Model-Based First-Order Policy Gradient for Contact Dynamics

Shenao Zhang [1] **Wanxin Jin** [2] **Zhaoran Wang** [3]

Abstract

In the realm of model-based reinforcement learning, the learned models are typically smooth approximators of the environment dynamics. However, this can present challenges in robotic systems that experience hard contact and have non-smooth or even discontinuous local behaviors. Despite the vast amount of data required to fit these behaviors, the resulting inaccurate model gradient can lead to poor performance when utilizing the First-Order Policy Gradient (FOPG) approach. Therefore, in this work, we investigate physics-guided models constructed using the complementarity problem that underlies contact events. Unfortunately, our theory indicates that the complementarity-based models tend to be stiff, which causes the FOPG gradient variance to explode, resulting in optimization difficulties such as chaotic and non-smooth loss landscapes. To address this issue, we introduce a class of softened complementarity models that correspond to the barrier-smoothed objectives. We propose *Analytical Barrier Smoothing* for the reduction of the large FOPG gradient variance, with a *contact-sensitive* central-path parameter to control the gradient bias. By establishing the upper bound of the gradient variance and bias, we are able to characterize the convergence of the proposed method. Experimental results are also provided to support our theory and method.

1 Introduction

Model-Based Reinforcement Learning (MBRL) has achieved success in sequential decision-making applications such as games and robotics (Schrittwieser et al., 2020; Wang et al., 2019b). First-Order Policy Gradient (FOPG) by back-propagating through the computational path of cumulative rewards, is the most straightforward method for controlling differentiable simulations (Xu et al., 2022; Freeman et al., 2021) and has demonstrated great potential for general non-differentiable tasks in model-based settings (Clavera et al., 2020; Li et al., 2021; Amos et al., 2021). However, physical systems such as robotic locomotion and manipulation are often characterized by stiff dynamics with extreme curvatures (Parmar et al., 2021; Anitescu & Potra, 2002) due to geometrical constraints and contact events. Most modern MBRL algorithms that fit the dynamics with universal function approximators, such as neural networks (Nagabandi et al., 2018; Chua et al., 2018), tend to select the smoothest interpolators as the simplest explanation of the environment transitions (Belkin et al., 2019; Pfrommer et al., 2021). As a result, these black-box models typically require a large amount of data to learn the contact behaviors while still struggling with inaccurate first-order gradient estimation in long-horizon problems (Hochlehnert et al., 2021).

In this work, we study the physics-informed model built upon the complementarity problem, which serves as the foundation for hard-contact simulations (Geilinger et al., 2020; Howell et al., 2022; Werling et al., 2021). The complementarity problem, from which the impact and frictional contact forces are solved using *Interior-Point Method* (IPM), ensures non-penetration and maximum dissipation. Unfortunately, despite the ability of the complementarity-based model to accurately approximate contact dynamics by learning physical parameters, we demonstrate that its stiffness can lead to optimization challenges when performing FOPG. Specifically, we first establish the convergence of model-based FOPG that depends on the gradient variance and bias. We then prove that the upper bound of the gradient variance has polynomial dependencies on the Lipschitz continuity of the model, where the degrees are linear in the task horizon. When the model is stiff, long chains of nonlinear mappings result in slow convergence due to the large gradient variance and chaotic (Bollt, 2000) optimization procedure, a phenomenon that is also observed in experiments (Parmas et al., 2018; Metz et al., 2021).

To combat the aforementioned problem, we introduce a class of μ-softened *Linear Complementarity Systems* (LCS) with the central-path parameter μ. We prove that the Lipschitz upper bound of the μ-softened LCS is inversely proportional to μ. Therefore, a natural method to avoid the large gradient variance is to differentiate through the softened complemen-

[1]Georgia Institute of Technology, Atlanta, GA, USA [2]University of Pennsylvania, Philadelphia, PA, USA [3]Northwestern University, Evanston, IL, USA.

tarity system by setting stopping criteria in the IPM solver. As the softened LCS can be linked to the optimality condition of a barrier-smoothed objective, we term this basic approach *Analytical Barrier Smoothing* (ABS). However, it is important to note that indiscriminately applying analytical smoothing can lead to unrealistic simulation and a significant gradient bias.

To achieve the best trade-off between the gradient variance and bias, we propose utilizing a *contact-sensitive* central-path parameter that decreases with the minimum distance-to-obstacle of the inactive impact contacts. Intuitively, since the root of stiffness is the sudden change of the impact contact force, smoothing is only needed locally near the impact contact, while the gradients are globally accurate by solving the exact complementarity problem elsewhere. By drawing on the equivalence between ABS and randomized smoothing (Suh et al., 2022b;a), we show that the proposed method minimizes the linearization residual in frictionless single-contact settings. Based on this result, the gradient bias of contact-sensitive ABS can be upper bounded. We also present experimental results in ball-bouncing and robotic locomotion tasks to support our theory and method.

2 Background

2.1 Reinforcement Learning

Consider learning to optimize a finite H-horizon Markov Decision Process (MDP) over repeated episodes of interaction. Denote the state space and action space as $\mathcal{X}$ and $\mathcal{U}$, respectively. When taking action $u \in \mathcal{U}$ at state $x \in \mathcal{X}$, the agent receives reward $r(x, u)$ and the MDP transitions to a new state according to probability $x' \sim f^*(\cdot \,|\, x, u)$.

We are interested in controlling the system by finding a policy π_θ that maximizes the expected cumulative reward. Denote by ζ the initial state distribution. The objective is:

$$\mathcal{J}(\pi) = \mathbb{E}_{x_0 \sim \zeta}\big[V_0^\pi(x_0)\big] = \mathbb{E}_{p_\pi(\alpha)}\left[\sum_{t=0}^{H-1} r(x_t, u_t)\right],$$

where V_0^π is the state value at the initial timestep, and $p_\pi(\alpha)$ is the distribution over rollouts $\alpha := ((x_0, u_0), \cdots, (x_{H-1}, u_{H-1}))$ when executing π, formally, $x_0 \sim \zeta(\cdot)$, $u_i \sim \pi(\cdot \,|\, s_i)$, and $x_{i+1} \sim f^*(\cdot \,|\, x_i, u_i)$.

2.2 Stochastic Gradient Estimation

The underlying problem of policy gradient is determining the gradient of a probabilistic objective with respect to the parameters of the sampling distribution. This is represented by the equation $\nabla_\theta \mathbb{E}_{p(z;\theta)}[y(z)]$. In RL, we view $p(z;\theta)$ as the trajectory distribution conditioned on the policy parameter θ, and $y(z)$ as the cumulative reward. In the sequel, we introduce two commonly used gradient estimators in RL.

Zeroth-Order (Likelihood Ratio) Gradient. By leveraging the *score function*, zeroth-order gradient estimators only require samples of the function values. In particular, as the score function satisfies $\nabla_\theta \log p(z;\theta) = \nabla_\theta p(z;\theta)/p(z;\theta)$, the zeroth-order gradient has the following form:

$$\nabla_\theta \mathbb{E}_{p(z;\theta)}\big[y(x)\big] = \mathbb{E}_{p(z;\theta)}\big[y(z)\nabla_\theta \log p(z;\theta)\big]. \quad (2.1)$$

First-Order (Reparameterization) Gradient. First-order 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:

$$\widehat{z} \sim p(z;\theta) \;\equiv\; \widehat{z} = g(\epsilon;\theta), \quad \epsilon \sim p. \quad (2.2)$$

Derived from the *law of the unconscious statistician* (LOTUS) (Grimmett & Stirzaker, 2020), i.e., $\mathbb{E}_{p(x;\theta)}[y(z)] = \mathbb{E}_{p(\epsilon)}[y(g(\epsilon;\theta))]$, the first-order gradient takes the form:

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

2.3 Bundled Gradient via Randomized Smoothing

When dealing with non-smooth functions with extreme curvatures, such as objectives for contact dynamics, the gradient can be prone to large jumps. The first-order bundled gradient is proposed by (Suh et al., 2022b;a; Pang et al., 2022) to solve this issue. Consider a deterministic objective $y(x)$. Differentiating through the randomized smoothed objective $\overline{y}(x) := \mathbb{E}_{w\sim\rho(w)}[y(x+w)]$ gives the bundled gradient:

$$\nabla \overline{y}(x) = \mathbb{E}_{w\sim\rho(w)}[\nabla y(x+w)].$$

2.4 Rigid-Body Dynamics

The standard approach for modeling robotic systems involves utilizing the framework of rigid-body systems with contacts. Adhering to Newton's laws, the continuous-time equation of motion is formulated as follows:

$$\mathcal{M}(q)\mathrm{d}v = \big(n(q,v) + u\big)\mathrm{d}t + J(q)^\top \lambda,$$

where we let q denote the generalized coordinates, v the generalized velocities, $u \in \mathbb{R}^{n_u}$ the applied control force, $\mathcal{M}(q)$ the generalized inertia matrix, $n(q,v)$ the passive forces (e.g., Coriolis, centrifugal, and gravity), and $J(q)$ the Jacobian of the active contacts. Here, we define $\lambda := (\gamma^{(1)}, \beta^{(1)}, \cdots, \gamma^{(c)}, \beta^{(c)}) \in \mathbb{R}^{n_\lambda}$ as the (unknown) contact space force, where γ and β are the normal *impact* forces and *frictional* forces, respectively, and c denotes the number of contact points. The state x usually contains q and v.

Using Euler approximation and multiplying by $\mathcal{M}_t^{-1}$, the discrete-time dynamics can be modeled in contact space by:

$$\begin{aligned} v_{t+1} &= v_t + \mathcal{M}_t^{-1}(n_t + u_t)h + \mathcal{M}_t^{-1} J_t^\top \lambda_t, \\ q_{t+1} &= q_t + h v_{t+1} \end{aligned} \quad (2.3)$$

where h is the discretization step size and t is the timestep.

The frictional and impact contact forces are constrained by the system's configuration. Specifically, the impact problem is encoded with the following constraints:

$$\gamma_{t+1} \circ \phi(q_{t+1}) = \vec{0}, \qquad \gamma_{t+1}, \phi(q_{t+1}) \geq \vec{0}, \tag{2.4}$$

where $\circ$ is the element-wise (Hadamard) product, the Signed Distance Function (SDF) $\phi(q_{t+1}) = \phi(x_t, u_t)$ measures the distance from the contact points to obstacles, $\vec{0}$ is the zero vector, and the equality, inequality are element-wise. Eq. (2.4) states that the magnitude of the normal impact forces must be non-negative and can only be non-zero to maintain non-negative gaps (non-penetration) when contact is active.

Moreover, the Coulomb friction can be modeled using the maximum-dissipation principle and a linearized friction cone, which has the following set of constraints:

$$\begin{aligned} &\beta_{t+1} \circ \xi_{t+1} = \vec{0}, \qquad \beta_{t+1}, \xi_{t+1} \geq \vec{0}, \\ &B(q_{t+1}) v_{t+1} + \omega_{t+1} \vec{1} - \xi_{t+1} = \vec{0}, \\ &\omega_{t+1} \cdot (\alpha_f \gamma_{t+1} - \beta_{t+1}) = \vec{0}, \end{aligned} \tag{2.5}$$

where $\alpha_f \geq 0$ is the friction coefficient, matrix B maps from the generalized coordinate velocity to tangential velocity in the contact frame, and ω_{t+1}, ξ_{t+1} are dual variables associated with the linearized friction cone and non-negative constraint, respectively.

3 Complementarity-Based Contact Models

3.1 Softened Linear Complementarity Systems

The dynamic (2.3) describes a hybrid system where different modes are controlled by the contact force λ under the nonlinear complementarity problem (2.4) and (2.5). To simplify our analysis, in the following sections, we study the *Linear Complementarity Systems* (LCS), which effectively capture the local behaviors of the state transitions and are widespread in robotics research (Aydinoglu et al., 2021; Tassa & Todorov, 2010; Drumwright & Shell, 2012).

