MODEL-BASED REPARAMETERIZATION POLICY GRADIENT METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

1

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), by finding actions that maximize the accumulated long-term reward. As one of the most pupular algorithms, the policy gradient method (PGM) (Sutton et al., 1999; Konda & Tsitsiklis, 1999; Silver et al., 2014) seeks to search for the optimal policy by iteratively computing and following a stochastic gradient direction with respect to the policy parameters. Thus, the [estimation accuracy] of the stochastic gradient naturally play crucial roles in the performance of PGMs.

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., 2022a; Xu et al., 2022) (see Section **??**). In a nutshell, the LR estimator performs zeroth-order estimation as it only requires sampling of function values, while the RP estimator exploits the differentiability of the estimated function, which offers a unique advantage when one can obtain accurate function approximation. In stochastic optimization, the theoretical benefits of using RP estimators have mainly been understood as enjoying smaller variance (Mohamed et al., 2020). Yet, we show that the opposite is true in RL objectives involving long-horizon sequential decision-making: RP-based PGMs have exponentially exploding gradient variance, which is the key obstacle to leveraging a model.

In fact, 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. A 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 to 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 unrolling step.

The lack of fundamental analysis for RP PGMs also has far-reaching effects beyond theoretical interest. As model-based RP gradients are experimentally observed to have 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 affect its convergence could significantly improve the current algorithmic design of model-based RP PGMs.

Our contributions mainly reside in the following aspects:

- We present a unified framework that can be instantiated to multiple RP-based PGM algorithms. We prove the non-asymptotic global convergence of this framework and 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 results also suggest 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 the model and critic.
- We also show that the initial states in the RP gradient estimators should be sampled from the mixture of the MDP initial distribution and the state visitation, suggesting new sampling schemes.
- We experimentally evaluate several instantiations of RP PGMs and demonstrate the explosion of gradient variance that leads to highly non-smooth loss landscape. We also investigate the effect of smooth proxy models.

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 7.

2 BACKGROUND

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 $r(s,a)$ and the MDP transitions 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

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

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

where the expectation $\mathbb{E}^{\pi,f}[\,\cdot\,]$ is taken with respect to the dynamic induced by the policy π and the transition model f.

Define 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

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

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

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

Stochastic Gradient Estimation. Consider the general problem form underlying policy gradient, i.e., computing the gradient of a probabilistic objective with respect to 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: By leveraging the *score function*, LR gradient estimators only require samples of the function values. With $\nabla_\theta \log p(x;\theta) = \nabla_\theta p(x;\theta)/p(x;\theta)$, the LR gradient is

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

Reparameterization (RP) Gradient: RP gradient benefits 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 indirect ways of drawing samples:

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

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

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

3 Reparameterization Gradient in RL

In this section, we first present a general form of the reparameterization gradient in RL. We then provide a dynamical expression of the model-based RP gradient.

3.1 Reparameterization Policy-Value Gradient

We consider the stochastic policy $a = \pi_\theta(s, \varsigma)$ with noise variable ς in continuous action spaces.

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

We make the above assumption to ensure that the first-order gradient through value is well-defined. The general form of the reparameterization policy-value gradient is as follows:

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

By performing sequential decision-making, any immediate action could lead to changes in 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

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

where $\nabla_a Q^\pi$ can be estimated by the derivative of a learned critic, leading to model-free frameworks (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:

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

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. It is thus natural to ask how the model properties (e.g., prediction accuracy and 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 introduce model value expansion and present the resulting RP PG frameworks.

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 be sampled from fixed distributions or inferred from real samples, which we will discuss in more detail.

4.2 MODEL-BASED GRADIENT ESTIMATION

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

Model Derivatives on 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 $\widehat{s}_{0,n}\sim\mu(\cdot)$, $\widehat{a}_{i,n} = \pi(\widehat{s}_{i,n},\varsigma_n;\theta)$ and $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n;\psi)$ with $\varsigma_n\sim p(\varsigma)$, $\xi_n\sim p(\xi)$. Here, $\mu(\cdot)$ is the distribution where the starting states of the simulated trajectories are sampled. We will show in Section 5 that μ can be specified as a mixture of ζ and σ.

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). For example, RP(0) (Amos et al., 2021) that can be viewed as the stochastic counterpart of DPG (Silver et al., 2014; Lillicrap et al., 2015). When $h=\infty$, the resulting algorithm is BPTT (Grzeszczuk 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,

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

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

The recursion terminates at the h-th timestep with $\widehat{\nabla} V^\pi(\widehat{s}_{h,n}) = \nabla \widehat{V}_\omega(\widehat{s}_{h,n})$ if $h < \infty$, and $\widehat{\nabla} V^\pi(\widehat{s}_{h,n}) = 0$ if $h = \infty$. The RP gradient estimator takes the form of

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

where $\widehat{s}_{0,n} \sim \mu(\cdot)$, $\widehat{a}_{i,n} = \pi(\widehat{s}_{i,n}, \varsigma_n; \theta)$, $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n}, \widehat{a}_{i,n}, \xi_n; \psi)$, and ς_n, ξ_n are inferred from the data sample $(s_{i,n}, a_{i,n}, s_{i+1,n})$ such that $\widehat{a}_{i,n} = a_{i,n}$ and $\widehat{s}_{i+1,n} = s_{i+1,n}$. Example algorithms include SVG (Heess et al., 2015) and its variants (Kumpati et al., 1990; Abbeel et al., 2006).

4.3 Algorithmic Framework

Based on the model-based RP gradient estimators, 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:

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

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

Model Update. Canonical model-based RL learns a forward model that predicts how the system evolves when applying action a at state s, by predicting the mean of transition with minimized mean squared error (MSE) or fitting a probabilistic function with maximum likelihood estimation (MLE).

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

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

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

Since objective mismatches between minimizing the state prediction error and gradient error, Li et al. (2021) proposed to learn the model whose directional derivative is consistent with the samples. However, we observe in experiments that state-predictive models suffice to give good results. This is because if the forward model learned in visited regions can extrapolate, the gradient error can be bounded with finite difference approximation. Thus, ϵ_f might be expressed as MSE with an additional measure of the 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, i.e., $Q = \mathcal{T}^\pi Q \implies Q = Q^\pi$. The Bellman operator $\mathcal{T}^\pi$ is defined as

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

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 error ϵ_v as

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

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).

Algorithm 1 Model-Based Reparameterization Policy Gradient Methods

Input: Number of iterations T, learning rate η, batch size N, state distribution $\mu(\cdot)$
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: Sample states from $\mu(\cdot)$ and estimate $\widehat{\nabla}_\theta J(\pi_{\theta_t}) = \widehat{\nabla}^{\mathrm{DP}}_\theta J(\pi_\theta)$ (4.2) or $\widehat{\nabla}^{\mathrm{DR}}_\theta J(\pi_\theta)$ (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 establish the convergence of model-based RP PGMs. More importantly, we study the relationship between the training error, model smoothness, and gradient bias, variance, which suggest several potential algorithmic designs (in the remarks).

To begin with, we impose a regularity condition on the smoothness of the objective $J(\pi_\theta)$. 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; Zhang et al., 2020).

Assumption 5.1 (Lipschitz Smooth Objective)**.** Assume $J(\pi_\theta)$ is L-smooth in θ, such that $\|\nabla_\theta J(\pi_{\theta_1}) - \nabla_\theta J(\pi_{\theta_2})\|_2 \le L\|\theta_1 - \theta_2\|_2$. See A.1 for an alternative low-level assumption.

We characterize the convergence of RP PGMs by first providing the following proposition.

