right] \\ &\quad=\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\widehat{x}_{i,n}-\nabla r(\overline{x}_i)\nabla_\theta\overline{x}_i\right\|_2\right] \\ &\quad\leq\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\widehat{x}_{i,n}-\nabla r(\widehat{x}_{i,n})\nabla_\theta\overline{x}_i\right\|_2\right]+\mathbb{E}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\overline{x}_i-\nabla r(\overline{x}_i)\nabla_\theta\overline{x}_i\right\|_2\right] \\ &\quad\leq L_r\cdot\left(\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{d\widehat{s}_{i,n}}{d\theta}-\frac{d\overline{s}_i}{d\theta}\right\|_2\right]+\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta}-\frac{d\overline{a}_i}{d\theta}\right\|_2\right]\right)+2L_r\cdot\widehat{K}(i). \end{aligned} \tag{A.17}$$

Plugging (A.16) and (A.17) into (A.10) gives us

$$
\begin{aligned}
v_t &= \mathbb{E}\left[\left\|\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\right\|_2^2\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[\left\|g - \widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2\right]^2\right] \qquad \text{(A.18)} \\
&\le \mathbb{E}\left[\Big(\sum_{i=0}^{h-1}\gamma^i \mathbb{E}_{\overline{x}_i}\left[\left\|\nabla_\theta r(\widehat{x}_i) - \nabla_\theta r(\overline{x}_i)\right\|_2\right] + \gamma^h \mathbb{E}_{\overline{x}_h}\left[\left\|\nabla \widehat{Q}(\widehat{x}_h)\nabla_\theta \widehat{x}_h - \nabla \widehat{Q}(\overline{x}_h)\nabla_\theta \overline{x}_h\right\|_2\right]\Big)^2\right] \qquad \text{(A.19)} \\
&\le \left((h\cdot L_r + \gamma^h \cdot L_{\widehat{Q}})\cdot\left(\mathbb{E}_{\overline{s}_h}\left[\left\|\frac{d\widehat{s}_{h,n}}{d\theta} - \frac{d\overline{s}_h}{d\theta}\right\|_2\right] + \mathbb{E}_{\overline{a}_h}\left[\left\|\frac{d\widehat{a}_{h,n}}{d\theta} - \frac{d\overline{a}_h}{d\theta}\right\|_2\right] + 2\widehat{K}(h)\right)\right)^2 \\
&\le O\Big(h^8 \widetilde{L}_{\widehat{f}}^{6h}\widetilde{L}_\pi^{4h} + \gamma^{2h}h^6\widetilde{L}_{\widehat{f}}^{6h}\widetilde{L}_\pi^{4h}\Big). \qquad \text{(A.20)}
\end{aligned}
$$

Since the analysis above considers batch size $N = 1$, the bound of gradient variance v_t is established by dividing N. □

Lemma A.1. Denote $e := \sup \mathbb{E}_{\overline{a}_0}\left[\left\|\frac{d\widehat{a}_{0,n}}{d\theta} - \frac{d\overline{a}_0}{d\theta}\right\|_2\right]$, which is a constant that only depends on the initial state distribution and $e = 0$ when the initial state is deterministic. For any $\widehat{s}_{i,n}$ and $i \ge 1$, we have the following results for the two gradient estimators with form (4.2) and (4.5):

$$
\begin{aligned}
\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta} - \frac{d\overline{a}_i}{d\theta}\right\|_2\right] &\le \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i \cdot \big(e + 2L_\theta \cdot i + 4\widetilde{L}_\theta \cdot i^2 \cdot \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i\big), \\
\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{d\widehat{s}_{i,n}}{d\theta} - \frac{d\overline{s}_i}{d\theta}\right\|_2\right] &\le \Big(2L_{\widehat{f}}\cdot \widehat{K}(i-1) + \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i \cdot \big(e + 2L_\theta \cdot i + 4L_\theta \cdot i^2 \cdot \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i\big)\Big)\cdot i \cdot \widetilde{L}_{\widehat{f}}^i.
\end{aligned}
$$

Proof. Firstly, we have

$$
\frac{d\widehat{a}_{i,n}}{d\theta} = \frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{d\widehat{a}_{i-1,n}}{d\theta} + \frac{\partial \widehat{a}_{i,n}}{\partial \theta}. \qquad \text{(A.21)}
$$

Firstly, we obtain from (A.11) that $\forall i \ge 1$,

$$
\begin{aligned}
&\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta} - \frac{d\overline{a}_i}{d\theta}\right\|_2\right] \\
&= \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{d\widehat{a}_{i-1,n}}{d\theta} + \frac{\partial \widehat{a}_{i,n}}{\partial \theta} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \overline{s}_i}{\partial \overline{a}_{i-1}} \cdot \frac{d\overline{a}_{i-1}}{d\theta} - \frac{\partial \overline{a}_i}{\partial \theta}\right\|_2\right] \\
&\le \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{d\widehat{a}_{i-1,n}}{d\theta} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \overline{s}_i}{\partial \overline{a}_{i-1}} \cdot \frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \overline{s}_i}{\partial \overline{a}_{i-1}} \cdot \frac{d\widehat{a}_{i-1,n}}{d\theta} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \overline{s}_i}{\partial \overline{a}_{i-1}} \cdot \frac{d\overline{a}_{i-1}}{d\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \theta} - \frac{\partial \overline{a}_i}{\partial \theta}\right\|_2\right] \\
&\le \left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2 \cdot \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}}\right\|_2 + \left\|\frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \overline{s}_i}{\partial \overline{a}_{i-1}}\right\|_2\right] \\
&\quad + L_\pi L_{\widehat{f}} \cdot \mathbb{E}_{\overline{a}_{i-1}}\left[\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta} - \frac{d\overline{a}_{i-1}}{d\theta}\right\|_2\right] + 2L_\theta \\
&\le 4L_{\widehat{f}} L_\pi \cdot \left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2 + L_\pi L_{\widehat{f}} \cdot \mathbb{E}_{\overline{a}_{i-1}}\left[\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta} - \frac{d\overline{a}_{i-1}}{d\theta}\right\|_2\right] + 2L_\theta. \qquad \text{(A.22)}
\end{aligned}
$$

