MODEL-BASED REPARAMETERIZATION POLICY GRADIENT METHODS

Shenao Zhang[1], **Boyi Liu**[2], **Yan Li**[1], **Zhaoran Wang**[2], **Tuo Zhao**[1]

[1]Georgia Tech [2]Northwestern University

ABSTRACT

ReParameterization (RP) Policy Gradient Methods (PGMs) have been widely adopted on continuous control tasks with applications in robotics and computer graphics. However, when applied to long-horizon reinforcement learning problems, some recent works (Parmas et al., 2018; Metz et al., 2021) reported that model-based RP PGMs often suffer from chaotic and non-smooth optimization landscapes with exploding gradient variance, leading to slow convergence. This is in sharp contrast to our conventional wisdom that the reparameterization methods enjoy small variance of gradient estimation in other problems such as training deep generative models. Such an intriguing phenomenon motivates us to investigate the theoretical properties of model-based RP PGMs and seek the cure for the exploding gradient variance. Specifically, we establish the first convergenve analysis result for model-based RP PGMs, and our theory identifies the smoothness of the function approximators as a major determining factor that affects the quality of gradient estimation. Based on our theory, we further propose a spectral normalization method, which can effectively mitigate the exploding variance due to the long model unrolls. Numerical experiments are provided to support our proposed theory and method: With a proper normalization, we can significantly reduce the gradient variance of model-based RP PGMs and improve their convergence, leading to equally or better performance than their counterparts based on other gradient estimators, e.g., Likelihood Ratio (LR) gradient estimator.

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) 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 popular methodologies, the policy gradient method (PGM) (Sutton et al., 1999; Kakade, 2001; 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 quality of the stochastic gradient naturally plays a crucial role in the performance of PGMs.

Most existing stochastic gradient estimation schemes fall into two categories: (I) Likelihood Ratio (LR) estimators, which perform zeroth-order estimation through sampling of function values (Williams, 1992; Konda & Tsitsiklis, 1999; Kakade, 2001; Degris et al., 2012); (II) ReParameterization (RP) gradient estimators, which exploit the differentiability of the learned value with function approximation (Figurnov et al., 2018; Ruiz et al., 2016; Clavera et al., 2020; Suh et al., 2022a).

Though both of LR and RP PGMs are popular in practice, existing literature on theoretical properties of PGMs mainly focuses on the LR PGMs: Their global convergence analysis has been established under various settings, and the estimation quality of the LR gradient estimators have been heavily investigated (Agarwal et al., 2021; Wang et al., 2019a; Bhandari & Russo, 2019). This is in sharp contrast to that theories on RP PGMs are largely missing: The properties of RP gradient estimators are not well understood and no convergence analysis of RP PGMs has been established.

The RP gradient estimators are well-known for training deep generative models such as variational autoencoders (Figurnov et al., 2018). From a stochastic optimization perspective, previous analysis (Ruiz et al., 2016; Mohamed et al., 2020) showed that RP gradient methods enjoy small variance,

which leads to better convergence performance. However, some recent works (Parmas et al., 2018; Metz et al., 2021) reported an opposite observation: When applied to long-horizon reinforcement learning problems, model-based RP PGMs often suffer from chaotic and non-smooth optimization landscapes with exploding gradient variance, which leads to slow convergence.

Such an intriguing phenomenon motivates us to revisit the theoretical properties of RP gradient estimators, and seek the cure for the exploding gradient variance in model-based RP PGMs. Specifically, we develop a unified theoretical framework for analyzing the properties of model-based RP PGMs and establish their first convergence result. Our analysis suggests that the smoothness and accuracy of the learned model and policy are major determining factors for the exploding variance of RP gradients: (1) Both the gradient variance and bias have polynomial dependence on the Lipschitz continuity of the learned model and policy w.r.t. the input state, where the degrees are linear in the steps of model value expansion; (2) The bias also depends on the error of the estimated model and value.

Our results imply that one can significantly reduce the variance of RP gradient estimators by enforcing the smoothness of the model and policy. Note that by doing so, we can improve the algorithm performance without introducing much bias when the underlying (ground truth) transition kernel is indeed smooth. However, for non-smooth transition kernels, enforcing the smoothness may introduce bias, which increases the the estimation error of the learned model. Therefore, given such a situation, we need to trade off between the model bias and gradient variance.

As a concrete application of our theoretical discovery, we propose a spectral normalization method to enforce the smoothness of the learned model and policy. Our numerical experiments show that with a proper normalization, we can significantly reduce the gradient variance of model-based RP PGMs and improve their convergence, leading to equal or better performance than the vanilla implementation or their Likelihood Ratio (LR) gradient estimator counterparts.

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

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

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

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

We denote by ζ the initial state distribution. 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

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

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

$$J(\pi) = \mathbb{E}_{s_0 \sim \zeta}\big[V^\pi(s_0)\big] = \mathbb{E}_{(s,a)\sim\sigma_\pi}\big[r(s,a)\big]. \tag{2.1}$$

Stochastic Gradient Estimation. The general underlying problem of policy gradient, i.e., computing the gradient of a probabilistic objective with respect to the parameters of sampling distribution, takes the form $\nabla_\theta \mathbb{E}_{p(x;\theta)}[y(x)]$. In RL, we set $p(x;\theta)$ as the trajectory distribution conditioned on policy parameter θ, and $y(x)$ as the cumulative reward. In the sequel, we introduce two commonly used gradient estimators in RL.

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

$$\nabla_\theta \mathbb{E}_{p(x;\theta)}\big[y(x)\big] = \int y(x)\nabla_\theta p(x;\theta)\mathrm{d}x = \mathbb{E}_{p(x;\theta)}\big[y(x)\nabla_\theta \log p(x;\theta)\big]. \tag{2.2}$$

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:

$$\widehat{x} \sim p(x;\theta) \;\equiv\; \widehat{x} = g(\epsilon;\theta), \quad \epsilon \sim p(\epsilon) \tag{2.3}$$

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

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

3 Analytic Reparameterization Gradient in RL

In this section, we present two basic *analytic* forms of the RP gradient in RL. We first consider the Policy-Value Gradient (PVG) method, which is model-free and can be expanded sequentially to obtain the Analytic Policy Gradient (APG) method. Then we discuss some potential problems when developing practical algorithms.

Consider a policy $\pi_\theta(s,\varsigma)$ with noise ς in continuous action spaces. We make the following assumption to ensure that the first-order gradient through the value is well-defined.

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'.

Policy-Value Gradient. The general form of the reparameterization Policy-Value Gradient (PVG) is as follows:

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

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

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

where $\nabla_a Q^\pi$ can be estimated using a 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.

By sequentially expanding the model-free PVG, we have the following dynamical representation of the policy gradient, which are commonly known as Analytic Policy Gradient (APG).

Analytic Policy Gradient. Due to the simulation property of continuous distributions in (2.3), we interchangeably write $a \sim \pi(\cdot \mid s)$, $a = \pi(s,\varsigma)$ and $s' \sim f(\cdot \mid s,a)$, $s' = f(s,a,\xi^*)$, where ξ^* is sampled from the unknown distribution $p(\xi^*)$. From the Bellman equation $V^\pi(s) = \mathbb{E}_\varsigma[(1-\gamma)\cdot r(s,\pi(s,\varsigma)) + \gamma\cdot\mathbb{E}_{\xi^*}[V^\pi(f(s,\pi(s,\varsigma),\xi^*))]]$, we obtain the backward recursions of gradient:

$$\nabla_\theta V^\pi(s) = \mathbb{E}_\varsigma\Big[(1-\gamma)\nabla_a r\nabla_\theta\pi + \gamma\mathbb{E}_{\xi^*}\big[\nabla_{s'}V^\pi(s')\nabla_a f\nabla_\theta\pi + \nabla_\theta V^\pi(s')\big]\Big], \tag{3.2}$$

$$\nabla_s V^\pi(s) = \mathbb{E}_\varsigma\Big[(1-\gamma)(\nabla_s r + \nabla_a r\nabla_s\pi) + \gamma\mathbb{E}_{\xi^*}\big[\nabla_{s'}V^\pi(s')(\nabla_s f + \nabla_a f\nabla_s\pi)\big]\Big]. \tag{3.3}$$

The detailed derivation of (3.2) and (3.3) can be found in Appendix A. Now we have the RP gradient backpropagated through the transition path starting at s. By taking expectation over the initial state distribution, we obtain the Analytic Policy Gradient (APG) that takes the following form:

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

There remain problems when developing practical algorithms: (I) The above formulas require the gradient information of the dynamics function. In this work, however, we consider a common RL

setting where the MDP transition f and its derivatives are not known and need to be fitted by a model. 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 resulting algorithms. (II) Even if we have access to an accurate model, unrolling it over full sequences faces practical difficulties: the memory and computational cost scale linearly with the unroll length; long chains of nonlinear mappings can also lead to exploding or vanishing gradients and, even worse, chaotic phenomenons and difficulty in optimization (Pascanu et al., 2013; Maclaurin et al., 2015; Vicol et al., 2021; Metz et al., 2019), which demand some form of truncation.

4 Model-Based RP Policy Gradient Methods

In this section, we introduce the Model Value Expansion (MVE) technique for model truncation, and present two model-based reparameterization policy gradient frameworks built upon MVE.