We first define a class of softened linear complementarity systems f_μ as the approximations of the exact LCS $f_{\mu=0}$.

Definition 3.1 (Softened LCS)**.** A model $x_{t+1} = f_\mu(x_t, u_t)$ is a softened LCS if the evolution of state $x \in \mathbb{R}^{d_x}$ is governed by a linear dynamics and a μ-complementarity problem (the last two lines of (3.1)):

$$\begin{aligned} &x_{t+1} = Ax_t + Bu_t + C\lambda_t + c, \\ &\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu \vec{1}, \\ &\lambda_t \geq \vec{0}, \quad Dx_t + Eu_t + F\lambda_t + d \geq \vec{0}, \end{aligned} \tag{3.1}$$

where $A \in \mathbb{R}^{d_x \times d_x}, B \in \mathbb{R}^{d_x \times d_u}, C \in \mathbb{R}^{d_x \times d_\lambda}, D \in \mathbb{R}^{d_\lambda \times d_x}, E \in \mathbb{R}^{d_\lambda \times d_u}, F \in \mathbb{R}^{d_\lambda \times d_\lambda}$, and $\mu \geq 0$. Denote the solver of the μ-complementarity problem as S_μ, which returns the solution $\lambda_t = S_\mu(Dx_t + Eu_t + d) \in \mathbb{R}^{d_\lambda}$.

When simulating, $\mu = 0$ corresponds to the exact *Linear Complementarity Problem* (LCP) and the LCS $f_{\mu=0}$ resembles the reality. Obviously, solving the contact space force λ_t is our primary goal, since x_{t+1} is readily obtained from the dynamics once λ_t is available. To accomplish this, we introduce the assumption and method for solving the LCP.

Assumption 3.2 (P-Matrix)**.** Assume F in the LCS (3.1) is a P-matrix, defined as a matrix whose principal minors are all positive, i.e., the determinants of its principal submatrices $\det(F_{\alpha\alpha}) > 0$, $\forall \alpha \subseteq \{1, \cdots, d_\lambda\}$.

Assumption 3.2 guarantees that the solution λ_t exists and is unique, which is commonly upheld in the study of contact dynamics problems (Aydinoglu et al., 2020; Jin et al., 2022).

3.2 Smoothed Objective with Barrier Function

To effectively and accurately solve the convex constrained optimization problem (3.1), we adopt the *Interior-Point Method* (IPM) (Wright et al., 1999) that solves a sequence of relaxed problems with decreasing $\mu > 0$ to reliably converge to a solution of the exact LCS $f_{\mu=0}$.

We show that the softened LCS is the optimality condition of a barrier-smoothed objective with the following lemma. We defer all the proofs in this paper to Appendix A.

Lemma 3.3 (Primal Problem with Log-Barrier Function)**.** The softened LCS (3.1) with $\mu \geq 0$ is the first-order optimality condition of the following program:

$$\begin{aligned} \min_{\lambda_t \geq \vec{0}, \epsilon_t \geq \vec{0}} \quad & \lambda_t^\top \epsilon_t - \mu \sum_{i=1}^{d_\lambda} \left(\log \lambda_t^{(i)} + \log \epsilon_t^{(i)} \right) \\ \text{s.t.} \quad & Dx_t + Eu_t + F\lambda_t + d = \epsilon_t, \\ & Ax_t + Bu_t + C\lambda_t + c = x_{t+1}, \end{aligned} \tag{3.2}$$

where $\lambda_t^{(i)}, \epsilon_t^{(i)}$ are the i-th elements of vector $\lambda_t, \epsilon_t \in \mathbb{R}^{d_\lambda}$.

Lemma 3.3 reveals that the softened LCS is in fact the perturbed Karush–Kuhn–Tucker (KKT) conditions, where the perturbation corresponds to smoothing the objective with barrier functions. The utilization of logarithmic barrier functions in (3.2) serves to discourage solutions from approaching the boundaries of the polytope formed by the hard constraints. As such, μ acts as a restraint, confining the solution within the analytical center of the constraint polytope and is considered a central-path parameter.

The barrier terms can be viewed as the potential of a force field whose strength is inversely proportional to the distance to the constraint boundary (Boyd et al., 2004). When applying IPM with central-path parameters, the intermediate problems with $\mu > 0$ achieve a smoothing effect akin to the "force-at-a-distance" relaxation of the complementarity constraints (Pang et al., 2022; Howell et al., 2022). In other words, μ controls both the *stiffness* and the *accuracy* of the

softened LCS model f_μ. In what follows, we will show that both properties are determining factors for the quality of first-order gradient estimation and the convergence of the resulting policy gradient algorithm.

4 Model-Based First-Order Policy Gradient

In this section, we first present an overview of model-based First-Order Policy Gradient (FOPG). Then we delve into the convergence properties of model-based FOPG and study the correlation between its convergence rate and the gradient bias, variance. Additionally, we investigate the connection between the gradient variance and the model stiffness, as well as the stiffness of complementarity-based models. Through our analysis, we find that non-smooth behaviors of contact events can impede optimization, motivating us to create an analytical method for smoothing the system.

4.1 Framework

In Algorithm 1, we provide the pseudocode of model-based FOPG, where two update procedures are performed iteratively. Namely, the model and the policy are updated in each iteration $n \in [N]$, which gives us sequences of $\{f_{\psi_n}\}_{n\in[N]}$ and $\{\pi_{\theta_t}\}_{t\in[T+1]}$, respectively.

Algorithm 1 Model-Based First-Order Policy Gradient

Input: Number of iterations N, transition data set $\mathcal{D} = \varnothing$
1: **for** iteration $n \in [N]$ **do**
2: Update the model parameter ψ_n by minimizing (4.1)
3: Update the policy parameter θ_n by (4.3)
4: Execute $\pi_{\theta_{n+1}}$ and update $\mathcal{D}$
5: **end for**
6: **Output:** $\{\pi_{\theta_n}\}_{n\in[N]}$

Model Update. A forward state-predictive model is learned from data $\mathcal{D} = \{(x_t^*, u_t^*, x_{t+1}^*)\}_{t=1}^T$, where the state $x_t \in \mathbb{R}^{d_x}$ is the system's configuration (including velocity v_t, coordinate q_t, etc.). For hard-contact rigid-body systems, we use a physically grounded model $x_{t+1} = f(x_t, u_t; \psi)$ where f returns the solution of (2.3) constrained by (2.4), (2.5). Instead of being parameterized by a black-box neural network, ψ contains all *estimated* physics data, such as friction coefficient and parameters of each body. The model training loss is as follows, minimized by random search:

$$L(\psi; \mathcal{D}) = \sum_{t=1}^{T} \frac{1}{2} \left\| f(x_t^*, u_t^*; \psi) - x_{t+1}^* \right\|_2^2. \tag{4.1}$$

Policy Update. Consider optimizing a stochastic policy $u \sim \pi_\theta(\cdot|x)$ in continuous action spaces, or equivalently $u = \pi_\theta(x, \varsigma)$ with noise $\varsigma \sim p(\varsigma)$. The first-order policy gradient at iteration n is given by linking together the reward, model, policy, and differentiating through the model-generated trajectories:

$$\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) = \frac{1}{M} \sum_{m=1}^{M} \nabla_\theta \left(\sum_{t=0}^{H-1} \gamma^t \cdot r(x_{t,m}, u_{t,m}) \right), \tag{4.2}$$

where M is the batch size, $x_{0,m} \sim \zeta$, $u_{t,m} = \pi(x_{t,m}, \varsigma_m)$, $\varsigma_m \sim p(\varsigma)$, and $x_{t+1,m} = f(x_{t,m}, u_{t,m})$.

The update rule for the policy parameter θ with learning rate η is as follows:

$$\theta_{n+1} \leftarrow \theta_n + \eta \cdot \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}). \tag{4.3}$$

4.2 Convergence of Model-Based FOPG

To begin our analysis, 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 $\mathcal{J}(\pi_\theta)$, which is required by most existing studies on the policy gradient methods (Wang et al., 2019a; Bastani, 2020; Agarwal et al., 2020).

Assumption 4.1 (Lipschitz Continuous Policy Gradient)**.** Assume that $\nabla_\theta \mathcal{J}(\pi_\theta)$ is L-Lipschitz continuous in θ, such that $\|\nabla_\theta \mathcal{J}(\pi_{\theta_1}) - \nabla_\theta \mathcal{J}(\pi_{\theta_2})\|_2 \le L\|\theta_1 - \theta_2\|_2$.

We characterize the convergence of model-based FOPG by first providing the following proposition.

Theorem 4.2 (Convergence to Stationary Points)**.** Define the gradient bias b_n and variance v_n at iteration n as

$$b_n := \left\| \nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\left[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\right] \right\|_2,$$
$$v_n := \mathbb{E}\left[\left\| \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\left[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\right] \right\|_2^2 \right].$$

Denote $\delta := \sup \|\theta\|_2$ and $c := (\eta - L\eta^2)^{-1}$. It then holds for $N \ge 4L^2$ that

$$\min_{n\in[N]} \mathbb{E}\left[\left\| \nabla_\theta \mathcal{J}(\pi_{\theta_n}) \right\|_2^2 \right] \le \frac{4c}{N} \cdot \mathbb{E}\left[\mathcal{J}(\pi_{\theta_N}) - \mathcal{J}(\pi_{\theta_1}) \right] + \frac{4}{N} \left(\sum_{n=0}^{N-1} c(2\delta \cdot b_n + \frac{\eta}{2} \cdot v_n) + b_n^2 + v_n \right).$$

Theorem 4.2 illustrates the reliance between the convergence and the variance, bias of the gradient estimators. In general, to guarantee the convergence of model-based FOPG, we have to control both the variance and the bias to the sublinear growth rate. Before studying the upper bound of b_n and v_n, we make the following Lipschitz assumption, which is adopted in various previous works (Pirotta et al., 2015; Clavera et al., 2020; Li et al., 2021).

Assumption 4.3 (Lipschitz Continuity)**.** Assume the policy, model, and reward are L_π, L_f, L_r Lipschitz continuous.

4.3 Gradient Variance and LCS Stiffness

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

Theorem 4.4 (Gradient Variance)**.** Under Assumption 4.3, at iteration $n \in [N]$, the gradient variance of FOPG satisfies:

$$v_n = O\Big(H^4 \widetilde{L}_f^{4H} \widetilde{L}_\pi^{4H} / M\Big). \tag{4.4}$$

We observe that the upper bound on variance is dependent on the Lipschitz of the model and policy in a polynomial manner, with degrees that are linear in relation to the effective horizon. This makes intuitive sense — when the system is chaotic (Bollt, 2000), as measured by the Jacobian of the dynamical system, the stochasticity during training can lead to diverging trajectories and gradient directions, causing large gradient variance. The optimization difficulties imposed by non-smooth models, such as hard contact models, result in slow convergence or training failure even in simple tasks (Parmas et al., 2018; Suh et al., 2022a).

The above analysis applies to model-based FOPG in general. When adopting the complementarity-based contact model f_μ, studying its stiffness, i.e. the Lipschitz L_{f_μ}, is especially important since they are inherently highly non-smooth at local mode-switching points. We characterize the stiffness of the softened LCS using the following theorem.

