Adaptive Barrier Smoothing for First-Order Policy Gradient with Contact Dynamics

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

Abstract

Differentiable physics-based simulators have witnessed remarkable success in robot learning involving contact dynamics, benefiting from their improved accuracy and efficiency in solving the underlying complementarity problem. However, when utilizing the First-Order Policy Gradient (FOPG) method, our theory indicates that the complementarity-based systems suffer from stiffness, leading to an explosion in the gradient variance of FOPG. As a result, optimization becomes challenging due to chaotic and non-smooth loss landscapes. To tackle this issue, we propose a novel approach called *Adaptive Barrier Smoothing* (ABS), which introduces a class of softened complementarity systems that correspond to barrier-smoothed objectives. With a contact-aware adaptive central-path parameter, ABS reduces the FOPG gradient variance while controlling the gradient bias. We justify the adaptive design by analyzing the roots of the system's stiffness. Additionally, we establish the convergence of FOPG and show that ABS achieves a reasonable trade-off between the gradient variance and bias by providing their upper bounds. Moreover, we present a variant of FOPG based on complementarity modeling that efficiently fits the contact dynamics by learning the physical parameters. Experimental results on various robotic tasks are provided to support our theory and method.

1. Introduction

The recent advancements in learning to control and plan for robotic systems can be largely attributed to the exceptional accuracy of physics-based simulators (Todorov et al., 2012; Degrave et al., 2019; de Avila Belbute-Peres et al., 2018; Howell et al., 2022). When it comes to optimizing control policies in differentiable simulators, the most straightforward approach is First-Order Policy Gradient (FOPG), which involves backpropagating through the computational path of cumulative rewards (Xu et al., 2022; Freeman et al., 2021) and has demonstrated great potential even for general non-differentiable tasks in model-based settings where the dynamics are unknown (Clavera et al., 2020; Li et al., 2021; Amos et al., 2021).

The fundamental behaviors underlying common robotic tasks, such as locomotion and manipulation, rely heavily on the intricate interactions between the robot and its environment (Aydinoglu et al., 2020; Lidec et al., 2023). To model the robotic contact dynamics, complementarity systems have emerged as the de-facto approaches. These systems are rooted in the complementarity problem, which forms the basis for simulating hard-contact scenarios (Geilinger et al., 2020; Howell et al., 2022; Werling et al., 2021). By employing the *Interior-Point Method* (IPM) (Mehrotra, 1992), the complementarity problem efficiently solves for impact and frictional contact forces, ensuring non-penetration and maximum dissipation.

Unfortunately, even though complementarity-based contact systems can provide accurate physical behaviors, they often exhibit stiff dynamics with extreme curvatures (Parmar et al., 2021; Anitescu & Potra, 2002) due to geometrical constraints and contact events. In this work, we demonstrate that such stiffness can lead to optimization challenges when performing FOPG. Specifically, we first establish the convergence of FOPG that depends on the gradient variance and bias. Then we 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 dealing with a stiff model, the presence of long chains of nonlinear mappings results in slow convergence due to the large gradient variance and chaotic (Bollt, 2000) optimization procedures, a phenomenon also observed in previous experimental studies (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

[1]Northwestern University, Evanston, IL, USA [2]Arizona State University, Tempe, AZ, USA.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

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 complementarity system by setting a universal stopping criteria in the IPM solver. As the softened LCS can be shown to be the optimality condition of a barrier-smoothed objective, we refer to this vanilla approach as *Barrier Smoothing* (BS). However, indiscriminately applying barrier smoothing can lead to unrealistic simulation and a significant gradient bias.

To balance the variance and bias, we propose *Adaptive Barrier Smoothing* (ABS) by utilizing a *contact-aware* central-path parameter that decreases with the minimum distance-to-obstacle of the inactive impact contact points. This design is justified by our result, which shows that the stiffness and variance are governed by inactive near-obstacle contacts. By drawing on the equivalence between BS and randomized smoothing (Suh et al., 2022b;a) in single-contact settings, we show that ABS minimizes the linearization residual and provide the upper bound of its gradient bias. We also present experimental results to support our theory and method.

2. Background

2.1. Reinforcement Learning

Consider learning to optimize a finite H-horizon Markov Decision Process (MDP). Denote the state space and action space as $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ and $\mathcal{U} \subseteq \mathbb{R}^{n_u}$, 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

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

where $V_0^{\pi_\theta}$ is the state value at the initial timestep, and $p_{\pi_\theta}(\alpha)$ is the distribution over rollouts $\alpha = ((x_0, u_0), \cdots, (x_{H-1}, u_{H-1}))$ when executing π_θ, formally, $x_0 \sim \zeta(\cdot)$, $u_i \sim \pi_\theta(\cdot \,|\, s_i)$, 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,

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

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 of 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 the applied control force (action), $\mathcal{M}(q)$ 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

$$\begin{aligned} v_{t+1} &= v_t + \mathcal{M}_t^{-1}(n_t + u_t)h + \mathcal{M}_t^{-1} J_t^\top \lambda_t, \\ q_{t+1} &= q_t + h v_{t+1} \end{aligned} \quad (2.2)$$

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

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

$$\gamma_t \circ \phi(x_t, u_t) = \vec{0}, \qquad \gamma_t, \phi(x_t, u_t) \geq \vec{0}, \tag{2.3}$$

where $\circ$ is the element-wise (Hadamard) product, the Signed Distance Function (SDF) $\phi(x_t, u_t) : \mathbb{R}^{n_x \times n_u} \to \mathbb{R}^c$ measures the distance from each contact point to the obstacles, $\vec{0}$ is the zero vector, and the equality, inequality are element-wise. Eq. (2.3) states that the impact forces must be non-negative and can only be non-zero to maintain non-negative gaps (non-penetration) when contact is active.

The friction can be modeled using the maximum-dissipation principle (Howell et al., 2022) to obtain a very similar complementarity problem as in (2.3). For analysis simplicity, the systems we consider in this work are frictionless.

3. Complementarity-Based Contact Models

3.1. Softened Linear Complementarity Systems

The dynamic (2.2) describes a hybrid system where different modes are controlled by the contact force λ under the nonlinear complementarity problem (2.3). 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 and its Solution)**.** 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 gives the solution $\lambda_t = S_\mu(Dx_t + Eu_t + d) \in \mathbb{R}^{d_\lambda}$.

Here, $\mu = 0$ corresponds to the exact *Linear Complementarity Problem* (LCP) and $f_{\mu=0}$ resembles the reality when simulating. 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)**.** We assume that in (3.1), F is a P-matrix, defined as a matrix whose principal minors are all positive, i.e., the determinants of its principal sub-matrices $\det(F_{\alpha\alpha}) > 0, \forall \alpha \subseteq \{1, \cdots, d_\lambda\}$.

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 in the following lemma that the softened LCS is the optimality condition of a barrier-smoothed objective. 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 a sequence of path-centering μ, 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. First-Order Policy Gradient

In this section, we first present an overview of the First-Order Policy Gradient (FOPG) framework. Then we delve into the convergence properties of 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 the non-smooth behaviors of contact events can impede optimization, motivating us to develop smoothing techniques.

4.1. Framework

The First-Order Policy Gradient (FOPG) method updates the policy π_{θ_n} by (4.2) in each iteration $n \in [N]$ and returns π_{θ_N} at the end of training in a total of N iterations.