4.1 h-Step Model Value Expansion

As a common technique for alleviating the challenges brought by full unrolls, algorithms with direct truncation split and backpropagate through the shorter sub-sequences (e.g., Truncated BPTT, Werbos (1990)). However, this naive scheme prioritizes short-term dependencies and yields biased gradients.

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

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

where $\widehat{s}_0 = s$, $\widehat{a}_i = \pi(\widehat{s}_i,\varsigma;\theta)$, $\widehat{s}_{i+1} = \widehat{f}(\widehat{s}_i,\widehat{a}_i,\xi;\psi)$, and $\widehat{Q}_\omega$ is a trainable critic. Here, the noises ς and ξ can be sampled from fixed distributions or inferred from real samples, which we will discuss in the following section.

4.2 Model-Based RP Gradient Estimation

By taking the pathwise gradient with respect to θ in (4.1), we obtain the following two frameworks.

Model Derivatives on Predictions. One intuitive way to estimate the first-order 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* state prediction and derivative calculation. The RP 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 $\widehat{s}_{0,n}\sim\mu_\pi$, $\widehat{a}_{i,n} = \pi(\widehat{s}_{i,n},\varsigma_n;\theta)$, and $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n;\psi)$ with $\varsigma_n\sim p(\cdot)$, $\xi_n\sim p$. Here μ_π is the distribution where the starting states of the simulated trajectories are sampled. We will show in Section 5 that μ_π can be specified as a mixture of ζ and σ_π.

Various algorithms can be instantiated with different choices of h. When $h=0$, the framework reduces to the model-free policy gradient, e.g. RP(0) (Amos et al., 2021) and the variants of DDPG (Lillicrap et al., 2015) such as SAC (Haarnoja et al., 2018). When $h\to\infty$, the resulting algorithm is BPTT (Grzeszczuk et al., 1998; Mozer, 1995; Degrave et al., 2019; Bastani, 2020) where only the model is learned. Recent model-based approaches, e.g. MAAC (Clavera et al., 2020) and related algorithms (Parmas et al., 2018; Amos et al., 2021; Li et al., 2021), require a carefully selected h.

Model Derivatives on Real Samples. Alternatively, we may also replace the ∇f terms in (3.2) and (3.3) with $\nabla\widehat{f}$. By doing so, the learned differentiable model is *only* used for derivative calculation and the Monte-Carlo estimates are computed on *real* samples. We defer the formulas to Appendix A.

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

$$\widehat{\nabla}^{\text{DR}}_\theta J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\nabla}_\theta V^\pi(\widehat{s}_{0,n}) = \frac{1}{N}\sum_{n=1}^{N}\nabla_\theta\left(\sum_{i=0}^{h-1}\gamma^i r(\widehat{s}_{i,n},\widehat{a}_{i,n})+\gamma^h\widehat{Q}_\omega(\widehat{s}_{h,n},\widehat{a}_{h,n})\right), \tag{4.3}$$

where $\widehat{s}_{0,n} \sim \mu_\pi$, $\widehat{a}_{i,n} = \pi(\widehat{s}_{i,n}, \varsigma_n)$, $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n}, \widehat{a}_{i,n}, \xi_n)$, and ς_n, ξ_n are inferred from the real data sample (s_i, a_i, s_{i+1}) generated via $a_i = \pi(s_i, \varsigma_n)$ and $s_{i+1} = \widehat{f}(s_i, a_i, \xi_n)$. Example algorithms include SVG (Heess et al., 2015) and its variants (Abbeel et al., 2006; Atkeson, 2012).

4.3 Algorithmic Framework

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

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 $\mu_{\pi_t}(\cdot)$ and estimate $\widehat{\nabla}_\theta J(\pi_{\theta_t}) = \widehat{\nabla}^{\mathrm{DP}}_\theta J(\pi_{\theta_t})$ (4.2) or $\widehat{\nabla}^{\mathrm{DR}}_\theta J(\pi_{\theta_t})$ (4.3)
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:** $\{\pi_{\theta_t}\}_{t\in[T]}$

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

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

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

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. Thus, we adopt the notation $\epsilon_f(t)$ to represent the model error at iteration t:

$$\epsilon_f(t) := \max_{i\in[h]} \mathbb{E}_{\mathbb{P}(s_i,a_i),\mathbb{P}(\widehat{s}_i,\widehat{a}_i)} \left[\left\| \frac{\partial s_i}{\partial s_{i-1}} - \frac{\partial \widehat{s}_i}{\partial \widehat{s}_{i-1}} \right\|_2 + \left\| \frac{\partial s_i}{\partial a_{i-1}} - \frac{\partial \widehat{s}_i}{\partial \widehat{a}_{i-1}} \right\|_2 \right]. \tag{4.4}$$

Here $\mathbb{P}(s_i, a_i)$ is the state-action distribution at step i, where $s_0 \sim \nu_{\pi_t}$, $a_j \sim \pi_t(\cdot \,|\, s_j)$, 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 $\widehat{s}_0 \sim \nu_{\pi_t}$, $\widehat{a}_j \sim \pi_t(\cdot \,|\, \widehat{s}_j)$, $\widehat{s}_{j+1} \sim \widehat{f}(\cdot \,|\, \widehat{s}_j, \widehat{a}_j)$ when $\widehat{\nabla}_\theta J(\pi_\theta) = \widehat{\nabla}^{\mathrm{DP}}_\theta J(\pi_\theta)$.

In model-based RL, it is popular to learn a (multi-step) *state*-predictive model, where the objective mismatch issue arises between minimizing the state prediction error and the model gradient error in (4.4). Although it is natural to explicitly regularize the models' directional derivatives to be consistent with the samples (Li et al., 2021), we argue that adopting state-predictive models does *not* cripple our analysis based on ϵ_f: For learned models that extrapolate beyond the visited regions, the gradient error can still be bounded via finite difference. In other words, ϵ_f can be expressed as the mean squared training error with an additional measure of the model class complexity to capture its generalizability. A similar argument also holds for learning a critic by temporal difference.

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

$$\epsilon_v(t) := \mathbb{E}_{\mathbb{P}(s_h,a_h),\mathbb{P}(\widehat{s}_h,\widehat{a}_h)} \left[\left\| \frac{\partial Q^{\pi_t}}{\partial s} - \frac{\partial \widehat{Q}_t}{\partial \widehat{s}} \right\|_2 + \left\| \frac{\partial Q^{\pi_t}}{\partial a} - \frac{\partial \widehat{Q}_t}{\partial \widehat{a}} \right\|_2 \right], \tag{4.5}$$

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 (4.4).

5 Main Results

We presents our main theoretical results in this section, with the proofs deferred to Appendix C. Specifically, we establish the convergence of model-based RP PGMs. More importantly, we study the relationship between the convergence rate, gradient bias, variance, and the model smoothness, approximation error. Based on our theory, we propose several potential algorithmic designs.

To begin with, we impose a common regularity condition on the policy functions following previous works (Xu et al., 2019; Pirotta et al., 2015; Zhang et al., 2020; Agarwal et al., 2021). The assumption below essentially ensures the smoothness of the objective $J(\pi_\theta)$, which is required by most existing analysis of policy gradient methods (Wang et al., 2019a; Bastani, 2020; Agarwal et al., 2020).

Assumption 5.1 (Lipschitz Score Function and Boundedness)**.** Assume that the score function of policy π_θ is Lipschitz continuous and has bounded norm for $(s, a) \in \mathcal{S} \times \mathcal{A}$, formally,

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

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

Proposition 5.2 (Convergence to Stationary Points)**.** Define the gradient bias b_t and variance v_t as

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

Suppose the absolute value of the reward $r(s, a)$ is bounded by $|r(s, a)| \le r_{\mathrm{m}}$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\|\theta\|_2 \le \delta$. Denote $L := r_{\mathrm{m}} \cdot L_1/(1-\gamma)^2 + (1+\gamma) \cdot r_{\mathrm{m}} \cdot B_\theta^2/(1-\gamma)^3$ and $c := (\eta - L\eta^2)^{-1}$. It then holds for $T \ge 4L^2$ that

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

Proposition 5.2 shows the reliance of the convergence error and the variance and bias of the gradient estimators. In general, to guarantee the convergence of model-based RP PGMs, we have to control both the variance and the bias to sublinear growth rate. Before studying the upper bound of b_t and v_t, we make the following Lipschitz assumption, which is adopted in various previous works (Pirotta et al., 2015; Clavera et al., 2020; Li et al., 2021).

Assumption 5.3 (Lipschitz Continuous Dynamics)**.** Assume $r(s, a)$, $f(s, a, \xi^*)$, $\widehat{f}_\psi(s, a, \xi)$, $\pi_\theta(s, \varsigma)$, $\widehat{Q}_\omega(s, a)$ are $L_r, L_f, L_{\widehat{f}}, L_\pi, L_{\widehat{Q}}$ Lipschitz continuous (we defer the details to Appendix B).

Denote $\widetilde{L}_g := \max\{L_g, 1\}$ for some function g. 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 $\widehat{\nabla}_\theta^{\mathrm{DP}} J(\pi_\theta)$ or $\widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta)$ can be bounded by

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

We observe that the variance upper bound has polynomial dependence on the Lipschitz continuity of the model and policy, where the degrees are linear in the steps of model value expansion. This makes intuitive sense as the dynamics can be chaotic (Bollt, 2000) when $L_{\widehat{f}} > 1$ and $L_\pi > 1$. As a result, the stochasticity during training can lead to diverging trajectories and stochastic gradient directions, causing large gradient variance.