By iterating the above formula, we have

$$
\begin{aligned}
\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta} - \frac{d\overline{a}_i}{d\theta}\right\|_2\right] &\le e \cdot L_\pi^i L_{\widehat{f}}^i + \Big(4L_{\widehat{f}} L_\pi \cdot \big(L_\theta \cdot i \cdot \widetilde{L}_{\widehat{f}}^i \widetilde{L}_\pi^i\big) + 2L_\theta\Big) \cdot \sum_{j=0}^{i-1} L_\pi^j L_{\widehat{f}}^j \\
&\le \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i \cdot \big(e + 2L_\theta \cdot i + 4L_\theta \cdot i^2 \cdot \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i\big), \qquad \text{(A.23)}
\end{aligned}
$$

where the first inequality follows from (A.13) and applying the bound of $\left\|\frac{d\widehat{a}_{i,n}}{d\theta}\right\|_2$ given in (A.12).

Using similar techniques we have $\forall i \geq 1$ that

$$
\begin{aligned}
&\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{d\widehat{s}_{i,n}}{d\theta}-\frac{d\overline{s}_i}{d\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}+\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{d\overline{s}_{i-1}}{d\theta}-\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{d\overline{a}_{i-1}}{d\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{d\overline{s}_{i-1}}{d\theta}\right\|_2\right] \\
&\qquad+\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{d\overline{a}_{i-1}}{d\theta}\right\|_2\right] \\
&\quad\leq 2L_{\widehat{f}}\cdot\left(\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}\right\|_2+\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\right)+L_{\widehat{f}}\cdot\mathbb{E}_{\overline{s}_{i-1}}\left[\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{d\overline{s}_{i-1}}{d\theta}\right\|_2\right] \\
&\qquad+L_{\widehat{f}}\cdot\mathbb{E}_{\overline{a}_{i-1}}\left[\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{d\overline{a}_{i-1}}{d\theta}\right\|_2\right] \\
&\quad\leq 2L_{\widehat{f}}\cdot\widehat{K}(i-1)+\widetilde{L}_\pi^i\widetilde{L}_{\widehat{f}}^i\cdot\left(e+2L_\theta\cdot i+4L_\theta\cdot i^2\cdot\widetilde{L}_\pi^i\widetilde{L}_{\widehat{f}}^i\right)+L_{\widehat{f}}\cdot\mathbb{E}_{\overline{s}_{i-1}}\left[\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{d\overline{s}_{i-1}}{d\theta}\right\|_2\right] \\
&\quad\leq\left(2L_{\widehat{f}}\cdot\widehat{K}(i-1)+\widetilde{L}_\pi^i\widetilde{L}_{\widehat{f}}^i\cdot\left(e+2L_\theta\cdot i+4L_\theta\cdot i^2\cdot\widetilde{L}_\pi^i\widetilde{L}_{\widehat{f}}^i\right)\right)\cdot\sum_{j=0}^{i-1}L_{\widehat{f}}^j \\
&\quad\leq\left(2L_{\widehat{f}}\cdot\widehat{K}(i-1)+\widetilde{L}_\pi^i\widetilde{L}_{\widehat{f}}^i\cdot\left(e+2L_\theta\cdot i+4L_\theta\cdot i^2\cdot\widetilde{L}_\pi^i\widetilde{L}_{\widehat{f}}^i\right)\right)\cdot i\cdot\widetilde{L}_{\widehat{f}}^i,
\end{aligned}
$$

where we plug (A.23) and the definition (A.15) of $\widehat{K}(i)$ to obtain the second inequality, and the third inequality holds due to (A.13).

□

A.3 Proof of Proposition 5.7

Proof. To begin with, we first provide an upper bound of the gradient bias as follows.

$$
\begin{aligned}
b_t&=\left\|\nabla_\theta J(\pi_{\theta_t})-\mathbb{E}\left[\widehat{\nabla}_\theta J(\pi_{\theta_t})\right]\right\|_2 \\
&=\left\|\mathbb{E}\left[\nabla_\theta J(\pi_{\theta_t})-\widehat{\nabla}_\theta J(\pi_{\theta_t})\right]\right\|_2 \\
&\leq\mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t})-\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2\right],
\end{aligned}
\tag{A.24}
$$

where the expectation is taken over the randomness in the policy and model dynamics. The inequality follows from (A.8).

We have from the h-step model value expansion that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t})-\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\sum_{i=0}^{h-1}\nabla_\theta r(s_i,a_i)+\nabla_\theta V^\pi(s_h)-\sum_{i=0}^{h-1}\nabla_\theta r(\widehat{s}_{i,n},\widehat{a}_{i,n})+\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right\|_2\right].
\end{aligned}
\tag{A.25}
$$

Here, $\widehat{s}_{i+1,n}=\widehat{f}(\widehat{s}_{i,n},\xi_n)$ where $\xi_n\sim p(\xi)$ when $\widehat{\nabla}_\theta J(\pi_\theta)=\widehat{\nabla}_\theta^{\mathrm{DP}}J(\pi_\theta)$, and ξ_n is inferred from real samples (s_i,a_i,s_{i+1}) when $\widehat{\nabla}_\theta J(\pi_\theta)=\widehat{\nabla}_\theta^{\mathrm{DR}}J(\pi_\theta)$.