Specifically, consider optimizing a stochastic policy $u \sim \pi_\theta(\cdot|x)$ in continuous action spaces, or equivalently $u = \pi_\theta(x, \varsigma)$ with noise $\varsigma \sim p(\varsigma)$. With batch size M, the policy gradient at iteration n is given by the chain rule as

$$\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) = \frac{1}{M}\sum_{m=1}^{M}\left(\sum_{t=0}^{H-1}\frac{\partial r_{t,m}}{\partial u_{t,m}}\frac{\partial u_{t,m}}{\partial \theta} + \frac{\partial r_{t,m}}{\partial x_{t,m}}\frac{\mathrm{d}x_{t,m}}{\mathrm{d}\theta}\right),$$
$$\text{where } \frac{\mathrm{d}x_{t+1,m}}{\mathrm{d}\theta} = \frac{\partial x_{t+1,m}}{\partial x_{t,m}}\frac{\mathrm{d}x_{t,m}}{\mathrm{d}\theta} + \frac{\partial x_{t+1,m}}{\partial u_{t,m}}\frac{\partial u_{t,m}}{\partial \theta}. \tag{4.1}$$

Here, $x_{0,m} \sim \zeta$, $u_{t,m} = \pi_{\theta_n}(x_{t,m}, \varsigma_m)$, $\varsigma_m \sim p(\varsigma)$, $x_{t+1,m} = f(x_{t,m}, u_{t,m})$, and $r_{t,m} = r(x_{t,m}, u_{t,m})$.

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

$$\theta_{n+1} \leftarrow \theta_n + \eta \cdot \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}). \tag{4.2}$$

4.2. Convergence of FOPG

To begin, we assume that the objective is smooth, which is required by most studies on policy gradient methods (Agarwal et al., 2021; Pirotta et al., 2015; Wang et al., 2019).

Assumption 4.1 (Lipschitz Continuous Policy Gradient)**.** We 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 FOPG by first providing the following theorem.

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

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

Let $\delta = \sup \|\theta\|_2$ and $c = (\eta - L\eta^2)^{-1}$. Under Assumption 4.1, it 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 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)**.** We assume that the policy, model, and reward functions are L_π, L_f, and L_r Lipschitz continuous, respectively (see App. A.3 for details).

4.3. Gradient Variance and LCS Stiffness

Let $\widetilde{L}_g = \max\{L_g, 1\}$, where L_g is the Lipschitz constant of function g. We have the following result for the variance.

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

$$v_n = O\left(H^4 \widetilde{L}_f^{4H} \widetilde{L}_\pi^{4H}/M\right). \tag{4.3}$$

We observe that the variance upper bound 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 the FOPG method 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 and for $\mu > 0$, 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 characterized by μ and is lower bounded by

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

Theorem 4.5 highlights the crucial role of the central-path parameter μ in determining the model stiffness, as $l(\mu)$ is at least inversely proportional to μ. To obtain a reasonable upper bound of L_{f_μ}, and thus of the gradient variance v_n in (4.3), it is necessary to prevent μ from reaching zero because $\mu \to 0$ implies $l(\mu) \to \infty$.

Differentiating through the stiff models, such as the exact LCS $f_{\mu=0}$ or the softened LCS $f_{\mu>0}$ when μ is small, can lead to the explosion of gradient variance and optimization obstacles, e.g., chaotic optimization procedures and highly non-smooth loss landscapes, which remain present even when contact events are occasional in a full model unroll. This motivates us to develop smoothing techniques for FOPG algorithms in order to achieve a quick convergence.

5. Adaptive Barrier Smoothing

5.1. Method

According to Theorem 4.4 and 4.5, a natural idea to alleviate the exploding FOPG variance issue is to leverage the gradients from the μ-softened complementarity-based system with a large μ. This can be accomplished by implementing a stopping criterion for the IPM iterations that terminates when the decreasing sequence of μ' reaches μ. This results in smoothed gradients $\partial x_{t+1}/\partial x_t$ and $\partial x_{t+1}/\partial u_t$ for calculating $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})$ in (4.1). As Lemma 3.3 indicates, μ-softened systems are equivalent to smoothed objectives with log-barrier functions. Therefore, we refer to this vanilla approach as *Barrier Smoothing* (BS).

However, universally applying BS with a fixed μ can lead to significant *bias* since the trajectories will *not* adhere to physics laws. Fortunately, our findings suggest that it is viable to *adaptively* apply BS. Specifically, we prove that the primary cause of stiffness and high variance is the presence of contact points that are in close proximity to obstacles.

Theorem 5.1 (Stiffness from Contact)**.** For contact point i, let $z_t^{(i)} = D^{(i)\top} x_t + E^{(i)\top} u_t + d^{(i)}$. When $z_t^{(i)} > 0$, the stiffness of the μ-softened LCS is governed by contact points whose $z_t^{(i)}$ is small, with the norm of Jacobian satisfying

$$\left\| \frac{\partial x_{t+1}}{\partial x_t} \right\|_2 \leq \|A\|_F + \|C\|_F \|D\|_F d_\lambda \mu \sum_{i=1}^{d_\lambda} 1/(z_t^{(i)})^2,$$

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

In Thm. 5.1, $z_t^{(i)}$ is identified as the distance $\phi(x_t, u_t)^{(i)}$ in LCS, as evidenced by comparing (2.3) and (3.1). A positive distance indicates an *inactive* contact since the contact force will be zero in the exact LCS. Thus, the theorem provides an answer to the question of when it is necessary to apply BS — if the distances to obstacles are small for some inactive contact points, as they are the roots of the system's stiffness.

Based on this result, we propose utilizing an *adaptive* central-path parameter $\mu(x_t, u_t)$. Specifically, instead of a fixed μ, we set $\mu(x_t, u_t) = g(d(x_t, u_t))$, where $g : \mathbb{R} \to \mathbb{R}$ is a non-increasing function. Thus, $\mu(x_t, u_t)$ is *contact-aware* as it scales *inversely* with $d(x_t, u_t)$, the minimum distance-to-obstacle of the inactive contact, defined as $d(x_t, u_t) = \min_{i \in \mathcal{I}} |\phi(x_t, u_t)^{(i)}|$, where $\mathcal{I} = \{1 \leq i \leq c \,|\, \gamma_t^{(i)} = 0\}$ represents the set of inactive contact points.

Intuitively, the stiffness or extreme curvature of contact dynamics is a result of the sudden change in impact force when penetration first arises. For example, in 1D systems depicted in Figure 1, the velocity is continuous everywhere except at $z = 0$. Therefore, when the velocity information is included in states, the transitions become stiff around $z = 0$ (see Section 8.1 for an illustration). As a result, we only need to apply BS when the inactive impact contact points are near the obstacles to avoid large variance, while still obtaining accurate gradients with minimal bias.