Proposition 5.2 (Convergence to Stationary Points)**.** Suppose the policy parameter space Θ satisfies that $\|\theta\|_2 \le \delta$ for any $\theta \in \Theta$. Denote $c := (\eta - L\eta^2)^{-1}$. It then holds for $T \ge 4L^2$ that

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

where the gradient bias b_t and gradient variance v_t is defined as

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

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 $L_r, L_{\widehat{f}}, L_\pi, L_{\widehat{Q}}$ Lipschitz continuous (we defer the details to Appendix A.2).

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

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

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

Therefore, when the underlying dynamic of the MDP is complex and contact-rich (Suh et al., 2022a; Xu et al., 2022), model-based RP PGMs have their own advantages by leveraging smooth proxy models to significantly reduce the gradient variance.

Remark 5.5. With a non-smooth model and policy, the loss landscapes can be highly non-smooth, which together with large gradient bias results in slow convergence and failure of training even in

simple toy examples (Parmas et al., 2018; Metz et al., 2021; Suh et al., 2022a). The results suggest that one can add smoothness regularization (e.g. spectral normalization (Miyato et al., 2018; Bjorck et al., 2021) or adversarial regularization (Shen et al., 2020)) to the model and policy to avoid exponentially increasing gradient variance and bias.

To restrain the bias brought by the smoothness regularization, we assume the system is controllable with Lipschitz continuous value functions.

Assumption 5.6 (Lipschitz Q-Value)**.** Assume the state-action value is L_Q Lipschitz continuous.

Besides, since policy actions can lead to changes in future states and rewards, unless we know the exact state-action value function which gives us accurate $\nabla_\theta Q^{\pi_\theta}$, 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 impose a regularity condition on the discrepancy between the initial distribution ζ and the state visitation ν_π.

Assumption 5.7 (Regularity Condition)**.** Assume there exists $\kappa > 0$ such that for any π_t, $t \in [1, T]$,

$$\mathbb{E}_{\nu_{\pi_t}}\left[\left(\frac{\mathrm{d}\zeta}{\mathrm{d}\nu_{\pi_t}}(s)\right)^2\right]^{1/2} \leq \kappa, \tag{5.2}$$

where $\mathrm{d}\zeta/\mathrm{d}\nu_{\pi_t}$ is the Radon-Nikodym derivative of ζ with respect to ν_{π_t}.

Consider the initial state distribution μ of the RP gradient estimator as a mixture of the MDP initial distribution ζ and the state visitation ν: $\mu(s) = \beta \cdot \nu(s) + (1-\beta) \cdot \zeta(s)$, where $\beta \in [0, 1]$.

Proposition 5.8 (Gradient Bias)**.** Denote $\kappa' := \beta + \kappa \cdot (1-\beta)$. Under Assumption 5.3, 5.6, and 5.7, for any $t \in [T]$, the gradient bias is bounded by

$$b_t \leq O\Big(\kappa'\kappa h^4 \widetilde{L}_{\widehat{f}}^{3h} \widetilde{L}_\pi^{2h} \epsilon_f + \kappa' h^2 \gamma^h \widetilde{L}_{\widehat{f}}^{2h} \widetilde{L}_\pi^h \epsilon_v\Big). \tag{5.3}$$

By setting $\beta = 1$ when $h = 0$ and $\beta = 0$ when $h = \infty$, the bound holds without Assumption 5.7.

Remark 5.9. This result suggests different state sampling schemes when updating policy: from ν_π when not using the model (e.g. SVG(0) (Heess et al., 2015; Amos et al., 2021) and DPG (Silver et al., 2014; Lillicrap et al., 2015)); from the initial distribution when unrolling the model over full sequences (e.g. BPTT (Kurutach et al., 2018; Curi et al., 2020; Xu et al., 2022)); from the mixture of ζ and ν_π when unrolling finite timesteps of the model (e.g. SVG and MAAC Clavera et al. (2020)).

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 5.10 (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 $W(x \cdot e^x) = x$.

Remark 5.11. 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.

Corollary 5.12 (Convergence Rate)**.** Denote $\varepsilon(T) = \sum_{t=0}^{T-1} b_t$. We have for $T \geq 4L^2$ that

$$\min_{t\in[T]} \mathbb{E}\Big[\big\|\nabla_\theta J(\pi_{\theta_t})\big\|_2^2\Big] \leq 16\delta \cdot \varepsilon(T)/\sqrt{T} + 4\varepsilon^2(T)/T + O\big(1/\sqrt{T}\big).$$

The convergence rate can be further specified by characterizing how fast the model and critic error goes to zero, i.e., $\sum_{t=0}^{T-1} \epsilon_f(t) + \epsilon_v(t)$. Such results can be shown by a more fine-grained investigation of the model, critic function class, e.g. adopting overparameterized neural nets with width scaling with T for bounding the prediction error and introducing a measure of the model complexity for bounding the derivative error ϵ_f, ϵ_v.

6 Related Work

Policy Gradient Methods. In Section **??** we discussed the likelihood ratio (LR) and reparameterization (RP) gradient estimators. In the context of RL, the LR estimator corresponds to the zeroth-order policy gradient that can be estimated only using samples of the function values. The LR gradient is the basis of most policy gradient algorithms, e.g. REINFORCE (Williams, 1992) and actor-critic methods (Sutton et al., 1999; Kakade, 2001; Kakade & Langford, 2002; Degris et al., 2012). Recent work (Agarwal et al., 2021; Wang et al., 2019a; Bhandari & Russo, 2019; Liu et al., 2019) has shown the global convergence of LR policy gradient under certain conditions, while less attention has been focused on RP PGMs. Remarkably, the analysis in (Li et al., 2021) is based on the strong assumptions on the *chained* gradient and ignores the effect of value approximation, which significantly simplifies the problem by reducing the h-step model value expansion to single-step model unrolls. Besides, Clavera et al. (2020) only focused on the gradient bias while still neglecting the necessary analysis needed by visitation distributions.

Differentiable Simulation. In this paper, we consider the model-based setting where a model fits the underlying transition of the MDP and is used to train a control policy. Recent approaches (Huang et al., 2021; Mora et al., 2021; Suh et al., 2022a; Xu et al., 2022) based on differentiable simulators (Freeman et al., 2021; Heiden et al., 2021) assume that gradients of simulation outcomes w.r.t. control actions are explicitly given. As a result, the length of simulator unrolls is typically larger compared to model-based approaches (Clavera et al., 2020; Amos et al., 2021). To deal with the non-smoothness and discontinuities in the differentiable simulation caused by contact dynamics and geometrical constraints, previous work proposed to use penalty-based contact formulation (Geilinger et al., 2020; Xu et al., 2021) or adopt bundled gradient with randomized smoothing (Suh et al., 2022b;a). However, these are complementary to our analysis based on function approximators.

7 Experiments

7.1 Different Instantiations of RP Policy Gradient Methods

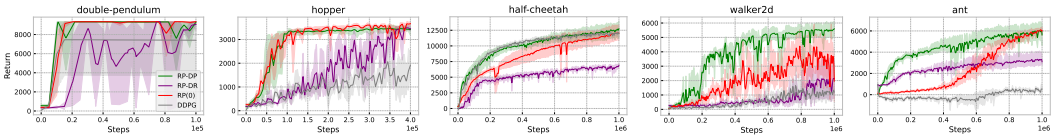


Figure 1: Evaluation of different instantiations of RP PGMs in several MuJoCo tasks.