Theorem 4.5 (Stiffness of the Softened LCS)**.** Let $\|\cdot\|_F$ denote the matrix Frobenius norm and define $\varepsilon := \sup \|Dx_t + Eu_t + d\|_2^2/(2\|F\|_F^2)$. Under Assumption 3.2, the Lipschitz L_{f_μ} of the model f_μ defined in (3.1) satisfies

$$L_{f_\mu} \le (\|A\|_F + \|B\|_F) + d_\lambda^2 \|C\|_F (\|D\|_F + \|E\|_F) \cdot l(\mu),$$

where $l(\mu)$ is determined by μ and is lower bounded by:

$$l(\mu) \ge \frac{\varepsilon}{\mu} + \frac{1}{\|F\|_F} + \varepsilon \sqrt{\frac{1}{\mu^2} + \frac{2}{\varepsilon \mu \|F\|_F}}.$$

Theorem 4.5 highlights the crucial role of the central-path parameter μ in determining the model stiffness. Specifically, the upper bound of L_{f_μ} (and thus of the variance (4.4)) is at least inversely proportional to μ. This poses challenges when implementing FOPG based on f_μ, as achieving accurate dynamics necessitates solving the exact LCP ($\mu \to 0$), which unfortunately leads to a significant increase in gradient variance as $l(\mu) \to \infty$. The optimization obstacles, such as chaotic and non-smooth landscapes, remain present even when contact events are occasional in a full model unroll.

5 Contact-Sensitive Analytical Barrier Smoothing

5.1 Method

A natural idea to alleviate the exploding FOPG variance issue is to differentiate through the intermediate solution of IPM when calculating the gradient in (4.2). This can be achieved by setting a positive stopping criteria, μ_{sc}, that terminates the IPM iterations when μ is approximately equal to μ_{sc}. The specific implementation details, including pseudocode, can be found in Appendix B. According to Lemma 3.3, μ-softened systems correspond to the smoothed objectives with log-barrier functions. Therefore, this method is referred to as *Analytical Barrier Smoothing*.

Unfortunately, vanilla analytical (barrier) smoothing with a constantly large μ can lead to a huge gradient bias since the generated trajectories will *not* adhere to physics laws. Consequently, to attain successful convergence in Theorem 4.2, additional care must be taken to take the trade-off between gradient variance and gradient bias.

To control the FOPG gradient bias, we propose utilizing *contact-sensitive* analytical barrier smoothing. Specifically, instead of a fixed μ_{sc}, we use an *adaptive* $\mu_{\text{sc}} = \mu(x_t, u_t) > 0$ that adjusts *inversely* with the minimum distance-to-obstacle $|\phi(x_t, u_t)|$ of the *inactive impact contacts*.

Our contact-sensitive design is based on the fact that the stiffness of complementarity-based models is a result of the sudden change in impact force when penetration first arises. For example, in systems depicted in Fig. 1(a) and 1(b), the velocity is continuous everywhere except at the hard contact location $z = 0$. Thus, when states contain the velocity information, the transitions are stiff around $z = 0$ (see Section 7.1 for an illustration). As a result, when performing FOPG, we only need *local* smoothing near the inactive impact contact to avoid large variance, while still achieving *globally* accurate gradients with minimal bias.

The selection of the contact-sensitive $\mu(x_t, u_t)$ can be problem-specific, as long as it decreases with $|\phi(x_t, u_t)|$ (e.g., we use (7.1) in our Dojo experiments). In what follows, we show that analytical barrier smoothing has a strong correlation with randomized smoothing and, when a suitable form of $\mu(x_t, u_t)$ is taken, enjoys a small gradient bias.

5.2 Analysis of Gradient Bias

Since we are interested in controlling the system stiffness and the large variance caused by *impact* contact, our focus in this section is on frictionless systems with a single point of contact. This streamlines our analysis by reducing d_λ to 1. While the findings may be applicable to more extensive scenarios, their forms are beyond the scope of this paper.

As a first step, we build the connection between the proposed *analytical barrier smoothing* and *randomized smoothing* (Suh et al., 2022a;b; Pang et al., 2022), which samples and averages the stochastic gradient. We show that these two smoothing techniques are essentially identical.

Proposition 5.1 (Equivalence between ABS and Randomized Smoothing)**.** Denote $z_t := Dx_t + Eu_t + d \in \mathbb{R}$. Recall that the solution of the exact LCP is $S_{\mu=0}(z_t)$ and the analytically smoothed LCP solution is $S_{\mu(z_t)}(z_t)$ (see Defn. 3.1). For any centering function $\mu(z_t)$, analytical smooth-

ing is equivalent to randomized smoothing: $S_{\mu(z_t)}(z_t) = \mathbb{E}_{w\sim\rho(w)}[S_{\mu=0}(z_t + w)]$, where $\rho(w) = \nabla_w^2 S_{\mu(z_t)}(w)$.

The above proposition shows that analytical barrier smoothing inherently smoothens the contact force λ_t (w.r.t. z_t), and as a result, also smoothens the dynamics $x_{t+1} = f_\mu(x_t, u_t)$ because x_t, u_t are prefixed. More importantly, by choosing a proper adaptive central-path parameter $\mu(z_t)$, the proposed method can accommodate any randomized smoothing method, while avoiding its drawbacks when calculating first-order gradients, which we will discuss in more detail.

As a consequence of Proposition 5.1, we are able to work directly on the randomization-smoothed gradient when studying the bias of analytical smoothing. This gives the following results adapted from the analysis on randomized smoothing presented in (Pang et al., 2022).

Proposition 5.2 (Smoothing as Linearization Minimizer)**.** Define the error function as the σ-Gaussian tail integral $\text{erf}(y;\sigma^2) := \int_y^\infty 1/(\sqrt{2\pi}\sigma)e^{-y^2/\sigma^2}$. Set the z_t-adaptive central-path parameter as $\mu(z_t) = \kappa \cdot (z_t + F\kappa)$, where

$$\kappa := z_t \cdot \text{erf}(z_t, \sigma) + e^{-z_t^2/(2\sigma)}/\sqrt{\pi} + c_1 z_t + c_2, \quad (5.1)$$

and $c_1, c_2 \in \mathbb{R}$ are tunable constants. Consider the problem of regressing the exact LCP solution $S_{\mu=0}$ with parameters (K, W) such that the residual around z_t distributed according to Gaussian is minimized, formally:

$$\delta = \min_{K,W} \mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}\Big[\big|S_{\mu=0}(z_t + w) - Ww - K\big|\Big].$$

The K^*, W^* that achieve the minimum are the analytically smoothed solution and its gradient, respectively:

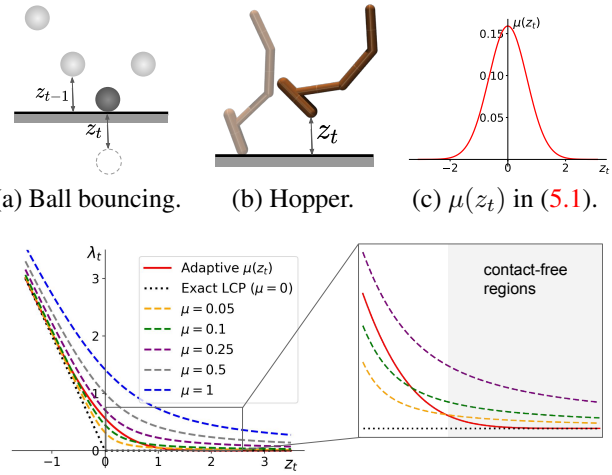
$$K^* = S_{\mu(z_t)}(z_t), \quad W^* = \nabla_z S_{\mu(z_t)}(z_t).$$

The above proposition shows that the solution of analytically smoothed LCP is the best linear approximation of the exact LCP solution around z_t. Thus, with a small linearization residual, we can conclude a small *gradient* bias.

Figure 1(c) demonstrates that the $\mu(z_t)$ defined in (5.1) is sensitive to contact as it is only positive when in the vicinity of contact $z_t = 0$. This design supports our intuition — when the body is away from contact, we can safely solve the LCP and get accurate simulations; when experiencing contact, the proposed method smoothens the LCP to obtain non-stiff local dynamics. This is also evident from Figure 1(d): λ_t is more accurate at contact-free regions while achieving the "force-at-a-distance" relaxation around $z_t = 0$.

Theorem 5.3 (Bias of Analytical Barrier Smoothing)**.** With the same definition of $\mu(z_t)$ in Proposition 5.2, the gradient of the softened LCS model $f_{\mu(z_t)}$ approximately matches the gradient of LCS $f_{\mu=0}$, with the bias upper bounded by:

$$\big\|\nabla f_{\mu=0} - \nabla f_{\mu(z_t)}\big\|_2 \leq \|C\|_F(\|D\|_F + \|E\|_F) \cdot \Big(\frac{\sigma F^2 \mathcal{Q}(3/4)}{2} + \frac{12\delta + \varsigma}{\sigma \mathcal{Q}(2/3)}\Big),$$



(a) Ball bouncing. (b) Hopper. (c) $\mu(z_t)$ in (5.1).

(d) Contact impulses λ_t of the exact and the softened LCP, including constant μ and contact-sensitive adaptive $\mu(z_t)$.

Figure 1. 1(a), 1(b): Example systems. The dashed circle in 1(a) arises penetration $z_t < 0$, where the contact force $\lambda_t > 0$ pushes the ball to be above the ground. 1(c): Plot of the proposed adaptive $\mu(z_t)$. 1(d): Contact force comparison. The adaptive $\mu(z_t)$ is contact-sensitive and has a better trade-off between controlling the stiffness and reducing the bias: $\mu(z_t)$ not only gives smoother dynamics (compared to $\mu \leq 0.1$) around the contact $z = 0$, but also best approximates the exact LCP solution at contact-free regions.

where we define $\varsigma := 1/\sqrt{\pi} + c_2$ and $\mathcal{Q} : [0, 1] \to \mathbb{R}$ is the inverse of the cumulative distribution function (or quantile function) of the standard normal distribution, $\mathcal{Q}(3/4) \approx 0.67$, $\mathcal{Q}(2/3) \approx 0.43$.

Theorem 5.3 establishes a bound on the gradient bias of analytical barrier smoothing when the contact-sensitive $\mu(z_t)$ conforms to certain forms. If the model parameters ψ are accurately fitted with supervised learning, i.e. $f_{\mu=0} \approx f^*$, the softened LCS $f_{\mu(z_t)}$ and its gradient achieves the best linearization error and has small gradient bias b_n.

Discussion on Randomized Smoothing. Despite the fact that analytical smoothing and randomized smoothing can ultimately be shown to be equivalent in the event of an infinite number of samples, the latter is plagued by both empirical bias (Suh et al., 2022b;a) and the presence of noisy gradients (Howell et al., 2022). The empirical bias phenomenon happens under discontinuities or stiffness (see Fig. 7(a)). Even if the system is non-stiff, sampling and averaging the noise-induced gradients is noisy and computationally expensive. In contrast, analytical smoothing by directly differentiating through the softened system $f_{\mu>0}$ prevents the above issues.

6 Related Work

Differentiable Simulation. The physics-guided (Jiang et al., 2018; Pizzuto & Mistry, 2021) complementarity-based model that is used in this work is adopted in various differentiable hard-contact engines, such as Dojo (Howell

et al., 2022), DART (Werling et al., 2021), and Bullet (Heiden et al., 2021). These simulators provide readily available gradients of simulation outcomes w.r.t. control actions. However, the extreme curvatures of contact events prevent the (sub-)gradients from being effective when performing FOPG, and our method offers a potential solution. On the other hand, simulators like MuJoCo (Todorov et al., 2012) and PhysX implement soft contacts and can generate physics-violated behaviors (Howell et al., 2022). Besides, their non-differentiable nature necessitates expensive finite-difference to obtain the first-order gradients. Our analysis of the gradient bias is most closely related to (Pang et al., 2022). However, they analyzed randomized smoothed gradients of the proposed quasi-dynamic differentiable model, while we study the general form where the equivalence between ABS and randomized smoothing holds, with the ultimate goal to bound the gradient bias. Moreover, (Suh et al., 2022a;b) aim to address the empirical bias issue of randomized smoothing, but we focus primarily on the trade-off between gradient variance and bias when applying analytical smoothing.