Algorithm 1 Adaptive Barrier Smoothing (ABS)

Input: State x_t, action u_t, contact-aware adaptive $\mu(\cdot, \cdot)$
for a sequence of decreasing $\mu' > 0$ **do**
  Solve the μ'-softened complementarity problem
  if $\mu' < \mu(x_t, u_t)$ **then**
    Compute the gradients $\partial x_{t+1}/\partial x_t$, $\partial x_{t+1}/\partial u_t$
  end if and **break**
end for
Output: $\partial x_{t+1}/\partial x_t$, $\partial x_{t+1}/\partial u_t$ (for computing (4.1))

We provide the pseudocode of Adaptive Barrier Smoothing in Algorithm 1. By adopting ABS to compute the policy gradient in (4.1) and following Algorithm 2 as the main training loop, we obtain the FOPG-ABS method. Specific implementation details can be found in Appendix B.

The choice of g can be problem-specific, as long as it is non-increasing (e.g., we use (C.2) in our Dojo experiments). In what follows, we show that BS is closely related to randomized smoothing and, when certain contact-aware forms of the adaptive $\mu(x_t, u_t)$ are taken, has a small gradient bias.

5.2. Analysis of Gradient Bias

In this section, we focus on 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.

We first build connections between BS and Randomized Smoothing (RS) (Suh et al., 2022a;b; Pang et al., 2022), which averages the stochastic gradient (see Sec. 2.3).

Proposition 5.2 (Equivalence between BS and RS)**.** Let $z_t = Dx_t + Eu_t + d \in \mathbb{R}$. Recall that the solutions of the exact LCP and the softened LCP are $S_{\mu=0}(z_t)$ and $S_{\mu(z_t)}(z_t)$, respectively (see Definition 3.1). For any central-path parameter $\mu(z_t)$, barrier smoothing 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 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 5.2, we are able to work directly on the randomization-smoothed gradient when studying the bias of barrier smoothing. This gives the following results adapted from the analysis on randomized smoothing presented in (Pang et al., 2022).

Proposition 5.3 (ABS Minimizes the Linearization Residual)**.** Let the error function be the σ-Gaussian tail integral $\text{erf}(y;\sigma^2) = \int_y^\infty 1/(\sqrt{2\pi}\sigma)e^{-y^2/\sigma^2}$. We set the 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 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, with the minimal residual error δ, i.e.,

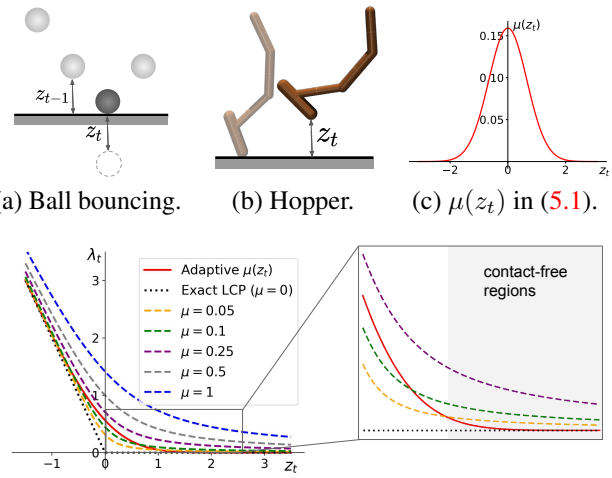
$$\delta = \min_{K,W} \mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}\Big[\big|S_{\mu=0}(z_t + w) - Ww - K\big|\Big].$$

Then K^*, W^* that achieve the minimum are the barrier-smoothed solution and its gradient, respectively. Formally,

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

The above proposition shows that the solution of barrier-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 contact-aware as it is large only 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



(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-aware 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-aware 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.

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 5.4 (Bias of Adaptive Barrier Smoothing)**.** With the same definition of $\mu(z_t)$ in Proposition 5.3, 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

$$\big\|\nabla f_{\mu=0} - \nabla f_{\mu(z_t)}\big\|_2 \leq \|C\|_F(\|D\|_F + \|E\|_F) \cdot \Big(\frac{1}{F} + \frac{12\delta + \varsigma}{\sigma\mathcal{Q}(2/3)}\Big),$$

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

Theorem 5.4 establishes a bound on the gradient bias of barrier smoothing when the contact-aware $\mu(z_t)$ conforms to certain forms. It indicates that the softened LCS $f_{\mu(z_t)}$ and its gradient achieves the best linearization error and has small gradient bias b_n.

Discussion on BS and RS. Although equivalence can be proven between Barrier Smoothing (BS) and Randomized Smoothing (RS) in the event of an infinite number of samples, RS is plagued by both empirical bias (Suh et al., 2022b;a) and the presence of noisy gradients. The empirical bias phenomenon of RS happens under discontinuities

or stiffness (see Fig. 6(a)). Besides, RS gradient is noisy and expensive to compute due to its sampling procedure. In contrast, BS avoids these issues by differentiating directly through the softened system $f_{\mu>0}$. However, it should be noted that unlike RS, vanilla BS does *not* automatically adjust the degree of smoothing based on the dynamics, causing significant bias when applied universally (Pang et al., 2022).

6. Complementarity-Model-Based FOPG

As the vanilla FOPG framework described in Section 4 assumes access to a differentiable simulator, in this section, we suggest a more general complementarity-model-based FOPG approach that works when the dynamics is unknown and needs to be learned.

In Alg. 2, we provide the pseudocode of complementarity-model-based FOPG methods, 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 $\{\psi_n\}_{n\in[N]}$ and $\{\theta_n\}_{n\in[N]}$, respectively.

Algorithm 2 Complementarity-Model-Based FOPG

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 (6.1)
3: Update the policy parameter θ_n by (4.2)
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 complementarity model $x_{t+1} = f_\psi(x_t, u_t)$ is learned from the training dataset $\mathcal{D} = \{(x_t^*, u_t^*, x_{t+1}^*)\}_t$, where the state $x_t^* \in \mathbb{R}^{d_x}$ is the system's configuration (including the velocity v_t, coordinate q_t, etc.). Instead of being parameterized by a black-box neural network, f_ψ returns the solution of (2.2) constrained by (2.3), where the ground-truth physics data is replaced by the estimated one contained in ψ, such as the parameters of each body. The model training loss is as follows, minimized by random search,

$$L(\psi; \mathcal{D}) = \sum_{t=1}^{|\mathcal{D}|} \frac{1}{2} \left\| f_\psi(x_t^*, u_t^*) - x_{t+1}^* \right\|_2^2. \tag{6.1}$$

Policy Update. The policy update procedure is the same as the vanilla FOPG except that $x_{t+1,m} = f_\psi(x_{t,m}, u_{t,m})$ in (4.1), as opposed to f. Applying ABS on the complementarity-model-based FOPG also follows a similar recipe to replace the $\partial x_{t+1}/\partial x_t$, $\partial x_{t+1}/\partial u_t$ terms in (4.1) by the outputs of Algorithm 1.

Discussion on the Model Representations. Most modern model-based RL 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 stiff contact behaviors while still struggling with inaccurate first-order gradient estimation in long-horizon problems (Hochlehnert et al., 2021). On the contrary, the complementarity models have the potential to efficiently fit the contact dynamics as they are physics-informed and only the necessary physical parameters are needed to be learned.

7. Related Work

Differentiable Simulation. Physics-informed (Jiang et al., 2018; Pizzuto & Mistry, 2021) complementarity-based models are 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. On the other hand, simulators like MuJoCo (Todorov et al., 2012) and PhysX implement soft contacts and can generate physics-violated behaviors. Their non-differentiable nature also necessitates expensive finite-difference to obtain the first-order gradients.