According to the recursive expression of $\nabla_\theta s$ and $\nabla_\theta \widehat{s}$ by the chain rule, we have

$$
\begin{aligned}
&\mathbb{E}_{a_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta}-\frac{da_i}{d\theta}\right\|_2\right] \\
&=\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}+\frac{\partial\widehat{a}_{i,n}}{\partial\theta}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{da_{i-1}}{d\theta}-\frac{\partial a_i}{\partial\theta}\right\|_2\right] \\
&\le\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{da_{i-1}}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\theta}-\frac{\partial a_i}{\partial\theta}\right\|_2\right] \\
&\le\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\cdot\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\right\|_2+\left\|\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\right\|_2\right] \\
&\quad+L_\pi L_{\widehat{f}}\cdot\mathbb{E}_{a_{i-1}}\left[\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{da_{i-1}}{d\theta}\right\|_2\right]+2L_\theta \\
&\le L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2+L_\pi L_{\widehat{f}}\cdot\mathbb{E}_{a_{i-1}}\left[\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{da_{i-1}}{d\theta}\right\|_2\right]+2L_\theta
\end{aligned}
\tag{A.26}
$$

Since we proved in (A.12) that $\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\le L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi$, we have

$$
\begin{aligned}
\mathbb{E}_{a_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta}-\frac{da_i}{d\theta}\right\|_2\right]&\le\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big)\cdot\sum_{j=0}^{i-1}L^j_\pi L^j_{\widehat{f}} \\
&\le i\cdot\widetilde{L}^i_\pi\widetilde{L}^i_{\widehat{f}}\cdot\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big),
\end{aligned}
\tag{A.27}
$$

where the first step follows from (A.13) and the fact that $\mathbb{E}_{a_i}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta}-\frac{da_i}{d\theta}\right\|_2\right]=0$ since the initial states are sampled from the same distribution ζ.

Similarly,

$$
\begin{aligned}
&\mathbb{E}_{s_i}\left[\left\|\frac{d\widehat{s}_{i,n}}{d\theta}-\frac{ds_i}{d\theta}\right\|_2\right] \\
&=\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}+\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{ds_{i-1}}{d\theta}-\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{da_{i-1}}{d\theta}\right\|_2\right] \\
&=\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{ds_{i-1}}{d\theta}\right\|_2\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{da_{i-1}}{d\theta}\right\|_2\right] \\
&\le\epsilon_f\cdot\mathbb{E}\left[\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}\right\|_2+\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}\right\|_2\right]+L_f\cdot\mathbb{E}_{s_{i-1}}\left[\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{ds_{i-1}}{d\theta}\right\|_2\right] \\
&\quad+L_f\cdot\mathbb{E}_{a_{i-1}}\left[\left\|\frac{d\widehat{a}_{i-1,n}}{d\theta}-\frac{da_{i-1}}{d\theta}\right\|_2\right] \\
&\le\epsilon_f\cdot\widehat{K}(i-1)+i\cdot\widetilde{L}^i_\pi\widetilde{L}^i_{\widehat{f}}\cdot\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big) \\
&\quad+L_f\cdot\mathbb{E}_{s_{i-1}}\left[\left\|\frac{d\widehat{s}_{i-1,n}}{d\theta}-\frac{ds_{i-1}}{d\theta}\right\|_2\right] \\
&\le\Big(\epsilon_f\cdot\widehat{K}(i-1)+i\cdot\widetilde{L}^i_\pi\widetilde{L}^i_{\widehat{f}}\cdot\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big)\Big)\cdot i\cdot\widetilde{L}^i_{\widehat{f}},
\end{aligned}
\tag{A.28}
$$

where the first inequality follows from the definition of ϵ_f in (4.7), the second inequality holds due to (A.27) and the definition of $\widehat{K}$ in (A.15), the last inequality follows from (A.13).

Therefore, the gradient bias of the reward at timestep i satisfies:

$$\begin{aligned}
&\mathbb{E}_{\widehat{x}_{i,n},x_i}\left[\left\|\frac{dr(\widehat{x}_{i,n})}{d\theta}-\frac{dr(x_i)}{d\theta}\right\|_2\right]\\
&=\mathbb{E}\left[\left\|\frac{dr}{d\widehat{x}_{i,n}}\cdot\frac{d\widehat{x}_{i,n}}{d\theta}-\frac{dr}{dx_i}\cdot\frac{dx_i}{d\theta}\right\|_2\right]\\
&\le\mathbb{E}\left[\left\|\frac{dr}{d\widehat{x}_{i,n}}\cdot\frac{d\widehat{x}_{i,n}}{d\theta}-\frac{dr}{dx_i}\cdot\frac{d\widehat{x}_{i,n}}{d\theta}\right\|_2+\left\|\frac{dr}{dx_i}\cdot\frac{d\widehat{x}_{i,n}}{d\theta}-\frac{dr}{dx_i}\cdot\frac{dx_i}{d\theta}\right\|_2\right]\\
&\le 2L_r\cdot\widehat{K}(i)+L_r\cdot\left(\mathbb{E}\left[\left\|\frac{d\widehat{s}_{i,n}}{d\theta}-\frac{ds_i}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{d\widehat{a}_{i,n}}{d\theta}-\frac{da_i}{d\theta}\right\|_2\right]\right).
\end{aligned}\tag{A.29}$$

Similar with the definition in (5.2), we define $\widehat{\nu}_2(s):=\mathbb{P}(\widehat{s}_h=s)$ and $\widehat{\sigma}_2(s):=\mathbb{P}(\widehat{s}_h=s,\widehat{a}_h=a)$ where $\widehat{s}_0\sim\zeta$, $\widehat{a}_i\sim\pi(\cdot\,|\,\widehat{s}_i)$, and $\widehat{s}_{i+1}\sim f(\cdot\,|\,\widehat{s}_i,\widehat{a}_i)$.