We first evaluate several algorithms instantiated from the reparameterization policy gradient methods in two MuJoCo (Todorov et al., 2012) locomotion tasks, Hopper and Walker2d. For model-based RL PGMs, we test the two types of gradient estimators in Section 4.2, where we use RP-DP and RP-DR to distinguish whether the model derivatives are calculated on predictions (4.2) or real samples (4.5). Specifically, RP-DP is implemented as MAAC (Clavera et al., 2020) with entropy regularization, as suggested by Amos et al. (2021); and RP-DR is implemented as the on-policy SVG (Heess et al., 2015). For comparison, we also evaluate the performance of the model-free PGMs, including DDPG (Lillicrap et al., 2015) and RP(0) with the form in Section 3.1, which is equivalent to setting the model unrolling step $h = 0$ and can be seen as the stochastic counterpart of DDPG.

7.2 Gradient Variance and Loss Landscape

Our previous results show that vanilla model-based RP PGMs can have highly non-smooth landscapes due to the exponentially increasing gradient variance. We now conduct experiments to validate this phenomenon. In Fig. 2, we plot the mean gradient variance of the vanilla RP-DP algorithm (the solid lines) during training. To visualize the loss landscapes, we plot in Fig. 3 the negative value estimate along two directions that are randomly selected in the policy parameter space of a training policy.

We can observe that for vanilla RP policy gradient algorithms, the gradient variance explodes in exponential rate with respect to the unrolling step. As a result, the loss landscape for a large unrolling step is highly non-smooth compared to a small one. This renders the importance of the smoothness regularization: when the model and policy neural nets are equipped with spectral normalization (SN)

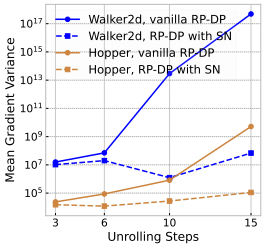
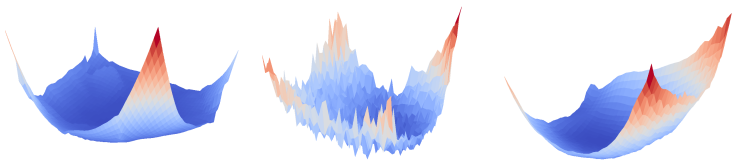


Figure 2: Gradient variance.



(a) Vanilla RP-DP, unrolling step $h = 3$. (b) Vanilla RP-DP, unrolling step $h = 15$. (c) RP-DP with SN, unrolling step $h = 15$.

Figure 3: 2D projection of the loss surface in the hopper environment.

(Miyato et al., 2018), the mean gradient variance is much lower for all settings of unrolling lengths, and the loss surface is smoother compared to its vanilla counterpart.

7.3 Benefit of Smoothness Regularization

In this part, we investigate the effect of smoothness regularization to support our theorem: the gradient variance scales exponentially with the Lipschitz constants of the function approximations and is a contributing factor to the quality of training. We defer the details of the setting to Appendix B. The results in Figure 4 show that adding SN will at least not sacrifice the performance. For long model unrolls (e.g. 10 in walker2d and 15 in hopper), vanilla RP PGMs fail to reach reliable performance, while SN significantly boosts training.

Figure 4: Performance of the RP policy gradient methods with and without spectral normalization.

By plotting the gradient variance of RP-DP during training in Figure 5, we observe that the failure of vanilla RP-DP for walker $h = 10$ and hopper $h = 15$ is mainly due to the exponentially exploding gradient variance. On the contrary, applying SN to the model and the policy leads to better training performance as a result of the drastically reduced variance.

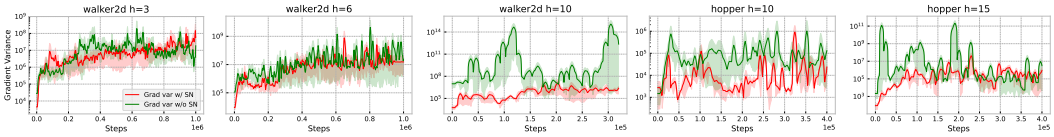


Figure 5: Gradient variance of RP policy gradient methods with and without spectral normalization.

8 Conclusion

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.

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.

C Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax–a differentiable physics engine for large scale rigid body simulation. *arXiv preprint arXiv:2106.13281*, 2021.

Moritz Geilinger, David Hahn, Jonas Zehnder, Moritz Bächer, Bernhard Thomaszewski, and Stelian Coros. Add: Analytically differentiable dynamics for multi-body systems with frictional contact. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

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.

Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S Sukhatme. Neuralsim: Augmenting differentiable simulators with neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9474–9481. IEEE, 2021.

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.

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

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.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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.

Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. Deep reinforcement learning with robust and smooth policy. In *International Conference on Machine Learning*, pp. 8707–8718. PMLR, 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*, 2022a.

Hyung Ju Terry Suh, Tao Pang, and Russ Tedrake. Bundled gradients through contact via randomized smoothing. *IEEE Robotics and Automation Letters*, 7(2):4000–4007, 2022b.

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.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

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, Tao Chen, Lara Zlokapa, Michael Foshey, Wojciech Matusik, Shinjiro Sueda, and Pulkit Agrawal. An end-to-end differentiable framework for contact-aware robot design. *arXiv preprint arXiv:2107.07501*, 2021.

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.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020.

A Assumption Clarification

A.1 Smooth Objective Assumption

We assume the objective is Lipschitz smooth in Assumption 5.1. This assumption can be equivalently stated in the following as a lower-level alternative.

Assumption A.1 (Lipschitz Smooth Objective)**.** Assume the absolute value of the reward $r(s,a)$ is bounded by $|r(s,a)| \leq r_{\max}$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$. Besides, assume that the score function of policy π_θ is Lipschitz continuous and has bounded norm for $(s,a) \in \mathcal{S} \times \mathcal{A}$, specifically,

$$\left\| \log \pi_{\theta_1}(a|s) - \log \pi_{\theta_2}(a|s) \right\|_2 \leq L_1 \cdot \|\theta_1 - \theta_2\|, \quad \left\| \log \pi_\theta(a|s) \right\|_2 \leq B_\theta.$$

Then $J(\pi_\theta)$ is L-smooth in θ, such that $\|\nabla_\theta J(\pi_{\theta_1}) - \nabla_\theta J(\pi_{\theta_2})\|_2 \leq L \left\|\theta_1 - \theta_2\right\|_2$, where

$$L = \frac{r_{\max} \cdot L_1}{(1-\gamma)^2} + \frac{(1+\gamma) \cdot r_{\max} \cdot B_\theta^2}{(1-\gamma)^3}.$$

We refer to Lemma 3.2 in Zhang et al. (2020) for a detailed proof.

A.2 Lipschitz Function Assumption

The full statement of the Lipschitz Assumption 5.3 is as follows.

Assumption A.2 (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 (c.f. Appendix A.2 for details). We assume that $r(s,a)$, $\widehat{f}_\psi(s,a,\xi)$, $\pi_\theta(s,\varsigma)$ $\widehat{Q}_\omega(s,a)$ are Lipschitz continuous (c.f. Appendix A.2 for details). such that

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

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

B Experiment Details

B.1 Experimental Settings and Spectral Normalization

We first provide the necessary background of spectral normalization to understand how it works and why we prefer it. By definition, the Lipschitz constant L_g of a function g satisfies $L_g = \sup_x \sigma_{\max}(\nabla g(x))$, where $\sigma_{\max}(W)$ denotes the largest singular value of the matrix W, defined as $\sigma_{\max}(W) := \max_{\|x\|_2 \leq 1} \|Wx\|_2$. Therefore, for neural network f with linear layers $g(x) = W_i x$ and 1-Lipschitz activation (e.g. ReLU and leaky ReLU), we have $L_g = \sigma_{\max}(W_i)$ and $L_{\widehat{f}} \leq \prod_i \sigma_{\max}(W_i)$. By normalizing the spectral norm of the W_i with $W_i^{\text{SN}} := W_i / \sigma_{\max}(W_i)$, SN guarantees that the Lipschitz of f is bounded by 1. For this reason, we adopt SN as the smoothness regularization.