Smoothing Techniques. In our analysis, we extend the results in (Pang et al., 2022) that connect vanilla BS and randomized smoothing to obtain a general equivalence that also holds for Adaptive BS, with the ultimate goal to bound the ABS gradient bias. Besides, (Pang et al., 2022) studied smoothing that adds *fixed* log-barrier terms to dynamics, while our implementation and analysis are based on the complementarity problems and we justify the bias issue of the universal application of BS, which motivates the *adaptive* utilization of BS. Moreover, (Suh et al., 2022a;b) aimed to address the empirical bias issue of randomized smoothing, while we focus primarily on the trade-off between gradient variance and bias when applying barrier smoothing. Previous works (Parmas et al., 2018; Metz et al., 2019) proposed to reduce the variance by combining the first- and zeroth-order gradients, but still experiencing similar issues of RS.

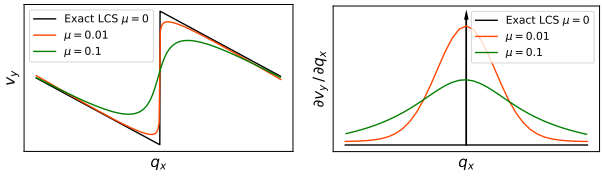
Policy Gradient Methods. The zeroth-order policy gradient methods include REINFORCE (Williams, 1992) and actor-critic (Sutton et al., 1999; Kakade, 2001; Kakade & Langford, 2002; Degris et al., 2012; Zhang et al., 2021), where the convergence results are established in recent works (Agarwal et al., 2021; Wang et al., 2019; 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). Our work is also related to the model-based RL literature, where a predictive model is learned from data for policy optimization (Janner et al., 2019; Feinberg et al., 2018; Zhang, 2022; Amos et al., 2021) or planning (Wang & Ba, 2019; Schrittwieser et al., 2020; Curi et al., 2020). The vast amount of data required to fit the stiff robotic contact behaviors using universal dynamics approximators highlights the significance of studying complementarity models and FOPG (Pfrommer et al., 2021).

8. Experiments

8.1. Contact Behaviors and System Stiffness

To begin, we examine how the contact events lead to the system's stiffness. In Figure 2, we plot the dynamics and derivatives of the velocity w.r.t. coordinate in the ball-bouncing example depicted in Figure 1(a), where the ball is thrown with an initial velocity and subsequently experiences impact contact upon hitting the ground. These contact events, which serve as the foundation for complex behaviors, are prevalent in nearly all robotic tasks.



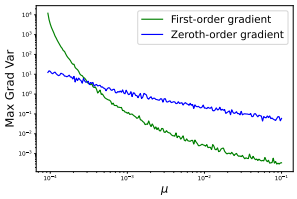
(a) Dynamics. (b) Derivative.

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.

In Figure 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 barrier smoothing with a larger value of μ, the dynamics of the system become less stiff.

8.2. 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, which we will discuss in more detail.



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 during the FOPG training in locomotion tasks is illustrated in Figure 3.

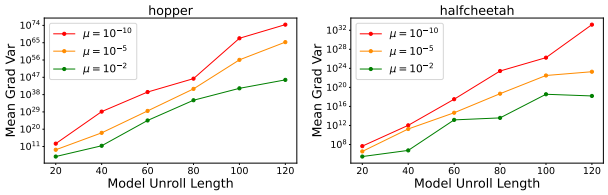
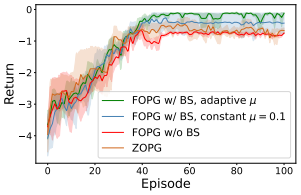


Figure 3. The mean gradient variance with different model unroll lengths when changing the value of μ.

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 theoretical findings in Theorem 4.4.

8.3. Performance of Adaptive Barrier Smoothing

In this section, we evaluate the proposed Adaptive Barrier Smoothing (ABS) mechanism applied to the complementarity-model-based FOPG algorithm.



We begin 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, see Appendix C.2 for an illustration.

We observe that the application of Barrier Smoothing (BS)

to FOPG results in a higher asymptotic return and faster convergence than both ZOPG and vanilla FOPG methods. Besides, Adaptive BS achieves superior performance compared to all the other variants.

In Figure 4, we compare FOPG-ABS (ours) and several (model-based) zeroth- and first-order policy gradient methods in Dojo locomotion tasks. For MBPO (Janner et al., 2019), we use neural network (NN) models that are trained by minimizing the mean squared error. For all other algorithms, we use complementarity-based models.

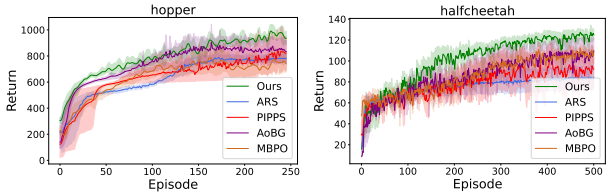
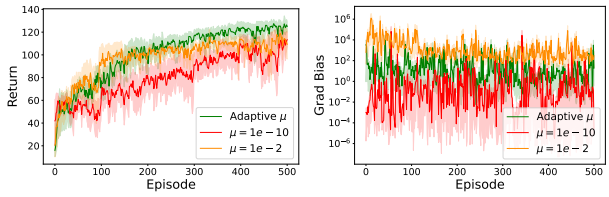


Figure 4. Comparison between FOPG-ABS (Ours) and ARS (Mania et al., 2018), PIPPS (Parmas et al., 2018), AoBG (Suh et al., 2022a), MBPO.

8.4. Ablation Studies

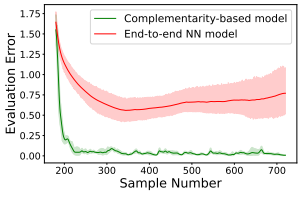
Gradient Bias. In Figure 5, we compare the performance of FOPG in the half-cheetah task when choosing different μ for barrier smoothing. As we are using ground-truth physics parameters, the only source of bias in this comparison is the central-path parameter. Our results show that $\mu = 10^{-10}$ results 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-aware adaptive design offers a more favorable balance between gradient bias and variance.



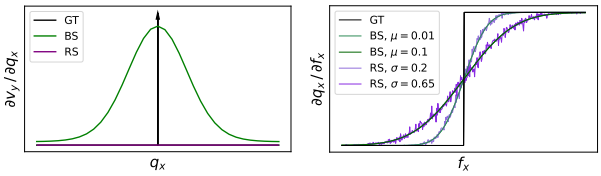
(a) Performance. (b) Gradient bias.

Figure 5. Barrier Smoothing applied to FOPG when equipped with different central-path parameter μ.

Model Learning. In order to showcase the effectiveness and efficiency of complementarity-based models in approximating contact behaviors compared to universal black-box neural network (NN) models, we plot the mean state prediction error in the hopper task tested on an evaluation transition dataset collected by random policies.



Different Smoothing Mechanisms. Figure 6(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 (RS). Specifically, the Barrier Smoothed (BS) gradient successfully approximates the unit impulse at contact, i.e., the ground-truth (GT), while the RS gradient is constant and exhibits a large bias. For frictional contact behaviors, such as pushing a box on a frictional ground, the RS gradients in Figure 6(b) are both noisy and computationally costly.



(a) Impact contact. (b) Frictional contact.