Remark 5.5. Non-smooth models and policies can lead to highly non-smooth loss landscapes, which result in slow convergence or failure of training even in simple toy examples (Parmas et al., 2018; Metz et al., 2021; Suh et al., 2022a). Proposition 5.4 suggests 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 the policy to avoid exploding gradient variance.

Model-based RP PGMs have their own advantages by leveraging smooth proxy models for variance reduction. By enforcing the smooth of , the algorithm performance can be improved without introducing much bias when the underlying transition is smooth. However, for non-smooth dynamics (e.g. complex and contact-rich environments (Suh et al., 2022a; Xu et al., 2022)), this may introduce additional bias due to the increased model estimation error, which demands the trade-off between the model bias and gradient variance. Interestingly, our empirical study shows that smoothness regularization improves the performance even with the cost of a higher bias.

Consider setting μ_π, the state distribution when estimating the RP gradient, as a mixture of the MDP initial distribution ζ and the state visitation ν_π: $\mu_\pi(s) = \beta \cdot \nu_\pi(s) + (1-\beta)\cdot\zeta(s)$, where $\beta \in [0,1]$. We study this form since it covers the different state sampling schemes when $h = 0$ and $h \to \infty$: States are sampled from ν_π when not using a model (e.g., SVG(0) (Heess et al., 2015; Amos et al., 2021) and DDPG (Silver et al., 2014; Lillicrap et al., 2015)); States are sampled 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)).

Since policy actions can lead to changes in all future states and rewards, unless we know the exact policy value function, its gradient $\nabla_\theta Q^{\pi_\theta}$ 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 account for the gradient bias brought by the critic. For this reason, we give the gradient bias bound below based on a measure of the discrepancy between the initial distribution ζ and the state visitation ν_π.

Proposition 5.6 (Gradient Bias)**.** Let $\kappa := \sup_\pi \mathbb{E}_{\nu_\pi}[(\mathrm{d}\zeta/\mathrm{d}\nu_\pi(s))^2]^{1/2}$, where $\mathrm{d}\zeta/\mathrm{d}\nu_\pi$ is the Radon-Nikodym derivative of ζ with respect to ν_π. Denote $\kappa' := \beta + \kappa\cdot(1-\beta)$. Under Assumption 5.3, for any $t \in [T]$, the gradient bias is bounded by

$$b_t \le O\Big(\kappa'\kappa h^3 \widetilde{L}_{\widehat{f}}^h \widetilde{L}_f^h \widetilde{L}_\pi^{2h} \epsilon_{f,t} + \kappa' h^2 \gamma^h \widetilde{L}_{\widehat{f}}^h \widetilde{L}_\pi^h \epsilon_{v,t}\Big), \tag{5.2}$$

where $\epsilon_{f,t}$ and $\epsilon_{v,t}$ is the shorthand notation of $\epsilon_f(t)$ and $\epsilon_v(t)$, defined in (4.4) and (4.5), respectively.

Besides, the analysis above naturally results in an optimal model expansion step $h^* \in [0,\infty)$ that achieves the best convergence rate. The expression of h^* is given by the following proposition.

Proposition 5.7 (Optimal Model Expansion Step)**.** Suppose $L_f \le 1$. If we regularize the model and policy such that $L_{\widehat{f}} \le 1$ and $L_\pi \le 1$, then the optimal model expansion step h^* at iteration t that minimizes the convergence rate upper bound satisfies:

$$h^* = O\Big(\frac{1}{\log(1/\gamma)} W\big(2(\log\gamma)^2\big)/N + \log\big((\epsilon_{v,t}/\epsilon_{f,t})(1/\log(1/\gamma))\big)\Big),$$

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

Remark 5.8. We observe that h^* is positive and increases in logarithmic rate with respect to the error scale $\epsilon_{v,t}/\epsilon_{f,t}$. This result can guide the algorithms to rely more on the model by performing longer unrolls when the model error $\epsilon_{f,t}$ is small; while avoiding long model unrolls when the critic error $\epsilon_{f,t}$ is relatively smaller.

In Proposition 5.7, the Lipschitz condition of the underlying dynamics, i.e. $L_f \le 1$, ensures the stability of the system. This is obvious by considering the linear system example: The transitions depend on the eigenspectrum of the family of transformations, causing the trajectories diverging exponentially w.r.t. the largest eigenvalue (Metz et al., 2021). In practical control systems where this condition is not met, we might need to empirically search for the best model unroll length. Fortunately, we observe in experiments that applying spectral normalization provides more training stability and offers a much wider range of the unroll lengths that achieve similarly good performance.

Corollary 5.9 (Convergence Rate)**.** Let $\varepsilon(T) = \sum_{t=0}^{T-1} b_t$. We have for $T \ge 4L^2$ that

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

The convergence rate can be further specified by characterizing how fast the model error and critic error go 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 to bound the prediction error that is minimized during training, and introducing the measure of the function (namely, model and critic) class complexity to bound the gradient error $\epsilon_f(t), \epsilon_v(t)$.

6 Related Work

Policy Gradient Methods. In the context of RL, the LR estimator 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., 2021b) assume that gradients of simulation outcomes w.r.t. control actions are explicitly given. To deal with the non-smoothness and discontinuities in the differentiable simulation caused by contact dynamics and geometrical constraints, previous works 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

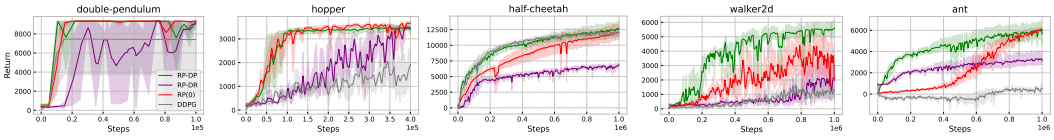


Figure 1: Evaluation of different instantiations of model-based RP PGMs in several MuJoCo tasks.

We first evaluate several algorithms instantiated from the reparameterization policy gradient methods in MuJoCo (Todorov et al., 2012) tasks. The two types of gradient estimators in Section 4.2 are tested: We use RP-DP and RP-DR to distinguish whether the model derivatives are calculated on predictions (4.2) or on real samples (4.3). 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 SVG (Heess et al., 2015). For completeness, we also evaluate model-free PGMs ($h = 0$), including RP(0) (Amos et al., 2021) and the DDPG (Lillicrap et al., 2015) baseline. The implementation details and the full results containing more RL baselines (e.g. LR PGMs) are deferred to Appendix D.1.

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.

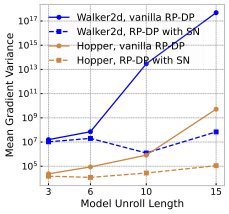
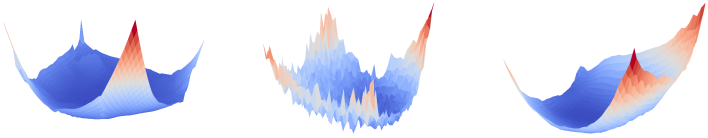


Figure 2: Gradient variance.



(a) Vanilla RP-DP, unroll length $h = 3$.

(b) Vanilla RP-DP, unroll length $h = 15$.

(c) RP-DP with SN, unroll length $h = 15$.

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

We can observe that for vanilla RP policy gradient algorithms, the gradient variance explodes in exponential rate with respect to the unroll length. 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) (Miyato et al., 2018), the mean gradient variance is much lower for all unroll length settings, 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 claim: The gradient variance has polynomial dependence on the Lipschitz continuity of the model and policy and is a contributing factor to training. The results in Fig. 4 show that adding SN achieves equal or better performance compared to the vanilla implementation. Importantly, for longer model unrolls (e.g. 10 in walker2d and 15 in hopper), vanilla RP PGMs fail to reach reliable performance, while SN significantly boosts training. Due to space limit, we defer the details of SN and the complete comparisons to App. D.2 and D.3. We also refer the readers to App. D.5 for a larger size of Figure 4, 5, and 6.

Figure 4: Performance of model-based RP PG methods with and without spectral normalization.

Ablation on variance. 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 model-based RP PG methods with and without spectral normalization.

Ablation on bias. When the underlying MDP is itself contact-rich and has non-smooth or even discontinuous dynamics, explicitly regularizing the Lipschitz of the transition model may lead to large error ϵ_f and thus large gradient bias. Therefore, it is also crucial to study the negative effect smoothness regularization may bring. To efficiently obtain accurate first-order gradient (instead of via finite difference in MuJoCo), we conduct ablation based on the *differentiable* simulator dFlex (Heiden et al., 2021a; Xu et al., 2022), where Analytic Policy Gradient (APG) can be implemented.

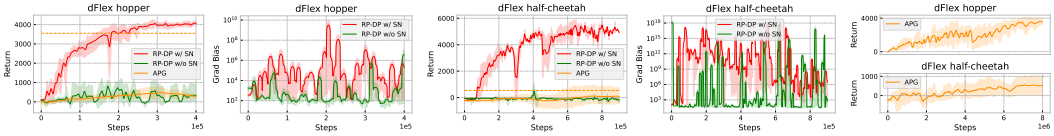


Figure 6: Performance and gradient bias in differentiable simulation. The last column is the full curves of APG, which needs 20 times more steps in hopper to reach a comparable return with RP-DP-SN.

