Model-Based Reparameterization Policy Gradient Methods: Theory and Practical Algorithms

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

Abstract

ReParameterization (RP) Policy Gradient Methods (PGMs) have been widely adopted for continuous control tasks in robotics and computer graphics. However, recent studies have revealed that when applied to long-term reinforcement learning problems, model-based RP PGMs may experience chaotic and non-smooth optimization landscapes with exploding gradient variance, leading to slow convergence. This is in contrast to the conventional belief that reparameterization methods have low gradient estimation variance in problems such as training deep generative models. To comprehend this phenomenon, we conduct a theoretical examination of model-based RP PGMs and search for solutions to the optimization difficulties. Specifically, we establish a convergence analysis for model-based RP PGMs and pinpoint the smoothness of function approximators as a major factor that affects gradient estimation quality. Based on our analysis, we propose a spectral normalization method to mitigate the exploding variance caused by long model unrolls. Our experiments support our theory and method, demonstrating that proper normalization can significantly reduce the gradient variance of model-based RP PGMs and improve their convergence, resulting in performance that is comparable or superior to other gradient estimators, such as the Likelihood Ratio (LR) gradient estimator.

1 Introduction

Reinforcement Learning (RL) has seen tremendous success in a variety of sequential decision-making applications, such as strategy games (Silver et al., 2017; Vinyals et al., 2019) and robotics (Duan et al., 2016; Wang et al., 2019b), by identifying actions that maximize long-term accumulated rewards. As one of the most popular methodologies, the policy gradient methods (PGM) (Sutton et al., 1999; Kakade, 2001; Silver et al., 2014) seek to search for the optimal policy by iteratively computing and following a stochastic gradient direction with respect to the policy parameters. Therefore, the quality of the stochastic gradient estimation is essential for the effectiveness of PGMs.

Two main categories have emerged in the realm of stochastic gradient estimation: (1) Likelihood Ratio (LR) estimators, which perform zeroth-order estimation through the sampling of function evaluations (Williams, 1992; Konda & Tsitsiklis, 1999; Kakade, 2001); (2) ReParameterization (RP) gradient estimators, which harness 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).

Despite the widespread use of both LR and RP PGMs in practical applications, the majority of research on the theoretical properties of PGMs has centered on LR PGMs: Their global convergence analysis has been established under various settings, and the estimation quality of the LR gradient estimators has been heavily investigated (Agarwal et al., 2021; Wang et al., 2019a; Bhandari & Russo, 2019). Conversely, the theoretical underpinnings of RP PGMs have yet to be fully explored, with a dearth of research on the properties of RP policy gradient estimators and no established convergence analysis.

RP gradient estimators have established themselves as a reliable technique for training deep generative models such as variational autoencoders (Figurnov et al., 2018). From a stochastic optimization perspective, previous studies (Ruiz et al., 2016; Mohamed et al., 2020) have shown that RP gradient methods enjoy small variance, which leads to better convergence and performance. However, recent research (Parmas et al., 2018; Metz et al., 2021) has reported an opposite observation: When applied to long-horizon reinforcement learning problems, model-based RP PGMs tend to encounter chaotic and non-smooth optimization landscapes with exploding gradient variance, causing slow convergence.

Such an intriguing phenomenon inspires us to delve deeper into the theoretical properties of RP gradient estimators in search of a remedy for the issue of exploding gradient vari-

[1]Georgia Tech, Atlanta, GA, USA [2]Northwestern University, Evanston, IL, USA.

ance in model-based RP PGMs. To this end, we develop a unified theoretical framework for the examination of model-based RP PGMs and establish their convergence results. Our analysis implies that the smoothness and accuracy of the learned model are crucial determinants of the exploding variance of RP gradients: (1) Both the gradient variance and bias exhibit a polynomial dependence on the Lipschitz continuity of the learned model and policy w.r.t. the input state, with degrees that increase linearly with the steps of model value expansion; (2) The bias also depends on the error of the estimated model and value.

Our findings suggest that imposing smoothness on the model and policy can greatly decrease the variance of RP gradient estimators. To put this discovery into practice, we propose a spectral normalization method to enforce the smoothness of the learned model and policy. It's worth noting that this method can enhance the algorithm's efficiency without substantially compromising accuracy when the underlying transition kernel is smooth. However, if the transition kernel is not smooth, enforcing smoothness may lead to increased error in the learned model and introduce bias. In such cases, a balance should be struck between model bias and gradient variance. Nonetheless, our empirical study demonstrates that the reduced gradient variance when applying spectral normalization leads to a significant performance boost, even with the cost of a higher bias. Furthermore, our results highlight the potential of investigating model-based RP PGMs, as they demonstrate superiority over other model-based and Likelihood Ratio (LR) gradient estimator alternatives.

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 $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$

$$V^\pi(s) = (1-\gamma) \cdot \mathbb{E}_{\pi,f}\Bigg[\sum_{i=0}^{\infty} \gamma^i \cdot r(s_i,a_i) \,\Bigg|\, s_0 = s\Bigg],$$

$$Q^\pi(s,a) = (1-\gamma) \cdot \mathbb{E}_{\pi,f}\Bigg[\sum_{i=0}^{\infty} \gamma^i \cdot r(s_i,a_i) \,\Bigg|\, s_0 = s, a_0 = a\Bigg],$$

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 and state-action visitation measure $\nu_\pi(s)$ over $\mathcal{S}$ and $\sigma_\pi(s,a)$ over $\mathcal{S} \times \mathcal{A}$ are defined respectively as

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

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

Here the summations are based on the trajectory following $s_0 \sim \zeta$, $a_i \sim \pi(\cdot \,|\, s_i)$, 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 underlying problem of policy gradient, i.e., computing the gradient of a probabilistic objective w.r.t. the parameters of the 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.

Likelihood Ratio (LR) Gradient: By leveraging the *score function*, LR gradients only require samples of the function values. Since $\nabla_\theta \log p(x;\theta) = \nabla_\theta p(x;\theta)/p(x;\theta)$, the LR gradient is

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

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:

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

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

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

3 Analytic Reparameterization Gradient in Reinforcement Learning

In this section, we present two fundamental *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 potential obstacles that may arise 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 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:

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

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

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

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 PVG, we have the following dynamical representation of the policy gradient.

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. 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:

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

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

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

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

There remain challenges when developing practical algorithms: (1) The above formulas require the gradient information of the dynamics function f. In this work, however, we consider a common RL setting where 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; (2) 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 (Bollt, 2000) 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

Through the application of Model Value Expansion (MVE) for model truncation, this section unveils two RP policy gradient frameworks constructed upon MVE.

4.1 h-Step Model Value Expansion

To combat the difficulties inherent in full unrolls, many algorithms employ direct truncation, where the longer sequence is broken down into shorter sub-sequences and backpropagation is applied accordingly (e.g., Truncated BPTT, (Werbos, 1990)). However, this simplistic scheme over-prioritize short-term dependencies, leading to biased gradients.

In the realm of model-based RL, one viable approach is to employ the h-step Model Value Expansion (Feinberg et al., 2018), which decomposes the value estimation $\widehat{V}^\pi(s)$ into the rewards gleaned from the learned model and the tail estimated by a critic $\widehat{Q}_\omega$:

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

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

4.2 Model-Based RP Gradient Estimation

Utilizing the pathwise gradient with respect to θ, we are able to establish the following two frameworks.

Model Derivatives on Predictions. A straightforward way to compute the first-order gradient is to link together the reward, model, policy, critic, and backpropagate through them. Specifically, the differentiation is carried out on the trajectories simulated by the model, which serves as a tool for *both* state prediction and derivative evaluation. The RP estimator of gradient $\nabla_\theta J(\pi_\theta)$ is represented in the form of

$$\begin{aligned}\widehat{\nabla}_\theta^{\text{DP}} J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^N \nabla_\theta\bigg(&\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n}) \\ &+ \gamma^h\cdot\widehat{Q}_\omega(\widehat{s}_{h,n},\widehat{a}_{h,n})\bigg),\end{aligned} \tag{4.1}$$

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 noise $\varsigma_n \sim p(\varsigma)$, $\xi_n \sim p(\xi)$. Here

μ_π is the distribution where the initial states of the simulated trajectories are sampled. We show in §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. An alternative approach is to use the learned differentiable model solely for the calculation of derivatives, with the aid of Monte-Carlo estimates obtained from *real* samples. By replacing the ∇f terms in (3.2), (3.3) with $\nabla \widehat{f}$ and setting the termination of backpropagation at the h-th step as $\widehat{\nabla} V^\pi(\widehat{s}_{h,n}) = \nabla \widehat{V}_\omega(\widehat{s}_{h,n})$, we are able to derive a dynamic representation of $\widehat{\nabla}_\theta V^\pi$, which we defer to §A.

The corresponding RP gradient estimator takes the form of

$$\widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta) = \frac{1}{N} \sum_{n=1}^{N} \widehat{\nabla}_\theta V^\pi(\widehat{s}_{0,n}), \tag{4.2}$$

where $\widehat{s}_{0,n} \sim \mu_\pi$. Equation (4.2) can be specified as (A.9), which is in the same format as (4.1), but with the noise variables ς_n, ξ_n inferred from the real data sample (s_i, a_i, s_{i+1}) via the relation $a_i = \pi(s_i, \varsigma_n)$ and $s_{i+1} = \widehat{f}(s_i, a_i, \xi_n)$ (see §A for details). Algorithms such as SVG (Heess et al., 2015) and its variants (Abbeel et al., 2006; Atkeson, 2012) are examples of this method.