Figure 6. Comparison between Barrier Smoothing (BS) and Randomized Smoothing (RS) in the two types of contact dynamics, namely impact contact and frictional contact.

9. Conclusion

In this work, we investigate the use of First-Order Policy Gradient (FOPG) methods for robotic hard-contact dynamics with extreme curvatures, specifically focusing on complementarity-based models. Our findings indicate that the convergence of FOPG is dependent on gradient variance and bias, and that stiff models can result in large gradient variance and optimization challenges. Although smoothing techniques can be applied to control the stiffness, the universal utilization can lead to significant bias. To achieve a balance between the gradient variance and bias, we propose *Adaptive Barrier Smoothing* that reduces the FOPG gradient variance while controlling the gradient bias using a contact-aware adaptive central-path parameter. Furthermore, we present a complementarity-model-based FOPG algorithm and conduct experiments to support our theory and method. In the future, it would be intriguing to explore the influence of *soft* contact dynamics on FOPG algorithms. Additionally, further research on the first-order policy gradient could be conducted in more general nonlinear frictional contact dynamics robotic systems.

Acknowledgements

Zhaoran Wang acknowledges National Science Foundation (Awards 2225087, 2211210, 2048075, 2015568, 2008827, 1934931/2216970), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, Two Sigma, Tencent for their supports.

References

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.

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.

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

de Avila Belbute-Peres, F., Smith, K., Allen, K., Tenenbaum, J., and Kolter, J. Z. End-to-end differentiable physics for learning and control. *Advances in neural information processing systems*, 31, 2018.

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

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

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.

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

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.

Lidec, Q. L., Jallet, W., Montaut, L., Laptev, I., Schmid, C., and Carpentier, J. Contact models in robotics: a comparative analysis. *arXiv preprint arXiv:2304.06372*, 2023.

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

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

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*, 2019.

Wang, T. and Ba, J. Exploring model-based planning with policy networks. *arXiv preprint arXiv:1906.08649*, 2019.

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.

Zhang, S. Conservative dual policy optimization for efficient model-based reinforcement learning. *Advances in neural information processing systems*, 2022.

Zhang, S., Shen, L., and Han, L. Learning meta representations for agents in multi-agent reinforcement learning. *arXiv preprint arXiv:2108.12988*, 2021.

Zhao, L., Xu, H., and Wong, L. L. Scaling up and stabilizing differentiable planning with implicit differentiation. *arXiv preprint arXiv:2210.13542*, 2022.

A. Proofs

A.1. Proof of Lemma 3.3

Proof. For the constrained optimization problem in (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 have $\epsilon_t^{(i)} \lambda_t^{(i)} = \mu$ 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 in (4.1), we know that $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) = (\theta_{n+1} - \theta_n)/\eta$. By Assumption 4.1, we have

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

By basic algebra, we have for $\nabla_\theta J(\pi_{\theta_n})^\top \widehat{\nabla}_\theta J(\pi_{\theta_n})$ that

$$\begin{aligned}
&\nabla_\theta J(\pi_{\theta_n})^\top \widehat{\nabla}_\theta J(\pi_{\theta_n}) \\
&= \Big(\nabla_\theta J(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_n})\big]\Big)^\top \widehat{\nabla}_\theta J(\pi_{\theta_n}) - \Big(\widehat{\nabla}_\theta J(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_n})\big]\Big)^\top \widehat{\nabla}_\theta J(\pi_{\theta_n}) + \widehat{\nabla}_\theta J(\pi_{\theta_n})^\top \widehat{\nabla}_\theta J(\pi_{\theta_n}) \\
&\geq \underbrace{-\Big|\Big(\nabla_\theta J(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_n})\big]\Big)^\top \widehat{\nabla}_\theta J(\pi_{\theta_n})\Big|}_{\text{(I)}} \underbrace{- \Big(\widehat{\nabla}_\theta J(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_n})\big]\Big)^\top \widehat{\nabla}_\theta J(\pi_{\theta_n})}_{\text{(II)}} + \underbrace{\widehat{\nabla}_\theta J(\pi_{\theta_n})^\top \widehat{\nabla}_\theta J(\pi_{\theta_n})}_{\text{(III)}}.
\end{aligned}$$

The resulting three terms can be bounded as follows,

$$\begin{aligned}
&\text{Term (I):} && \text{(I)} \leq \Big\|\widehat{\nabla}_\theta J(\pi_{\theta_n})\Big\|_2 \cdot \Big\|\nabla_\theta J(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_n})\big]\Big\|_2 = \Big\|\widehat{\nabla}_\theta J(\pi_{\theta_n})\Big\|_2 \cdot b_n, \\
&\text{Term (II):} && \text{(II)} \leq \frac{\Big\|\widehat{\nabla}_\theta J(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta J(\pi_{\theta_n})\big]\Big\|_2^2}{2} + \frac{\Big\|\widehat{\nabla}_\theta J(\pi_{\theta_n})\Big\|_2^2}{2}, \\
&\text{Term (III):} && \text{(III)} \geq \Big\|\widehat{\nabla}_\theta J(\pi_{\theta_n})\Big\|_2^2.
\end{aligned}$$

Thus, by plugging the above three inequalities into (A.4), we have

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

By taking expectation on both sides of (A.5), we obtain

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

Rearranging terms gives

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

By establishing the connection between the minimum expected gradient norm and the average norm over T iterations, we are able to obtain the following bound,

$$\begin{aligned}
\min_{t\in[T]} \mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] &\leq \frac{1}{N} \cdot \sum_{n=0}^{N-1} \mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] \\
&\leq \frac{2}{N} \cdot \sum_{n=0}^{N-1} \Big(\mathbb{E}\Big[\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\|_2^2\Big] + \mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] \Big),
\end{aligned} \tag{A.7}$$

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

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

The last term on the right-hand side of (A.7) can be characterized by

$$\begin{aligned}
\mathbb{E}\Big[\big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2^2\Big] &= \mathbb{E}\Big[\Big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big] + \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big] - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\Big\|_2^2\Big] \\
&\leq 2\Big\|\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big]\Big\|_2^2 + 2\mathbb{E}\Big[\Big\|\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}\big[\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big]\Big\|_2^2\Big] \\
&= 2b_n^2 + 2v_n,
\end{aligned} \tag{A.9}$$

For $N \geq 4L^2$, 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.9), we further have

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

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

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

where the second inequality holds since $\|\theta\|_2 \leq \delta$ for any θ. This concludes the proof. □

A.3. Proof of Theorem 4.4

We first provide the complete statement of Assumption 4.3.

Assumption A.1 (Lipschitz Continuous Functions)**.** We assume that the policy $\pi_\theta(x, \varsigma)$, model $f(x, u; \psi)$, and reward function $r(x, u)$ are L_π, L_f, and L_r Lipschitz continuous, respectively, that is, for any θ, ψ,

$$\begin{aligned}\big\|\pi_\theta(x_1, \varsigma_1) - \pi_\theta(x_2, \varsigma_2)\big\|_2 &\le L_\pi \cdot \big\|(x_1 - x_2, \varsigma_1 - \varsigma_2)\big\|_2,\\ \big\|f(x_1, u_1) - f(x_2, u_2; \psi)\big\|_2 &\le L_f \cdot \big\|(x_1 - x_2, u_1 - u_2)\big\|_2,\\ \big|r(x_1, u_1) - r(x_2, u_2)\big| &\le L_r \cdot \big\|(x_1 - x_2, u_1 - u_2)\big\|_2.\end{aligned}$$

Additionally, we assume that the policy $\pi_\theta(x, \varsigma)$ is L_θ-Lipschitz continuous in θ, which implies $\|\nabla_\theta \pi_\theta(x, \varsigma)\|_2 \le L_\theta$ for any state $x \in \mathcal{X}$.