In Lemma A.2, we deal with the misalignment between $\nabla_\theta V^{\pi_\theta}$ and $\nabla_\theta\widehat{V}_\omega$, specifically, the recursive structure of V^{π_θ} and the non-recursive value function approximation $\widehat{V}_\omega$. Now we are ready to bound the gradient bias brought by the critic with the following inequality:

$$\begin{aligned}
&\mathbb{E}_{s_h\sim\overline{\nu}_2,\widehat{s}_h\sim\widehat{\nu}_2}\left[\left\|\nabla_\theta V^\pi(s_h)-\nabla_\theta\widehat{V}(\widehat{s}_h)\right\|_2\right]\\
&\le\mathbb{E}_{(s,a)\sim\overline{\sigma}_2,(\widehat{s},\widehat{a})\sim\widehat{\sigma}_2}\left[\left\|\kappa\cdot\left(\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\right)-\frac{\partial\widehat{Q}}{\partial\widehat{a}}\cdot\frac{d\widehat{a}}{d\theta}-\frac{\partial\widehat{Q}}{\partial\widehat{s}}\cdot\frac{d\widehat{s}}{d\theta}\right\|_2\right]\\
&\le\kappa'\cdot\mathbb{E}_{(s,a)\sim\overline{\sigma}_2,(\widehat{s},\widehat{a})\sim\widehat{\sigma}_2}\left[\left\|\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}-\frac{\partial\widehat{Q}}{\partial\widehat{a}}\cdot\frac{d\widehat{a}}{d\theta}-\frac{\partial\widehat{Q}}{\partial\widehat{s}}\cdot\frac{d\widehat{s}}{d\theta}\right\|_2\right]\\
&\le\kappa'\cdot\mathbb{E}\left[\left\|\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta}-\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta}\right\|_2+\left\|\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta}-\frac{\partial\widehat{Q}}{\partial\widehat{a}}\cdot\frac{d\widehat{a}}{d\theta}\right\|_2\right.\\
&\qquad\left.+\left\|\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}-\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\right\|_2+\left\|\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}-\frac{\partial\widehat{Q}}{\partial\widehat{s}}\cdot\frac{d\widehat{s}}{d\theta}\right\|_2\right]\\
&\le\kappa'\cdot L_Q\cdot\left(\mathbb{E}\left[\left\|\frac{da}{d\theta}-\frac{d\widehat{a}}{d\theta}\right\|_2+\left\|\frac{ds}{d\theta}-\frac{d\widehat{s}}{d\theta}\right\|_2\right]\right)+\kappa'\cdot\widehat{K}(h)\cdot\epsilon_v,
\end{aligned}\tag{A.30}$$

where the first inequality uses the results in Lemma A.2, the last inequality follows from (A.15) and the definition of ϵ_v in (4.8).

Therefore, according to the gradient bias decomposition in (A.25), we obtain

$$\begin{aligned}
b_t&\le\mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t})-\widehat{\nabla}_\theta J(\pi_{\theta_t})\right\|_2\right]\\
&\le\sum_{i=0}^{h-1}\gamma^i\cdot\mathbb{E}\left[\left\|\nabla_\theta r(s_i,a_i)-\nabla_\theta r(\widehat{s}_i,\widehat{a}_i)\right\|_2\right]+\gamma^h\cdot\mathbb{E}_{s_h\sim\overline{\nu}_2,\widehat{s}_h\sim\widehat{\nu}_2}\left[\left\|\nabla_\theta V^\pi(s_h)-\nabla_\theta\widehat{V}(\widehat{s}_h)\right\|_2\right]\\
&\le h\cdot\left(L_r\cdot\left(\mathbb{E}\left[\left\|\frac{d\widehat{s}_{h,n}}{d\theta}-\frac{ds_h}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{d\widehat{a}_{h,n}}{d\theta}-\frac{da_h}{d\theta}\right\|_2\right]\right)+2L_r\cdot\widehat{K}(h)\right)\\
&\quad+\gamma^h\cdot\kappa'\cdot L_Q\cdot\left(\mathbb{E}\left[\left\|\frac{d\widehat{s}_{h,n}}{d\theta}-\frac{ds_h}{d\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{d\widehat{a}_{h,n}}{d\theta}-\frac{da_h}{d\theta}\right\|_2\right]\right)+\gamma^h\cdot\kappa'\cdot\widehat{K}(h)\cdot\epsilon_v,
\end{aligned}$$

where the last step follows from previous results, i.e. (A.29) and (A.30).