In experiments, we use multilayer perceptrons (MLPs) for the critic, policy, and the model. To test the benefit of smooth function approximations for the RP-DP algorithm, the spectral normalization is applied to all layers of the policy MLP and all except the final layers of the dynamics model MLP. The number of layers for the policy and the dynamics model is 4 and 5, respectively.

Our code is based on PyTorch (Paszke et al., 2019), which has out-of-the-shelf implementation of spectral normalization. Thus, applying SN to the MLP is pretty simple and no additional lines of code is needed. Specifically, we only need to import and add SN in each layer:

```
from torch.nn.utils.parametrizations import spectral_norm
layer = [spectral_norm(nn.Linear(in_dim, hidden_dim)), nn.ReLU()]
```

B.2 Different Model Learners

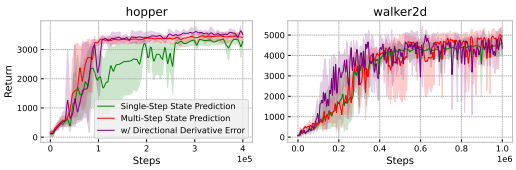


Figure 6: Comparison between model learners.

Our main results depend on the model error defined in 4.7, which, however, is not a training objective. In Figure 6, we evaluate different model learners: single- and multi-step (h-step) state predictive models and combined with directional derivative error Li et al. (2021). We observe that enlarging the prediction steps benefits training. The algorithm also converges faster in walker2d when considering derivative error, which approximately minimizes 4.7 and supports our analysis. However, calculating the directional derivative error by searching k nearest points in the buffer significantly increases the computational cost, for which reason we use h-step state predictive models as default in experiments.

C Proofs

C.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 the Lipschitz Assumption 5.3, we have

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

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

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

In order to lower-bound $\nabla_\theta J(\pi_{\theta_t})^\top \widehat{\nabla}_\theta J(\pi_{\theta_t})$, we turn to bound the resulting three terms:

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

Thus, we have the following inequality for (C.1):

$$\begin{aligned} J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) \geq \frac{\eta}{2} \cdot \Bigg(&-\Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\Big\|_2 \cdot 2b_t - \Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}[\widehat{\nabla}_\theta J(\pi_{\theta_t})]\Big\|_2^2 + \Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\Big\|_2^2\Bigg) \\ &- \frac{L\eta^2}{2} \cdot \Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\Big\|_2^2. \end{aligned} \tag{C.2}$$

By taking expectation in (C.2), we obtain

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

Rearranging terms gives us

$$\frac{\eta - L\eta^2}{2} \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\big\|_2^2\Big] \leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta\mathbb{E}[\|\widehat{\nabla}_\theta J(\pi_{\theta_t})\|_2]b_t + \frac{\eta}{2}v_t. \tag{C.3}$$

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

$$\begin{aligned}
\mathbb{E}\Big[\big\|\nabla_\theta J(\pi_{\theta_t}) - \widehat{\nabla}_\theta J(\pi_{\theta_t})\big\|_2^2\Big] &= \mathbb{E}\Big[\Big\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] + \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] - \widehat{\nabla}_\theta J(\pi_{\theta_t})\Big\|_2^2\Big] \\
&\leq 2\Big\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2^2 + 2\mathbb{E}\Big[\Big\|\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2^2\Big] \\
&= 2b_t^2 + 2v_t,
\end{aligned} \tag{C.4}$$

where the second inequality holds since for any vector $y, z \in \mathbb{R}^d$,

$$\|y + z\|_2^2 \leq \|y\|_2^2 + \|z\|_2^2 + 2\|y\|_2 \cdot \|z\|_2 \leq 2\|y\|_2^2 + 2\|z\|_2^2. \tag{C.5}$$

Then we are ready to bound the minimum expected gradient norm by relating it to the average norm over T iterations. Specifically,

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

where the second inequality follows from (C.5).

For $T \geq 4L^2$, by setting $\eta = 1/\sqrt{T}$, we have $\eta < 1/L$ and $(\eta - L\eta^2)/2 > 0$. Therefore, following the results in (C.3) and (C.4), we further have

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

where the last step holds due to the definition $c := \big(\eta - L\eta^2\big)^{-1}$.

By noting that $\eta\widehat{\nabla}_\theta J(\pi_{\theta_t}) = \theta_{t+1} - \theta_t$, we conclude that

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

where the second inequality holds since $\|\theta\|_2 \leq \delta$ for any $\theta \in \Theta$. □

C.2 Proof of Proposition 5.4

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

$$\big\|\mathbb{E}[y]\big\|_2 = \Big\|\sum_{k=1}^{\infty} y_k \cdot p(y_k)\Big\|_2 \leq \sum_{k=1}^{\infty} p(y_k) \cdot \|y_k\|_2 = \mathbb{E}\big[\|y\|_2\big]. \tag{C.6}$$

In order to upper-bound the gradient variance $v_t = \mathbb{E}[\|\widehat{\nabla}_\theta J(\pi_{\theta_t}) - \mathbb{E}[\widehat{\nabla}_\theta J(\pi_{\theta_t})]\|_2^2]$, we turn to find the supremum of the norm inside the outer expectation, which serves as a loose yet acceptable variance upper bound.

For now, we analyze the case when the sample size $N = 1$, which naturally generalizes to $N > 1$. Specifically, consider an *arbitrary* trajectory obtained by unrolling the model under policy π_{θ_t}. Denote the pathwise gradient of this trajectory as g'. Then we have

$$v_t \le \max_{g'} \left\| g' - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \right\|_2^2 = \left\| g - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \right\|_2^2,$$

where we let g denote the pathwise gradient of some *fixed* (but unknown) trajectory $(\widehat{s}_{0,n}, \widehat{a}_{0,n}, \widehat{s}_{1,n}, \widehat{a}_{1,n}, \cdots)$ such that the maximum is achieved.

Following (C.6), we further have

$$v_t \le \left\| \mathbb{E}\big[g - \widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \right\|_2^2 \le \mathbb{E}\Big[\big\| g - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \big\|_2 \Big]^2. \tag{C.7}$$

In what follows, the proof is established for the two gradient estimators simultaneously, 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 $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n}, \xi_n)$.

Denote $\widehat{x}_{i,n} := (\widehat{s}_{i,n}, \widehat{a}_{i,n})$. Then we have

$$\begin{aligned} \mathbb{E}\Big[\big\| g - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \big\|_2 \Big] \le \sum_{i=0}^{h-1} \gamma^i \cdot \mathbb{E}_{\overline{x}_i} \Big[\big\| \nabla_\theta r(\widehat{x}_{i,n}) - \nabla_\theta r(\overline{x}_i) \big\|_2 \Big] \\ + \gamma^h \cdot \mathbb{E}_{\overline{x}_h} \Big[\big\| \nabla \widehat{Q}(\widehat{x}_{h,n}) \nabla_\theta \widehat{x}_{h,n} - \nabla \widehat{Q}(\overline{x}_h) \nabla_\theta \overline{x}_h \big\|_2 \Big]. \end{aligned} \tag{C.8}$$

For $i \ge 1$, we have the following relationship according to the chain rule:

$$\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} = \frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{a}_{i,n}}{\partial \theta}. \tag{C.9}$$

By the Lipschitz Assumption 5.3, we have

$$\begin{aligned} \left\| \frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} \right\|_2 &= \left\| \frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{a}_{i,n}}{\partial \theta} \right\|_2 \\ &\le L_{\widehat{f}} L_\pi \cdot \left\| \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} \right\|_2 + L_\theta. \end{aligned}$$