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

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

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

$$v_n \le \mathbb{E}\Big[\big\|g - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2\Big]^2. \tag{A.10}$$

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

$$\mathbb{E}\Big[\big\|g - \widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})\big\|_2\Big] \le \sum_{t=0}^{H-1} \mathbb{E}_{\overline{y}_t}\Big[\big\|\nabla_\theta r(y_t) - \nabla_\theta r(\overline{y}_t)\big\|_2\Big]. \tag{A.11}$$

By the chain rule, we have for any $t \ge 1$ that

$$\frac{\mathrm{d}u_t}{\mathrm{d}\theta} = \frac{\partial u_t}{\partial x_t} \cdot \frac{\mathrm{d}x_t}{\mathrm{d}\theta} + \frac{\partial u_t}{\partial \theta}, \tag{A.12}$$

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

Plugging (A.12) at the $(t-1)$-th timestep, i.e., $\mathrm{d}u_{t-1}/\mathrm{d}\theta$, into (A.13), we get

$$\begin{aligned}\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta}\right\|_2 &= \left\|\left(\frac{\partial x_t}{\partial x_{t-1}} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\partial u_{t-1}}{\partial x_{t-1}}\right) \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\partial u_{t-1}}{\partial \theta}\right\|_2\\ &\le L_f \widetilde{L}_\pi \cdot \left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2 + L_f L_\theta,\end{aligned} \tag{A.14}$$

where the inequality follows from the Cauchy-Schwarz inequality and Assumption 4.3.

Recursively applying (A.14), we obtain for any $t \geq 1$ that

$$\left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 \leq L_f L_\theta \cdot \sum_{j=0}^{t-1} L_f^j \widetilde{L}_\pi^j \leq i \cdot L_\theta L_f^{t+1} \widetilde{L}_\pi^t, \tag{A.15}$$

where the first inequality follows from the induction

$$z_n = a z_{t-1} + b = a \cdot (a z_{i-2} + b) + b = a^t \cdot z_0 + b \cdot \sum_{j=0}^{t-1} a^j, \tag{A.16}$$

In (A.16), $\{z_j\}_{0 \leq j \leq i}$ is the real sequence satisfying $z_j = a z_{j-1} + b$. For $\mathrm{d}u_t/\mathrm{d}\theta$ defined in (A.12), we further have

$$\left\| \frac{\mathrm{d}u_t}{\mathrm{d}\theta} \right\|_2 \leq L_\pi \cdot \left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 + L_\theta \leq t \cdot L_\theta L_f^{t+1} \widetilde{L}_\pi^{t+1} + L_\theta. \tag{A.17}$$

Combining (A.15) and (A.17), we obtain

$$\left\| \frac{\mathrm{d}y_t}{\mathrm{d}\theta} \right\|_2 = \left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} \right\|_2 + \left\| \frac{\mathrm{d}u_t}{\mathrm{d}\theta} \right\|_2 \leq \underbrace{2t \cdot L_\theta L_f^{t+1} \widetilde{L}_\pi^{t+1} + L_\theta}_{K(t)}. \tag{A.18}$$

By the chain rule, (A.11) can be decomposed and bounded by

$$\begin{aligned}
&\mathbb{E}_{\overline{y}_t}\left[\left\| \nabla_\theta r(y_t) - \nabla_\theta r(\overline{y}_t) \right\|_2 \right] \\
&= \mathbb{E}_{\overline{y}_t}\left[\left\| \nabla r(y_t) \nabla_\theta y_t - \nabla r(\overline{y}_t) \nabla_\theta \overline{y}_t \right\|_2 \right] \\
&\leq \mathbb{E}_{\overline{y}_t}\left[\left\| \nabla r(y_t) \nabla_\theta y_t - \nabla r(y_t) \nabla_\theta \overline{y}_t \right\|_2 \right] + \mathbb{E}\left[\left\| \nabla r(y_t) \nabla_\theta \overline{y}_t - \nabla r(\overline{y}_t) \nabla_\theta \overline{y}_t \right\|_2 \right] \\
&\leq L_r \cdot \left(\mathbb{E}_{\overline{x}_n}\left[\left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta} \right\|_2 \right] + \mathbb{E}_{\overline{u}_n}\left[\left\| \frac{\mathrm{d}u_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta} \right\|_2 \right] \right) + 2L_r \cdot K(t),
\end{aligned} \tag{A.19}$$

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

Plugging (A.19) into (A.11) and (A.10), we have

$$\begin{aligned}
v_n &\leq L_r \cdot \left(\sum_{t=0}^{H-1} \left(\mathbb{E}_{\overline{x}_t}\left[\left\| \frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta} \right\|_2 \right] + \mathbb{E}_{\overline{u}_t}\left[\left\| \frac{\mathrm{d}u_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta} \right\|_2 \right] + 2K(t) \right) \right)^2 \\
&= O\left(\left(\sum_{t=0}^{H-1} t^2 \widetilde{L}_f^{2t} \widetilde{L}_\pi^{2t} \right)^2 \right) = O\left(H^4 \widetilde{L}_f^{4H} \widetilde{L}_\pi^{4H} \right),
\end{aligned} \tag{A.20}$$

where the second inequality follows from the results from Lemma A.2 and by plugging the definition of K in (A.18).

Note that the variance v_t scales with the batch size N at the rate of $1/M$. Since the analysis above is established for $M = 1$, the bound of the gradient variance v_t is established by dividing the right-hand side of (A.20) by M, which concludes the proof of Theorem 4.4.

□

Lemma A.2. 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 $t \geq 1$ and the corresponding state x_t, control input u_t, we have the following inequality results,

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

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

Proof. From (A.13), we obtain for any $t \geq 1$ that

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

According to the triangle inequality, we continue by

$$\begin{aligned}
&\leq \mathbb{E}\left[\left\|\frac{\partial x_t}{\partial x_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{x}_{t-1}} \cdot \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\partial \overline{x}_t}{\partial \overline{u}_{t-1}} \cdot \frac{\mathrm{d}\overline{u}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\leq 2L_f \cdot \left(\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2 + \left\|\frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}\right\|_2\right) + L_f \cdot \mathbb{E}_{\overline{x}_{t-1}}\left[\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad + L_f \cdot \mathbb{E}_{\overline{u}_{t-1}}\left[\left\|\frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_{t-1}}{\mathrm{d}\theta}\right\|_2\right]. \qquad \text{(A.21)}
\end{aligned}$$