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, the policy, model, and critic are updated at each iteration $t \in [T]$, generating sequences of $\{\pi_{\theta_t}\}_{t\in[T+1]}$, $\{\widehat{f}_{\psi_t}\}_{t\in[T]}$, and $\{\widehat{Q}_{\omega_t}\}_{t\in[T]}$, respectively.

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

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

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

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, accurate state predictions do not imply accurate RP gradient estimation. Thus, we introduce the notation

Algorithm 1 Model-Based RP Policy Gradient Method

Input: Number of iterations T, learning rate η, batch size N, state distribution $\mu(\cdot)$

for iteration $t \in [T]$ **do**
  Update the model parameter ψ_t by MSE or MLE
  Update the critic parameter ω_t by performing TD
  Sample states from μ_{π_t} and estimate $\widehat{\nabla}_\theta J(\pi_{\theta_t}) = \widehat{\nabla}_\theta^{\mathrm{DP}} J(\pi_{\theta_t})$ (4.1) or $\widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_{\theta_t})$ (4.2)
  Update the policy parameter θ_t by $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \widehat{\nabla}_\theta J(\pi_{\theta_t})$ and execute $\pi_{\theta_{t+1}}$
end for
Output: $\{\pi_{\theta_t}\}_{t\in[T]}$

$\epsilon_f(t)$ to represent the model (gradient) error at iteration t:

$$\epsilon_f(t) := \max_{i\in[h]} \mathbb{E}_{\mathbb{P}(s_i,a_i),\mathbb{P}(\widehat{s}_i,\widehat{a}_i)} \bigg[\bigg\| \frac{\partial s_i}{\partial s_{i-1}} - \frac{\partial \widehat{s}_i}{\partial \widehat{s}_{i-1}} \bigg\|_2 + \bigg\| \frac{\partial s_i}{\partial a_{i-1}} - \frac{\partial \widehat{s}_i}{\partial \widehat{a}_{i-1}} \bigg\|_2 \bigg], \tag{4.3}$$

where $\mathbb{P}(s_i, a_i)$ is the true state-action distribution at the i-th timestep by following $s_0 \sim \nu_{\pi_t}$, $a_j \sim \pi_t(\cdot \,|\, s_j)$, $s_{j+1} \sim f(\cdot \,|\, s_j, a_j)$, with policy and transition noise sampled from a fixed distribution. Similarly, $\mathbb{P}(\widehat{s}_i, \widehat{a}_i)$ is the model rollout distribution at the i-th timestep by following $\widehat{s}_0 \sim \nu_{\pi_t}$, $\widehat{a}_j \sim \pi_t(\cdot \,|\, \widehat{s}_j)$, $\widehat{s}_{j+1} \sim \widehat{f}(\cdot \,|\, \widehat{s}_j, \widehat{a}_j)$, where the noise is sampled when $\widehat{\nabla}_\theta J(\pi_\theta) = \widehat{\nabla}_\theta^{\mathrm{DP}} J(\pi_\theta)$, and is inferred when $\widehat{\nabla}_\theta J(\pi_\theta) = \widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta)$ (in this case $\mathbb{P}(\widehat{s}_i, \widehat{a}_i) = \mathbb{P}(s_i, a_i)$).

In model-based RL, it is common to learn a state-predictive model that can make multi-step predictions. However, this presents a challenge in reconciling the discrepancy between minimizing state prediction error and the gradient error of the model. Although it is natural to consider regularizing the models' directional derivatives to be consistent with the samples (Li et al., 2021), we contend that the use of state-predictive models does *not* cripple our analysis of gradient bias 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. This same argument can also be applied to the case of learning a critic through 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 $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ as

$$\mathcal{T}^\pi Q(s, a) = \mathbb{E}_{\pi,f}\big[(1-\gamma) \cdot r(s, a) + \gamma \cdot Q(s', a')\big].$$

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 at the t-th iteration as

$$\epsilon_v(t) := \alpha^2 \cdot \mathbb{E}_{\mathbb{P}(s_h,a_h),\mathbb{P}(\widehat{s}_h,\widehat{a}_h)}\Big[\Big\|\frac{\partial Q^{\pi_t}}{\partial s} - \frac{\partial \widehat{Q}_t}{\partial \widehat{s}}\Big\|_2 + \Big\|\frac{\partial Q^{\pi_t}}{\partial a} - \frac{\partial \widehat{Q}_t}{\partial \widehat{a}}\Big\|_2\Big], \tag{4.4}$$

where $\alpha := (1-\gamma)/\gamma^h$ and $\mathbb{P}(s_h, a_h)$, $\mathbb{P}(\widehat{s}_h, \widehat{a}_h)$ are distributions at timestep h with the same definition as in (4.3). The inclusion of α^2 ensures that the critic error remains in alignment with the single-step model error ϵ_f: (1) The critic estimates the tail terms that occur after h steps in the model expansion, therefore the step-average critic error should be inversely proportional to the tail discount summation $\sum_{i=h}^{\infty}\gamma^i = 1/\alpha$; (2) The quadratic form shares similarities with the canonical MBRL analysis – the cumulative error of the model trajectories scales linearly with the single-step prediction error and quadratically with the considered horizon (i.e. tail after the h-th step). This is because the cumulative error is linear in the considered horizon and the maximum state discrepancy, which is linear in the single-step error and, again, the horizon (Janner et al., 2019).

5 Main Results

We present our main theoretical results in this section, with the proofs deferred to §C. Specifically, we establish the convergence of model-based RP PGMs and, more importantly, study the correlation between the convergence rate, gradient bias, variance, and the model's smoothness and approximation error. Utilizing our theory, we suggest various algorithmic designs for model-based RP PGMs.

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 analyses 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 $(s,a) \in \mathcal{S}\times\mathcal{A}$:

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

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

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

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

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

Proposition 5.2 illustrates the interdependence between the convergence and the variance, bias of the gradient estimators. In order for model-based RP PGMs to converge, it is imperative to maintain both the variance and bias at sublinear growth rates. Prior to examining the upper bound of b_t and v_t, we make the following Lipschitz assumption, which has been implemented in a plethora of preceding studies (Pirotta et al., 2015; Clavera et al., 2020; Li et al., 2021).

Assumption 5.3 (Lipschitz Continuity)**.** We assume that $r(s,a)$, $f(s,a,\xi^*)$, $\widehat{f}_\psi(s,a,\xi)$, $\pi_\theta(s,\varsigma)$, $\widehat{Q}_\omega(s,a)$ are $L_r, L_f, L_{\widehat{f}}, L_\pi, L_{\widehat{Q}}$ Lipschitz continuous (details in §B).

Denote $\widetilde{L}_g := \max\{L_g, 1\}$ for any function g. We have the following result for the gradient variance.

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

$$v_t = O\Big(h^4\Big(\frac{1-\gamma^h}{1-\gamma}\Big)^2 \widetilde{L}_{\widehat{f}}^{4h}\widetilde{L}_\pi^{4h}/N + \gamma^{2h}h^4\widetilde{L}_{\widehat{f}}^{4h}\widetilde{L}_\pi^{4h}/N\Big).$$

We observe that the variance upper bound exhibits a polynomial dependence on the Lipschitz continuity of the model and policy, where the degrees are linear in the model unroll length. This makes sense intuitively, as the transition can be highly chaotic when $L_{\widehat{f}} > 1$ and $L_\pi > 1$. This can result in diverging trajectories and variable gradient directions during training, leading to significant variance in the gradients.

Remark 5.5. Model-based RP PGMs with non-smooth models and policies can suffer from large variance and highly non-smooth loss landscapes, which can lead to slow convergence or failure during 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 to avoid exploding gradient variance. See our discussion at the end of this section for more details.

Model-based RP PGMs possess unique advantages by utilizing proxy models for variance reduction. By enforcing the smoothness of the model, the gradient variance is reduced without a burden when the underlying transition is smooth. However, in cases of non-smooth dynamics, such as contact-rich tasks (Suh et al., 2022a; Pang et al., 2022), doing so may introduce additional bias due to increased model estimation error. This necessitates a trade-off between the model error and gradient variance. Nonetheless, our empirical study demonstrates that smoothness regularization improves performance, despite the cost of increased bias.

Next, we study the gradient bias. We consider the case where μ_π, the state distribution when estimating the RP gradient, is a mixture of the MDP initial distribution ζ and the state visitation ν_π, such that $\mu_\pi = \beta \cdot \nu_\pi + (1-\beta)\cdot\zeta$, where $\beta \in [0,1]$. This form is of particular interest as it encompasses various state sampling schemes that can be employed, such as when $h=0$ and $h\to\infty$: When not utilizing a model, such as in SVG(0) (Heess et al., 2015; Amos et al., 2021) and DDPG (Lillicrap et al., 2015), states are sampled from ν_π; while when unrolling the model over full sequences, as in BPTT, states are sampled from the initial distribution.