By plugging (A.27), (A.28) and $\widehat{K}$ in (A.15) into the above expression, we obtain

$$b_t\le O\Big(h^4L_{\widehat{f}}^{3h}L_\pi^h(\epsilon_f+\widetilde{L}_\pi^h)+\kappa'h^2\gamma^hL_{\widehat{f}}^{2h}L_\pi^h\epsilon_v\Big).$$

□

Lemma A.2. Define $\overline{\nu}_2=\mathbb{P}(s_h=s)$ where $s_0\sim\zeta$, $a_i\sim\pi(\cdot\,|\,s_i)$, and $s_{i+1}\sim f(\cdot\,|\,s_i,a_i)$. Under Assumption 5.6, the expected value gradient over state distribution at timestep h can be bounded by

$$\mathbb{E}_{s_h\sim\overline{\nu}_2}\left[\nabla_\theta V^{\pi_\theta}(s_h)\right]\le\kappa\cdot\mathbb{E}_{(s,a)\sim\overline{\sigma}_2}\left[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\right]$$

Proof. At state s_h, the true value gradient can be decomposed by

$$
\begin{aligned}
&\nabla_\theta V^{\pi_\theta}(s_h) \\
&= \nabla_\theta \mathbb{E}\Big[r(s_h,a_h) + \gamma \cdot \int_{\mathcal{S}} f\big(s_{h+1}|s_h,a_h\big)\cdot V^{\pi}(s_{h+1})ds_{h+1}\Big] \\
&= \nabla_\theta \mathbb{E}\Big[r(s_h,a_h)\Big] + \gamma \cdot \mathbb{E}\Big[\nabla_\theta \int_{\mathcal{S}} f\big(s_{h+1}|s_h,a_h\big)\cdot V^{\pi}(s_{h+1})ds_{h+1}\Big] \\
&= \mathbb{E}\Big[\frac{\partial r_h}{\partial a_h}\cdot\frac{da_h}{d\theta} + \frac{\partial r_h}{\partial s_h}\cdot\frac{ds_h}{d\theta} \\
&\qquad + \gamma\int_{\mathcal{S}}\Big(\nabla_\theta f\big(s_{h+1}|s_h,a_h\big)\cdot V^{\pi}(s_{h+1}) + f\big(s_{h+1}|s_h,a_h\big)\cdot\nabla_\theta V^{\pi}(s_{h+1})\Big)ds_{h+1}\Big] \\
&= \mathbb{E}\Big[\frac{\partial r_h}{\partial a_h}\cdot\frac{da_h}{d\theta} + \frac{\partial r_h}{\partial s_h}\cdot\frac{ds_h}{d\theta} + \gamma\int_{\mathcal{S}}\Big(\nabla_a f(s_{h+1}|s_h,a)\cdot\frac{da_h}{d\theta}\cdot V^{\pi}(s_{h+1}) \\
&\qquad + \nabla_s f(s_{h+1}|s_h,a_h)\cdot\frac{ds_h}{d\theta}\cdot V^{\pi}(s_{h+1}) + f\big(s_{h+1}|s_h,a_h\big)\cdot\nabla_\theta V^{\pi}(s_{h+1})\Big)ds_{h+1}\Big],
\end{aligned}
$$

where the first follows from Bellman equation and the remaining equations hold due to the chain rule. The continuity Assumption 3.1 allows us to use the Leibniz integral rule to exchange order of derivative and integration.

It is worth noting that when $h \geq 1$, both a_h and s_h have dependencies on all previous timesteps. For example, $\nabla_\theta r(s_h,a_h) = \frac{\partial r_h}{\partial a_h}\cdot\frac{da_h}{d\theta} + \frac{\partial r_h}{\partial s_h}\cdot\frac{ds_h}{d\theta}$ for $h \geq 1$. However, it differs from the case when $h = 0$, e.g. the Deterministic Policy Gradient theorem Silver et al. (2014) where we can simply write $\nabla_\theta r(s_h,a_h) = \frac{\partial r_h}{\partial a_h}\cdot\frac{\partial a_h}{\partial \theta}$.

By noting that $Q^{\pi_\theta}(s_h,a_h) = r(s_h,a_h) + \gamma\int_{\mathcal{S}} f(s_{h+1}|s_h,a_h)\cdot V^{\pi}(s_{h+1})ds_{h+1}$, we can combine the reward and value terms and proceed by

$$
\begin{aligned}
V^{\pi_\theta}(s_h) &= \mathbb{E}\Big[\nabla_a\Big(r(s_h,a_h) + \gamma\int_{\mathcal{S}} f(s_{h+1}|s_h,a_h)\cdot V^{\pi}(s_{h+1})ds_{h+1}\Big)\cdot\frac{da_h}{d\theta} \\
&\qquad + \nabla_s\Big(r(s_h,a_h) + \gamma\int_{\mathcal{S}} f(s_{h+1}|s_h,a_h)\cdot V^{\pi}(s_{h+1})ds_{h+1}\Big)\cdot\frac{ds_h}{d\theta} \\
&\qquad + \gamma\int_{\mathcal{S}} f\big(s_{h+1}|s_h,a_h\big)\cdot\nabla_\theta V^{\pi}(s_{h+1})ds_{h+1}\Big] \\
&= \mathbb{E}\Big[\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{da_h}{d\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{ds_h}{d\theta} + \gamma\int_{\mathcal{S}} f\big(s_{h+1}|s_h,a_h\big)\cdot\nabla_\theta V^{\pi}(s_{h+1})ds_{h+1}\Big],
\end{aligned}
$$

Iterating the above formula we obtain

$$
\nabla_\theta V^{\pi_\theta}(s_h) = \mathbb{E}\Big[\int_{\mathcal{S}}\sum_{i=h}^{\infty}\gamma^{i-h}\cdot f\big(s_{i+1}|s_i,a_i\big)\cdot\Big(\frac{\partial Q^{\pi_\theta}}{\partial a_i}\cdot\frac{da_i}{d\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_i}\cdot\frac{ds_i}{d\theta}\Big)ds_{i+1}\Big].
$$