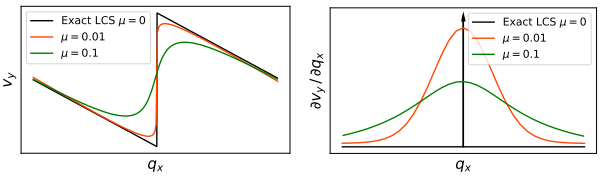
Policy Gradient Methods. The zeroth-order policy gradient methods include REINFORCE (Williams, 1992) and actor-critic (Sutton et al., 1999; Kakade, 2001; Kakade & Langford, 2002; Degris et al., 2012), where the convergence results are established in recent works (Agarwal et al., 2021; Wang et al., 2019a; Bhandari & Russo, 2019; Liu et al., 2019). However, first-order policy gradient methods have received less attention. Difficulties in optimization, such as discontinuous contact behaviors and the curse of chaos (Parmas et al., 2018; Metz et al., 2021; Xu et al., 2022), have hindered the widespread use of FOPG even in differentiable simulation. To alleviate this issue, (Xu et al., 2022) proposed to shorten the optimization horizon and (Clavera et al., 2020) proposed to leverage model-critic expanded values. In this work, we focus on the naive implementation of FOPG. Modifications from previous works can be naturally integrated, e.g. using an additional critic as the tail estimation (Clavera et al., 2020), minimizing the model gradient error (Li et al., 2021), or adding actor entropy loss (Amos et al., 2021).

7 Experiments

7.1 Contact Behaviors and System Stiffness

To begin, we aim to shed light on the insight behind Analytical Barrier Smoothing (ABS). In Figure 2, we plot the dynamics and derivatives of the velocity w.r.t. coordinate in the ball-bouncing example, where the ball is thrown with an initial velocity and subsequently experiences the effects of gravity and impact force upon hitting the ground. These contact events, which serve as the foundation for complex behaviors, are prevalent in nearly all robotic tasks.

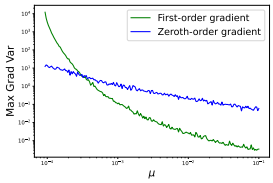
In Fig. 2(a), the velocity is discontinuous at contact due to the sudden shift in impact force γ from 0 to a positive value. This results in a stiff system $f_{\mu=0}(x)$, where the state $x := (q_x, v_y)$. By applying ABS with a larger value of μ, the dynamics of the system become less stiff.



(a) Dynamics. (b) Derivative.

Figure 2. Contact behaviors in the Fig. 1(a) ball-bouncing example. 2(a): The vertical velocity v_y w.r.t. the x-coordinate q_x in the exact LCS and in the μ-smoothed system. 2(b): Derivative of v_y w.r.t. q_x. The black arrow represents the impulse function, i.e. $\partial v_y / \partial q_x = \infty$ at the contact point and $= -g$ (gravity) elsewhere.

7.2 First-Order Gradient Variance



We now investigate the ball-bouncing dynamics with the inclusion of Gaussian noise. In the right figure, we plot the maximum variance of first-order reparameterization gradients and zeroth-order likelihood ratio (LR) gradients. We observe that stiff systems with small μ lead to large first-order gradient variance. This is a result of the curse of chaos, where non-smooth dynamics can cause gradients and trajectories to diverge due to the presence of stochasticity. Here, LR gradients are parameterized with Gaussian, following evolutionary strategies (Salimans et al., 2017; Mania et al., 2018). Despite exhibiting low variance in the presence of stiffness due to its sole reliance on function evaluations, the LR gradient is notorious for its poor scaling capabilities when the dimensionality increases (see Section 7.3).

We then conduct experiments in the Dojo (Howell et al., 2022) physics engine, which enables differentiable simulation with hard contact. For now, we use the ground-truth physics parameters. The mean gradient variance in FOPG training for Dojo locomotion tasks is illustrated in Figure 3.

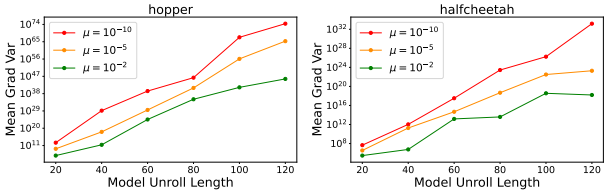


Figure 3. The mean gradient variance with different model unroll lengths when changing the value of μ.

We observe that the gradient variance of FOPG can explode with exponential order w.r.t. the horizon or model unroll length. As the value of μ increases, indicating a larger Lipschitz constant in the complementarity-based model, the variance decreases. This confirms our findings in Thm. 4.4.

7.3 Contact-Sensitive Analytical Barrier Smoothing

We evaluate the analytical barrier smoothing mechanism applied to first-order policy gradient algorithms by examining its effectiveness in optimizing the angle of throwing a ball to reach the goal. The initial speed and height are set to specific values in order to guarantee that contact takes place before the goal is reached.

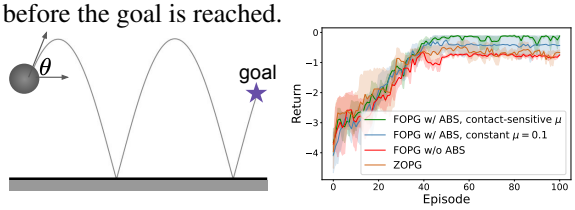


Figure 4. Throw a ball and optimize θ to reach the goal at a certain velocity. Return is the negative mean squared error at the final step.

We can see from Fig. 4 that the application of ABS to FOPG results in a higher asymptotic return and faster convergence compared to both ZOPG and vanilla FOPG. Besides, contact-sensitive ABS achieves superior performance.

Moreover, we report the results evaluated in Dojo locomotion tasks in Fig. 5. For MBPO, the NN models are trained by minimizing the mean squared error. For all other algorithms, we use complementarity-based models. In our method, we would like $\mu \to 0$ when all impact contacts are active or the distance between contact and obstacles is large, and $\mu \approx 10^{-2}$ when this distance approaches zero, based on the result in Fig. 3. To accomplish this, the contact-sensitive $\mu(x_t, u_t)$ is designed in the following manner:

$$\mu(x_t, u_t) = 10^{-2}\big(100 \min_{i \in \mathcal{I}} |\phi(x_t, u_t)^{(i)}|^2 + 1\big)^{-4}, \quad (7.1)$$

where $\mathcal{I} := \{1 \le i \le c \,|\, \gamma_{t-1}^{(i)} \le 1\}$ represents the set of inactive impact contact points and recall that ϕ is the SDF.

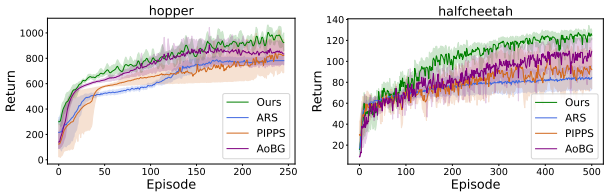
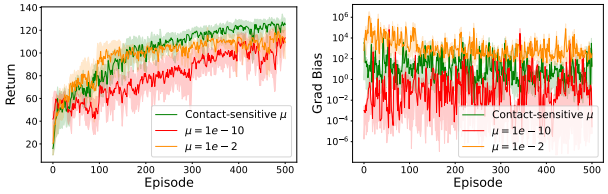


Figure 5. Comparison between contact-sensitive ABS applied to FOPG (Ours), ARS (Mania et al., 2018), PIPPS (Parmas et al., 2018), AoBG (Suh et al., 2022a), and MBPO (Janner et al., 2019).

7.4 Ablation Studies

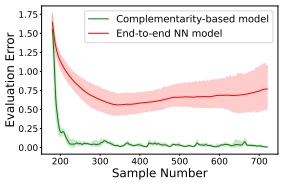
Gradient Bias. In Figure 6, we compare the performance of FOPG in the half-cheetah task when using different values of μ for ABS. As we are using ground-truth physics parameters, the only source of bias in this comparison is the smoothing parameter. Our results show that $\mu = 10^{-10}$ results in a small gradient bias but slow convergence due to high variance. On the other hand, $\mu = 10^{-2}$ leads to a larger bias and lower asymptotic return. Our proposed contact-sensitive design offers a more favorable balance between gradient bias and variance.



(a) Performance. (b) Gradient bias.

Figure 6. FOPG ABS equipped with different choices of μ.

Model Learning. To demonstrate that complementarity-based models are more effective at approximating contact behaviors than NN models, we plot the mean state prediction error in the hopper task evaluated on an evaluation transition data set collected by random policies.



Different Smoothing Mechanisms. Figure 7(a) illustrates the derivatives of the *impact* contact dynamics in the ball-bouncing system, from which we observe the empirical bias phenomenon of randomized smoothing. Specifically, the AS gradient successfully approximates the unit impulse at contact, while the RS gradient is constant and has a large bias. For frictional contact 7(b), the RS gradient is both noisy and computationally costly due to the sampling process.

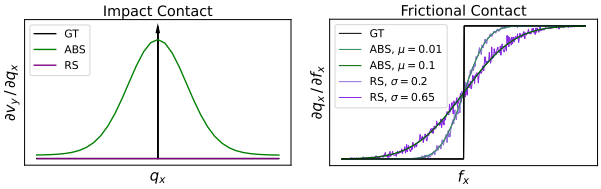


Figure 7. Comparison between Analytical Barrier Smoothing and Randomized Smoothing (RS) in the two types of contact dynamics.

8 Conclusion

In this work, we study the model-based First-Order Policy Gradient (FOPG) methods in robotic contact dynamics with extreme curvatures, focusing on complementarity-based models. We find that the convergence of FOPG relies on the gradient variance and bias, and stiff models can lead to large gradient variance and optimization difficulties. To fix this issue, we propose Analytical Barrier Smoothing to reduce the model stiffness, and control the gradient bias with a contact-sensitive central-path parameter.

Our work also opens some new problems. It would be interesting to investigate how the *soft* contact dynamics in systems like MuJoCo affect the model learning and policy gradients. Besides, our study on the gradient bias is conducted in simplified settings. We leave the analysis of general contact systems for future work.

References

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.

Anitescu, M. and Potra, F. A. A time-stepping method for stiff multibody dynamics with contact and friction. *International journal for numerical methods in engineering*, 55(7):753–784, 2002.

Aydinoglu, A., Preciado, V. M., and Posa, M. Contact-aware controller design for complementarity systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1525–1531. IEEE, 2020.

Aydinoglu, A., Sieg, P., Preciado, V. M., and Posa, M. Stabilization of complementarity systems via contact-aware controllers. *IEEE Transactions on Robotics*, 2021.

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.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

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.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

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

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

Drumwright, E. and Shell, D. A. Extensive analysis of linear complementarity problem (lcp) solver performance on randomly generated rigid body contact problems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5034–5039. IEEE, 2012.

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.

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.

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, 2021.

Hochlehnert, A., Terenin, A., Sæmundsson, S., and Deisenroth, M. Learning contact dynamics using physically structured neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 2152–2160. PMLR, 2021.

Howell, T. A., Cleac'h, S. L., Kolter, J. Z., Schwager, M., and Manchester, Z. Dojo: A differentiable simulator for robotics. *arXiv preprint arXiv:2203.00806*, 2022.

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.

Jiang, Y., Sun, J., and Liu, C. K. Data-augmented contact model for rigid body simulation. *arXiv preprint arXiv:1803.04019*, 2018.

Jin, W., Aydinoglu, A., Halm, M., and Posa, M. Learning linear complementarity systems. In *Learning for Dynamics and Control Conference*, pp. 1137–1149. PMLR, 2022.