Given that the effects of policy actions extend to 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. Hence, differentiating through a single critic function requires extra attention, as the true value gradient has recursive structures. To tackle this issue, we provide the gradient bias bound below that is based on the measure of discrepancy between the initial distribution ζ and the state visitation ν_π.

Proposition 5.6 (Gradient Bias)**.** We 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

$$b_t \le O\Big(\kappa\kappa' h^2 \frac{1-\gamma^h}{1-\gamma}\widetilde{L}_{\widehat{f}}^h \widetilde{L}_f^h \widetilde{L}_\pi^{2h}\epsilon_{f,t} + \kappa' h\gamma^h\Big(\frac{\gamma^h}{1-\gamma}\Big)^2 \widetilde{L}_{\widehat{f}}^h \widetilde{L}_\pi^h \epsilon_{v,t}\Big),$$

where $\epsilon_{f,t}$ and $\epsilon_{v,t}$ is the shorthand notation of $\epsilon_f(t)$ and $\epsilon_v(t)$, defined in (4.3) and (4.4), respectively.

The analysis above yields the identification of an optimal model expansion step h^* that achieves the best convergence rate, whose form is presented by the following proposition.

Proposition 5.7 (Optimal Model Expansion Step)**.** Given $L_f \le 1$, if we regularize the model and policy so that $L_{\widehat{f}} \le 1$ and $L_\pi \le 1$, then when $\gamma\approx 1$, the optimal model expansion step h^* at iteration t that minimizes the convergence rate upper bound satisfies $h^* = \max\{h'^*, 0\}$, where $h'^* = O\big(\epsilon_{v,t}/((1-\gamma)(\epsilon_{f,t}+\epsilon_{v,t}))\big)$ scales linearly with $\epsilon_{v,t}/(\epsilon_{f,t}+\epsilon_{v,t})$ and the effective task horizon $1/(1-\gamma)$.

In Proposition 5.7, the Lipschitz condition of the underlying dynamics, i.e. $L_f \le 1$, ensures the stability of the system. This can be seen in the linear system example, where the transitions are determined by the eigenspectrum of the family of transformations, leading to exponential divergence of trajectories w.r.t. the largest eigenvalue. In cases where this condition is not met in practical control systems, finding the best model unroll length may require trial and error. Fortunately, we have observed through experimentation that enforcing smoothness offers a wider range of unrolling lengths that still provide satisfactory results.

Remark 5.8. Our analysis reveals that as the error scale $\epsilon_{v,t}/(\epsilon_{f,t}+\epsilon_{v,t})$ increases, so too does the value of h^*. This finding can inform the practical algorithms to rely more on the model by performing longer unrolls when the model error $\epsilon_{f,t}$ is small; while avoiding long unrolls when the critic error $\epsilon_{v,t}$ is small.

Finally, we characterize the algorithm convergence rate.

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 clarified by determining how quickly the errors of model and critic approach zero, i.e., $\sum_{t=0}^{T-1}\epsilon_f(t)+\epsilon_v(t)$. Such results can be accomplished by conducting a more fine-grained investigation of the model and critic function classes, such as utilizing overparameterized neural nets with width scaling with T to bound the training error, as done in (Cai et al., 2019; Liu et al., 2019), and incorporating complexity measures of the model and critic function classes to bound $\epsilon_f(t), \epsilon_v(t)$. This, however, is beyond the scope of this paper.

A Spectral Normalization Method. To ensure a smooth transition and faster convergence, we propose using a Spectral Normalization (SN) (Miyato et al., 2018) model-based RP PGM that applies SN to all layers of the deep model network and policy network. While other techniques, such as adversarial regularization (Shen et al., 2020), exist, we focus primarily on SN as it directly regulates the Lipschitz constant of the function. Specifically, 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 $\sigma_{\max}(W) := \max_{\|x\|_2\le 1}\|Wx\|_2$. For neural network f with linear layers $g(x) = W_i x$ and 1-Lipschitz activation (e.g. ReLU and leaky ReLU), we have $L_g = \sigma_{\max}(W_i)$ and $L_{\widehat{f}} \le \prod_i \sigma_{\max}(W_i)$. By normalizing the spectral norm of W_i with $W_i^{\mathrm{SN}} := W_i/\sigma_{\max}(W_i)$, SN guarantees that the Lipschitz of f is upper-bounded by 1.

6 Related Work

Policy Gradient Methods. Within the RL field, 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 works (Agarwal et al., 2021; Wang et al., 2019a; Bhandari & Russo, 2019; Liu et al., 2019) have 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 impact of value approximation, which oversimplifies 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 visitation analysis.

Differentiable Simulation. This paper delves into the model-based setting, where a model that captures the transition of an MDP is employed to train a control policy. Recent approaches (Mora et al., 2021; Suh et al., 2022a;b; 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. actions are explicitly given. To deal with the discontinuities and empirical bias phenomenon in the differentiable simulation caused by contact dynamics, previous works proposed using penalty-based contact formulations (Geilinger et al., 2020; Xu et al., 2021) or adopting bundled gradient with randomized smoothing (Suh et al., 2022a;b). However, these are not in direct comparison to our analysis, which relies on model function approximators.

7 Experiments

7.1 Instantiations and Comparisons of RP PGMs

We begin by evaluating several algorithms originating from the RP policy gradient methods in several MuJoCo (Todorov et al., 2012) tasks. We use RP-DP and RP-DR to distinguish whether the model derivatives are calculated on predictions (4.1) or on real samples (4.2). Specifically, RP-DP is implemented as MAAC (Clavera et al., 2020) with entropy regularization, as suggested by (Amos et al., 2021); RP-DR is implemented as SVG (Heess et al., 2015). We also evaluate model-free PGMs, including RP(0) (Amos et al., 2021), DDPG (Lillicrap et al., 2015), and its variants such as SAC (Haarnoja et al., 2018) and TD3 (Fujimoto et al., 2018).

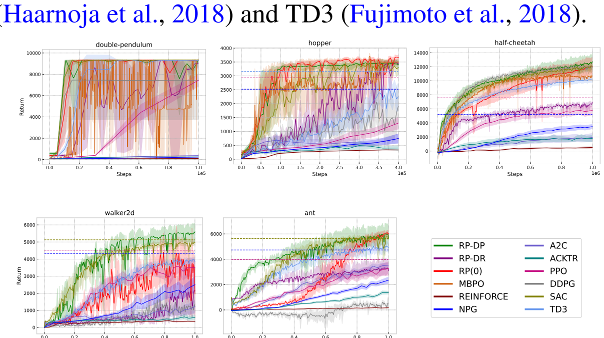


Figure 1. Evaluation of model-based RP PGMs in MuJoCo tasks. The dashed lines represent the return at convergence.

The results indicate that RP-DP consistently outperforms or matches the performance of existing methods such as MBPO (Janner et al., 2019) and LR PGMs, including REINFORCE (Sutton et al., 1999), NPG (Kakade, 2001), ACKTR (Wu et al., 2017), and PPO (Schulman et al., 2017). This highlights the significance and potential of model-based RP PGMs. Due to space limitations, we refer readers to §D.4 for larger versions of the figures. Further implementation details and discussions can be found in §D.1 and §D.2.

7.2 Gradient Variance and Loss Landscape

Our prior investigations have revealed that vanilla model-based RP PGMs tend to have highly non-smooth landscapes due to the significant increase in gradient variance. We now conduct experiments to validate this phenomenon in practice. In Fig. 2, we plot the mean gradient variance of the vanilla RP-DP algorithm 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.

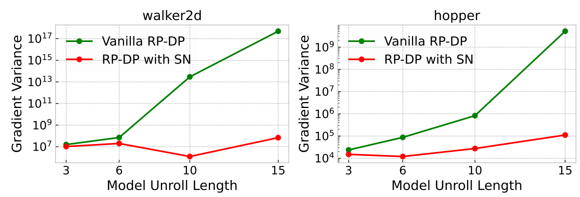
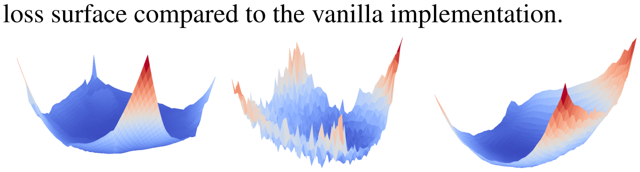


Figure 2. The gradient variance of the vanilla RP-DP explodes while adding spectral normalization solves this issue.

We can observe that for vanilla RP policy gradient algorithms, the gradient variance explodes in exponential rate with respect to the model unroll length. This results in a loss landscape that is highly non-smooth for larger unrolling steps. This renders the importance of smoothness regularization. Specifically, incorporating Spectral Normalization (SN) (Miyato et al., 2018) in the model and policy neural nets leads to a marked reduction in mean gradient variance for all unroll length settings, resulting in a much smoother loss surface compared to the vanilla implementation.



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

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

7.3 Benefit of Smoothness Regularization

In this section, 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, which is a contributing factor to training. Our results in Fig. 4 show that SN-based RP PGMs achieve equivalent or superior 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 produce reliable performance. SN-based methods, on the other hand, significantly boost training.