Taking the expectation over s_h we have

$$
\mathbb{E}_{s_h\sim\overline{\nu}_2}\big[\nabla_\theta V^{\pi_\theta}(s_h)\big] = \mathbb{E}_{(s,a)\sim\overline{\sigma}_1}\Big[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\Big]. \tag{A.31}
$$

From Assumption 5.6, it holds that

$$
\begin{aligned}
\mathbb{E}_{s_h\sim\overline{\nu}_2}\big[\nabla_\theta V^{\pi_\theta}(s_h)\big] &= \mathbb{E}_{(s,a)\sim\overline{\sigma}_1}\Big[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\Big] \\
&\leq \mathbb{E}_{(s,a)\sim\overline{\sigma}_2}\Big[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\Big]\cdot\Big\{\mathbb{E}_{\overline{\sigma}_1}\Big[\Big(\frac{d\overline{\sigma}_1}{d\overline{\sigma}_2}(s,a)\Big)^2\Big]\Big\}^{1/2} \\
&\leq \kappa\cdot\mathbb{E}_{(s,a)\sim\overline{\sigma}_2}\Big[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{da}{d\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{ds}{d\theta}\Big]
\end{aligned} \tag{A.32}
$$

□

A.4 PROOF OF PROPOSITION 5.8

Proof. To solve the problem of finding the optimal model expansion step h^*, we define $g(h)$ as follows:

$$g(h) := c \cdot (\delta \cdot b_t + \frac{\eta}{2} \cdot v_t) + b_t^2 + v_t, \tag{A.33}$$

where the bound of the gradient bias b_t and gradient variance v_t are given in (**??**) and (A.18), respectively.

Thus, h^* is given by $h^* = \operatorname{argmin}_h g(h)$. It then holds for the optimal $h = h^*$ that

$$\frac{\partial}{\partial h} g(h) = O\big((h^7 + h^6\gamma^{2h} \cdot \log\gamma)/N + h^7 \cdot \epsilon_f^2 + h^4\gamma^{2h} \log\gamma \cdot \epsilon_v^2\big) = 0. \tag{A.34}$$

For notation simplicity, we define

$$\frac{\partial}{\partial h} g_1(h) := (h^7 + h^6\gamma^{2h} \cdot \log\gamma)/N, \ \ \frac{\partial}{\partial h} g_2(h) := h^7 \cdot \epsilon_f^2 + h^4\gamma^{2h} \log\gamma \cdot \epsilon_v^2. \tag{A.35}$$

By solving $\frac{\partial}{\partial h} g_1(h) + \frac{\partial}{\partial h} g_2(h) = 0$, we can represent the optimal h^* in terms of the solution using the big-O notation.

Next, we show that both $g_1(h)$ and $g_2(h)$ have unique optima in the domain $h \in (0, \infty)$. By setting $\frac{\partial}{\partial h} g_2(h) = 0$, we have

$$h^3 \cdot \epsilon_f^2 = \gamma^{2h} \log\frac{1}{\gamma} \cdot \epsilon_v^2. \tag{A.36}$$

Taking the natural logarithm on both sides gives us

$$3\log h + 2\log\epsilon_f = 2h \cdot \log\gamma + \log\log\frac{1}{\gamma} + 2\log\epsilon_v. \tag{A.37}$$

Rearranging terms we obtain

$$\log h - \frac{2}{3} h \cdot \log\gamma = \frac{1}{3}\log\log\frac{1}{\gamma} + \frac{2}{3}\log\epsilon_v. \tag{A.38}$$

We have from the exponential of both sides that

$$h \cdot \exp\big(-\frac{2}{3} h \cdot \log\gamma\big) = \big(\log\frac{1}{\gamma}\big)^{\frac{1}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}. \tag{A.39}$$

Therefore, by multiplying $-2/5 \cdot \log\gamma$ in both the LHS and the RHS, it holds that

$$-\frac{2}{3} h \log\gamma \cdot \exp\big(-\frac{2}{3} h \cdot \log\gamma\big) = \frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}. \tag{A.40}$$

Recall the definition of Lambert W function that $W(x \cdot e^x) = x$, we can simplify the above equation by

$$-\frac{2}{3} h \cdot \log\gamma = W\big(\frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}\big). \tag{A.41}$$

The unique optima of $g_2(h)$ is thus

$$h = \frac{3}{2\log\frac{1}{\gamma}} \cdot W\big(\frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}\big). \tag{A.42}$$

Using similar techniques, we can solve $\frac{\partial}{\partial h} g_1(h) = 0$ by

$$\begin{aligned}
h &= \gamma^{2h} \cdot \log \frac{1}{\gamma} \\
\log h &= 2h \log \gamma + \log \log \frac{1}{\gamma} \\
\log h - 2h \log \gamma &= \log \log \frac{1}{\gamma} \\
h \cdot \exp\left(-2h \cdot \log \gamma\right) &= \log \frac{1}{\gamma} \\
-2h \log \gamma \cdot \exp\left(-2h \cdot \log \gamma\right) &= 2(\log \gamma)^2 \\
-2h \cdot \log \gamma &= W\Big(2(\log \gamma)^2\Big) \\
h &= \frac{1}{2 \log \frac{1}{\gamma}} \cdot W\Big(2(\log \gamma)^2\Big).
\end{aligned} \tag{A.43}$$