We observe in Figure 6 that spectral normalization also plays an important role in dFlex locomotion tasks. More importantly, SN with higher bias does *not* lead to performance degradation, but rather in the opposite way benefiting from the reduced variance. This result suggests that even in differentiable simulation, one may still use a smooth proxy model when the dynamics has bumps or discontinuous jumps, sharing similarities with gradient smoothing applied to APG (Suh et al., 2022a;b).

8 Conclusion & Future Work

In this work, we study the convergence of model-based reparameterization policy gradient methods and identifies the determining factors that affect the quality of gradient estimation. Based on our theory, we propose a spectral normalization (SN) method to mitigate the exploding gradient variance issue. Our experimental results also support the proposed theory and method. Since adding SN allows longer model unrolls, learning an accurate multi-step model to fully leverage its gradient information should be a fruitful future direction. It will also be interesting to explore different smoothness regularization designs and applying them to a broader range of algorithms, e.g. using proxy models in differentiable simulation to obtain smooth policy gradient, which we would like to leave as future work.

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. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

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

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

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

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

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

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

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

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

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

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

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

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

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

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 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.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

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.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

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, Miles Macklin, Yashraj Narang, Dieter Fox, Animesh Garg, and Fabio Ramos. Disect: A differentiable simulation engine for autonomous robotic cutting. *arXiv preprint arXiv:2105.12244*, 2021a.

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, 2021b.

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.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

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.

Luis Pineda, Brandon Amos, Amy Zhang, Nathan O Lambert, and Roberto Calandra. Mbrl-lib: A modular library for model-based reinforcement learning. *arXiv preprint arXiv:2104.10159*, 2021.

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

Emmanuel Rachelson and Michail G Lagoudakis. On the locality of action domination in sequential decision making. 2010.

Francisco R Ruiz, Titsias RC AUEB, David Blei, et al. The generalized reparameterization gradient. *Advances in neural information processing systems*, 29, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

J Weng et al. A highly modularized deep reinforcement learning library. *arXiv preprint arXiv:2107.14171*, 2021.

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.

Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30, 2017.

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.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

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 Recursive Expression of Analytic Policy Gradient

In this part, we give the derivation of (3.2) and (3.3), i.e. the backward recursions of gradient in APG:

Following Heess et al. (2015), we define the operator

$$\nabla_\theta^i := \sum_{j\geq i} \frac{\mathrm{d}a_j}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial a_j} + \sum_{j>i}\frac{\mathrm{d}s_j}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial s_j}. \tag{A.1}$$

In general, we can expand the total derivative by chain rule as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta} &= \sum_{i\geq 0}\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial a_i} + \sum_{i>0}\frac{\mathrm{d}s_i}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial s_i}\\ &= \frac{\mathrm{d}a_0}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial a_0} + \frac{\mathrm{d}s_1}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial s_1} + \sum_{i\geq 1}\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial a_i} + \sum_{i>1}\frac{\mathrm{d}s_i}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial s_i}.\end{aligned} \tag{A.2}$$

This shows that the operator ∇_θ^i obeys the recursive formula:

$$\nabla_\theta^i = \frac{\mathrm{d}a_i}{\mathrm{d}\theta}\cdot\frac{\partial}{\partial a_i} + \frac{\mathrm{d}a_t}{\mathrm{d}\theta}\cdot\frac{\mathrm{d}s_{i+1}}{\mathrm{d}a_t}\cdot\frac{\partial}{\partial s_{i+1}} + \nabla_\theta^{i+1}. \tag{A.3}$$

Besides, we have from the Bellman equation that

$$V^\pi(s) = \mathbb{E}_\varsigma\Big[(1-\gamma)\cdot r(s,\pi(s,\varsigma)) + \gamma\cdot\mathbb{E}_{\xi^*}\Big[V^\pi\Big(f\big(s,\pi(s,\varsigma),\xi^*\big)\Big)\Big]\Big]. \tag{A.4}$$

By applying the recursive formula (A.3) to the Bellman equation (A.4), we obtain

$$\begin{aligned}\frac{\mathrm{d}V^\pi(s)}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\varsigma\Big[(1-\gamma)\cdot r(s,\pi(s,\varsigma)) + \gamma\cdot\mathbb{E}_{\xi^*}\Big[V^\pi\Big(f\big(s,\pi(s,\varsigma),\xi^*\big)\Big)\Big]\Big]\\ &= \mathbb{E}_\varsigma\Big[(1-\gamma)\cdot\frac{\partial r}{\partial a}\cdot\frac{\mathrm{d}a}{\mathrm{d}\theta} + \gamma\cdot\mathbb{E}_{\xi^*}\Big[\frac{\mathrm{d}a}{\mathrm{d}\theta}\cdot\frac{\mathrm{d}s'}{\mathrm{d}a}\cdot\frac{\mathrm{d}V^\pi(s')}{\mathrm{d}s'} + \frac{\mathrm{d}V^\pi(s')}{\mathrm{d}\theta}\Big]\Big].\end{aligned} \tag{A.5}$$

For the $\mathrm{d}V^\pi(s)/\mathrm{d}s$ term, we have the following recursion:

$$\begin{aligned}\frac{\mathrm{d}V^\pi(s)}{\mathrm{d}s} &= \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}_\varsigma\Big[(1-\gamma)\cdot r(s,\pi(s,\varsigma)) + \gamma\cdot\mathbb{E}_{\xi^*}\Big[V^\pi\Big(f\big(s,\pi(s,\varsigma),\xi^*\big)\Big)\Big]\Big]\\ &= \mathbb{E}_\varsigma\Big[(1-\gamma)\cdot\Big(\frac{\partial r}{\partial s} + \frac{\partial r}{\partial a}\cdot\frac{\partial a}{\partial s}\Big) + \gamma\cdot\mathbb{E}_{\xi^*}\Big[\frac{\partial s'}{\partial s}\cdot\frac{\mathrm{d}V^\pi(s')}{\mathrm{d}s'} + \frac{\partial s'}{\partial a}\cdot\frac{\partial a}{\partial s}\cdot\frac{\mathrm{d}V^\pi(s')}{\mathrm{d}s'}\Big]\Big].\end{aligned} \tag{A.6}$$

The above two equations correspond to (3.2) and (3.3), respectively, which completes the derivation.

By replacing the $\nabla_a f$ and $\nabla_s f$ terms in the gradient recursions with $\nabla_a \widehat{f}$ and $\nabla_s \widehat{f}$, we obtain the missing equations in Section 4.2 as follows:

$$\begin{aligned}\widehat{\nabla}_\theta V^\pi(\widehat{s}_{i,n}) &= (1-\gamma)\nabla_a r(\widehat{s}_{i,n},\widehat{a}_{i,n})\nabla_\theta\pi(\widehat{s}_{i,n},\varsigma_n)\\ &\quad + \gamma\widehat{\nabla}_s V^\pi(\widehat{s}_{i+1,n})\nabla_a\widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n)\nabla_\theta\pi(\widehat{s}_{i,n},\varsigma_n) + \gamma\widehat{\nabla}_\theta V^\pi(\widehat{s}_{i+1,n}),\end{aligned} \tag{A.7}$$

$$\begin{aligned}\widehat{\nabla}_s V^\pi(\widehat{s}_{i,n}) &= (1-\gamma)\big(\nabla_s r(\widehat{s}_{i,n},\widehat{a}_{i,n}) + \nabla_a r(\widehat{s}_{i,n},\widehat{a}_{i,n})\nabla_s\pi(\widehat{s}_{i,n},\varsigma_n)\big)\\ &\quad + \gamma\widehat{\nabla}_s V^\pi(\widehat{s}_{i+1,n})\big(\nabla_s\widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n) + \nabla_a\widehat{f}(\widehat{s}_{i,n},\widehat{a}_{i,n},\xi_n)\nabla_s\pi(\widehat{s}_{i,n},\varsigma_n)\big).\end{aligned} \tag{A.8}$$

In (4.3), the noise ς_n, ξ_n are inferred from the real data samples (s_i, a_i, s_{i+1}) by $a_i = \pi(s_i,\varsigma_n)$ and $s_{i+1} = \widehat{f}(s_i,a_i,\xi_n)$, such that $\widehat{a}_{i,n} = a_i$ and $\widehat{s}_{i+1,n} = s_{i+1}$. For example, for state s_{i+1} sampled from a one-dimensional Gaussian transition model $s_{i+1}\sim\mathcal{N}(\phi(s_i,a_i),\sigma^2)$ with the ϕ-parameterized mean and fixed variance σ, the noise ξ_n can be inferred as $\xi_n = (s_{i+1}-\phi(s_i,a_i))/\sigma$.