Applying the above recursion gives us

$$\left\| \frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} \right\|_2 \le L_{\widehat{f}}^i L_\pi^i \cdot L_\theta + L_\theta \cdot \sum_{j=0}^{i-1} L_{\widehat{f}}^j L_\pi^j \le L_\theta \cdot (i+1) \cdot \widetilde{L}_{\widehat{f}}^i \widetilde{L}_\pi^i, \tag{C.10}$$

where the first inequality holds due to the fact that for a sequence $z_0, \cdots, z_i \in \mathbb{R}$ with linear transformation $z_i = a z_{i-1} + b$, it holds that

$$z_i = a z_{i-1} + b = a \cdot (a z_{i-2} + b) + b = a^i \cdot z_0 + b \cdot \sum_{j=0}^{i-1} a^j. \tag{C.11}$$

Similarly, we can bound $\|\nabla_\theta \widehat{s}_{i,n}\|_2$ iteratively as follows:

$$\begin{aligned} \left\| \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} \right\|_2 &= \left\| \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{s}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} \right\|_2 \\ &\le L_{\widehat{f}} \cdot \left\| \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} \right\|_2 + L_\theta \cdot i \cdot \widetilde{L}_{\widehat{f}}^i \widetilde{L}_\pi^i \\ &= L_\theta \cdot \widetilde{L}_{\widehat{f}}^i \cdot \widetilde{L}_\pi + L_\theta \cdot i \cdot \widetilde{L}_{\widehat{f}}^i \widetilde{L}_\pi^i \cdot \sum_{j=0}^{i-1} L_{\widehat{f}}^j \\ &\le L_\theta \cdot (i^2 + 1) \cdot \widetilde{L}_{\widehat{f}}^{2i} \widetilde{L}_\pi^i. \end{aligned} \tag{C.12}$$

Combining (C.10) and (C.12) we obtain

$$\left\|\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}\right\|_2 \leq \widehat{K}(i) := L_\theta \cdot (i+1) \cdot \widetilde{L}^i_{\widehat{f}} \widetilde{L}^i_\pi + L_\theta \cdot (i^2+1) \cdot \widetilde{L}^{2i}_{\widehat{f}} \widetilde{L}^i_\pi, \tag{C.13}$$

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

Therefore, the second term of (C.8) can be decomposed and bounded by

$$\begin{aligned}
&\mathbb{E}_{\overline{x}_h}\left[\left\|\nabla\widehat{Q}(\widehat{x}_{h,n})\nabla_\theta\widehat{x}_{h,n} - \nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\overline{x}_h\right\|_2\right] \\
&\quad\leq \mathbb{E}_{\overline{x}_h}\left[\left\|\nabla\widehat{Q}(\widehat{x}_{h,n})\nabla_\theta\widehat{x}_{h,n} - \nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\widehat{x}_{h,n}\right\|_2\right] + \mathbb{E}_{\overline{x}_h}\left[\left\|\nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\widehat{x}_{h,n} - \nabla\widehat{Q}(\overline{x}_h)\nabla_\theta\overline{x}_h\right\|_2\right] \\
&\quad\leq 2L_{\widehat{Q}} \cdot \widehat{K}(i) + L_{\widehat{Q}} \cdot \left(\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right]\right),
\end{aligned} \tag{C.14}$$

where the last step follows from the Cauchy–Schwarz inequality and the Lipschitz critic assumption.

By the chain rule, we have a similar result for the first term of (C.8) as follows:

$$\begin{aligned}
&\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla_\theta r(\widehat{x}_{i,n}) - \nabla_\theta r(\overline{x}_i)\right\|_2\right] \\
&\quad= \mathbb{E}_{\overline{x}_i}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\widehat{x}_{i,n} - \nabla r(\overline{x}_i)\nabla_\theta\overline{x}_i\right\|_2\right] \\
&\quad\leq \mathbb{E}_{\overline{x}_i}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\widehat{x}_{i,n} - \nabla r(\widehat{x}_{i,n})\nabla_\theta\overline{x}_i\right\|_2\right] + \mathbb{E}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\overline{x}_i - \nabla r(\overline{x}_i)\nabla_\theta\overline{x}_i\right\|_2\right] \\
&\quad\leq L_r \cdot \left(\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right]\right) + 2L_r \cdot \widehat{K}(i).
\end{aligned} \tag{C.15}$$

Plugging (C.14), (C.15) into (C.8) and (C.7) gives us

$$\begin{aligned}
v_t &\leq \left((h \cdot L_r + \gamma^h \cdot L_{\widehat{Q}}) \cdot \left(\mathbb{E}_{\overline{s}_h}\left[\left\|\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{s}_h}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}_{\overline{a}_h}\left[\left\|\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_h}{\mathrm{d}\theta}\right\|_2\right] + 2\widehat{K}(h)\right)\right)^2 \\
&\leq O\left(h^8 \widetilde{L}^{6h}_{\widehat{f}} \widetilde{L}^{4h}_\pi + \gamma^{2h} h^6 \widetilde{L}^{6h}_{\widehat{f}} \widetilde{L}^{4h}_\pi\right).
\end{aligned} \tag{C.16}$$

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

Lemma C.1. Denote $e := \sup \mathbb{E}_{\overline{a}_0}[\|\mathrm{d}\widehat{a}_{0,n}/\mathrm{d}\theta - \mathrm{d}\overline{a}_0/\mathrm{d}\theta\|_2]$, which is a constant that only depends on the initial state distribution[1]. For any $i \geq 1$ and the corresponding $\widehat{s}_{i,n}$, we have the following inequality results:

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

Proof. Firstly, we obtain from (C.9) that $\forall i \geq 1$,

$$\begin{aligned}
&\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad= \mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}} \cdot \frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} + \frac{\partial\widehat{a}_{i,n}}{\partial\theta} - \frac{\partial\overline{a}_i}{\partial\overline{s}_i} \cdot \frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}} \cdot \frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta} - \frac{\partial\overline{a}_i}{\partial\theta}\right\|_2\right] \\
&\quad\leq \mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}} \cdot \frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial\overline{a}_i}{\partial\overline{s}_i} \cdot \frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right] \\
&\qquad+ \mathbb{E}\left[\left\|\frac{\partial\overline{a}_i}{\partial\overline{s}_i} \cdot \frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial\overline{a}_i}{\partial\overline{s}_i} \cdot \frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}} \cdot \frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\theta} - \frac{\partial\overline{a}_i}{\partial\theta}\right\|_2\right].
\end{aligned}$$

[1] We define e to account for the stochasticity in the initial state distribution. $e = 0$ when the initial state is deterministic.

After decomposing the intricate terms to simpler ones, we obtain the following recursive expression with the Lipschitz assumption:

$$
\begin{aligned}
&\leq \left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2 \cdot \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}}\right\|_2 + \left\|\frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} - \frac{\partial \overline{a}_i}{\partial \overline{s}_i} \cdot \frac{\partial \overline{s}_i}{\partial \overline{a}_{i-1}}\right\|_2\right] \\
&\quad + L_\pi L_{\widehat{f}} \cdot \mathbb{E}_{\overline{a}_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\right\|_2\right] + 2L_\theta \\
&\leq 4L_{\widehat{f}} L_\pi \cdot \left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2 + L_\pi L_{\widehat{f}} \cdot \mathbb{E}_{\overline{a}_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\right\|_2\right] + 2L_\theta. \qquad \text{(C.17)}
\end{aligned}
$$

By iterating over the above recursion, we have

$$
\begin{aligned}
\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right] &\leq e \cdot L_\pi^i L_{\widehat{f}}^i + \left(4L_{\widehat{f}} L_\pi \cdot \left(L_\theta \cdot i \cdot \widetilde{L}_{\widehat{f}}^i \widetilde{L}_\pi^i\right) + 2L_\theta\right) \cdot \sum_{j=0}^{i-1} L_\pi^j L_{\widehat{f}}^j \\
&\leq \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i \cdot \left(e + 2L_\theta \cdot i + 4L_\theta \cdot i^2 \cdot \widetilde{L}_\pi^i \widetilde{L}_{\widehat{f}}^i\right), \qquad \text{(C.18)}
\end{aligned}
$$

where the first inequality follows from (C.11) and applying the bound of $\|\mathrm{d}\widehat{a}_{i,n}/\mathrm{d}\theta\|_2$ given in (C.10).