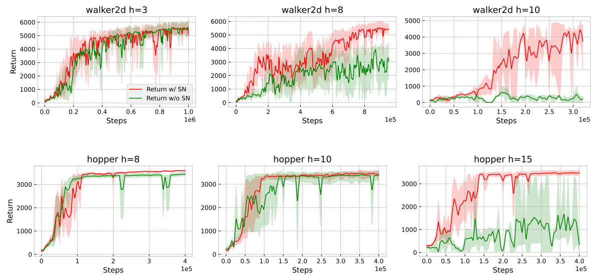


Figure 4. Performance of model-based RP PG methods with and without spectral normalization.

Additionally, we explore different choices of model unroll lengths and examine the impact of spectral normalization, with results shown in Fig. 5. We find that by utilizing SN, the curse of chaos can be mitigated, allowing for longer model unrolls. This is crucial for practical algorithmic designs: The most popular model-based RP PGMs such as (Clavera et al., 2020; Amos et al., 2021) often rely on a carefully chosen (small) h (e.g. $h = 3$). When the model is good enough, a small h may not fully leverage the accurate gradient information. As evidence, approaches (Xu et al., 2022; Mora et al., 2021) based on differentiable simulators typically adopt longer unrolls compared to model-based approaches. Therefore, with SN, more accurate multi-step predictions should enable more efficient learning without making the underlying optimization process harder. SN-based approaches also provide more robustness since the return is insensitive to h and the variance of return is smaller compared to the vanilla implementation when h is large.

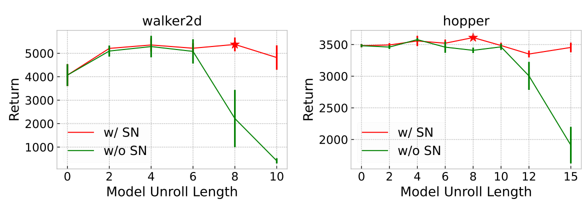


Figure 5. Spectral normalization with different model unrolls.

Ablation on Variance. By plotting the gradient variance of RP-DP during training in Figure 6, we can discern that for walker $h = 10$ and hopper $h = 15$a key contributor to the failure of vanilla RP-DP is the exploding gradient variances. On the contrary, the SN-based approach excels in training performance as a result of the drastically reduced variance.

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 important to study if SN causes such a negative effect and if it does, how to trade off between the model bias and gradient variance. To efficiently obtain an

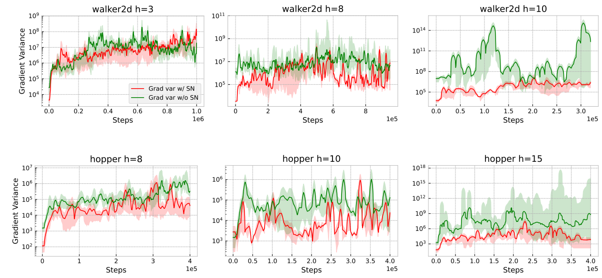


Figure 6. Gradient variance of RP PGMs during training.

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.

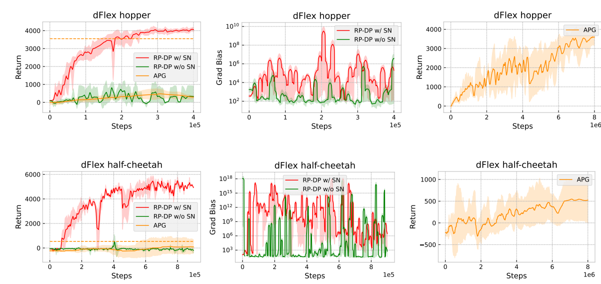


Figure 7. Performance and gradient bias in differentiable simulation. The last column is the full training curves of APG, which needs 20 times more steps in the hopper task to reach a comparable return with RP-DP-SN in the first column.

Figure 7 illustrates the crucial role that SN plays in dFlex locomotion tasks. It is noteworthy that the higher bias of the SN method does *not* impede performance, but rather improves it, indicating that the primary obstacle in training RP PGMs is the large variance in gradients. This suggests that even in differentiable simulations, one may still use a smooth proxy model when the dynamics have bumps or discontinuous jumps, sharing similarities with the gradient smoothing techniques (Suh et al., 2022a;b) applied to APG.

8 Conclusion & Future Work

In this work, we study the convergence of model-based reparameterization policy gradient methods and identify 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 SN-based RP PGMs allow longer model unrolls without introducing additional optimization hardness, learning more accurate multi-step models to fully leverage their gradient information should be a fruitful future direction. It will also be interesting to explore different smoothness regularization designs and apply them to a broader range of algorithms, such as using proxy models in differentiable simulation to obtain smooth policy gradients, which we would like to leave as future work.

References

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

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

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

Amos, B., Stanton, S., Yarats, D., and Wilson, A. G. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, pp. 6–20. PMLR, 2021.

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

Bastani, O. 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.

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

Bollt, E. M. 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.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

Clavera, I., Fu, V., and Abbeel, P. Model-augmented actor-critic: Backpropagating through paths. *arXiv preprint arXiv:2005.08068*, 2020.

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

Degris, T., White, M., and Sutton, R. S. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Feinberg, V., Wan, A., Stoica, I., Jordan, M. I., Gonzalez, J. E., and Levine, S. Model-based value expansion for efficient model-free reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Freeman, C. D., Frey, E., Raichuk, A., Girgin, S., Mordatch, I., and Bachem, O. Brax–a differentiable physics engine for large scale rigid body simulation. *arXiv preprint arXiv:2106.13281*, 2021.

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

Geilinger, M., Hahn, D., Zehnder, J., Bächer, M., Thomaszewski, B., and Coros, S. Add: Analytically differentiable dynamics for multi-body systems with frictional contact. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

Grimmett, G. and Stirzaker, D. *Probability and random processes*. Oxford university press, 2020.

Grzeszczuk, R., Terzopoulos, D., and Hinton, G. 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.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. 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.

Heess, N., Wayne, G., Silver, D., Lillicrap, T., Erez, T., and Tassa, Y. Learning continuous control policies by stochastic value gradients. *Advances in neural information processing systems*, 28, 2015.

Heiden, E., Macklin, M., Narang, Y., Fox, D., Garg, A., and Ramos, F. Disect: A differentiable simulation engine for autonomous robotic cutting. *arXiv preprint arXiv:2105.12244*, 2021a.

Heiden, E., Millard, D., Coumans, E., Sheng, Y., and Sukhatme, G. S. Neuralsim: Augmenting differentiable simulators with neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9474–9481. IEEE, 2021b.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

Li, C., Wang, Y., Chen, W., Liu, Y., Ma, Z.-M., and Liu, T.-Y. Gradient information matters in policy optimization by back-propagating through model. In *International Conference on Learning Representations*, 2021.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.

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

Metz, L., Maheswaranathan, N., Nixon, J., Freeman, D., and Sohl-Dickstein, J. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*, pp. 4556–4565. PMLR, 2019.

Metz, L., Freeman, C. D., Schoenholz, S. S., and Kachman, T. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.

Mora, M. A. Z., Peychev, M. P., Ha, S., Vechev, M., and Coros, S. Pods: Policy optimization via differentiable simulation. In *International Conference on Machine Learning*, pp. 7805–7817. PMLR, 2021.

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

Pang, T., Suh, H., Yang, L., and Tedrake, R. Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models. *arXiv preprint arXiv:2206.10787*, 2022.

Parmas, P., Rasmussen, C. E., Peters, J., and Doya, K. Pipps: Flexible model-based policy search robust to the curse of chaos. In *International Conference on Machine Learning*, pp. 4065–4074. PMLR, 2018.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pineda, L., Amos, B., Zhang, A., Lambert, N. O., and Calandra, R. Mbrl-lib: A modular library for model-based reinforcement learning. *arXiv preprint arXiv:2104.10159*, 2021.

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

Rachelson, E. and Lagoudakis, M. G. On the locality of action domination in sequential decision making. 2010.

Ruiz, F. R., AUEB, T. R., Blei, D., et al. The generalized reparameterization gradient. *Advances in neural information processing systems*, 29, 2016.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shen, Q., Li, Y., Jiang, H., Wang, Z., and Zhao, T. Deep reinforcement learning with robust and smooth policy. In *International Conference on Machine Learning*, pp. 8707–8718. PMLR, 2020.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Suh, H., Simchowitz, M., Zhang, K., and Tedrake, R. Do differentiable simulators give better policy gradients? *arXiv preprint arXiv:2202.00817*, 2022a.

Suh, H. J. T., Pang, T., and Tedrake, R. Bundled gradients through contact via randomized smoothing. *IEEE Robotics and Automation Letters*, 7(2):4000–4007, 2022b.

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Todorov, E., Erez, T., and Tassa, Y. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

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

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019a.

Wang, T., Bao, X., Clavera, I., Hoang, J., Wen, Y., Langlois, E., Zhang, S., Zhang, G., Abbeel, P., and Ba, J. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019b.

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

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

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

Wu, Y., Mansimov, E., Grosse, R. B., Liao, S., and Ba, J. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30, 2017.

Xu, J., Chen, T., Zlokapa, L., Foshey, M., Matusik, W., Sueda, S., and Agrawal, P. An end-to-end differentiable framework for contact-aware robot design. *arXiv preprint arXiv:2107.07501*, 2021.