B Assumption Clarification

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

Assumption B.1 (Lipschitz Continuous Functions)**.** We assume that $r(s,a)$, $f_\psi(s,a,\xi^*)$, $\widehat{f}_\psi(s,a,\xi)$, $\pi_\theta(s,\varsigma)$, $\widehat{Q}_\omega(s,a)$ are $L_r, L_f, L_{\widehat{f}}, L_\pi, L_{\widehat{Q}}$ Lipschitz continuous such that

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

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

C Proofs

C.1 Proof of Proposition 5.2

As a preparation before proving Proposition 5.2, we first present the following lemma stating that the objective in (2.1) is Lipschitz smooth under Assumption 5.1.

Lemma C.1 (Smooth Objective)**.** The objective $J(\pi_\theta)$ is L-smooth in θ, such that $\|\nabla_\theta J(\pi_{\theta_1})-\nabla_\theta J(\pi_{\theta_2})\|_2\le L\|\theta_1-\theta_2\|_2$, where

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

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

Then we are ready to prove Proposition 5.2.

Proof of Proposition 5.2. 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

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

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:

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

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

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

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

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

By rearranging terms,

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

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,

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

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

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

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 the proof by

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

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

C.2 Proof of Proposition 5.4

In what follows, we interchangeably write $\nabla_a x$ and $\mathrm{d}x/\mathrm{d}a$ as the gradient, and use the notation $\partial x/\partial a$ to denote the partial derivative.

Proof. 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.

We start with 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 $\widehat{\nabla}_\theta J(\pi_{\theta_t})$ of this trajectory as g'. Then we have

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

where we let g denote the pathwise gradient $\widehat{\nabla}_\theta J(\pi_{\theta_t})$ of a *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.

Using the fact that $\|\mathbb{E}[\cdot]\|_2 \le \mathbb{E}[\|\cdot\|_2]$, we further obtain

$$v_t \le \mathbb{E}\Big[\big\| g - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \big\|_2 \Big]^2. \tag{C.6}$$

In what follows, the proof is established for the two gradient estimators simultaneously, i.e., when $\widehat{\nabla}_\theta J(\pi_\theta) = \widehat{\nabla}_\theta^{\mathrm{DP}} J(\pi_\theta)$ as in (4.2) and when $\widehat{\nabla}_\theta J(\pi_\theta) = \widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta)$ as in (4.3). Under both frameworks, it holds that $\widehat{s}_{i+1,n} = \widehat{f}(\widehat{s}_{i,n}, \xi_n)$.

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

$$\begin{aligned}\mathbb{E}\Big[\big\| g - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \big\|_2 \Big] \le \sum_{i=0}^{h-1} \gamma^i \cdot \mathbb{E}_{\overline{x}_i} \Big[\big\| \nabla_\theta r(\widehat{x}_{i,n}) - \nabla_\theta r(\overline{x}_i) \big\|_2 \Big] \\ + \gamma^h \cdot \mathbb{E}_{\overline{x}_h} \Big[\big\| \nabla \widehat{Q}(\widehat{x}_{h,n}) \nabla_\theta \widehat{x}_{h,n} - \nabla \widehat{Q}(\overline{x}_h) \nabla_\theta \overline{x}_h \big\|_2 \Big].\end{aligned} \tag{C.7}$$

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

$$\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} = \frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{a}_{i,n}}{\partial \theta}, \tag{C.8}$$

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

Plugging $\mathrm{d}\widehat{a}_{i-1,n}/\mathrm{d}\theta$ in (C.8) into (C.9), we get

$$\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} = \left(\frac{\partial \widehat{s}_{i,n}}{\partial \widehat{s}_{i-1,n}} + \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\partial \widehat{a}_{i-1,n}}{\partial \widehat{s}_{i-1,n}} \right) \cdot \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\partial \widehat{a}_{i-1,n}}{\partial \theta}. \tag{C.10}$$

By the Cauchy-Schwarz inequality and the Lipschitz Assumption 5.3, we have

$$\left\| \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} \right\|_2 \le L_{\widehat{f}} \widetilde{L}_\pi \cdot \left\| \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} \right\|_2 + L_{\widehat{f}} L_\theta.$$

Applying the above recursion gives us

$$\left\| \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} \right\|_2 \le L_{\widehat{f}} L_\theta \cdot \sum_{j=0}^{i-1} L_{\widehat{f}}^j \widetilde{L}_\pi^j \le i \cdot L_\theta L_{\widehat{f}}^{i+1} \widetilde{L}_\pi^i, \tag{C.11}$$

where the first inequality follows from the induction

$$z_i = a z_{i-1} + b = a \cdot (a z_{i-2} + b) + b = a^i \cdot z_0 + b \cdot \sum_{j=0}^{i-1} a^j, \tag{C.12}$$

for the real sequence $\{z_j\}_{0 \le j \le i}$ satisfying $z_j = a z_{j-1} + b$. For $\mathrm{d}\widehat{a}_{i,n}/\mathrm{d}\theta$ defined in (C.8), we further have

$$\left\| \frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} \right\|_2 \le L_\pi \cdot \left\| \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} \right\|_2 + L_\theta \le i \cdot L_\theta L_{\widehat{f}}^{i+1} \widetilde{L}_\pi^{i+1} + L_\theta. \tag{C.13}$$

Combining (C.11) and (C.13), we obtain

$$\left\| \frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta} \right\|_2 = \left\| \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} \right\|_2 + \left\| \frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} \right\|_2 \le \widehat{K}(i) := 2i \cdot L_\theta L_{\widehat{f}}^{i+1} \widetilde{L}_\pi^{i+1} + L_\theta, \tag{C.14}$$

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

Therefore, the second term of (C.7) can be decomposed and bounded by

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

where the last step follows from the Cauchy-Schwartz inequality and the Lipschitz critic assumption.

By the chain rule, a similar result also holds for the first term of (C.7):

$$\begin{aligned}
&\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla_\theta r(\widehat{x}_{i,n})-\nabla_\theta r(\overline{x}_i)\right\|_2\right]\\
&\quad=\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\widehat{x}_{i,n}-\nabla r(\overline{x}_i)\nabla_\theta\overline{x}_i\right\|_2\right]\\
&\quad\le\mathbb{E}_{\overline{x}_i}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\widehat{x}_{i,n}-\nabla r(\widehat{x}_{i,n})\nabla_\theta\overline{x}_i\right\|_2\right]+\mathbb{E}\left[\left\|\nabla r(\widehat{x}_{i,n})\nabla_\theta\overline{x}_i-\nabla r(\overline{x}_i)\nabla_\theta\overline{x}_i\right\|_2\right]\\
&\quad\le L_r\cdot\left(\mathbb{E}_{\overline{s}_i}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right]\right)+2L_r\cdot\widehat{K}(i).
\end{aligned}\tag{C.16}$$

Plugging (C.15), (C.16) into (C.7) and (C.6) gives us

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

where the second inequality follows from the results from Lemma C.2 and by plugging the definition of $\widehat{K}$ in (C.14). Since the analysis above considers batch size $N=1$, the bound of gradient variance v_t is established by dividing N, which concludes the proof. □

Lemma C.2. Denote $e:=\sup\mathbb{E}_{\overline{s}_0}[\|\mathrm{d}\widehat{s}_{0,n}/\mathrm{d}\theta-\mathrm{d}\overline{s}_0/\mathrm{d}\theta\|_2]$, which is a constant that only depends on the initial state distribution[1]. For any $i\ge1$ and the corresponding $\widehat{s}_{i,n}$, we have the following inequality results:

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

Proof. Firstly, we obtain from (C.9) that $\forall i\ge1$,

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

According to the triangle inequality, we continue with

$$\begin{aligned}
&\le\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{\mathrm{d}\overline{s}_{i-1}}{\mathrm{d}\theta}\right\|_2\right]\\
&\quad+\mathbb{E}\left[\left\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\right\|_2\right]\\
&\le2L_{\widehat{f}}\cdot\left(\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right)+L_{\widehat{f}}\cdot\mathbb{E}_{\overline{s}_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{s}_{i-1}}{\mathrm{d}\theta}\right\|_2\right]\\
&\quad+L_{\widehat{f}}\cdot\mathbb{E}_{\overline{a}_{i-1}}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\right\|_2\right].
\end{aligned}\tag{C.17}$$

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

Similarly, we have from (C.8) that

$$
\begin{aligned}
&\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}+\frac{\partial\widehat{a}_{i,n}}{\partial\theta}-\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}-\frac{\partial\overline{a}_i}{\partial\theta}\right\|_2\right] \\
&\quad\le\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\theta}-\frac{\partial\overline{a}_i}{\partial\theta}\right\|_2\right] \\
&\quad\le 2L_\pi\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\right\|\right]+L_\pi\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right]+2L_\theta. \qquad \text{(C.18)}
\end{aligned}
$$

Plugging (C.18) back to (C.17),

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

where the last inequality follows from the definition of $\widehat{K}$ in (C.14).

Applying this recursion gives us

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

where the first equality follows from (C.12).

As a consequence, we have from (C.18) that

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

This concludes the proof. □

C.3 Proof of Proposition 5.6

Proof. Different from the gradient variance where the analysis is solely based on the distribution of the approximated states, additional care must be taken when dealing with the gradient bias where the true value has recurrent dependencies on the timesteps.

In what follows, we first use similar techniques that show up in the previous section to upper-bound the decomposed reward terms in the gradient bias. Then we deal with the distribution mismatch problem between the gradient of the true value and the estimated one, specifically, the recursive structure of V^{π_θ} and the non-recursive value approximation $\widehat{V}_{\omega_t}$.