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

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

By plugging (C.18) and the definition (C.13) of $\widehat{K}(i)$, we continue with

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

where the last equality follows from (C.11).

□

C.3 Proof of Proposition 5.8

Proof. Different from the gradient variance where the proof is solely based on the distribution generated by function approximators, additional care must be taken when dealing with the gradient bias where the ground-truth distribution also appears.

The chain rule gives us the recursive expression of $\nabla_\theta s$ and $\nabla_\theta \widehat{s}$, based on which we have for any $(\widehat{s}_{i,n}, \widehat{a}_{i,n}) \sim \mathbb{P}(\widehat{s}_i, \widehat{a}_i)$ that

$$
\begin{aligned}
&\mathbb{E}_{(s_i,a_i)\sim\mathbb{P}(s_i,a_i)}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\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{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{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{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}-\frac{\partial a_i}{\partial\theta}\right\|_2\right] \\
&\quad\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{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right] \\
&\qquad+\mathbb{E}\left[\left\|\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial a_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}a_{i-1}}{\mathrm{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] \\
&\quad\le\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{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] \\
&\qquad+L_\pi L_{\widehat{f}}\cdot\mathbb{E}_{a_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right]+2L_\theta \\
&\quad\le L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2+L_\pi L_{\widehat{f}}\cdot\mathbb{E}_{a_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right]+2L_\theta,
\end{aligned}
\tag{C.19}
$$

where recall that $\mathbb{P}(s_i, a_i)$ and $\mathbb{P}(\widehat{s}_i, \widehat{a}_i)$ are defined in (4.7) with respect to $s_0 \sim \nu$, $\widehat{s}_0 \sim \nu$.

Since we proved in (C.10) that $\|\mathrm{d}\widehat{a}_{i-1,n}/\mathrm{d}\theta\|_2 \le L_\theta \cdot i \cdot \widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi$, we have

$$
\begin{aligned}
\mathbb{E}_{(s_i,a_i)\sim\mathbb{P}(s_i,a_i)}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\right\|_2\right] &\le\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big)\cdot\sum_{j=0}^{i-1}L^j_\pi L^j_{\widehat{f}} \\
&\le i\cdot\widetilde{L}^i_\pi\widetilde{L}^i_{\widehat{f}}\cdot\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big),
\end{aligned}
\tag{C.20}
$$

where the first step follows from (C.11) and the fact that $\mathbb{E}_{a_i}[\|\mathrm{d}\widehat{a}_{i,n}/\mathrm{d}\theta - \mathrm{d}a_i/\mathrm{d}\theta\|_2] = 0$ since the initial states are sampled from the same distribution ζ.

Similarly, we have

$$
\begin{aligned}
&\mathbb{E}_{(s_i,a_i)\sim\mathbb{P}(s_i,a_i)}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}+\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}-\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial s_i}{\partial s_{i-1}}\cdot\frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\qquad+\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial s_i}{\partial a_{i-1}}\cdot\frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad\le\epsilon_f\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right]+L_f\cdot\mathbb{E}_{s_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\qquad+L_f\cdot\mathbb{E}_{a_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right]
\end{aligned}
$$

where the inequality follows from the definition of ϵ_f in (4.7). Following (C.20) and the definition of $\widehat{K}$ in (C.13), we obtain

$$
\begin{aligned}
&\le\epsilon_f\cdot\widehat{K}(i-1)+i\cdot\widetilde{L}^i_\pi\widetilde{L}^i_{\widehat{f}}\cdot\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big) \\
&\qquad+L_f\cdot\mathbb{E}_{s_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\le\Big(\epsilon_f\cdot\widehat{K}(i-1)+i\cdot\widetilde{L}^i_\pi\widetilde{L}^i_{\widehat{f}}\cdot\Big(L_\pi(2L_{\widehat{f}}+\epsilon_f)\cdot\big(L_\theta\cdot i\cdot\widetilde{L}^i_{\widehat{f}}\widetilde{L}^i_\pi\big)+2L_\theta\Big)\Big)\cdot i\cdot\widetilde{L}^i_{\widehat{f}},
\end{aligned}
\tag{C.21}
$$

where the inequality holds due to (C.11).

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

$$\begin{aligned}
&\mathbb{E}_{(s_i,a_i)\sim\mathbb{P}(s_i,a_i),(\widehat{s}_i,\widehat{a}_i)\sim\mathbb{P}(\widehat{s}_i,\widehat{a}_i)}\left[\left\|\frac{\mathrm{d}r(\widehat{x}_{i,n})}{\mathrm{d}\theta}-\frac{\mathrm{d}r(x_i)}{\mathrm{d}\theta}\right\|_2\right]\\
&=\mathbb{E}\left[\left\|\frac{\mathrm{d}r}{\mathrm{d}\widehat{x}_{i,n}}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}x_i}{\mathrm{d}\theta}\right\|_2\right]\\
&\le\mathbb{E}\left[\left\|\frac{\mathrm{d}r}{\mathrm{d}\widehat{x}_{i,n}}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}x_i}{\mathrm{d}\theta}\right\|_2\right]\\
&\le 2L_r\cdot\widehat{K}(i)+L_r\cdot\left(\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\right\|_2\right]\right).
\end{aligned}\tag{C.22}$$

We define $\overline{\sigma}_1(s,a):=\mathbb{P}(s_h=s,a_h=a)$ where $s_0\sim\nu$, $a_i\sim\pi(\cdot\,|\,s_i)$, and $s_{i+1}\sim f(\cdot\,|\,s_i,a_i)$. In a similar way, we define $\widehat{\sigma}_1(s,a):=\mathbb{P}(\widehat{s}_h=s,\widehat{a}_h=a)$ where $\widehat{s}_0\sim\nu$, $\widehat{a}_i\sim\pi(\cdot\,|\,\widehat{s}_i)$, and $\widehat{s}_{i+1}\sim\widehat{f}(\cdot\,|\,\widehat{s}_i,\widehat{a}_i)$.