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.

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.

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.

Mania, H., Guy, A., and Recht, B. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

Mehrotra, S. On the implementation of a primal-dual interior point method. *SIAM Journal on optimization*, 2(4):575–601, 1992.

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

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.

Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566. IEEE, 2018.

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.

Parmar, M., Halm, M., and Posa, M. Fundamental challenges in deep learning for stiff contact dynamics. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5181–5188. IEEE, 2021.

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.

Pfrommer, S., Halm, M., and Posa, M. Contactnets: Learning discontinuous contact dynamics with smooth, implicit representations. In *Conference on Robot Learning*, pp. 2279–2291. PMLR, 2021.

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

Pizzuto, G. and Mistry, M. Physics-penalised regularisation for learning dynamics models with contact. In *Learning for Dynamics and Control*, pp. 611–622. PMLR, 2021.

Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

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

Tassa, Y. and Todorov, E. Stochastic complementarity for local control of discontinuous dynamics. 2010.

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.

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.

Werling, K., Omens, D., Lee, J., Exarchos, I., and Liu, C. K. Fast and feature-complete differentiable physics engine for articulated rigid bodies with contact constraints. In *Robotics: Science and Systems*, 2021.

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

Wright, S., Nocedal, J., et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

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 Proofs

A.1 Proof of Lemma 3.3

Proof. Corresponding to the constrained optimization problem (3.2) we can introduce the multipliers ι and form the Lagrangian function by

$$L(\lambda_t, \epsilon_t, \iota) = \lambda_t^\top \epsilon_t - \mu \sum_{i=1}^{n_\lambda} \big(\log \lambda_t^{(i)} + \log \epsilon_t^{(i)}\big) + \iota^\top (Dx_t + Eu_t + F\lambda_t + d - \epsilon_t).$$

Here, we omit the last equality constraint in (3.2) since x_{t+1} can be directly calculated when λ_t is obtained.

We have from the Karush–Kuhn–Tucker (KKT) conditions that the optimal solution must satisfy

$$\frac{\partial}{\partial \lambda_t^{(i)}} L(\lambda_t, \epsilon_t, \iota) = \epsilon_t^{(i)} - \mu \cdot \frac{1}{\lambda_t^{(i)}} + (\iota^\top F)^{(i)} - \iota_2^{(i)} = 0, \tag{A.1}$$

$$\frac{\partial}{\partial \epsilon_t^{(i)}} L(\lambda_t, \epsilon_t, \iota) = \lambda_t^{(i)} - \mu \cdot \frac{1}{\epsilon_t^{(i)}} - \iota_1^{(i)} - \iota_3^{(i)} = 0, \tag{A.2}$$

$$Dx_t + Eu_t + F\lambda_t + d = \epsilon_t, \tag{A.3}$$

where (A.1), (A.2) follow from the stationarity of the optimal solution, and (A.3) follows from the primal feasibility.

Combining the above equations, we know that $\epsilon_t^{(i)} \lambda_t^{(i)} = \mu$ and $\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu \vec{1}$. □

A.2 Proof of Theorem 4.2

Proof. From the policy update rule, we know that $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) = (\theta_{n+1} - \theta_n)/\eta$. By the Lipschitz Assumption 4.3, we have

$$\begin{aligned}
\mathcal{J}(\pi_{\theta_{n+1}}) - \mathcal{J}(\pi_{\theta_n}) &\geq \nabla_\theta \mathcal{J}(\pi_{\theta_n})^\top (\theta_{n+1} - \theta_n) - \frac{L}{2} \|\theta_{n+1} - \theta_n\|_2^2 \\
&= \eta \nabla_\theta \mathcal{J}(\pi_{\theta_n})^\top \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \frac{L\eta^2}{2} \big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2.
\end{aligned} \tag{A.4}$$

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

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

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

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

Thus, we have the following inequality for (A.4):

$$\begin{aligned}
\mathcal{J}(\pi_{\theta_{n+1}}) - \mathcal{J}(\pi_{\theta_n}) \geq \frac{\eta}{2} \cdot \Big(&-\left\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\right\|_2 \cdot 2b_n - \left\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})]\right\|_2^2 + \left\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\right\|_2^2\Big) \\
&- \frac{L\eta^2}{2} \cdot \left\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\right\|_2^2.
\end{aligned} \tag{A.5}$$

By taking expectation in (A.5), we obtain

$$\mathbb{E}[\mathcal{J}(\pi_{\theta_{n+1}}) - \mathcal{J}(\pi_{\theta_n})] \geq -\eta \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2\Big] \cdot b_n - \frac{\eta}{2} \cdot v_n + \frac{\eta - L\eta^2}{2} \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big].$$

By rearranging terms,

$$\frac{\eta - L\eta^2}{2} \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] \le \mathbb{E}[\mathcal{J}(\pi_{\theta_{n+1}}) - \mathcal{J}(\pi_{\theta_n})] + \eta \mathbb{E}[\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\|_2] b_n + \frac{\eta}{2} v_n. \tag{A.6}$$

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

$$\begin{aligned}
\mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] &= \mathbb{E}\Big[\Big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big] + \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big] - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\Big\|_2^2\Big] \\
&\le 2\Big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big]\Big\|_2^2 + 2\mathbb{E}\Big[\Big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big]\Big\|_2^2\Big] \\
&= 2b_n^2 + 2v_n,
\end{aligned} \tag{A.7}$$

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

$$\|y + z\|_2^2 \le \|y\|_2^2 + \|z\|_2^2 + 2\|y\|_2 \cdot \|z\|_2 \le 2\|y\|_2^2 + 2\|z\|_2^2. \tag{A.8}$$

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

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

where the second inequality follows from (A.8).

For $N \ge 4L^2$, by setting $\eta = 1/\sqrt{N}$, we have $\eta < 1/L$ and $(\eta - L\eta^2)/2 > 0$. Therefore, following the results in (A.6) and (A.7), we further have

$$\begin{aligned}
&\min_{n\in[N]} \mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] \\
&\le \frac{4c}{N} \cdot \left(\mathbb{E}[\mathcal{J}(\pi_{\theta_N}) - \mathcal{J}(\pi_{\theta_1})] + \sum_{n=0}^{N-1} \left(\eta \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2\Big] \cdot b_n + \frac{\eta}{2} \cdot v_n\right)\right) + \frac{4}{N} \cdot \sum_{n=0}^{N-1} (b_n^2 + v_n) \\
&= \frac{4}{N} \cdot \left(\sum_{n=0}^{N-1} c \cdot \left(\eta \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2\Big] \cdot b_n + \frac{\eta}{2} \cdot v_n\right) + b_n^2 + v_n\right) + \frac{4c}{N} \cdot \mathbb{E}\big[\mathcal{J}(\pi_{\theta_N}) - \mathcal{J}(\pi_{\theta_1})\big],
\end{aligned}$$

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

By noting that $\eta \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) = \theta_{n+1} - \theta_n$, we conclude the proof by

$$\begin{aligned}
&\min_{n\in[N]} \mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] \\
&\le \frac{4}{N} \cdot \left(\sum_{n=0}^{N-1} c \cdot \left(\mathbb{E}\Big[\big\|\theta_{n+1} - \theta_n\big\|_2\Big] \cdot b_n + \frac{\eta}{2} \cdot v_n\right) + b_n^2 + v_n\right) + \frac{4c}{N} \cdot \mathbb{E}\big[\mathcal{J}(\pi_{\theta_N}) - \mathcal{J}(\pi_{\theta_1})\big] \\
&\le \frac{4}{N} \cdot \left(\sum_{n=0}^{N-1} c \cdot (2\delta \cdot b_n + \frac{\eta}{2} \cdot v_n) + b_n^2 + v_n\right) + \frac{4c}{N} \cdot \mathbb{E}\big[\mathcal{J}(\pi_{\theta_N}) - \mathcal{J}(\pi_{\theta_1})\big].
\end{aligned}$$

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

A.3 Proof of Theorem 4.4

In what follows, we interchangeably write $\nabla_a b$ and $\mathrm{d}b/\mathrm{d}a$ as the derivative, and use the notation $\partial b/\partial a$ to denote the partial derivative. With slight abuse of notation, for vector s and vector w, we denote the Jacobian matrix consisting of entries $\partial s^{(i)}/\partial w^{(j)}$ as $\partial s/\partial w$.

Proof. In order to upper-bound the gradient variance $v_n = \mathbb{E}[\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})]\|_2^2]$, 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 $M = 1$, which can naturally generalize to $N > 1$. Specifically, consider an *arbitrary* trajectory obtained by unrolling the model under policy π_{θ_n}. Denote the pathwise gradient $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})$ of this trajectory as g'. Then we have

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

where we let g denote the pathwise gradient $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})$ of a *fixed* (but unknown) trajectory $(x_0, u_0, x_1, u_1, \cdots)$ such that the maximum is achieved.

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

$$v_n \leq \mathbb{E}\Big[\big\| g - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) \big\|_2 \Big]^2. \tag{A.9}$$

Denote $y_t := (x_t, u_t)$. By triangular inequality, we have

$$\mathbb{E}\Big[\big\| g - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) \big\|_2 \Big] \leq \sum_{t=0}^{H-1} \mathbb{E}_{\overline{y}_t} \Big[\big\| \nabla_\theta r(y_t) - \nabla_\theta r(\overline{y}_t) \big\|_2 \Big]. \tag{A.10}$$

For $t \geq 1$, we have the following relationship according to the chain rule:

$$\frac{\mathrm{d}u_t}{\mathrm{d}\theta} = \frac{\partial u_t}{\partial x_t} \cdot \frac{\mathrm{d}x_t}{\mathrm{d}\theta} + \frac{\partial u_t}{\partial \theta}, \tag{A.11}$$

$$\frac{\mathrm{d}x_t}{\mathrm{d}\theta} = \frac{\partial x_t}{\partial x_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}. \tag{A.12}$$

Plugging $\mathrm{d}u_{t-1}/\mathrm{d}\theta$ in (A.11) into (A.12), we get

$$\frac{\mathrm{d}x_t}{\mathrm{d}\theta} = \left(\frac{\partial x_t}{\partial x_{t-1}} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\partial u_{t-1}}{\partial x_{t-1}} \right) \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\partial u_{t-1}}{\partial \theta}. \tag{A.13}$$

By the Cauchy-Schwarz inequality and the Lipschitz Assumption 4.3, we have

$$\left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 \leq L_f \widetilde{L}_\pi \cdot \left\| \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} \right\|_2 + L_f L_\theta.$$

Applying the above recursion gives us

$$\left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 \leq L_f L_\theta \cdot \sum_{j=0}^{t-1} L_f^j \widetilde{L}_\pi^j \leq i \cdot L_\theta L_f^{t+1} \widetilde{L}_\pi^t, \tag{A.14}$$

where the first inequality follows from the induction

$$z_n = a z_{t-1} + b = a \cdot (a z_{i-2} + b) + b = a^t \cdot z_0 + b \cdot \sum_{j=0}^{t-1} a^j, \tag{A.15}$$

for the real sequence $\{z_j\}_{0 \leq j \leq i}$ satisfying $z_j = a z_{j-1} + b$. For $\mathrm{d}u_t/\mathrm{d}\theta$ defined in (A.11), we further have

$$\left\| \frac{\mathrm{d}u_t}{\mathrm{d}\theta} \right\|_2 \leq L_\pi \cdot \left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 + L_\theta \leq t \cdot L_\theta L_f^{t+1} \widetilde{L}_\pi^{t+1} + L_\theta. \tag{A.16}$$

Combining (A.14) and (A.16), we obtain

$$\left\| \frac{\mathrm{d}y_t}{\mathrm{d}\theta} \right\|_2 = \left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 + \left\| \frac{\mathrm{d}u_t}{\mathrm{d}\theta} \right\|_2 \leq K(t) := 2t \cdot L_\theta L_f^{t+1} \widetilde{L}_\pi^{t+1} + L_\theta, \tag{A.17}$$

where $K(t)$ is introduced for notation simplicity.