Xu, J., Makoviychuk, V., Narang, Y., Ramos, F., Matusik, W., Garg, A., and Macklin, M. Accelerated policy learning with parallel differentiable simulation. *arXiv preprint arXiv:2204.07137*, 2022.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

Zhang, K., Koppel, A., Zhu, H., and Basar, T. 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:

$$\begin{aligned} \widehat{\nabla}_\theta V^\pi(\widehat{s}_{i,n}) &= (1-\gamma)\nabla_a r(\widehat{s}_{i,n}, \widehat{a}_{i,n})\nabla_\theta \pi(\widehat{s}_{i,n}, \varsigma_n) \qquad\qquad (A.7)\\ &\qquad + \gamma\widehat{\nabla}_s V^\pi(\widehat{s}_{i+1,n})\nabla_a \widehat{f}(\widehat{s}_{i,n}, \widehat{a}_{i,n}, \xi_n)\nabla_\theta\pi(\widehat{s}_{i,n}, \varsigma_n) + \gamma\widehat{\nabla}_\theta V^\pi(\widehat{s}_{i+1,n}), \\ \widehat{\nabla}_s V^\pi(\widehat{s}_{i,n}) &= (1-\gamma)\big(\nabla_s r(\widehat{s}_{i,n}, \widehat{a}_{i,n}) + \nabla_a r(\widehat{s}_{i,n}, \widehat{a}_{i,n})\nabla_s\pi(\widehat{s}_{i,n}, \varsigma_n)\big) \qquad\qquad (A.8)\\ &\qquad + \gamma\widehat{\nabla}_s V^\pi(\widehat{s}_{i+1,n})\big(\nabla_s \widehat{f}(\widehat{s}_{i,n}, \widehat{a}_{i,n}, \xi_n) + \nabla_a \widehat{f}(\widehat{s}_{i,n}, \widehat{a}_{i,n}, \xi_n)\nabla_s\pi(\widehat{s}_{i,n}, \varsigma_n)\big), \end{aligned}$$

where the termination of backpropagation at the h-th step is $\widehat{\nabla} V^\pi(\widehat{s}_{h,n}) = \nabla \widehat{V}_\omega(\widehat{s}_{h,n})$.

The resulting RP gradient estimator has the following form:

$$\widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^N \widehat{\nabla}_\theta V^\pi(\widehat{s}_{0,n}) = \frac{1}{N}\sum_{n=1}^N \nabla_\theta\Big(\sum_{i=0}^{h-1}\gamma^i r(\widehat{s}_{i,n}, \widehat{a}_{i,n}) + \gamma^h \widehat{Q}_\omega(\widehat{s}_{h,n}, \widehat{a}_{h,n})\Big), \tag{A.9}$$

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}) 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

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

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:

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

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

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

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,

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

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

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

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

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

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

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

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

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 (4.1) and when $\widehat{\nabla}_\theta J(\pi_\theta) = \widehat{\nabla}_\theta^{\mathrm{DR}} J(\pi_\theta)$ as in (4.2). 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

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

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

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

Applying the above recursion gives us

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

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

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

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

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

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

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

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

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):

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

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

$$\begin{aligned}
v_t &\le \bigg[\bigg(L_r\cdot\sum_{i=0}^{h-1}\gamma^i + \gamma^h\cdot L_{\widehat{Q}}\bigg)\cdot\bigg(\mathbb{E}_{\overline{s}_h}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{s}_h}{\mathrm{d}\theta}\bigg\|_2\bigg] + \mathbb{E}_{\overline{a}_h}\bigg[\bigg\|\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_h}{\mathrm{d}\theta}\bigg\|_2\bigg] + 2\widehat{K}(h)\bigg)\bigg]^2 \\
&\le O\Big(h^4\Big(\frac{1-\gamma^h}{1-\gamma}\Big)^2\widetilde{L}_{\widehat{f}}^{4h}\widetilde{L}_\pi^{4h} + \gamma^{2h}h^4\widetilde{L}_{\widehat{f}}^{4h}\widetilde{L}_\pi^{4h}\Big),
\end{aligned} \tag{C.17}$$

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 timestep $i \ge 1$ and the corresponding state, action, we have the following inequality results:

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

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

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

According to the triangle inequality, we continue with

$$\begin{aligned}
&\le \mathbb{E}\bigg[\bigg\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{s}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\bigg\|_2\bigg] + \mathbb{E}\bigg[\bigg\|\frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial\overline{s}_i}{\partial\overline{s}_{i-1}}\cdot\frac{\mathrm{d}\overline{s}_{i-1}}{\mathrm{d}\theta}\bigg\|_2\bigg] \\
&\quad+ \mathbb{E}\bigg[\bigg\|\frac{\partial\widehat{s}_{i,n}}{\partial\widehat{a}_{i-1,n}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\bigg\|_2\bigg] + \mathbb{E}\bigg[\bigg\|\frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial\overline{s}_i}{\partial\overline{a}_{i-1}}\cdot\frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\bigg\|_2\bigg] \\
&\le 2L_{\widehat{f}}\cdot\bigg(\bigg\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta}\bigg\|_2 + \bigg\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta}\bigg\|_2\bigg) + L_{\widehat{f}}\cdot\mathbb{E}_{\overline{s}_{i-1}}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{s}_{i-1}}{\mathrm{d}\theta}\bigg\|_2\bigg] \\
&\quad+ L_{\widehat{f}}\cdot\mathbb{E}_{\overline{a}_{i-1}}\bigg[\bigg\|\frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{a}_{i-1}}{\mathrm{d}\theta}\bigg\|_2\bigg].
\end{aligned} \tag{C.18}$$

[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

$$\begin{aligned}
&\mathbb{E}_{\overline{a}_i}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{a}_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}+\frac{\partial\widehat{a}_{i,n}}{\partial\theta}-\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}-\frac{\partial\overline{a}_i}{\partial\theta}\right\|_2\right] \\
&\quad\le\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\widehat{s}_{i,n}}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\partial\overline{a}_i}{\partial\overline{s}_i}\cdot\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\widehat{a}_{i,n}}{\partial\theta}-\frac{\partial\overline{a}_i}{\partial\theta}\right\|_2\right] \\
&\quad\le 2L_\pi\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\right\|\right]+L_\pi\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{s}_i}{\mathrm{d}\theta}\right\|_2\right]+2L_\theta.
\end{aligned}\tag{C.19}$$

Plugging (C.19) back to (C.18),

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

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

Applying this recursion gives us

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

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

As a consequence, we have from (C.19) that

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

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:

$$\begin{aligned}
&\mathbb{E}_{(s_i,a_i)\sim\mathbb{P}(s_i,a_i),(\widehat{s}_{i,n},\widehat{a}_{i,n})\sim\mathbb{P}(\widehat{s}_i,\widehat{a}_i)}\left[\left\|\frac{\mathrm{d}r(\widehat{x}_{i,n})}{\mathrm{d}\theta}-\frac{\mathrm{d}r(x_i)}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\mathrm{d}r}{\mathrm{d}\widehat{x}_{i,n}}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}x_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad\le\mathbb{E}\left[\left\|\frac{\mathrm{d}r}{\mathrm{d}\widehat{x}_{i,n}}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}\widehat{x}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}r}{\mathrm{d}x_i}\cdot\frac{\mathrm{d}x_i}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad\le 2L_r\cdot\widehat{K}(i)+L_r\cdot\left(\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_i}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_i}{\mathrm{d}\theta}\right\|_2\right]\right),
\end{aligned}\tag{C.20}$$

where $\mathbb{P}(s_i, a_i)$ and $\mathbb{P}(\widehat{s}_i, \widehat{a}_i)$ are defined in (4.3) with respect to $s_0 \sim \nu_\pi$, $\widehat{s}_0 \sim \nu_\pi$.

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

$$\mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}a_i}{\mathrm{d}\theta}\bigg\|_2\bigg]$$

$$= \mathbb{E}\bigg[\bigg\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{a}_{i,n}}{\partial \theta} - \frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}s_i}{\mathrm{d}\theta} - \frac{\partial a_i}{\partial \theta}\bigg\|_2\bigg]$$

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

$$\leq \mathbb{E}\bigg[\bigg\|\frac{\partial \widehat{a}_{i,n}}{\partial \widehat{s}_{i,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\bigg\|_2\bigg] + \mathbb{E}\bigg[\bigg\|\frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\partial a_i}{\partial s_i} \cdot \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\bigg\|_2\bigg] + \mathbb{E}\bigg[\bigg\|\frac{\partial \widehat{a}_{i,n}}{\partial \theta} - \frac{\partial a_i}{\partial \theta}\bigg\|_2\bigg]$$

$$\leq 2L_\pi \cdot \mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta}\bigg\|_2\bigg] + L_\pi \cdot \mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\bigg\|_2\bigg] + 2L_\theta. \tag{C.21}$$

Similarly, we have from (C.9) that

$$\mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\bigg\|_2\bigg]$$

$$= \mathbb{E}\bigg[\bigg\|\frac{\partial \widehat{s}_{i,n}}{\partial \widehat{s}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} + \frac{\partial \widehat{s}_{i,n}}{\partial \widehat{a}_{i-1,n}} \cdot \frac{\mathrm{d}\widehat{a}_{i-1,n}}{\mathrm{d}\theta} - \frac{\partial s_i}{\partial s_{i-1}} \cdot \frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta} - \frac{\partial s_i}{\partial a_{i-1}} \cdot \frac{\mathrm{d}a_{i-1}}{\mathrm{d}\theta}\bigg\|_2\bigg]$$