Step 1: Upper-bound the cumulative reward term in the gradient bias.

To begin with, we decompose the bias of the reward gradient at timestep i as follows:

$$
\begin{aligned}
&\mathbb{E}_{(s_i,a_i)\sim\mathbb{P}(s_i,a_i),(\widehat{s}_{i,n},\widehat{a}_{i,n})\sim\mathbb{P}(\widehat{s}_i,\widehat{a}_i)}\left[\left\|\frac{\mathrm{d}r(\widehat{x}_{i,n})}{\mathrm{d}\theta}-\frac{\mathrm{d}r(x_i)}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\mathrm{d}r}{\mathrm{d}\widehat{x}_{i,n}}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}x_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad\le\mathbb{E}\left[\left\|\frac{\mathrm{d}r}{\mathrm{d}\widehat{x}_{i,n}}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}x_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad\le 2L_r\cdot\widehat{K}(i)+L_r\cdot\left(\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\right\|_2\right]\right), \qquad \text{(C.19)}
\end{aligned}
$$

where $\mathbb{P}(s_i, a_i)$ and $\mathbb{P}(\widehat{s}_i, \widehat{a}_i)$ are defined in (4.4) with respect to $s_0 \sim \nu_\pi$, $\widehat{s}_0 \sim \nu_\pi$.

From (C.8) that is given by the chain rule,

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

By the triangle inequality and the Lipschitz assumption, it then follows that

$$\begin{aligned}
&\leq \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \widehat{a}_{i,n}}{\partial \theta} - \frac{\partial a_i}{\partial \theta}\right\|_2\right] \\
&\leq 2L_\pi \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\right\|_2\right] + L_\pi \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right] + 2L_\theta. \qquad \text{(C.20)}
\end{aligned}$$

Similarly, we have from (C.9) that

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

We proceed by applying the triangle inequality to extract the ϵ_f term defined in (4.4):

$$\begin{aligned}
&\leq \mathbb{E}\left[\left\|\frac{\partial \widehat{s}_{i,n}}{\partial \widehat{s}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial s_i}{\partial s_{i-1}} \cdot \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial s_i}{\partial s_{i-1}} \cdot \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial s_i}{\partial s_{i-1}} \cdot \frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial s_i}{\partial a_{i-1}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial s_i}{\partial a_{i-1}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial s_i}{\partial a_{i-1}} \cdot \frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\leq \epsilon_f \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\right\|_2 + \left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\right\|_2\right] + L_f \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad + L_f \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\right\|_2\right].
\end{aligned}$$

Combining with the result in (C.20), we have the following recursive expression:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right] &\lesssim (\epsilon_f + 2L_f L_\pi) \cdot \widehat{K}(i-1) + L_f \widetilde{L}_\pi \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\right\|_2\right] + 2L_f L_\theta \\
&= \left((\epsilon_f + 2L_f L_\pi) \cdot \widehat{K}(i-1) + 2L_f L_\theta\right) \cdot \sum_{j=0}^{i-1} L_f^j \widetilde{L}_\pi^j \\
&\leq \left((\epsilon_f + 2L_f L_\pi) \cdot \widehat{K}(i-1) + 2L_f L_\theta\right) \cdot i \cdot \widetilde{L}_f^i \widetilde{L}_\pi^i. \qquad \text{(C.21)}
\end{aligned}$$

where the inequality holds due to (C.12) and the fact that $s_0, \widehat{s}_{0,n}$ are sampled from the same initial distribution.

Plugging (C.21) into (C.20), we obtain

$$\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}a_i}{\mathrm{d}\theta}\right\|_2\right] \leq \left[(\epsilon_f + 2L_f L_\pi) \cdot \widehat{K}(i-1) + 2L_f L_\theta\right] \cdot i \cdot \widetilde{L}_f^i \widetilde{L}_\pi^{i+1} + 2L_\pi \widehat{K}(i) + 2L_\theta. \qquad \text{(C.22)}$$

Step 2: Deal with the recursive value function and the state distribution mismatch.

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

Now we are ready to bound the gradient bias. From Lemma C.4, we know that

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

For notation convenience, we denote $L_Q := L_r/(1-\gamma L_f(1+L_\pi))$ such that the state-action value function is L_Q-Lipschitz continuous (Rachelson & Lagoudakis, 2010; Pirotta et al., 2015).

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]\\
&\quad=\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]\\
&\quad\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.\\
&\qquad\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]\\
&\quad\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.14) and the definition of ϵ_v in (4.5).

Using the results in (C.19) and (C.23), we obtain

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

Plugging (C.21), (C.22), and $\widehat{K}$ in (C.14) into the above expression, we conclude the proof by obtaining

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

□

Lemma C.3. 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

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

where the first step follows from the 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 obtain

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

Iterating the above formula we obtain

$$\nabla_\theta V^{\pi_\theta}(s_h) = \mathbb{E}\Bigg[\int_{\mathcal{S}} \sum_{i=h}^{\infty} \gamma^{i-h} \cdot f\big(s_{i+1}|s_i, a_i\big) \cdot \Big(\frac{\partial Q^{\pi_\theta}}{\partial a_i} \cdot \frac{\mathrm{d}a_i}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_i} \cdot \frac{\mathrm{d}s_i}{\mathrm{d}\theta} \Big) \mathrm{d}s_{i+1} \Bigg]. \tag{C.24}$$

Define $\overline{\sigma}_2(s, a) = \sum_{i=h}^{\infty} \gamma^{i-h} \cdot \mathbb{P}(s_i = s, a_i = a)$, where $s_0 \sim \zeta$, $a_i \sim \pi(\cdot \,|\, s_i)$, and $s_{i+1} \sim f(\cdot \,|\, s_i, a_i)$. By definition we have

$$\sum_{i=0}^{h-1} \gamma^i \cdot \mathbb{P}(s_i = s, a_i = a) + \gamma^h \cdot \overline{\sigma}_1(s, a) = \sigma(s, a) = \sum_{i=0}^{h-1} \gamma^i \cdot \mathbb{P}(s_i = s, a_i = a) + \gamma^h \cdot \overline{\sigma}_2(s, a).$$

Therefore we have the equivalence $\overline{\sigma}_1(s, a) = \overline{\sigma}_2(s, a)$.

By taking the expectation over s_h in (C.24), we have the stated result

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

□

Lemma C.4. Recall that $\mu_\pi(s) := \beta \cdot \nu_\pi(s) + (1 - \beta) \cdot \zeta(s)$. The gradient of the h-step Model Value Expansion satisfies

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

Proof. To begin with, we decompose the gradient bias by

$$\begin{aligned}
b_t &= \Big\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \Big\|_2 \\
&= \Big\| \mathbb{E}\big[\nabla_\theta J(\pi_{\theta_t}) - \widehat{\nabla}_\theta J(\pi_{\theta_t})\big] \Big\|_2 \\
&= \left\| \mathbb{E}_{s_0 \sim \zeta, \widehat{s}_{0,n} \sim \mu_\pi}\left[\nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(s_i, a_i) + \gamma^h \cdot \nabla_\theta V^{\pi_\theta}(s_h) - \nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(\widehat{s}_{i,n}, \widehat{a}_{i,n}) - \gamma^h \cdot \nabla_\theta \widehat{V}_t(\widehat{s}_{h,n}) \right] \right\|_2
\end{aligned}$$

Now we notice the initial state distribution mismatch between s_0 and $\widehat{s}_{0,n}$, which leads to

$$
\begin{aligned}
&\leq \left\| \mathbb{E}_{s_0\sim\zeta, \widehat{s}_{0,n}\sim\mu_\pi} \left[\nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(s_i, a_i) - \nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(\widehat{s}_{i,n}, \widehat{a}_{i,n}) \right] \right\|_2 \\
&\quad + \left\| \mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1, \widehat{s}_{0,n}\sim\mu_\pi} \left[\gamma^h \cdot \frac{\partial Q^{\pi_\theta}}{\partial a_h} \cdot \frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h} \cdot \frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h \cdot \nabla_\theta \widehat{V}_t(\widehat{s}_{h,n}) \right] \right\|_2 \\
&\leq \mathbb{E}_{\widehat{s}_{0,n}\sim\mu_\pi} \left[\left\| \mathbb{E}_{s_0\sim\zeta} \left[\nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(s_i, a_i) - \nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(\widehat{s}_{i,n}, \widehat{a}_{i,n}) \right] \right\|_2 \right] \\
&\quad + \mathbb{E}_{\widehat{s}_{0,n}\sim\mu_\pi} \left[\left\| \mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1} \left[\gamma^h \cdot \frac{\partial Q^{\pi_\theta}}{\partial a_h} \cdot \frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h} \cdot \frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \gamma^h \cdot \nabla_\theta \widehat{V}_t(\widehat{s}_{h,n}) \right] \right\|_2 \right],
\end{aligned}
\tag{C.25}
$$

where we plug in the result from Lemma C.3 in the first inequality and the second inequality follows from $\|\mathbb{E}[\cdot]\|_2 \leq \mathbb{E}[\|\cdot\|_2]$.