Now we have shown that the minima of $g_1(h)$ and $g_2(h)$ is unique (i.e. (A.43) and (A.42)). Therefore, $g_1(h) + g_2(h)$ also has a unique minima within the range of

$$\begin{aligned}
\Bigg[&\min\Big\{ \frac{1}{2 \log \frac{1}{\gamma}} \cdot W\big(2(\log \gamma)^2\big),\ \frac{5}{2 \log \frac{1}{\gamma}} \cdot W\big(\frac{2}{5} (\log \gamma)^{\frac{6}{5}} \cdot (\epsilon_v / \epsilon_f)^{\frac{2}{5}}\big) \Big\}, \\
&\max\Big\{ \frac{1}{2 \log \frac{1}{\gamma}} \cdot W\big(2(\log \gamma)^2\big),\ \frac{5}{2 \log \frac{1}{\gamma}} \cdot W\big(\frac{2}{5} (\log \gamma)^{\frac{6}{5}} \cdot (\epsilon_v / \epsilon_f)^{\frac{2}{3}}\big) \Big\} \Bigg].
\end{aligned} \tag{A.44}$$

We conclude that the optimal expansion step has the following expression

$$h^* = O\Big(W\big(2(\log \gamma)^2\big)/N + W\big(\frac{2}{3} (\log \gamma)^{\frac{4}{3}} \cdot (\epsilon_v / \epsilon_f)^{\frac{2}{3}}\big)\Big).$$

□

A.5 Proof of Proposition 5.9

Proof. According to the definition of stationary point $\widehat{\theta}$ that $J(\pi_{\widehat{\theta}}) = 0$, we have

$$\nabla_\theta J(\pi_{\widehat{\theta}})^\top \theta = 0,\ \forall \theta. \tag{A.45}$$

Define the advantage function as $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$. By the policy gradient theorem, it holds that

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim \sigma_\pi} \big[A^\pi(s,a) \cdot \nabla_\theta \log \pi(a|s) \big]. \tag{A.46}$$

Therefore, combining (A.45) and (A.46) we obtain that

$$\mathbb{E}_{(s,a) \sim \sigma_{\pi_{\widehat{\theta}}}} \Big[A^{\pi_{\widehat{\theta}}}(s,a) \cdot \big(\nabla_\theta \log \pi_{\widehat{\theta}}(a|s) \big)^\top \theta \Big] = 0,\ \forall \theta. \tag{A.47}$$

By the Performance Difference Lemma A.3, we have

$$\begin{aligned}
(1-\gamma) \cdot \big(J(\pi^*) - J(\pi_{\widehat{\theta}}) \big) &= \mathbb{E}_{(s,a) \sim \sigma_{\pi^*}} \big[A^{\pi_{\widehat{\theta}}}(s,a) \big] \\
&= \mathbb{E}_{s \sim \nu_{\pi^*}} \Big[\mathbb{E}_{a \sim \pi^*} \big[A^{\pi_{\widehat{\theta}}}(s,a) \big] \Big] \\
&= \mathbb{E}_{s \sim \nu_{\pi^*}} \Big[\mathbb{E}_{a \sim \pi^*} \big[Q^{\pi_{\widehat{\theta}}}(s,a) \big] - V^{\pi_{\widehat{\theta}}}(s) \Big] \\
&= \mathbb{E}_{s \sim \nu_{\pi^*}} \Big[\big\langle Q^{\pi_{\widehat{\theta}}}(s,\cdot), \pi^*(\cdot|s) - \pi_{\widehat{\theta}}(\cdot|s) \big\rangle \Big] \\
&= \mathbb{E}_{s \sim \nu_{\pi^*}} \Big[\big\langle A^{\pi_{\widehat{\theta}}}(s,\cdot), \pi^*(\cdot|s) - \pi_{\widehat{\theta}}(\cdot|s) \big\rangle \Big].
\end{aligned} \tag{A.48}$$

Combining the above equation with (A.47) implies that

$$
\begin{aligned}
&(1-\gamma)\cdot\big(J(\pi^*)-J(\pi_{\widehat{\theta}})\big)\\
&\quad=\mathbb{E}_{s\sim\nu_{\pi^*}}\Big[\big\langle A^{\pi_{\widehat{\theta}}}(s,\cdot),\pi^*(\cdot|s)-\pi_{\widehat{\theta}}(\cdot|s)\big\rangle\Big]-\mathbb{E}_{(s,a)\sim\sigma_{\pi_{\widehat{\theta}}}}\Big[A^{\pi_{\widehat{\theta}}}(s,a)\cdot\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\Big]\\
&\quad=\mathbb{E}_{s\sim\nu_{\pi^*}}\Big[\big\langle A^{\pi_{\widehat{\theta}}}(s,\cdot),\pi^*(\cdot|s)-\pi_{\widehat{\theta}}(\cdot|s)\big\rangle\Big]-\mathbb{E}_{s\sim\nu_{\pi_{\widehat{\theta}}}}\Big[\big\langle A^{\pi_{\widehat{\theta}}}(s,\cdot),\big(\nabla_\theta\log\pi_{\widehat{\theta}}(\cdot|s)\big)^\top\theta\cdot\pi_{\widehat{\theta}}(\cdot|s)\big\rangle\Big]\\
&\quad=\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}A^{\pi_{\widehat{\theta}}}(s,a)\Big(\big(\pi^*(a|s)-\pi_{\widehat{\theta}}(a|s)\big)d\nu_{\pi^*}(s)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\cdot\pi_{\widehat{\theta}}(a|s)d\nu_{\pi_{\widehat{\theta}}}(s)\Big),\forall\theta,
\end{aligned}
\tag{A.49}
$$

where the first equality holds since $\sigma(\cdot,\cdot)=\nu_\pi(\cdot)\cdot\pi(\cdot|\cdot)$.