We proceed by applying the triangle inequality to extract the $\epsilon_{f,t}$ term defined in (4.3):

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

Combining with the result in (C.21), we have the following recursive expression:

$$\begin{aligned}
\mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_i}{\mathrm{d}\theta}\bigg\|_2\bigg] &\lesssim (\epsilon_{f,t} + 2L_f L_\pi) \cdot \widehat{K}(i-1) + L_f \widetilde{L}_\pi \cdot \mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{i-1,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}s_{i-1}}{\mathrm{d}\theta}\bigg\|_2\bigg] + 2L_f L_\theta \\
&= \Big((\epsilon_{f,t} + 2L_f L_\pi) \cdot \widehat{K}(i-1) + 2L_f L_\theta\Big) \cdot \sum_{j=0}^{i-1} L_f^j \widetilde{L}_\pi^j \\
&\leq \Big((\epsilon_{f,t} + 2L_f L_\pi) \cdot \widehat{K}(i-1) + 2L_f L_\theta\Big) \cdot i \cdot \widetilde{L}_f^i \widetilde{L}_\pi^i.
\end{aligned} \tag{C.22}$$

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.22) into (C.21), we obtain

$$\mathbb{E}\bigg[\bigg\|\frac{\mathrm{d}\widehat{a}_{i,n}}{\mathrm{d}\theta} - \frac{\mathrm{d}a_i}{\mathrm{d}\theta}\bigg\|_2\bigg] \leq \big[(\epsilon_{f,t} + 2L_f L_\pi) \cdot \widehat{K}(i-1) + 2L_f L_\theta\big] \cdot i \cdot \widetilde{L}_f^i \widetilde{L}_\pi^{i+1} + 2L_\pi \widehat{K}(i) + 2L_\theta. \tag{C.23}$$

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

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

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}\bigg[\bigg\|\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})\bigg\|_2\bigg]\\
&\quad=\gamma^h\cdot\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\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}-\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}\bigg\|_2\bigg]\\
&\quad\le\gamma^h\cdot\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\bigg\|\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}\bigg\|_2+\bigg\|\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}\bigg\|_2\\
&\qquad+\bigg\|\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}\bigg\|_2+\bigg\|\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}\bigg\|_2\bigg]\\
&\quad\le\gamma^h\cdot L_Q\cdot\bigg(\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\bigg\|\frac{\mathrm{d}a_h}{\mathrm{d}\theta}-\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}\bigg\|_2+\bigg\|\frac{\mathrm{d}s_h}{\mathrm{d}\theta}-\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}\bigg\|_2\bigg]\bigg)+\gamma^h\cdot\Big(\frac{\gamma^h}{1-\gamma}\Big)^2\widehat{K}(h)\cdot\epsilon_{v,t},
\end{aligned}\tag{C.24}$$

where the last inequality follows from (C.14) and the definition of $\epsilon_{v,t}$ in (4.4).

Using the results in (C.20) and (C.24), we obtain

$$\begin{aligned}
b_t\le\kappa\kappa'\cdot h\cdot\bigg(&L_r\cdot\bigg(\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_h}{\mathrm{d}\theta}\bigg\|_2\bigg]+\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\bigg\|\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_h}{\mathrm{d}\theta}\bigg\|_2\bigg]\bigg)+2L_r\cdot\widehat{K}(h)\bigg)\\
&+\kappa'\gamma^h\cdot\bigg(L_Q\cdot\bigg(\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\bigg\|\frac{\mathrm{d}\widehat{s}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}s_h}{\mathrm{d}\theta}\bigg\|_2\bigg]+\mathbb{E}_{\overline{\sigma}_1,\widehat{\sigma}_1}\bigg[\bigg\|\frac{\mathrm{d}\widehat{a}_{h,n}}{\mathrm{d}\theta}-\frac{\mathrm{d}a_h}{\mathrm{d}\theta}\bigg\|_2\bigg]\bigg)+\widehat{K}(h)\cdot\Big(\frac{\gamma^h}{1-\gamma}\Big)^2\epsilon_{v,t}\bigg),
\end{aligned}$$

Plugging (C.22), (C.23), and $\widehat{K}$ in (C.14) into the above expression, we conclude the proof by obtaining

$$b_t\le O\Big(\kappa\kappa'h^2\frac{1-\gamma^h}{1-\gamma}\widetilde{L}_{\widehat{f}}^h\widetilde{L}_f^h\widetilde{L}_\pi^{2h}\epsilon_{f,t}+\kappa'h\gamma^h\Big(\frac{\gamma^h}{1-\gamma}\Big)^2\widetilde{L}_{\widehat{f}}^h\widetilde{L}_\pi^h\epsilon_{v,t}\Big).\tag{C.25}$$

□

Lemma C.3. The expected value gradient over state distribution at timestep h can be represented by

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

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

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

where the first 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

$$
\nabla_\theta V^{\pi_\theta}(s_h) = \mathbb{E}\bigg[\int_{\mathcal{S}} \sum_{i=h}^{\infty} \gamma^{i-h} \cdot f\big(s_{i+1} | s_i, a_i\big) \cdot \Big(\frac{\partial Q^{\pi_\theta}}{\partial a_i} \cdot \frac{\mathrm{d}a_i}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_i} \cdot \frac{\mathrm{d}s_i}{\mathrm{d}\theta} \Big) \mathrm{d}s_{i+1} \bigg]. \tag{C.26}
$$

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

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

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

By taking the expectation over s_h in (C.26), we have the stated result:

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

□

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

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

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

$$
\begin{aligned}
b_t &= \Big\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_t}) \big] \Big\|_2 \\
&= \Big\| \mathbb{E}\big[\nabla_\theta J(\pi_{\theta_t}) - \widehat{\nabla}_\theta J(\pi_{\theta_t}) \big] \Big\|_2 \\
&= \bigg\| \mathbb{E}_{s_0 \sim \zeta, \widehat{s}_{0,n} \sim \mu_\pi} \bigg[\nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(s_i, a_i) + \gamma^h \cdot \nabla_\theta V^{\pi_\theta}(s_h) - \nabla_\theta \sum_{i=0}^{h-1} \gamma^i \cdot r(\widehat{s}_{i,n}, \widehat{a}_{i,n}) - \gamma^h \cdot \nabla_\theta \widehat{V}_t(\widehat{s}_{h,n}) \bigg] \bigg\|_2,
\end{aligned} \tag{C.27}
$$

where note that s_0 and $\widehat{s}_{0,n}$ are sampled from ζ and μ_π following the definition of the RL objective and the form of gradient estimator, respectively.

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. For any random variable Y, following the law of total expectation, we know that

$$
\begin{aligned}
\mathbb{E}_{\mu_\pi}[Y] &= \mathbb{E}[\mathbb{E}[Y|Z]] = \mathbb{E}[Y|Z=0]\mathbb{P}(Z=0) + \mathbb{E}[Y|Z=1]\mathbb{P}(Z=1)\\
&= \beta\mathbb{E}[Y|Z=0] + (1-\beta)\mathbb{E}[Y|Z=1]\\
&= \beta\mathbb{E}_{\nu_\pi}[Y] + (1-\beta)\mathbb{E}_\zeta[Y].
\end{aligned}
\tag{C.28}
$$

Therefore, we have from (C.27) that

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

where the first inequality holds since $\|\mathbb{E}[\cdot]\|_2 \le \mathbb{E}[\|\cdot\|_2]$ and the second inequality holds due to (C.28).

Using the result from Lemma C.3, we know that

$$
\begin{aligned}
&\mathbb{E}_{s_0\sim\zeta}\bigg[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i) + \gamma^h\cdot\nabla_\theta V^{\pi_\theta}(s_h) - \nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n}) - \gamma^h\cdot\nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\bigg]\\
&\quad = \underbrace{\mathbb{E}_{s_0\sim\zeta}\bigg[\nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(s_i,a_i) - \nabla_\theta\sum_{i=0}^{h-1}\gamma^i\cdot r(\widehat{s}_{i,n},\widehat{a}_{i,n})\bigg]}_{:=B_r} + \underbrace{\gamma^h\mathbb{E}_{(s_h,a_h)\sim\widetilde{\sigma}_1}\bigg[\frac{\partial Q^{\pi_\theta}}{\partial a_h}\cdot\frac{\mathrm{d}a_h}{\mathrm{d}\theta} + \frac{\partial Q^{\pi_\theta}}{\partial s_h}\cdot\frac{\mathrm{d}s_h}{\mathrm{d}\theta} - \nabla_\theta\widehat{V}_t(\widehat{s}_{h,n})\bigg]}_{:=B_v}.
\end{aligned}
$$

Therefore, with the shorthand notation B_r and B_v, we may rewrite the upper bound of b_t as

$$
\begin{aligned}
b_t &\le \beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu_\pi}\big[\|B_r+B_v\|_2\big] + (1-\beta)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\zeta}\big[\|B_r+B_v\|_2\big]\\
&\le \Big(\beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu_\pi}\big[\|B_r\|_2\big] + (1-\beta)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\zeta}\big[\|B_r\|_2\big]\Big) + \Big(\beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu_\pi}\big[\|B_v\|_2\big] + (1-\beta)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\zeta}\big[\|B_v\|_2\big]\Big).
\end{aligned}
\tag{C.29}
$$

Next, we separately bound the two terms on the right-hand side of (C.29), namely the bias introduced by the h-step model expansion and by the tail estimation using a critic.