Now we are ready to bound the gradient bias. In Lemma C.2 and C.3, we deal with the misalignment between $\nabla_\theta V^{\pi_\theta}$ and $\nabla_\theta\widehat{V}_t$, specifically, the recursive structure of V^{π_θ} and the non-recursive value function approximation $\widehat{V}_{\omega_t}$. From Lemma C.3, we have

$$\begin{aligned}
b_t\le{}&\kappa\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{s_0\sim\nu,\widehat{s}_{0,n}\sim\nu}\left[\left\|\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right\|_2\right]\\
&+\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1,(\widehat{s}_{h,n},\widehat{a}_{h,n})\sim\widehat{\sigma}_1}\left[\left\|\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\gamma^h\cdot\nabla_\theta\widehat{Q}_t(\widehat{s}_{h,n},\widehat{a}_{h,n})\right\|_2\right].
\end{aligned}$$

The bias brought by the critic, i.e. the last term, can be further bounded by

$$\begin{aligned}
&\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1,(\widehat{s}_{h,n},\widehat{a}_{h,n})\sim\widehat{\sigma}_1}\left[\left\|\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\gamma^h\cdot\nabla_\theta\widehat{Q}_t(\widehat{s}_{h,n},\widehat{a}_{h,n})\right\|_2\right]\\
&=\gamma^h\cdot\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\frac{\partial\widehat{Q}_t}{\partial\widehat{a}_{h,n}}\cdot\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}-\frac{\partial\widehat{Q}_t}{\partial\widehat{s}_{h,n}}\cdot\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}\right\|_2\right]\\
&\le\gamma^h\cdot\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}-\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}-\frac{\partial\widehat{Q}_t}{\partial\widehat{a}_{h,n}}\cdot\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}\right\|_2\right.\\
&\quad\left.+\left\|\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}-\frac{\partial\widehat{Q}_t}{\partial\widehat{s}_{h,n}}\cdot\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}\right\|_2\right]\\
&\le\gamma^h\cdot L_Q\cdot\left(\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\mathrm{d}a_h}{\mathrm{d}\theta}-\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}\right\|_2\right]\right)+\gamma^h\cdot\widehat{K}(h)\cdot\epsilon_v,
\end{aligned}\tag{C.23}$$

where the last inequality follows from (C.13) and the definition of ϵ_v in (4.8).

Using the results in (C.22) and (C.23), we obtain

$$\begin{aligned}
b_t\le{}&\kappa\big(\beta+\kappa\cdot(1-\beta)\big)\cdot h\cdot\left(L_r\cdot\left(\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_h}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_h}{\mathrm{d}\theta}\right\|_2\right]\right)+2L_r\cdot\widehat{K}(h)\right)\\
&+\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\gamma^h\cdot\left(L_Q\cdot\left(\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_h}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\left[\left\|\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_h}{\mathrm{d}\theta}\right\|_2\right]\right)+\widehat{K}(h)\cdot\epsilon_v\right),
\end{aligned}\tag{C.24}$$

By plugging (C.20), (C.21) and $\widehat{K}$ in (C.13) into the above expression, we obtain

$$b_t\le O\Big(\kappa\big(\beta+\kappa\cdot(1-\beta)\big)h^4\widetilde{L}_{\widehat{f}}^{3h}\widetilde{L}_\pi^{2h}\epsilon_f+\big(\beta+\kappa\cdot(1-\beta)\big)h^2\gamma^h\widetilde{L}_{\widehat{f}}^{2h}\widetilde{L}_\pi^h\epsilon_v\Big).$$

□

Lemma C.2. Under Assumption 5.7, the expected value gradient over state distribution at timestep h can be represented by

$$\mathbb{E}_{s_h\sim\mathbb{P}(s_h)}\left[\nabla_\theta V^{\pi_\theta}(s_h)\right] = \mathbb{E}_{(s,a)\sim\overline{\sigma}_1}\left[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{\mathrm{d}a}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{\mathrm{d}s}{\mathrm{d}\theta}\right],$$

where $\mathbb{P}(s_h)$ is the state distribution at timestep h where $s_0\sim\zeta$, $a_i\sim\pi(\cdot\,|\,s_i)$, and $s_{i+1}\sim f(\cdot\,|\,s_i,a_i)$.

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

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

where the first follows from Bellman equation and the remaining equations hold due to the chain rule.

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

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

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

Iterating the above formula we obtain

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

Define $\overline{\sigma}_2(s,a)=\sum_{i=h}^{\infty}\gamma^{i-h}\cdot\mathbb{P}(s_i=s,a_i=a)$. By the definition of the state-action visitation $\sigma(s,a)$, we have

$$\sigma(s,a)=\gamma^h\cdot\overline{\sigma}_1(s,a)+\sum_{i=0}^{h-1}\gamma^i\cdot\mathbb{P}(s_i=s,a_i=a), \tag{C.25}$$

$$\sigma(s,a)+\gamma^h\cdot\overline{\sigma}_2(s,a)=\gamma^h\cdot\overline{\sigma}_1(s,a)+\sigma(s,a). \tag{C.26}$$

Therefore, we conclude that $\overline{\sigma}_1(s,a) = \overline{\sigma}_2(s,a)$.

Taking the expectation over s_h we have

$$\begin{aligned}\mathbb{E}_{s_h\sim\mathbb{P}(s_h)}\big[\nabla_\theta V^{\pi_\theta}(s_h)\big] &= \mathbb{E}_{(s,a)\sim\overline{\sigma}_2}\left[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{\mathrm{d}a}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{\mathrm{d}s}{\mathrm{d}\theta}\right]\\ &= \mathbb{E}_{(s,a)\sim\overline{\sigma}_1}\left[\frac{\partial Q^{\pi_\theta}}{\partial a}\cdot\frac{\mathrm{d}a}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s}\cdot\frac{\mathrm{d}s}{\mathrm{d}\theta}\right]. \qquad \text{(C.27)}\end{aligned}$$

□

Lemma C.3. Recall that $\mu(s) = \beta\cdot\nu(s) + (1-\beta)\cdot\zeta(s)$. The gradient of the h-step Model Value Expansion satisfies

$$\begin{aligned}b_t \le{}& \kappa\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{s_0\sim\nu,\widehat{s}_{0,n}\sim\nu}\left[\left\|\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right\|_2\right]\\ &+\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1,(\widehat{s}_{h,n},\widehat{a}_{h,n})\sim\widehat{\sigma}_1}\left[\left\|\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\gamma^h\cdot\nabla_\theta\widehat{Q}_t(\widehat{s}_{h,n},\widehat{a}_{h,n})\right\|_2\right].\end{aligned}$$

Proof. To begin with, we upper bound the gradient bias by

$$\begin{aligned}b_t &= \Big\|\nabla_\theta J(\pi_{\theta_t})-\mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2\\ &= \Big\|\mathbb{E}\big[\nabla_\theta J(\pi_{\theta_t})-\widehat{\nabla}_\theta J(\pi_{\theta_t})\big]\Big\|_2\\ &= \left\|\mathbb{E}_{s_0\sim\zeta,\widehat{s}_{0,n}\sim\mu}\left[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)+\gamma^h\cdot\nabla_\theta V^{\pi_\theta}(s_h)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})-\gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\\ &\le \left\|\mathbb{E}_{s_0\sim\zeta,\widehat{s}_{0,n}\sim\mu}\left[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right\|_2\right.\\ &\quad\left.+\left\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1,\widehat{s}_{0,n}\sim\mu}\left[\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\right.\\ &\le \mathbb{E}_{\widehat{s}_{0,n}\sim\mu}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right]\\ &\quad+\mathbb{E}_{\widehat{s}_{0,n}\sim\mu}\left[\left\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\left[\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta}+\frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\right],\end{aligned}$$

where the expectation is taken over the randomness in the policy and model dynamics. We plug in the result in Lemma C.2 in the first inequality and the second inequality follows from (C.6).