For $\mu_\pi(s) = \beta \cdot \nu_\pi(s) + (1-\beta)\cdot\zeta(s)$, let Z be the random variable satisfying $\mathbb{P}(Z=0) = \beta$ and $\mathbb{P}(Z=1) = 1-\beta$, i.e., the event $Z=0$ and $Z=1$ corresponds to that the state s is sampled from ν_π and ζ, respectively. Therefore, for any random variable Y, following the law of total expectation, we know that

$$
\begin{aligned}
\mathbb{E}_{\mu_\pi}[Y] &= \mathbb{E}[\mathbb{E}[Y|Z]] = \mathbb{E}[Y|Z=0]\mathbb{P}(Z=0) + \mathbb{E}[Y|Z=1]\mathbb{P}(Z=1) \\
&= \beta\mathbb{E}[Y|Z=0] + (1-\beta)\mathbb{E}[Y|Z=1] \\
&= \beta\mathbb{E}_{\nu_\pi}[Y] + (1-\beta)\mathbb{E}_{\zeta}[Y].
\end{aligned}
\tag{C.26}
$$

We separately upper-bound the two terms on the right-hand side of (C.25) as follows:

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

where the equality follows from (C.26).

Following the definition of κ, we can proceed to bound the above terms by

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

Similarly, for the critic bias, we have

$$
\begin{aligned}
&\mathbb{E}_{\widehat{s}_{0,n}\sim\mu_\pi}\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=\beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu_\pi}\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]\\
&\qquad+(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]\\
&\quad\le\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu_\pi}\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=\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}
$$

Combining the above inequalities completes the proof. □

C.4 Proof of Proposition 5.7

Proof. To find the optimal unroll length h^* that minimizes the convergence rate upper bound, we define $g(h)$ as follows:

$$g(h):=c\cdot(2\delta\cdot b'_t+\frac{\eta}{2}\cdot v'_t)+b'^2_t+v'_t.$$

where v'_t and b'_t are the terms in the variance, bias bound (i.e., (5.1) and (5.2)) when L_f, $L_{\widehat{f}}$, and L_π are less than or equal to 1. Formally, $v'_t:=h^6/N+\gamma^{2h}h^4/N$ and $b'_t:=\kappa'\kappa h^3\epsilon_f+\kappa' h^2\gamma^h\epsilon_v$.

The problem is to find h^* given by $h^*=\mathrm{argmin}_h\, g(h)$. By plugging in the bound of the gradient bias, variance and setting the derivative to zero, we obtain

$$\frac{\partial}{\partial h}g(h)=O\big((h^5+h^4\gamma^{2h}\cdot\log\gamma)/N+h\cdot\epsilon_f^2+h^4\gamma^{2h}\log\gamma\cdot\epsilon_v^2\big)=0,$$

where we omit the terms that are not dependant on h, γ, ϵ_f, and ϵ_v.

For notation simplicity, we define

$$\frac{\partial}{\partial h}g_1(h):=(h^5+h^4\gamma^{2h}\cdot\log\gamma)/N,\quad\frac{\partial}{\partial h}g_2(h):=h^5\cdot\epsilon_f^2+h^4\gamma^{2h}\log\gamma\cdot\epsilon_v^2.$$

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 the unique optima in the domain $h\in(0,\infty)$, which we denote as h_1 and h_2. By setting $\frac{\partial}{\partial h}g_1(h)=0$, we have

$$h_1=\gamma^{2h_1}\cdot\log\frac{1}{\gamma}.$$

Taking the natural logarithm on both sides gives us

$$\log h_1=2h_1\log\gamma+\log\log\frac{1}{\gamma}.$$

We obtain by rearranging terms that

$$\log h_1-2h_1\log\gamma=\log\log\frac{1}{\gamma}.$$

By taking exponential on both sides, we have

$$h_1\cdot\exp\left(-2h_1\cdot\log\gamma\right)=\log\frac{1}{\gamma}.$$

Therefore, by multiplying $-2\log\gamma$ on both sides, it follows that

$$-2h_1\log\gamma\cdot\exp\left(-2h_1\cdot\log\gamma\right)=2(\log\gamma)^2.$$

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

$$-2h_1\cdot\log\gamma=W\Big(2(\log\gamma)^2\Big).$$

We now obtain the following form of the unique optima of $g_1(h)$:

$$h_1 = \frac{1}{2\log(1/\gamma)} \cdot W\Big(2(\log\gamma)^2\Big).$$

To find the optima h_2 of $g_2(h)$, the problem is reduced to solve $h_2^2\epsilon_f + h_2^2\gamma^{h_2}\log\gamma\epsilon_v = 0$.

By basic algebra, we get the solution

$$h_2 = O\big(\log\big((\epsilon_v/\epsilon_f)(1/\log(1/\gamma))\big)\big).$$

By omitting the unnecessary terms in h_1, we conclude our proof by writing the optimal model expansion step as

$$h^* = O\Big(\frac{1}{\log(1/\gamma)} W\big(2(\log\gamma)^2\big)/N + \log\big((\epsilon_v/\epsilon_f)(1/\log(1/\gamma))\big)\Big).$$

□

C.5 Proof of Corollary 5.9

Proof. We let $\eta = 1/\sqrt{T}$ 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

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

This concludes the proof. □

D Experimental Details

D.1 Implementations and Comparisons with More RL Baselines

In order to better understand the potential of model-based RP PGMs, we compare them with additional policy gradient baselines in several MuJoCo tasks. The results are shown in Figure 7. For the model-based baseline Model-Based Policy Optimization (MBPO) (Janner et al., 2019), we use the implementation in the Mbrl-lib (Pineda et al., 2021). For all other model-free baselines, we use the implementations in Tianshou (Weng et al., 2021) that has state-or-the-art results.

We observe that the RP-DP has competitive performance in all the evaluation tasks compared to the popular baselines, suggesting the importance of studying model-based RP PGMs. In experiments, we implement RP-DR as the on-policy SVG(1) (Heess et al., 2015). We observe that the training can be unstable when using the off-policy SVG implementation, which require a carefully chosen policy update rate as well as a proper size of the experience replay buffer. This is because that when the learning rate is large, the magnitude of the inferred policy noise (from the previous data samples in the experience replay) can be huge. Implementing an on-policy version of RP-DR can avoid such a issue, following Heess et al. (2015). This, however, can degrade the performance of RP-DR compared to the off-policy RP-DP algorithm in several tasks. We conjecture that implementing the off-policy version of RP-DR can boost its performance, which requires techniques to stabilize training and we leave it as future work. For RP-DP, we implement it as Model-Augmented Actor-Critic (MAAC) (Clavera et al., 2020) with entropy regularization (Haarnoja et al., 2018), as suggested by Amos et al. (2021). RP(0) represents setting $h = 0$ in the RP PGM formulas (Amos et al., 2021), which is a model-free algorithm that is a stochastic counterpart of deterministic policy gradients.

For model-free baselines, we compare with Likelihood Ratio (LR) policy gradient methods (c.f. (2.2)), including REINFORCE (Sutton et al., 1999), Natural Policy Gradient (NPG) (Kakade, 2001),

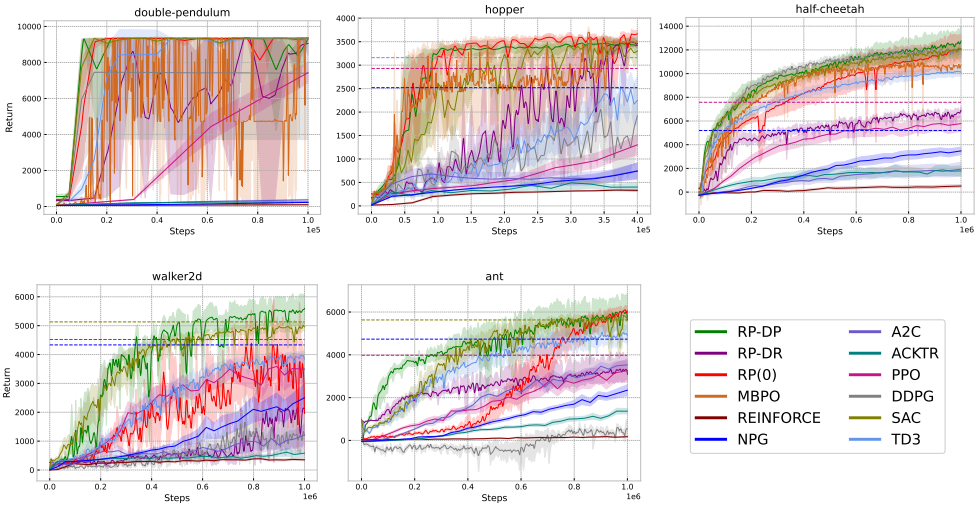


Figure 7: Full comparison results in the MuJoCo tasks. The dashed lines represent the value at convergence of the corresponding algorithms.

Advantage Actor Critic (A2C), Actor Critic using Kronecker-Factored Trust Region (ACKTR) (Wu et al., 2017), and Proximal Policy Optimization (PPO) (Schulman et al., 2017). We also evaluate algorithms that are built upon DDPG (Lillicrap et al., 2015), including Soft Actor-Critic (SAC) (Haarnoja et al., 2018) and Twin Delayed Deep Deterministic policy gradient (TD3) (Fujimoto et al., 2018).