By the chain rule, (A.10) can be decomposed and bounded by

$$\begin{aligned}
&\mathbb{E}_{\overline{y}_t}\left[\left\|\nabla_\theta r(y_t) - \nabla_\theta r(\overline{y}_t)\right\|_2\right] \\
&= \mathbb{E}_{\overline{y}_t}\left[\left\|\nabla r(y_t)\nabla_\theta y_t - \nabla r(\overline{y}_t)\nabla_\theta \overline{y}_t\right\|_2\right] \\
&\le \mathbb{E}_{\overline{y}_t}\left[\left\|\nabla r(y_t)\nabla_\theta y_t - \nabla r(y_t)\nabla_\theta \overline{y}_t\right\|_2\right] + \mathbb{E}\left[\left\|\nabla r(y_t)\nabla_\theta \overline{y}_t - \nabla r(\overline{y}_t)\nabla_\theta \overline{y}_t\right\|_2\right] \\
&\le L_r \cdot \left(\mathbb{E}_{\overline{x}_n}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}_{\overline{u}_n}\left[\left\|\frac{\mathrm{d}u_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta}\right\|_2\right]\right) + 2L_r \cdot K(t),
\end{aligned} \tag{A.18}$$

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

Plugging (A.18) into (A.10) and (A.9), we have

$$\begin{aligned}
v_n &\le L_r \cdot \left(\sum_{t=0}^{H-1}\left(\mathbb{E}_{\overline{x}_t}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}_{\overline{u}_t}\left[\left\|\frac{\mathrm{d}u_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta}\right\|_2\right] + 2K(t)\right)\right)^2 \\
&= O\left(\left(\sum_{t=0}^{H-1} t^2 \widetilde{L}_f^{2t}\widetilde{L}_\pi^{2t}\right)^2\right) = O\left(H^4 \widetilde{L}_f^{4H}\widetilde{L}_\pi^{4H}\right),
\end{aligned} \tag{A.19}$$

where the second inequality follows from the results from Lemma A.1 and by plugging the definition of K in (A.17). Since the analysis above considers batch size $M = 1$, the bound of gradient variance v_n is established by dividing M, which concludes the proof. □

Lemma A.1. Denote $e := \sup \mathbb{E}_{\overline{x}_0}[\|\mathrm{d}x_0/\mathrm{d}\theta - \mathrm{d}\overline{x}_0/\mathrm{d}\theta\|_2]$, which is a constant that only depends on the initial state distribution[1]. For any timestep $t \ge 1$ and the corresponding state x_t, control input u_t, we have the following inequality results:

$$\begin{aligned}
\mathbb{E}_{\overline{x}_t}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] &\le \widetilde{L}_f^t\widetilde{L}_\pi^t\left(e + 4t \cdot \widetilde{L}_f\widetilde{L}_\pi \cdot K(t-1) + 2t \cdot \widetilde{L}_f L_\theta\right), \\
\mathbb{E}_{\overline{u}_n}\left[\left\|\frac{\mathrm{d}u_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta}\right\|_2\right] &\le \widetilde{L}_f^t\widetilde{L}_\pi^{t+1}\left(e + 4i \cdot \widetilde{L}_f\widetilde{L}_\pi \cdot K(t-1) + 2t \cdot \widetilde{L}_f L_\theta\right) + 2L_\pi K(t) + 2L_\theta.
\end{aligned}$$

Proof. Firstly, we obtain from (A.12) that $\forall t \ge 1$,

$$\begin{aligned}
&\mathbb{E}_{\overline{x}_t}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] \\
&= \mathbb{E}\left[\left\|\frac{\partial x_t}{\partial x_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}\overline{u}_{t-1}}{\mathrm{d}\theta}\right\|_2\right]
\end{aligned}$$

According to the triangle inequality, we continue with

$$\begin{aligned}
&\le \mathbb{E}\left[\left\|\frac{\partial x_t}{\partial x_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}\overline{u}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\le 2L_f \cdot \left(\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2 + \left\|\frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}\right\|_2\right) + L_f \cdot \mathbb{E}_{\overline{x}_{t-1}}\left[\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad + L_f \cdot \mathbb{E}_{\overline{u}_{t-1}}\left[\left\|\frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_{t-1}}{\mathrm{d}\theta}\right\|_2\right].
\end{aligned} \tag{A.20}$$

[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 (A.11) that

$$\begin{aligned}
&\mathbb{E}_{\overline{u}_n}\left[\left\|\frac{\mathrm{d}u_t}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{\partial u_t}{\partial x_t}\cdot\frac{\mathrm{d}x_t}{\mathrm{d}\theta}+\frac{\partial u_t}{\partial\theta}-\frac{\partial\overline{u}_t}{\partial\overline{x}_t}\cdot\frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}-\frac{\partial\overline{u}_t}{\partial\theta}\right\|_2\right] \\
&\quad\le\mathbb{E}\left[\left\|\frac{\partial u_t}{\partial x_t}\cdot\frac{\mathrm{d}x_t}{\mathrm{d}\theta}-\frac{\partial\overline{u}_t}{\partial\overline{x}_t}\cdot\frac{\mathrm{d}x_t}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial\overline{u}_t}{\partial\overline{x}_t}\cdot\frac{\mathrm{d}x_t}{\mathrm{d}\theta}-\frac{\partial\overline{u}_t}{\partial\overline{x}_t}\cdot\frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right]+\mathbb{E}\left[\left\|\frac{\partial u_t}{\partial\theta}-\frac{\partial\overline{u}_t}{\partial\theta}\right\|_2\right] \\
&\quad\le 2L_\pi\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta}\right\|\right]+L_\pi\cdot\mathbb{E}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right]+2L_\theta. \qquad \text{(A.21)}
\end{aligned}$$

Plugging (A.21) back to (A.20),

$$\begin{aligned}
&\mathbb{E}_{\overline{x}_t}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad\lesssim 4L_f\widetilde{L}_\pi\cdot\left(\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2+\left\|\frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}\right\|_2\right)+L_f\widetilde{L}_\pi\cdot\mathbb{E}_{\overline{x}_{t-1}}\left[\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right]+2L_fL_\theta \\
&\quad\le 4L_f\widetilde{L}_\pi\cdot K(t-1)+L_f\widetilde{L}_\pi\cdot\mathbb{E}_{\overline{x}_{t-1}}\left[\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right]+2L_fL_\theta,
\end{aligned}$$

where the last inequality follows from the definition of K in (A.17).

Applying this recursion gives us

$$\begin{aligned}
\mathbb{E}_{\overline{x}_t}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right]&=e\big(L_f\widetilde{L}_\pi\big)^t+\big(4L_f\widetilde{L}_\pi\cdot K(t-1)+2\widetilde{L}_fL_\theta\big)\cdot\sum_{j=0}^{t-1}\big(\widetilde{L}_f\widetilde{L}_\pi\big)^j \\
&\le\widetilde{L}_f^t\widetilde{L}_\pi^t\Big(e+4t\cdot\widetilde{L}_f\widetilde{L}_\pi\cdot K(t-1)+2t\cdot\widetilde{L}_fL_\theta\Big),
\end{aligned}$$

where the first equality follows from (A.15).

As a consequence, we have from (A.21) that

$$\mathbb{E}_{\overline{u}_t}\left[\left\|\frac{\mathrm{d}u_t}{\mathrm{d}\theta}-\frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta}\right\|_2\right]\le\widetilde{L}_f^t\widetilde{L}_\pi^{t+1}\Big(e+4t\cdot\widetilde{L}_f\widetilde{L}_\pi\cdot K(t-1)+2t\cdot\widetilde{L}_fL_\theta\Big)+2L_\pi K(t)+2L_\theta.$$

This concludes the proof. □

A.4 Proof of Theorem 4.5

In the following proof, we use the notation $\|z\|_2$ to represent the Euclidean l_2 norm for vector z, and $\|Z\|_2$ to represent the induced 2-norm for matrix Z, i.e. $\|Z\|_2:=\max_{\|x\|_2=1}\|Zx\|_2$. Recall that $\|Z\|_F$ denotes the Frobenius norm of matrix Z, i.e. $\|Z\|_F=\sqrt{\mathrm{tr}(ZZ^\top)}$.

To characterize the Lipschitz of the LCS model, we need the partial derivatives of x_{t+1} with respect to x_t and u_t, which, however, further depend on the partial derivatives of λ_t with respect to x_t and u_t and cannot be expressed in closed form. Instead, they are implicitly defined by the LCP. Therefore, we introduce the following implicit function theorem.

Theorem A.2 (Implicit Function Theorem)**.** An implicit function $g:\mathbb{R}^{d_s}\times\mathbb{R}^{d_w}\to\mathbb{R}^{d_s}$ is defined as $g(s,w)=0$ for solution $s\in\mathbb{R}^{d_s}$ and problem data $w\in\mathbb{R}^{d_w}$. Then the Jacobian $\partial s/\partial w$, i.e. the sensitivity of the solution with respect to the problem data, is given by

$$\frac{\partial s}{\partial w}=-\Big(\frac{\partial g}{\partial s}\Big)^{-1}\frac{\partial g}{\partial w}.$$

Proof. Differentiating g with respect to the problem data w gives:

$$\frac{\mathrm{d}g}{\mathrm{d}w}=\frac{\partial g}{\partial w}+\frac{\partial g}{\partial s}\frac{\partial s}{\partial w}.$$

Since for any w, $g(s,w)=0$ always holds, the above total derivative is also always 0. This observation allows us to calculate the Jacobian

$$\frac{\partial s}{\partial w} = -\Big(\frac{\partial g}{\partial s}\Big)^{-1}\frac{\partial g}{\partial w}.$$

□

Proof of Theorem 4.5. To begin with, we first study the Jacobian $\partial x_{t+1}/\partial x_t$, and the Jacobian $\partial x_{t+1}/\partial u_t$ can be analyzed using similar techniques.

Denote $C^{(i)} \in \mathbb{R}^{d_x}$ as the i-th column of the matrix $C \in \mathbb{R}^{d_x \times d_\lambda}$. Similarly, denote $D^{(i)} \in \mathbb{R}^{d_x}$, $E^{(i)} \in \mathbb{R}^{d_u}$, $F^{(i)} \in \mathbb{R}^{d_\lambda}$ as the i-the rows of matrices D, E, F, respectively. Then we have the Jacobian with the form

$$\frac{\partial x_{t+1}}{\partial x_t} = A + \sum_{i=1}^{d_\lambda} C^{(i)} \frac{\partial \lambda^{(i)}}{\partial x_t}. \tag{A.22}$$

We rewrite the contact equation $\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu \vec{1}$ in (3.1) as

$$\lambda_t^{(i)}(D^{(i)\top}x_t + E^{(i)\top}u_t + F^{(i)\top}\lambda_t + d^{(i)}) = \mu, \ \ \forall i \in [1, d_\lambda]. \tag{A.23}$$

By the Implicit Function Theorem A.2, we have

$$\begin{aligned}\frac{\partial \lambda^{(i)}}{\partial x_t} &= -\Big(D^{(i)\top}x_t + E^{(i)\top}u_t + \frac{\partial}{\partial \lambda_t^{(i)}}\lambda_t^{(i)}F^{(i)\top}\lambda_t + d^{(i)}\Big)^{-1}\lambda_t^{(i)}D^{(i)\top} \\ &= -(D^{(i)\top}x_t + E^{(i)\top}u_t + F^{(i)\top}\lambda_t + \lambda_t^{(i)}F^{(i)(i)} + d^{(i)})^{-1}\lambda_t^{(i)}D^{(i)\top}, \ \ \forall i \in [1, d_\lambda],\end{aligned} \tag{A.24}$$

where $F^{(i)(i)} \in \mathbb{R}$ is the i-th element of $F^{(i)}$.