Similarly, we have from (A.12) that

$$\begin{aligned}
&\mathbb{E}_{\overline{u}_n}\left[\left\|\frac{\mathrm{d}u_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{u}_t}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad = \mathbb{E}\left[\left\|\frac{\partial u_t}{\partial x_t} \cdot \frac{\mathrm{d}x_t}{\mathrm{d}\theta} + \frac{\partial u_t}{\partial \theta} - \frac{\partial \overline{u}_t}{\partial \overline{x}_t} \cdot \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta} - \frac{\partial \overline{u}_t}{\partial \theta}\right\|_2\right] \\
&\quad \leq \mathbb{E}\left[\left\|\frac{\partial u_t}{\partial x_t} \cdot \frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\partial \overline{u}_t}{\partial \overline{x}_t} \cdot \frac{\mathrm{d}x_t}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial \overline{u}_t}{\partial \overline{x}_t} \cdot \frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\partial \overline{u}_t}{\partial \overline{x}_t} \cdot \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] + \mathbb{E}\left[\left\|\frac{\partial u_t}{\partial \theta} - \frac{\partial \overline{u}_t}{\partial \theta}\right\|_2\right] \\
&\quad \leq 2L_\pi \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta}\right\|\right] + L_\pi \cdot \mathbb{E}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] + 2L_\theta. \qquad \text{(A.22)}
\end{aligned}$$

Plugging (A.22) back to (A.21),

$$\begin{aligned}
&\mathbb{E}_{\overline{x}_t}\left[\left\|\frac{\mathrm{d}x_t}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_t}{\mathrm{d}\theta}\right\|_2\right] \\
&\quad \lesssim 4L_f\widetilde{L}_\pi \cdot \left(\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta}\right\|_2 + \left\|\frac{\mathrm{d}u_{t-1}}{\mathrm{d}\theta}\right\|_2\right) + L_f\widetilde{L}_\pi \cdot \mathbb{E}_{\overline{x}_{t-1}}\left[\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] + 2L_fL_\theta \\
&\quad \leq 4L_f\widetilde{L}_\pi \cdot K(t-1) + L_f\widetilde{L}_\pi \cdot \mathbb{E}_{\overline{x}_{t-1}}\left[\left\|\frac{\mathrm{d}x_{t-1}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{x}_{t-1}}{\mathrm{d}\theta}\right\|_2\right] + 2L_fL_\theta, \qquad \text{(A.23)}
\end{aligned}$$

where the last inequality follows from the definition of K in (A.18).

Recursively applying (A.23), we obtain

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

where the first equality follows from (A.16).

As a consequence, we have from (A.22) that

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

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., $\|Z\|_F = \sqrt{\text{tr}(ZZ^\top)}$.

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.3 (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 $\partial x_{t+1}/\partial x_t$, so 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

$$\frac{\partial x_{t+1}}{\partial x_t} = A + \sum_{i=1}^{d_\lambda} C^{(i)} \frac{\partial \lambda^{(i)}}{\partial x_t}. \tag{A.24}$$

We rewrite the contact equation $\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu \vec{1}$ in (3.1) as follows, such that for any $i \in [1, d_\lambda]$,

$$\lambda_t^{(i)} \big(D^{(i)\top} x_t + E^{(i)\top} u_t + F^{(i)\top} \lambda_t + d^{(i)}\big) = \mu. \tag{A.25}$$

By the Implicit Function Theorem A.3, we have for any $i \in [1, d_\lambda]$ that

$$\begin{aligned}\frac{\partial \lambda^{(i)}}{\partial x_t} &= -\Big(D^{(i)\top} x_t + E^{(i)\top} u_t + \frac{\partial}{\partial \lambda_t^{(i)}} \lambda_t^{(i)} F^{(i)\top} \lambda_t + d^{(i)}\Big)^{-1} \lambda_t^{(i)} D^{(i)\top} \\ &= -\big(D^{(i)\top} x_t + E^{(i)\top} u_t + F^{(i)\top} \lambda_t + \lambda_t^{(i)} F^{(i)(i)} + d^{(i)}\big)^{-1} \lambda_t^{(i)} D^{(i)\top},\end{aligned} \tag{A.26}$$

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

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

where the inequality holds due to the Cauchy–Schwarz inequality. From (A.25), we know that $D^{(i)\top}x_t + E^{(i)\top}u_t + F^{(i)\top}\lambda_t + d^{(i)} = \mu/\lambda_t^{(i)}$, it then follows that

$$\begin{aligned}
&= \|A\|_2 + \sum_{i=1}^{d_\lambda} \lambda_t^{(i)} \|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2 \cdot \left|\mu/\lambda_t^{(i)} + \lambda_t^{(i)} F^{(i)(i)}\right|^{-1} \\
&\le \|A\|_2 + \sum_{i=1}^{d_\lambda} \lambda_t^{(i)} \|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2 \cdot \left|\mu/\lambda_t^{(i)}\right|^{-1} \\
&= \|A\|_2 + \sum_{i=1}^{d_\lambda} \|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2 \cdot (\lambda_t^{(i)})^2/\mu, \qquad \text{(A.27)}
\end{aligned}$$

where the inequality holds since $F^{(i)(i)} > 0$, $\mu > 0$, and $\lambda_t^{(i)} > 0$.

By the definition of Frobenius norm, we know that

$$\begin{aligned}
\|C\|_F &= \sqrt{\sum_{i=1}^{d_\lambda} \|C^{(i)}\|_2^2} = \sqrt{d_\lambda} \cdot \sqrt{\sum_{i=1}^{d_\lambda} \frac{1}{d_\lambda} \|C^{(i)}\|_2^2} \\
&\ge \sqrt{d_\lambda} \cdot \sum_{i=1}^{d_\lambda} \frac{1}{d_\lambda} \sqrt{\|C^{(i)}\|_2^2} = \frac{1}{\sqrt{d_\lambda}} \sum_{i=1}^{d_\lambda} \|C^{(i)}\|_2, \qquad \text{(A.28)}
\end{aligned}$$

where we adopt Jensen's inequality in the inequality.

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

$$\|\lambda_t\|_2^2 = \sum_{i=1}^{d_\lambda} (\lambda_t^{(i)})^2 \le d_\lambda \cdot \|\Lambda_t\|_F^2. \qquad \text{(A.29)}$$

Therefore, we can further bound (A.27) by

$$\begin{aligned}
\left\| \frac{\partial x_{t+1}}{\partial x_t} \right\|_2 &\le \|A\|_2 + \frac{1}{\mu} \left(\sum_{i=1}^{d_\lambda} \|C^{(i)}\|_2 \right) \cdot \left(\sum_{i=1}^{d_\lambda} \|D^{(i)}\|_2 \right) \cdot \left(\sum_{i=1}^{d_\lambda} (\lambda_t^{(i)})^2 \right) \\
&\le \|A\|_2 + \frac{d_\lambda}{\mu} \|C\|_F \|D\|_F \|\lambda_t\|_2^2 \\
&\le \|A\|_F + \frac{d_\lambda^2}{\mu} \|C\|_F \|D\|_F \|\Lambda_t\|_F^2, \qquad \text{(A.30)}
\end{aligned}$$

where the first inequality holds since $\sum_i y_i \cdot z_i \le (\sum_i y_i) \cdot (\sum_i z_i)$ for any non-negative scalar sequences y_i, z_i and the second inequality follows from (A.28). The third inequality follows from (A.29) and the fact that $\|A\|_2 \le \|A\|_F$.