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

By the definition of κ, we further have

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

Similarly, for the critic bias, we have

$$\begin{aligned}
&\mathbb{E}_{\widehat{s}_{0,n}\sim\mu_\pi}\bigg[\bigg\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\bigg[\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})\bigg]\bigg\|_2\bigg]\\
&\quad=\beta\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\nu_\pi}\bigg[\bigg\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\bigg[\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})\bigg]\bigg\|_2\bigg]\\
&\qquad+(1-\beta)\cdot\mathbb{E}_{\widehat{s}_{0,n}\sim\zeta}\bigg[\bigg\|\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1}\bigg[\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})\bigg]\bigg\|_2\bigg]\\
&\quad\le\big(\beta+\kappa\cdot(1-\beta)\big)\cdot\mathbb{E}_{(s_h,a_h)\sim\overline{\sigma}_1,\widehat{s}_{0,n}\sim\nu_\pi}\bigg[\bigg\|\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})\bigg\|_2\bigg]\\
&\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\overline{\sigma}_1}\bigg[\bigg\|\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})\bigg\|_2\bigg].
\end{aligned}$$

Plugging the above two inequalities into (C.29) completes the proof. □

C.4 Proof of Proposition 5.7

Proof. When $\gamma\approx 1$, we have

$$\frac{1-\gamma^h}{1-\gamma}=\sum_{i=0}^{h-1}\gamma^i\approx h,\quad \frac{\gamma^h}{1-\gamma}=\frac{1}{1-\gamma}-\frac{1-\gamma^h}{1-\gamma}\approx\frac{1}{1-\gamma}-h.$$

We denote $H:=1/(1-\gamma)=\sum_{i=0}^\infty\gamma^i$ as the effective task horizon.

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 leading terms in the variance, bias bound (i.e., (C.17) and (C.25)) when L_f, $L_{\widehat{f}}$, and L_π are less than or equal to 1. Formally, $v'_t:=h^6$ and $b'_t:=h^3\epsilon_{f,t}+h(H-h)^2\epsilon_{v,t}$. Here, we consider the terms that are only dependent on h, H, $\epsilon_{f,t}$, and $\epsilon_{v,t}$ to simplify the analysis and determine the order of h^*.

The first problem is to find the optimal model unroll h'^* that minimizes $g(h)$. We notice that $g(h)$ increases monotonically with respect to b'_t and $\sqrt{v'_t}$ when they are non-negative. This further simplifies the problem to find

$$h'^*=\operatorname*{argmin}_h b'_t+c'\sqrt{v'_t}=\operatorname*{argmin}_h \underbrace{h^3(\epsilon_{f,t}+c')+h(H-h)^2\epsilon_{v,t}}_{:=g_1(h)},\tag{C.30}$$

where c' is some constant that does not affect the order or h'^*.

By taking the derivative of the right-hand side of (C.30) with respect to h and setting it to zero, we obtain

$$\frac{\partial}{\partial h}g_1(h)=3h^2\cdot(\epsilon_{f,t}+c')+(3h^2-4Hh+H^2)\cdot\epsilon_{v,t}=0.\tag{C.31}$$

Solve the above quadratic equation with respect to h, we have the following two non-negative roots:

$$h_1'^*=\frac{4H\epsilon_{v,t}+\sqrt{(4H\epsilon_{v,t})^2-12c_1\epsilon_{v,t}H^2}}{6c_1},\quad h_2'^*=\frac{4H\epsilon_{v,t}-\sqrt{(4H\epsilon_{v,t})^2-12c_1\epsilon_{v,t}H^2}}{6c_1},$$

where $c_1:=\epsilon_{f,t}+\epsilon_{v,t}+c'$. Without loss of generality, we assume $(4H\epsilon_{v,t})^2-12c_1\epsilon_{v,t}H^2>0$. When this condition is not met and (C.31) does not have a real solution h'^*, we set h^* to either 0 or H.

Now we verify that the second-order derivative at $h_1'^*$ is

$$\begin{aligned}
\frac{\partial^2 g_1(h_1'^*)}{\partial h^2}&=\frac{4H\epsilon_{v,t}+\sqrt{(4H\epsilon_{v,t})^2-4c_1\cdot H^2}}{6c_1}*6(\epsilon_{f,t}+\epsilon_{v,t}+c')-4H\epsilon_{v,t}\\
&=\sqrt{(4H\epsilon_{v,t})^2-4c_1\cdot H^2}>0.
\end{aligned}$$

This indicates that $h_1'^*$ is the minima.

By omitting the constant terms, we can write h'^* as

$$h'^*=O\big(\epsilon_{v,t}/(\epsilon_{f,t}+\epsilon_{v,t})\cdot H\big).$$

Constraining the optimal h^* within the non-negative range concludes the proof. □

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

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 have state-of-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 requires a carefully chosen policy update rate as well as a proper size of the experience replay buffer. This is because 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 an 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), 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 Implementation and Ablation of Spectral Normalization