D.2 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 $L_{\hat{f}} \leq \prod_i \sigma_{\max}(W_i)$. 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 upper-bounded by 1. For this reason, spectral normalization is a natural choice for smoothness regularization in model-based RP policy gradient methods, based on our analysis in Section 5.

In experiments, we use Multilayer Perceptrons (MLPs) for the critic, policy, and model. Besides, we adopt Gaussian models and policies for the stochasticity. To test the benefit of smooth function approximations for the RP-DP algorithm, 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 an out-of-the-shelf implementation of spectral normalization. Thus, applying SN to the MLP is pretty simple and no additional lines of code are needed. Specifically, we only need to import and apply SN to each layer:

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

D.3 Ablation on Spectral Normalization and Unroll Lengths

In this part, we conduct ablation studies on the choices of model unroll lengths and the effect of spectral normalization. We select hopper and walker2d in MuJoCo as the representative tasks. The results are shown in Figure 8. We find that the benefit of spectral normalization mainly resides in the following two aspects.

Firstly, spectral normalization avoids the chaos underlying the long chains of nonlinear mappings and allows longer model unroll lengths. It is obvious from Figure 8 that vanilla model-based RP PGMs without SN fail to provide satisfying results when the model unroll length h is large, while SN-based RP PGMs are still able to achieve competitive or even better performance when increasing h. This is important for algorithmic designs in practice: the most popular model-based RP policy gradient algorithms such as Clavera et al. (2020) and Amos et al. (2021), often rely on a carefully chosen (small) model unroll length (e.g. $h = 3$ in their implementations). When the model is learned well enough, a small h may not fully leverage the accurate gradient information provided by the model. As an evidence, approaches (Xu et al., 2022; Mora et al., 2021) based on differentiable simulators, where the true gradient is explicitly available, typically use longer length of simulator unrolls compared to model-based approaches. Therefore, a fruitful future direction is to learn models that are able to more accurately predict states and derivatives in multiple steps (thus more accurate gradient estimation), which along with spectral normalization can enable more efficient learning and achieve higher asymptotic value.

Secondly, applying spectral normalization to model-based RP PGMs provides more robustness. This is because that the training return is not very sensitive to the model unroll length and the variance is significantly smaller compared to the vanilla implementation when h is large.

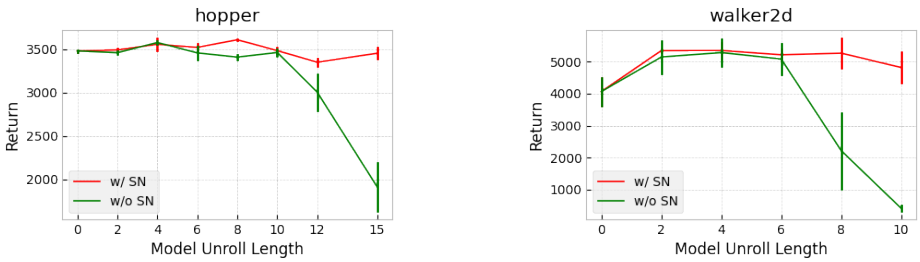


Figure 8: Ablation on the model unroll length and on the effect of Spectral Normalization (SN).

D.4 Ablation on Different Model Learners

Our main theoretical results in Section 5 depend on the model error defined in (4.4), which, however, cannot directly serve as the model training objective. For this reason, we evaluate different model learners: single- and multi-step (h-step) state prediction models, as well as multi-step predictive models integrated with the directional derivative error (Li et al., 2021). The results are reported in Figure 9. We observe that enlarging the prediction steps benefits training. The algorithm also converges faster in walker2d when considering derivative error, which approximately minimizes 4.4 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.

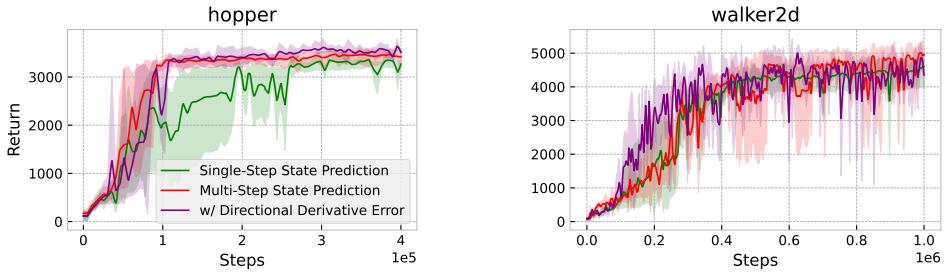


Figure 9: Ablation on different model learners: single-step and multi-step state prediction models, and multi-step state prediction models trained with an additional directional derivative error.

D.5 FIGURES IN THE MAIN TEXT IN LARGER SIZES

Here, we provide the identical figures that are larger in size. Figure 10, 11, 12 correspond to Figure 4, 5, 6 in the main text, respectively.

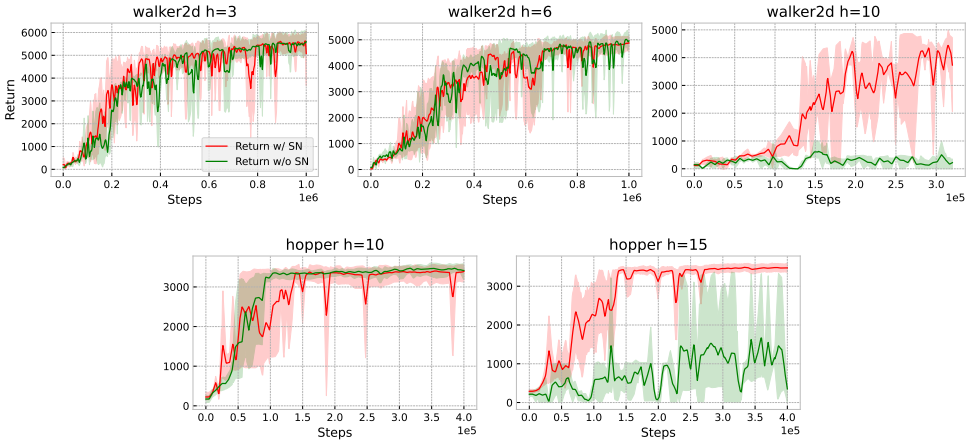


Figure 10: Performance of model-based RP PG methods with and without spectral normalization.

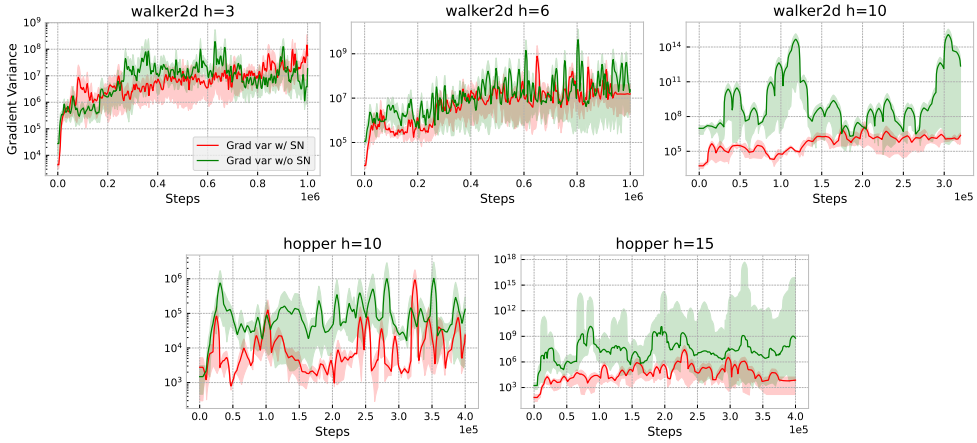


Figure 11: Gradient variance of model-based RP PG methods with and without spectral normalization.

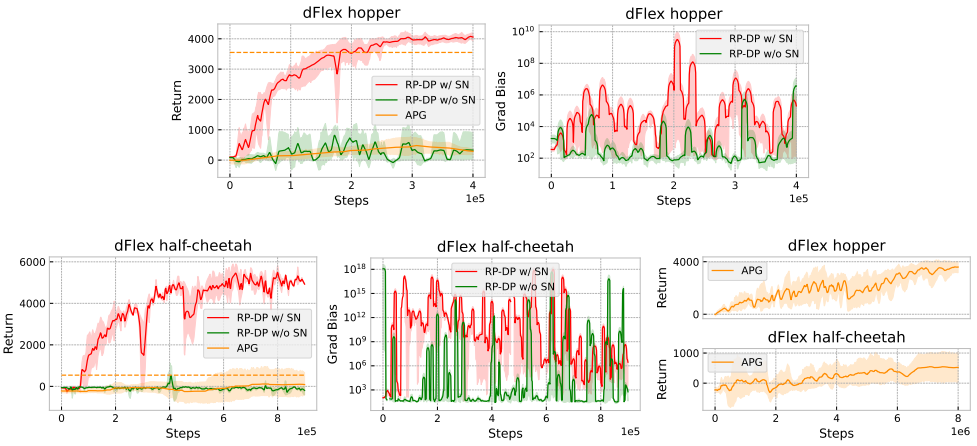


Figure 12: Performance and gradient bias in differentiable simulation. The rightmost column is the full curves of APG, which needs 20 times more steps in hopper to reach a comparable return with RP-DP-SN.