Since F is a P-matrix, we know that all its first order principal sub-matrices are positive, i.e., $F^{(i)(i)} > 0$.

Plugging (A.24) into (A.22) and take the induced 2-norm, we obtain

$$\begin{aligned}\left\|\frac{\partial x_{t+1}}{\partial x_t}\right\|_2 &= \left\|A - \sum_{i=1}^{d_\lambda} C^{(i)}\big(D^{(i)\top}x_t + E^{(i)\top}u_t + F^{(i)\top}\lambda_t + \lambda_t^{(i)}F^{(i)(i)} + d^{(i)}\big)^{-1}\lambda_t^{(i)}D^{(i)\top}\right\|_2 \\ &\le \|A\|_2 + \sum_{i=1}^{d_\lambda} \lambda_t^{(i)}\|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2 \cdot \big|D^{(i)\top}x_t + E^{(i)\top}u_t + F^{(i)\top}\lambda_t + \lambda_t^{(i)}F^{(i)(i)} + d^{(i)}\big|^{-1} \\ &\le \|A\|_2 + \sum_{i=1}^{d_\lambda} \|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2 \cdot (\lambda_t^{(i)})^2/\mu,\end{aligned} \tag{A.25}$$

where the first inequality holds due to the Cauchy–Schwarz inequality, the second inequality holds since $F^{(i)(i)} > 0$ and $D^{(i)\top}x_t + E^{(i)\top}u_t + F^{(i)\top}\lambda_t + d^{(i)} \ge 0$.

By the definition of Frobenius norm, we know that

$$\begin{aligned}\|C\|_F &= \sqrt{\sum_{i=1}^{d_\lambda}\|C^{(i)}\|_2^2} = \sqrt{d_\lambda}\cdot\sqrt{\sum_{i=1}^{d_\lambda}\frac{1}{d_\lambda}\|C^{(i)}\|_2^2} \\ &\ge \sqrt{d_\lambda}\cdot\sum_{i=1}^{d_\lambda}\frac{1}{d_\lambda}\sqrt{\|C^{(i)}\|_2^2} = \frac{1}{\sqrt{d_\lambda}}\sum_{i=1}^{d_\lambda}\|C^{(i)}\|_2,\end{aligned} \tag{A.26}$$

where we adopt the Jensen's inequality in the second line.

Besides, define the diagonal matrix $\Lambda_t := \mathrm{diag}(\lambda_t^{(1)}, \cdots, \lambda_t^{(d_\lambda)}) \in \mathbb{R}^{d_\lambda \times d_\lambda}$. By definition, $\|\Lambda_t\|_2 = \max_i \lambda^{(i)}$ and thus

$$\|\lambda_t\|_2^2 = \sum_{i=1}^{d_\lambda}(\lambda_t^{(i)})^2 \le d_\lambda \cdot \|\Lambda_t\|_F^2. \tag{A.27}$$

Therefore, we can further bound (A.25) by

$$\begin{aligned}\left\|\frac{\partial x_{t+1}}{\partial x_t}\right\|_2 &\leq \|A\|_2 + \frac{1}{\mu}\left(\sum_{i=1}^{d_\lambda}\|C^{(i)}\|_2\right)\cdot\left(\sum_{i=1}^{d_\lambda}\|D^{(i)}\|_2\right)\cdot\left(\sum_{i=1}^{d_\lambda}(\lambda_t^{(i)})^2\right)\\ &\leq \|A\|_2 + \frac{d_\lambda}{\mu}\|C\|_F\|D\|_F\|\lambda_t\|_2^2\\ &\leq \|A\|_F + \frac{d_\lambda^2}{\mu}\|C\|_F\|D\|_F\|\Lambda_t\|_F^2, \end{aligned} \tag{A.28}$$

where the first inequality holds since $\sum_i y_i \cdot z_i \leq (\sum_i y_i)\cdot(\sum_i z_i)$ for any non-negative scalar sequences y_i, z_i and the second inequality follows from (A.26). The third inequality follows from (A.27) and the fact that $\|A\|_2 \leq \|A\|_F$.

The final step is to characterize the magnitude of $\|\Lambda_t\|_F^2$. This can be done by rewriting the contact equation $\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu\vec{1}$ in (3.1) as

$$\Lambda_t(Dx_t + Eu_t + F\Lambda_t\vec{1} + d) = \mu\vec{1}$$

By the Cauchy-Schwartz inequality we have

$$\|\Lambda_t\|_F \cdot \left(\|Dx_t + Eu_t + d\|_2 + \|F\|_F\|\Lambda_t\|_F\right) \geq \mu.$$

Denote $e := \sup\|Dx_t + Eu_t + d\|_2$. The above inequality can be simplified as

$$\|F\|_F \cdot \|\Lambda_t\|_F^2 + e\cdot\|\Lambda_t\|_F - \mu \geq 0. \tag{A.29}$$

Solving (A.29) gives

$$\|\Lambda_t\|_F \geq \frac{\sqrt{e^2 + 4\mu\|F\|_F} - e}{2\|F\|_F}$$

Since $\varepsilon = e^2/(2\|F\|_F^2)$, we further have

$$\begin{aligned} l(\mu) :=& \frac{\|\Lambda_t\|_F^2}{\mu} \geq \frac{2e^2 + 4\mu\|F\|_F - 2e\sqrt{e^2 + 4\mu\|F\|_F}}{4\mu\|F\|_F^2}\\ =& \frac{e^2}{2\mu\|F\|_F^2} + \frac{1}{\|F\|_F} + \frac{e^2\sqrt{\frac{1}{\mu^2} + \frac{4\|F\|_F}{\mu e^2}}}{2\|F\|_F^2}\\ =& \frac{\varepsilon}{\mu} + \frac{1}{\|F\|_F} + \varepsilon\sqrt{\frac{1}{\mu^2} + \frac{2}{\varepsilon\mu\|F\|_F}}. \end{aligned} \tag{A.30}$$

Plug (A.30) into (A.28), we get the Jacobian norm

$$\left\|\frac{\partial x_{t+1}}{\partial x_t}\right\|_2 \leq \|A\|_F + d_\lambda^2\|C\|_F\|D\|_F\cdot l(\mu).$$

Using the same proof steps, the norm of Jacobian $\partial x_{t+1}/\partial u_t$ satisfies

$$\left\|\frac{\partial x_{t+1}}{\partial u_t}\right\|_2 \leq \|B\|_F + d_\lambda^2\|C\|_F\|E\|_F\cdot l(\mu).$$

We conclude the proof by noticing the relationship between the norm of Jacobian and the Lipschitz of the LCS model. □

A.5 Proof of Proposition 5.1

Proof. We first consider the original unsmoothed problem $\lambda_t(Dx_t + Eu_t + F\lambda_t + d) = 0$. Since $\lambda_t \geq 0$, we know that the solution λ_t is a piece-wise linear function with the form:

$$\lambda_t = \begin{cases} -(Dx_t + Eu_t + d)/F & \text{if } Dx_t + Eu_t + d \leq 0\\ 0 & \text{else} \end{cases}.$$

By rewriting the above formula as a function of $z_t := Dx_t + Eu_t + d$, we can express the solver $S_{\mu=0}$ of the exact LCP as follows:

$$S_{\mu=0}(z_t) = \begin{cases} -z_t/F & \text{if } z_t \leq 0 \\ 0 & \text{else} \end{cases}. \tag{A.31}$$

Now our goal is to find the noise distribution $\rho(w)$ such that the following holds:

$$S_{\mu(z_t)}(z_t) = \mathbb{E}_{w\sim\rho(w)}[S_{\mu=0}(z_t + w)] = \int S_{\mu=0}(z_t + w)\rho(w)\mathrm{d}w. \tag{A.32}$$

Define H(x) as a Heaviside-like step function:

$$H(x) := \begin{cases} -1/F & \text{if } x \leq 0 \\ 0 & \text{else} \end{cases}.$$

We observe that the derivative of $S_{\mu=0}(z_t)$ is in fact $H(z_t)$. This allows us to write

$$\begin{aligned} \nabla_{z_t} S_{\mu(z_t)}(z_t) &= \nabla_{z_t} \int S_{\mu=0}(z_t + w)\rho(w)\mathrm{d}w \\ &= \int \nabla_{z_t} S_{\mu=0}(z_t + w)\rho(w)\mathrm{d}w \\ &= \int H(z_t + w)\rho(w)\mathrm{d}w. \end{aligned}$$

Since the derivative of the Heaviside step function is the dirac delta function $\delta(\cdot)$, we have

$$\begin{aligned} \nabla^2_{z_t} S_{\mu(z_t)}(z_t) &= \nabla_{z_t} \int H(z_t + w)\rho(w)\mathrm{d}w \\ &= \int \delta(z_t + w)\rho(w)\mathrm{d}w = \rho(z_t). \end{aligned}$$

This concludes the proof. □

A.6 Proof of Proposition 5.2

Recall that Proposition 5.1 connects the proposed analytical barrier smoothing with the randomized smoothing. Therefore, we first provide the following lemma established in randomized smoothing as a preparation before proving Proposition 5.2.

Lemma A.3 (Randomized Smoothing as Linearization Minimizer (Pang et al., 2022))**.** Let $\rho(w) = \mathcal{N}(0, \Sigma)$ be a zero-mean, Σ-covariance Gaussian. Consider the problem of regressing a function g with parameters (K, W) such that the residual around $\overline{x}$ distributed according to ρ is minimized:

$$\mathcal{L}(K, W) = \min_{K,W} \frac{1}{2}\mathbb{E}_{w\sim\rho(w)}\left[\left\|g(\overline{x} + w) - Ww - K\right\|_2^2\right]. \tag{A.33}$$

The solution is the linearization of the smoothed surrogate:

$$\begin{aligned} K^* &= \mathbb{E}_{w\sim\rho(w)}[g(\overline{x} + w)], \\ W^* &= \frac{\partial}{\partial x}\mathbb{E}_{w\sim\rho(w)}[g(x + w)]|_{x=\overline{x}}. \end{aligned}$$

Proof. The proof is originally provided in (Pang et al., 2022), which is adapted here for completeness.

Since (A.33) is a linear regression problem and is convex, the first-order stationarity condition implies optimality. By calculating the gradients and setting them to zero, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial K} &= \mathbb{E}_{w\sim\rho(w)}[g(\overline{x} + w)] - K^* = 0 \\ \frac{\partial \mathcal{L}}{\partial W} &= \mathbb{E}_{w\sim\rho(w)}[ww^\top]W^* - \mathbb{E}_{w\sim\rho(w)}[g(\overline{x} + w)w^\top] = 0. \end{aligned}$$

Therefore, we obtain the solution

$$\begin{aligned} K^* &= \mathbb{E}_{w\sim\rho(w)}[g(\overline{x}+w)], \\ W^* &= \mathbb{E}_{w\sim\rho(w)}[ww^\top]^{-1}\mathbb{E}_{w\sim\rho(w)}[g(\overline{x}+w)w^\top] \\ &= \frac{\partial}{\partial x}\mathbb{E}_{w\sim\rho(w)}[g(x+w)]|_{x=\overline{x}}, \end{aligned} \tag{A.34}$$

where the last step follows from the likelihood ratio gradient with the form (2.1), as well as the fact that the score function of the Gaussian is $\Sigma^{-1}w$. □

Proof of Proposition 5.2. By applying Lemma A.3, we know that Proposition 5.2 holds once the following equivalence is established:

$$S_{\mu(z_t)}(z_t) = \mathbb{E}_{w\sim\rho(w)}[S_{\mu=0}(z_t+w)], \tag{A.35}$$

where $\rho(w)$ is any zero-mean Gaussian distribution.