For all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $\forall\theta$, it holds that

$$
\begin{aligned}
&\big(\pi^*(a|s)-\pi_{\widehat{\theta}}(a|s)\big)d\nu_{\pi^*}(s)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\cdot\pi_{\widehat{\theta}}(a|s)d\nu_{\pi_{\widehat{\theta}}}(s)\\
&\quad=\left(\frac{\pi^*(a|s)-\pi_{\widehat{\theta}}(a|s)}{\pi_{\widehat{\theta}}(a|s)}\cdot\frac{d\nu_{\pi^*}}{d\nu_{\pi_{\widehat{\theta}}}}(s)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\right)\cdot\pi_{\widehat{\theta}}(a|s)d\nu_{\pi_{\widehat{\theta}}}(s)\\
&\quad=\Big(u_{\widehat{\theta}}(s,a)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\Big)d\sigma_{\pi_{\widehat{\theta}}}(s,a).
\end{aligned}
\tag{A.50}
$$

Here, $u_{\widehat{\theta}}(s,a)$ is defined as

$$
u_{\widehat{\theta}}(s,a):=\frac{d\sigma_{\pi^*}}{d\sigma_{\pi_{\widehat{\theta}}}}(s,a)-\frac{d\nu_{\pi^*}}{d\nu_{\pi_{\widehat{\theta}}}}(s)
$$

where $d\sigma_{\pi^*}/d\sigma_{\pi_{\widehat{\theta}}}$ and $d\nu_{\pi^*}/d\nu_{\pi_{\widehat{\theta}}}$ are the Radon-Nikodym derivatives.

Therefore, by plugging (A.50) into (A.49), we have

$$
\begin{aligned}
(1-\gamma)\cdot\big(J(\pi^*)-J(\pi_{\widehat{\theta}})\big)&=\int_{\mathcal{S}\times\mathcal{A}}A^{\pi_{\widehat{\theta}}}(s,a)\cdot\Big(u_{\widehat{\theta}}(s,a)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\Big)d\sigma_{\pi_{\widehat{\theta}}}(s,a)\\
&\le\big\|A^{\pi_{\widehat{\theta}}}(\cdot,\cdot)\big\|_{\sigma_{\pi_{\widehat{\theta}}}}\cdot\Big\|u_{\widehat{\theta}}(s,a)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\Big\|_{\sigma_{\pi_{\widehat{\theta}}}}\\
&\le 2r_{\max}\cdot\Big\|u_{\widehat{\theta}}(s,a)-\big(\nabla_\theta\log\pi_{\widehat{\theta}}(a|s)\big)^\top\theta\Big\|_{\sigma_{\pi_{\widehat{\theta}}}},\ \forall\theta,
\end{aligned}
\tag{A.51}
$$

where the last inequality holds since $|A^{\pi_{\widehat{\theta}}}(\cdot,\cdot)|\le 2\max_{s,a}Q(s,a)\le 2r_{\max}$.

Taking the infimum with respect to θ concludes the proof.

□

Lemma A.3 (Performance Difference Lemma)**.** For all policies π, π^* and distribution μ over $\mathcal{S}$, we have

$$
J(\pi)-J(\pi')=\frac{1}{1-\gamma}\cdot\mathbb{E}_{(s,a)\sim\sigma_\pi}[A^{\pi'}(s,a)].
\tag{A.52}
$$

Proof. Let $\mathbb{P}^{\pi}(\tau|s_0 = s)$ denote the probability of observing trajectory τ starting at state s_0 and then following π. Then the value difference can be written as

$$\begin{aligned}
V^{\pi}(s) - V^{\pi'}(s) &= \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_0=s)}\Big[\sum_{h=0}^{\infty}\gamma^h r(s_h,a_h)\Big] - V^{\pi'}(s) \\
&= \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_0=s)}\Big[\sum_{h=0}^{\infty}\gamma^h \big(r(s_h,a_h) + V^{\pi'}(s_h) - V^{\pi'}(s_h)\big)\Big] - V^{\pi'}(s) \\
&= \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_0=s)}\Big[\sum_{h=0}^{\infty}\gamma^h \big(r(s_h,a_h) + \gamma V^{\pi'}(s_{h+1}) - V^{\pi'}(s_h)\big)\Big] \\
&= \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_0=s)}\Big[\sum_{h=0}^{\infty}\gamma^h \big(r(s_h,a_h) + \gamma\mathbb{E}[V^{\pi'}(s_{h+1})|s_h,a_h] - V^{\pi'}(s_h)\big)\Big] \\
&= \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_0=s)}\Big[\sum_{h=0}^{\infty}\gamma^h \big(Q^{\pi'}(s_h,a_h) - V^{\pi'}(s_h)\big)\Big] \\
&= \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_0=s)}\Big[\sum_{h=0}^{\infty}\gamma^h A^{\pi'}(s_h,a_h)\Big], \qquad \text{(A.53)}
\end{aligned}$$

where the third equation rearranges terms in the summation via telescoping, and the forth equality follows from the law of total expectation.

From the definition of objective $J(\pi)$ in (2.4), we obtain

$$\begin{aligned}
J(\pi) - J(\pi') &= \mathbb{E}_{s_0\sim\zeta}[V^{\pi}(s_0) - V^{\pi'}(s_0)] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\sigma_{\pi}}[A^{\pi'}(s,a)]. \qquad \text{(A.54)}
\end{aligned}$$

□