We separately upper-bound the two terms on the R.H.S. of the above inequality as follows:

$$\begin{aligned}&\mathbb{E}_{\widehat{s}_{0,n}\sim\mu}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right]\\ &\quad=\beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right]\\ &\qquad+(1-\beta)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\zeta}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i)-\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right].\end{aligned}$$

Following the regularity condition in Assumption 5.7, we have

$$
\begin{aligned}
&\leq \beta \cdot \mathbb{E}_{\widehat{s}_{0,n}\sim\nu}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(s_i,a_i) - \nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right] \\
&\quad + (1-\beta)\cdot \mathbb{E}_{\widehat{s}_{0,n}\sim\nu}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(s_i,a_i) - \nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right] \cdot \left\{\mathbb{E}_\nu\left[\left(\frac{\mathrm{d}\zeta}{\mathrm{d}\nu}(s)\right)^2\right]\right\}^{1/2} \\
&\leq \big(\beta + \kappa\cdot(1-\beta)\big)\cdot \mathbb{E}_{\widehat{s}_{0,n}\sim\nu}\left[\left\|\mathbb{E}_{s_0\sim\zeta}\left[\nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(s_i,a_i) - \nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right]\right\|_2\right] \\
&\leq \kappa\big(\beta + \kappa\cdot(1-\beta)\big)\cdot \mathbb{E}_{s_0\sim\nu,\widehat{s}_{0,n}\sim\nu}\left[\left\|\nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(s_i,a_i) - \nabla_\theta \sum_{i=0}^{h-1}\gamma^i \cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\right\|_2\right].
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{E}_{\widehat{s}_{0,n}\sim\mu}\left[\left\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\left[\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\right] \\
&= \beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu}\left[\left\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\left[\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\right] \\
&\quad + (1-\beta)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\zeta}\left[\left\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\left[\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\right] \\
&\leq \big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu}\left[\left\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\left[\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\right]\right\|_2\right] \\
&= \big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1,(\widehat{s}_{h,n},\widehat{a}_{h,n})\sim\widehat{\sigma}_1}\left[\left\|\gamma^h\cdot\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h\cdot\nabla_\theta\widehat{Q}_t(\widehat{s}_{h,n},\widehat{a}_{h,n})\right\|_2\right].
\end{aligned}
$$

We get the claimed result by combining the above inequalities together. □

C.4 Proof of Proposition 5.10

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

$$
g(h) := c\cdot(\delta\cdot b_t + \frac{\eta}{2}\cdot v_t) + b_t^2 + v_t, \tag{C.28}
$$

where the bound of the gradient bias b_t and gradient variance v_t are given in (C.24) and (C.16), respectively.

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

$$
\frac{\partial}{\partial h}g(h) = O\big((h^7 + h^6\gamma^{2h}\cdot\log\gamma)/N + h^7\cdot\epsilon_f^2 + h^4\gamma^{2h}\log\gamma\cdot\epsilon_v^2\big) = 0. \tag{C.29}
$$

For notation simplicity, we define

$$
\frac{\partial}{\partial h}g_1(h) := (h^7 + h^6\gamma^{2h}\cdot\log\gamma)/N,\ \ \frac{\partial}{\partial h}g_2(h) := h^7\cdot\epsilon_f^2 + h^4\gamma^{2h}\log\gamma\cdot\epsilon_v^2. \tag{C.30}
$$

By solving $\frac{\partial}{\partial h}g_1(h) + \frac{\partial}{\partial h}g_2(h) = 0$, we can represent the optimal h^* 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

$$
h^3\cdot\epsilon_f^2 = \gamma^{2h}\log\frac{1}{\gamma}\cdot\epsilon_v^2. \tag{C.31}
$$

Taking the natural logarithm on both sides gives us

$$
3\log h + 2\log\epsilon_f = 2h\cdot\log\gamma + \log\log\frac{1}{\gamma} + 2\log\epsilon_v. \tag{C.32}
$$

Rearranging terms we obtain

$$\log h - \frac{2}{3}h \cdot \log\gamma = \frac{1}{3}\log\log\frac{1}{\gamma} + \frac{2}{3}\log\epsilon_v. \tag{C.33}$$

We have from the exponential of both sides that

$$h \cdot \exp\left(-\frac{2}{3}h \cdot \log\gamma\right) = \left(\log\frac{1}{\gamma}\right)^{\frac{1}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}. \tag{C.34}$$

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

$$-\frac{2}{3}h\log\gamma \cdot \exp\left(-\frac{2}{3}h \cdot \log\gamma\right) = \frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}. \tag{C.35}$$

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

$$-\frac{2}{3}h \cdot \log\gamma = W\left(\frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}\right). \tag{C.36}$$

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

$$h = \frac{3}{2\log\frac{1}{\gamma}} \cdot W\left(\frac{2}{3}(\log\gamma)^{\frac{4}{3}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}\right). \tag{C.37}$$

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

$$\begin{aligned}
h &= \gamma^{2h} \cdot \log\frac{1}{\gamma} \\
\log h &= 2h\log\gamma + \log\log\frac{1}{\gamma} \\
\log h - 2h\log\gamma &= \log\log\frac{1}{\gamma} \\
h \cdot \exp\left(-2h \cdot \log\gamma\right) &= \log\frac{1}{\gamma} \\
-2h\log\gamma \cdot \exp\left(-2h \cdot \log\gamma\right) &= 2(\log\gamma)^2 \\
-2h \cdot \log\gamma &= W\Big(2(\log\gamma)^2\Big) \\
h &= \frac{1}{2\log\frac{1}{\gamma}} \cdot W\Big(2(\log\gamma)^2\Big).
\end{aligned} \tag{C.38}$$

Now we have shown that the minima of $g_1(h)$ and $g_2(h)$ is unique (i.e. (C.38) and (C.37)). Therefore, $g_1(h) + g_2(h)$ also has a unique minima within the range of

$$\begin{aligned}
\Bigg[\min\Big\{&\frac{1}{2\log\frac{1}{\gamma}} \cdot W\left(2(\log\gamma)^2\right), \frac{5}{2\log\frac{1}{\gamma}} \cdot W\left(\frac{2}{5}(\log\gamma)^{\frac{6}{5}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{5}}\right)\Big\}, \\
&\max\Big\{\frac{1}{2\log\frac{1}{\gamma}} \cdot W\left(2(\log\gamma)^2\right), \frac{5}{2\log\frac{1}{\gamma}} \cdot W\left(\frac{2}{5}(\log\gamma)^{\frac{6}{5}} \cdot (\epsilon_v/\epsilon_f)^{\frac{2}{3}}\right)\Big\}\Bigg].
\end{aligned} \tag{C.39}$$

We conclude that the optimal expansion step has the following expression

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

494 □

C.5 PROOF OF COROLLARY 5.12

Proof. We set $\eta = 1/\sqrt{T}$ and $T \geq 4L^2$, which gives us $c = (\eta - L\eta^2)^{-1} \leq 2\sqrt{T}$ and $L\eta \leq 1/2$. By setting $N = O(\sqrt{T})$, we have

$$\begin{aligned}
\min_{t\in[T]} \mathbb{E}\Big[\big\|\nabla_\theta J(\pi_{\theta_t})\big\|_2^2\Big] &\leq \frac{4}{T}\cdot\left(\sum_{t=0}^{T-1} c\cdot(2\delta\cdot b_t + \frac{\eta}{2}\cdot v_t) + b_t^2 + v_t\right) + \frac{4c}{T}\cdot\mathbb{E}\big[J(\pi_{\theta_T}) - J(\pi_{\theta_1})\big] \\
&\leq \frac{4}{T}\left(\sum_{t=0}^{T-1} 4\sqrt{T}\delta\cdot b_t + b_t^2 + 2v_t\right) + \frac{8}{\sqrt{T}}\cdot\mathbb{E}\big[J(\pi_{\theta_T}) - J(\pi_{\theta_1})\big] \\
&\leq \frac{4}{T}\left(\sum_{t=0}^{T-1} 4\sqrt{T}\delta\cdot b_t + b_t^2\right) + O\big(1/\sqrt{T}\big) \\
&\leq \frac{16\delta}{\sqrt{T}}\varepsilon(T) + \frac{4}{T}\varepsilon^2(T) + O\big(1/\sqrt{T}\big).
\end{aligned}$$

□