This is a direct result from Proposition 5.1. Specifically, when $\mu(z_t) = \kappa\cdot(z_t+F\kappa)$, the corresponding softened LCP is

$$\lambda_t(z_t+F\lambda_t) = \mu(z_t) = \kappa\cdot(z_t+F\kappa).$$

The solution of the above equation is given by

$$S_{\mu(z_t)}(z_t) = \lambda_t = \kappa = z_t\cdot\mathrm{erf}(z_t,\sigma) + e^{-z_t^2/(2\sigma)}/\sqrt{\pi} + c_1 z_t + c_2. \tag{A.36}$$

Proposition 5.1 states that when $\rho(w) = \nabla_w^2 S_{\mu(z_t)}(w)$, then $S_{\mu(z_t)}(z_t) = \mathbb{E}_{w\sim\rho(w)}[S_{\mu=0}(z_t+w)]$. For $S_{\mu(z_t)}(z_t)$ satisfying (A.36), its second-order derivative is the Gaussian $\mathcal{N}(0,\sigma)$, due to the definition of the error function. Therefore, $S_{\mu(z_t)}(z_t) = \mathbb{E}_{w\sim\mathcal{N}(w;0,\sigma)}[S_{\mu=0}(z_t+w)]$, which concludes the proof of (A.35) and the proposition. □

A.7 Proof of Theorem 5.3

Proof. According to Taylor's theorem, we know that

$$\left|\frac{S_{\mu=0}(z_r+w)-S_{\mu=0}(z_t)}{w} - \nabla_z S_{\mu=0}(z_t)\right| \le |w|\cdot\sup\frac{|\nabla_z^2 S_{\mu=0}(z_t)|}{2} = \frac{F^2|w|}{2}, \tag{A.37}$$

where the second inequality follows from (A.31).

We define the linearization residual at point z_t+w as

$$\nu(w) := \left|S_{\mu=0}(z_r+w) - \nabla_z S_{\mu(z_t)}(z_t)\cdot w - S_{\mu(z_t)}(z_t)\right|.$$

Then we have from (A.37) that

$$\left|\frac{\nu(w)+S_{\mu(z_t)}(z_t)-S_{\mu=0}(z_t)}{w} + \nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)\right| \le \frac{F^2|w|}{2}.$$

Since $|S_{\mu(z_t)}(z_t) - S_{\mu=0}(z_t)| \le 1/\sqrt{\pi}+c_2 := \varsigma$, achieved at $z=0$, we obtain from the triangle inequality that the bias of gradient satisfies

$$\left|\nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)\right| \le \frac{F^2|w|}{2} + \frac{\nu(w)+\varsigma}{|w|}. \tag{A.38}$$

From Proposition 5.2, we know that

$$\mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}[\nu(w)] = \delta. \tag{A.39}$$

We claim that there exists $\sigma\mathcal{Q}(2/3) \le w \le \sigma\mathcal{Q}(3/4)$ such that $\nu(w) \le 12\delta$.

This can be proved by contradiction: Suppose $\forall w \in [\sigma\mathcal{Q}(2/3), \sigma\mathcal{Q}(3/4)]$, $\nu(w) > 12\delta$. Then the expectation $\mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}[\nu(w)] > (3/4-2/3)\cdot 12\delta = \delta$. This contradicts with (A.39). Therefore, the claim is correct.

Using the above claim, we have from (A.38) that

$$\left|\nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)\right| \leq \frac{F^2 \sigma \mathcal{Q}(3/4)}{2} + \frac{12\delta + \varsigma}{\sigma \mathcal{Q}(2/3)}.$$

We conclude the proof by applying the chain rule in the LCS model (3.1):

$$\left\|\nabla_x f_{\mu=0} - \nabla_x f_{\mu(z_t)}\right\|_2 \leq \|C\|_F \|D\|_F \cdot \left(\frac{\sigma F^2 \mathcal{Q}(3/4)}{2} + \frac{12\delta + \varsigma}{\sigma \mathcal{Q}(2/3)}\right),$$
$$\left\|\nabla_u f_{\mu=0} - \nabla_u f_{\mu(z_t)}\right\|_2 \leq \|C\|_F \|E\|_F \cdot \left(\frac{\sigma F^2 \mathcal{Q}(3/4)}{2} + \frac{12\delta + \varsigma}{\sigma \mathcal{Q}(2/3)}\right).$$

□

B Interior-Point Solver

In this section, we describe the IPM solver that is used to solve the Nonlinear Complementarity Problem (NCP) in (2.4) and (2.5) (or the LCP in (3.1) that corresponds to $mu = 0$).

We adopt the primal-dual interior-point solver with Mehrotra correction (Mehrotra, 1992). Each iteration of the primal-dual interior-point solver consists of a predictor step that computes the affine search direction for zero complementarity violation, and a centering (with Mehrotra correction) step that computes a target relaxation for the search direction. For notation simplicity, we consider the problem of the following form:

$$\text{find } a, b, c$$
$$\text{subject to } E(a,b,c) = 0,\ b \circ c = \mu \vec{1},\ b \geq \vec{0},\ c \geq \vec{0},$$

where $a, b \in \mathbb{R}^{n\times 1}, c \in \mathbb{R}^{n\times 1}$ are the decision variables and E is the set of equality constraints. We denote $\omega := (a, b, c)$.

The solver aims to find a fixed point for the following residual:

$$\mathcal{R}(\omega; \mu) := [E(a,b,c), bc - \mu \vec{1}]^\top.$$

We denote the Jacobian of this residual with respect to the decision variables as

$$\mathcal{R}_J(\omega; \mu) := \partial \mathcal{R}(\omega; \mu)/\partial \omega,$$

where Δb_{aff} and Δc_{aff} are the corresponding elements in the affine scaling direction $\Delta_{\text{aff}} := -\mathcal{R}_J^{-1}(\omega; mu)\mathcal{R}(\omega; \mu)$.

With Mehrotra correction, we define

$$\overline{\mathcal{R}}(\omega; \mu) := [E(a,b,c), bc - \mu \vec{1} + \Delta b_{\text{aff}} \Delta c_{\text{aff}}]^\top.$$

Then the search direction Δ is given by Newton's method as

$$\Delta := -\mathcal{R}_J^{-1}(\omega; \mu)\overline{\mathcal{R}}(\omega; \mu). \tag{B.1}$$

The IPM solver adaptively relaxes the above problem by first computing the duality measure ϱ, the affine duality measure ϱ_{aff}, and the centering parameter σ:

$$\varrho := \frac{1}{n} b^\top c = \frac{1}{n} \sum_{i=1}^{n} b^{(i)} c^{(i)}, \tag{B.2}$$
$$\varrho_{\text{aff}} := \frac{1}{n} (b + \alpha_{\text{aff}}^{\text{pri}} \Delta b_{\text{aff}})^\top (c + \alpha_{\text{aff}}^{\text{dual}} \Delta c_{\text{aff}}), \tag{B.3}$$
$$\sigma := (\varrho_{\text{aff}}/\varrho)^3, \tag{B.4}$$

where $\alpha_{\text{aff}}^{\text{pri}}$ and $\alpha_{\text{aff}}^{\text{dual}}$ are the maximum step-sizes to the boundary, defined as

$$\alpha_{\text{aff}}^{\text{pri}} := \min\left(1, \min_{i:\Delta b_{\text{aff}}^{(i)} < 0} -\frac{b^{(i)}}{\Delta b_{\text{aff}}^{(i)}}\right),\quad \alpha_{\text{aff}}^{\text{dual}} := \min\left(1, \min_{i:\Delta c_{\text{aff}}^{(i)} < 0} -\frac{c^{(i)}}{\Delta c_{\text{aff}}^{(i)}}\right).$$

Algorithm 2 Primal-Dual Interior-Point Solver with Stopping Criteria $\texttt{SOLVER}(\mu_{\text{sc}})$

Input: Stopping criteria μ_{sc}
Initialize $a = a_0, b = b_0, c = c_0, \omega = (a, b, c)$
Update the complementarity violation $\mu_{\text{vio}} \leftarrow \max_i\{\|b^{(i)}c^{(i)}\|_\infty\}$
while $\mu_{\text{vio}} \leq \mu_{\text{sc}}$ **do**
  Calculate the duality measure ϱ, affine duality measure ϱ_{aff}, and the centering parameter σ by (B.2), (B.3), and (B.4)
  Update $\mu \leftarrow \sigma\varrho$
  Calculate the search direction Δ by (B.1), $\Delta = -\mathcal{R}_J^{-1}(\omega;\mu)\overline{\mathcal{R}}(\omega;\mu)$
  Update $\omega \leftarrow \omega + \alpha\Delta$
  Update the complementarity violation $\mu_{\text{vio}} \leftarrow \max_i\{\|b^{(i)}c^{(i)}\|_\infty\}$
end while
Output: ω

For a μ_{sc}-softened complementarity system, the predictor steps and the centering steps are performed iteratively until the complementarity violation is smaller than the stopping criteria (or tolerance threshold) μ_{sc}. Specifically, the pseudocode of the solver is provided in Algorithm 2.

In complementarity-model-based first-order policy gradient methods, the output $\omega = \texttt{SOLVER}(\mu_{\text{sc}})$ is used to replace the exact first-order policy gradient $\partial\texttt{SOLVER}(0)/\partial\theta$ by the gradient of the analytically smoothed solution $\partial\omega/\partial\theta$. In other words, for vanilla FOPG and analytically smoothed FOPG, the first-order gradient calculation in (4.2) is obtained by differentiating through the transitions governed by $\texttt{SOLVER}(0)$ and $\texttt{SOLVER}(\mu_{\text{sc}})$, respectively.

C Details of Experiments

C.1 Dynamics in the Ball Bouncing Example

In Section 7.1, we plot the dynamics and derivatives of the contact behavior. Here, we describe how they are generated using ordinary differential equations.

Without loss of generality, we assume that the discretization timestep size, the mass of the ball, and its initial velocity v_0 are all 1. Denote the initial vertical coordinate of the ball is q_0. Then the distance of the ball to the ground is given by $q_0 - \int\int(g - \gamma_t)\mathrm{d}t\mathrm{d}t$, where recall that γ_t is the normal impact contact force at timestep t.

Then we are able to obtain the μ-softened complementarity problem as

$$\gamma_t\Big(q_0 - \int\int(g - \gamma_t)\mathrm{d}t\,\mathrm{d}t\Big) = \mu.$$

This can be rewritten as

$$-(g - \gamma_t) = \frac{\partial^2}{\partial t^2}\frac{\mu}{\gamma_t} = \Big(-\mu\frac{\partial^2\gamma_t}{\partial t^2}\gamma_t + 2\mu\Big(\frac{\partial\gamma_t}{\partial t}\Big)^2\Big)/\gamma_t^3.$$

We simplify this second-order ODE by defining another variable e_t, such that

$$\begin{aligned}\frac{\partial\gamma_t}{\partial t} &= e_t,\\ \frac{\partial e_t}{\partial t} &= \frac{\gamma_t^4 - \gamma_t^3 g - 2\mu e_t^2}{-\mu\gamma_t}.\end{aligned} \tag{C.1}$$

Using Python to solve (C.1) gives us γ_t. Then the y-axis velocity is naturally obtained by $v_y = -\int(g - \gamma_t)\mathrm{d}t$. Since $q_x = v_0 t = t$, we get the relationship between v_y and q_x. The derivatives can be calculated using finite differences.