In experiments, we use Multilayer Perceptrons (MLPs) for the critic, policy, and model. Besides, we adopt Gaussian dynamical models and policies as the source of stochasticity. To test the benefit of smooth function approximations in model-based RP policy gradient algorithms, spectral normalization is applied to all layers of the policy MLP and all except the final layers of the 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()]
```

Moreover, we conduct ablation study on the functions that spectral normalization is applied to: Both the model and the policy (default setting); Only the model; Only the policy; No SN is applied (vanilla setting). The results are showsn in Figure 8.

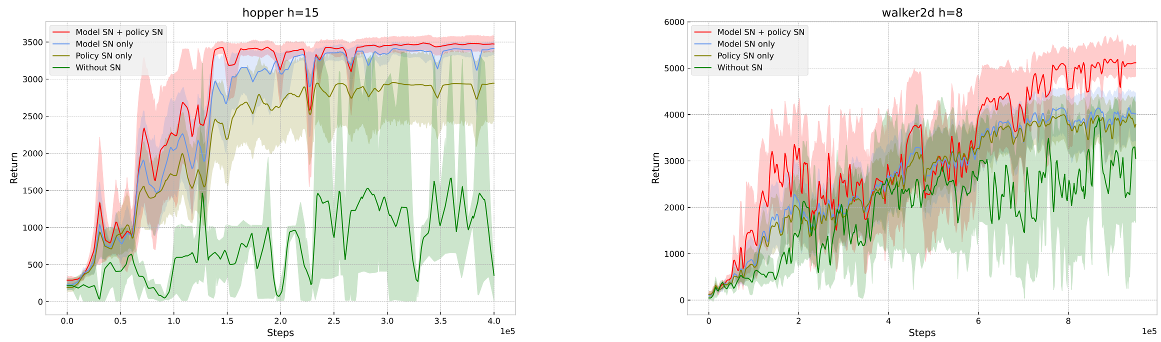


Figure 8. Ablation on different model learners: single-step and multi-step state prediction models, and multi-step state prediction models trained with an additional directional derivative error.

We observe in Fig. 8 that for both the hopper and the walker2d tasks, applying SN to the model and policy simultaneously achieves the best performance, which supports our theoretical results. Besides, learning a smooth transition kernel by applying SN to the neural network model only is slightly better than only applying SN to the policy. At the same time, the vanilla implementation of model-based RP PGM fails to give acceptable result.

D.3 Ablation on Different Model Learners

Our main theoretical results in Section 5 depend on the model error defined in (4.3), 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 4.3 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.

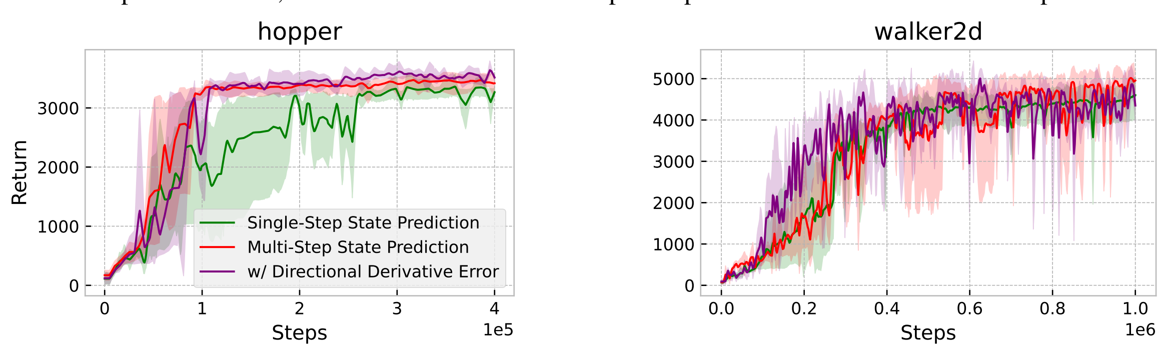


Figure 9. Ablation on different model learners: single-step and multi-step state prediction models, and multi-step state prediction models trained with an additional directional derivative error.

D.4 Figures in the Main Text in Larger Sizes

Here, we provide identical figures that are larger in size. Figure 10, 11, 12, 13 correspond to Figure 1, 4, 6, 7 in the main text, respectively.

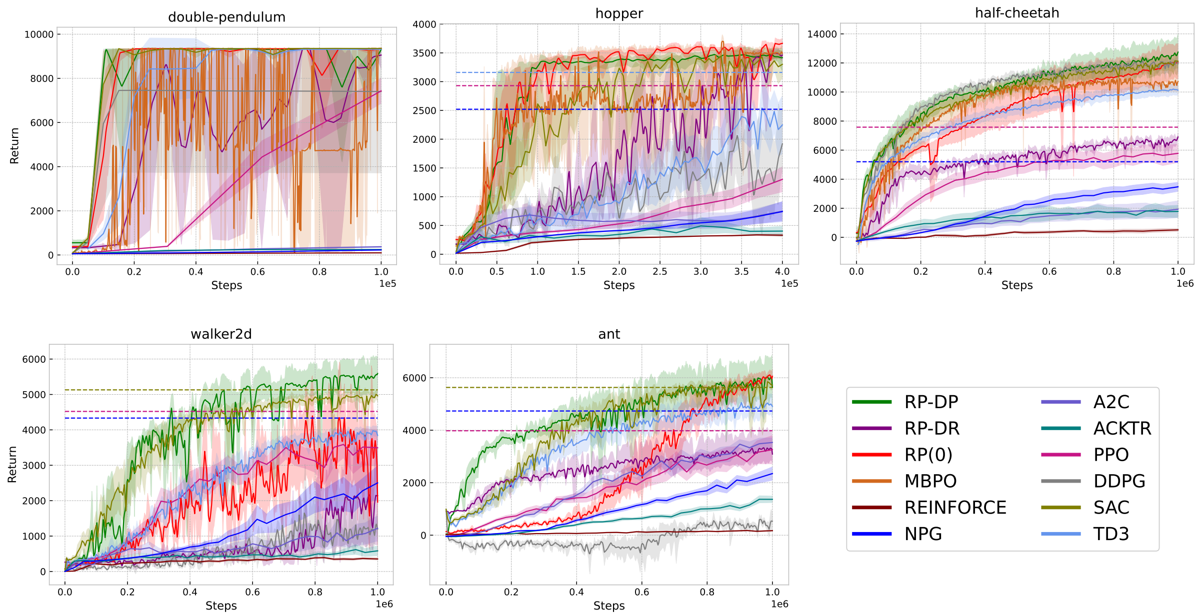


Figure 10. Evaluation of model-based RP PGMs in MuJoCo tasks. The dashed lines represent the value at the convergence of the corresponding model-free algorithms.

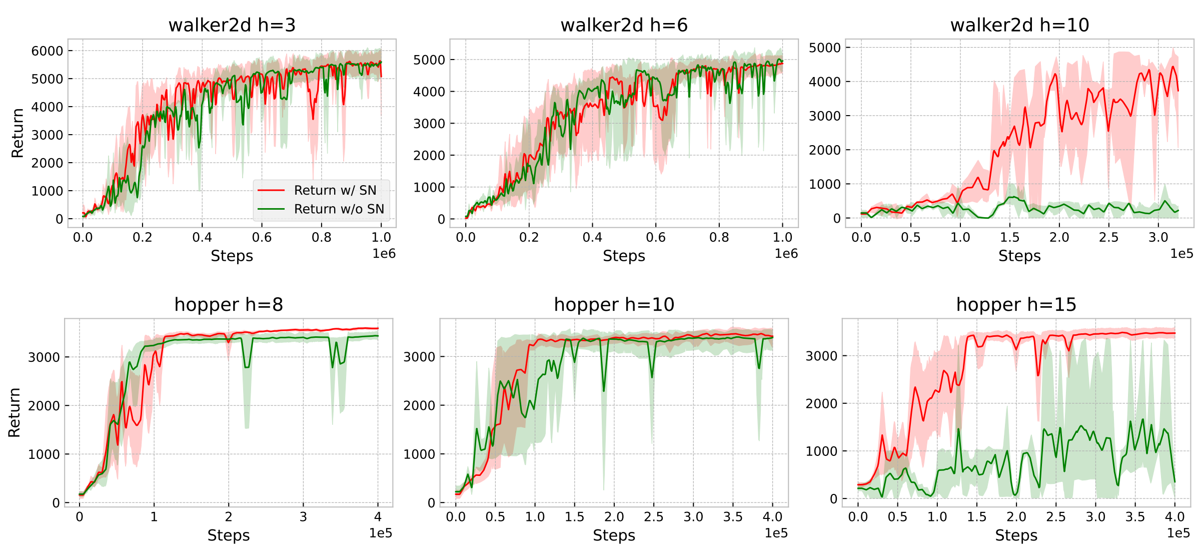


Figure 11. Performance of model-based RP PG methods with and without spectral normalization.

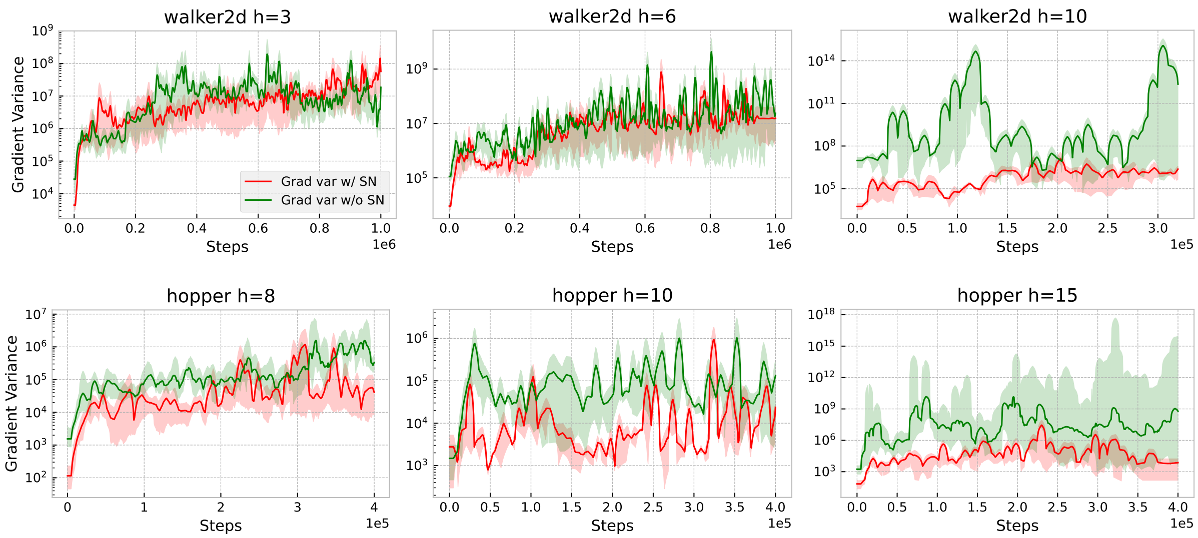


Figure 12. Ablation on the gradient variance.

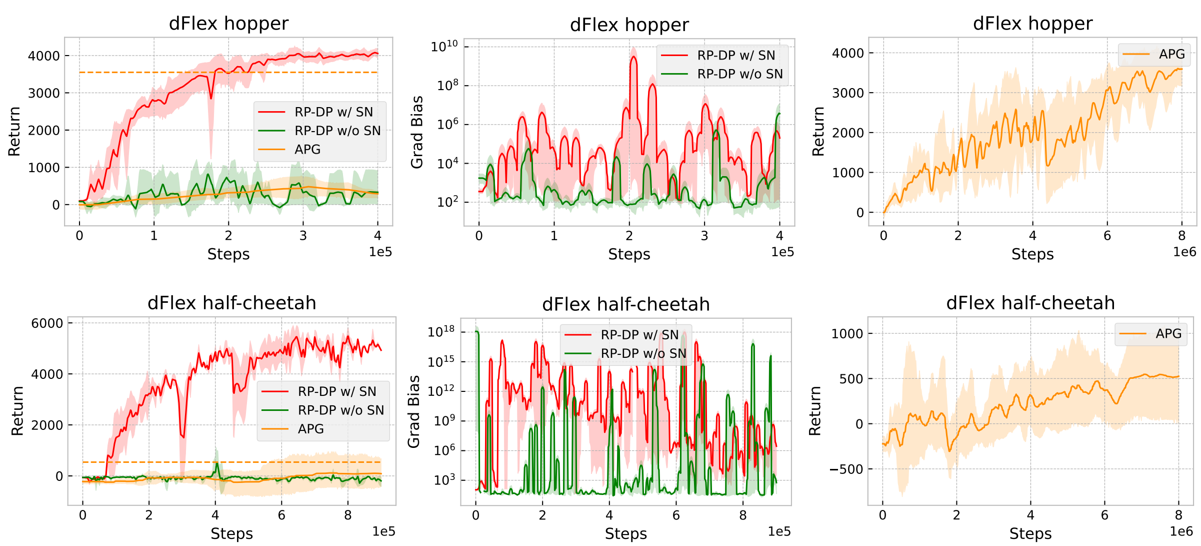


Figure 13. Performance and gradient bias in differentiable simulation. The last column is the full training curves of APG, which needs 20 times more steps in hopper to reach a comparable return with RP-DP-SN in the first column.