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

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

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

$$\|F\|_F \cdot \|\Lambda_t\|_F^2 + e \cdot \|\Lambda_t\|_F - \mu \ge 0. \qquad \text{(A.31)}$$

Solving (A.31) 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

$$\begin{aligned} l(\mu) =& \frac{\|\Lambda_t\|_F^2}{\mu} \geq \frac{2e^2+4\mu\|F\|_F-2e\sqrt{e^2+4\mu\|F\|_F}}{4\mu\|F\|_F^2} \\ =& \frac{e^2}{2\mu\|F\|_F^2}+\frac{1}{\|F\|_F}+\frac{e^2\sqrt{\frac{1}{\mu^2}+\frac{4\|F\|_F}{\mu e^2}}}{2\|F\|_F^2} \\ =& \frac{\varepsilon}{\mu}+\frac{1}{\|F\|_F}+\varepsilon\sqrt{\frac{1}{\mu^2}+\frac{2}{\varepsilon\mu\|F\|_F}}. \end{aligned} \tag{A.32}$$

Plug (A.32) into (A.30), 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 Theorem 5.1

Proof. From (A.25), we have for any $i \in [1, d_\lambda]$ that

$$\mu = \lambda_t^{(i)}\left(z_t^{(i)} + F^{(i)\top}\lambda_t\right) \geq \lambda_t^{(i)} \cdot z_t^{(i)},$$

where the inequality holds since $F^{(i)\top}\lambda_t > 0$.

Therefore, for $z_t^{(i)} > 0$, we obtain

$$\lambda_t^{(i)} \leq \mu / z_t^{(i)}. \tag{A.33}$$

Plugging (A.33) into (A.27), we have

$$\begin{aligned} \left\|\frac{\partial x_{t+1}}{\partial x_t}\right\|_2 &\leq \|A\|_2 + \sum_{i=1}^{d_\lambda}\|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2 \cdot (\lambda_t^{(i)})^2/\mu \\ &\leq \|A\|_2 + \mu\sum_{i=1}^{d_\lambda}\|C^{(i)}\|_2 \cdot \|D^{(i)}\|_2/(z_t^{(i)})^2. \end{aligned} \tag{A.34}$$

Since $\|A\|_2 \leq \|A\|_F$ and $\|C^{(i)}\|_2 \leq \sum_{i=1}^{d_\lambda}\|C^{(i)}\|_2 \leq \sqrt{d_\lambda}\|C\|_F$ following (A.28), we further have from (A.34) that

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

Similarly, we obtain

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

This completes the proof of Theorem 5.1. □

A.6. Proof of Proposition 5.2

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 following 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,

$$S_{\mu=0}(z_t) = \begin{cases} -z_t/F & \text{if } z_t \leq 0 \\ 0 & \text{else} \end{cases}. \tag{A.35}$$

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

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

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.7. Proof of Proposition 5.3

Recall that Proposition 5.2 connects the proposed barrier smoothing with the randomized smoothing. Therefore, we first provide the following lemma established in randomized smoothing as a preparation before proving Proposition 5.3.

Lemma A.4 (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, i.e.,

$$\mathcal{L}(K, W) = \min_{K,W} \frac{1}{2}\mathbb{E}_{w\sim\rho(w)}\left[\left\|g(\overline{x} + w) - Ww - K\right\|_2^2\right]. \tag{A.36}$$

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.36) 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

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

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 5.3. By applying Lemma A.4, we know that Proposition 5.3 holds once the following equivalence is established,

$$S_{\mu(z_t)}(z_t) = \mathbb{E}_{w\sim\rho(w)}[S_{\mu=0}(z_t+w)], \tag{A.37}$$

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

This is a direct result from Proposition 5.2. 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

$$S_{\mu(z_t)}(z_t) = \lambda_t = \kappa = z_t\cdot\mathrm{erf}(z_t,\sigma) + e^{-z_t^2/(2\sigma)}/\sqrt{\pi} + c_1 z_t + c_2. \tag{A.38}$$

Proposition 5.2 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.38), 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.37) and the proposition. □

A.8. Proof of Theorem 5.4

Proof. According to the definition of $S_{\mu=0}(z_t)$ in (A.35), we know that

$$\left|\frac{S_{\mu=0}(z_r+w) - S_{\mu=0}(z_t)}{w} - \nabla_z S_{\mu=0}(z_t)\right| \le \frac{1}{F}. \tag{A.39}$$

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.39) that

$$\left|\frac{\nu(w) + S_{\mu(z_t)}(z_t) - S_{\mu=0}(z_t)}{w} + \nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)\right| \le \frac{1}{F}.$$

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

$$\left|\nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)\right| \leq \frac{1}{F} + \frac{\nu(w) + \varsigma}{|w|}. \tag{A.40}$$

From Proposition 5.3, we know that

$$\mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}[\nu(w)] = \delta. \tag{A.41}$$

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

This can be proved by contradiction. Specifically, suppose for any $w \in [\sigma\mathcal{Q}(2/3), \sigma\mathcal{Q}(3/4)]$, $\nu(w) > 12\delta$. 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.41). Therefore, the claim is correct.

Using the above claim, we have from (A.40) that

$$\left|\nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)\right| \leq \frac{1}{F} + \frac{12\delta + \varsigma}{\sigma\mathcal{Q}(2/3)}.$$

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

$$\left\|\nabla_x f_{\mu=0} - \nabla_x f_{\mu(z_t)}\right\|_2 \leq \|C\|_F\|D\|_F \cdot \left(\frac{1}{F} + \frac{12\delta + \varsigma}{\sigma\mathcal{Q}(2/3)}\right),$$
$$\left\|\nabla_u f_{\mu=0} - \nabla_u f_{\mu(z_t)}\right\|_2 \leq \|C\|_F\|E\|_F \cdot \left(\frac{1}{F} + \frac{12\delta + \varsigma}{\sigma\mathcal{Q}(1/2)}\right).$$

□

B. Interior-Point Solver

In Algorithm 1, we provide the high-level framework of the proposed Adaptive Barrier Smoothing mechanism. In this section, we describe the IPM solver that is used to solve the complementarity problems, such as the Nonlinear Complementarity Problem (NCP) in (2.3) (or when $\mu = 0$ in the LCP in (3.1)). Besides, we depict how the gradients are calculated.

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. As we have discussed in Section 3.2, the interior-point method solves a sequence of μ-softened problems with decreasing $\mu > 0$. For notation simplicity, we consider the problem of the following form,

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

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

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

For a μ-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. Specifically, the pseudocode of the solver is provided in Algorithm 3.

Algorithm 3 Primal-Dual Interior-Point Solver with Stopping Criteria $\texttt{SOLVER}(\mu)$

Input: Stopping criteria μ
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 $\mu_{\text{vio}} \geq \mu$ **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

To obtain the gradients of the output with respect to the inputs, we adopt the Implicit Function Theorem in A.3 to obtain the implicit gradients, following (Howell et al., 2022; Geilinger et al., 2020; Zhao et al., 2022).

C. Details of Experiments

C.1. Dynamics in the Ball Bouncing Example

In Section 8.1, 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

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

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

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

Solving (C.1) with Python, we obtain 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.

C.2. Experimental Details

In our Dojo experiments in Section 8.3, we use a contact-aware central-path parameter for the proposed Adaptive Barrier Smoothing method. From the results in Figure 3, to balance the gradient variance and bias, we would like $\mu \to 0$ when all impact contacts are active or the distance-to-obstacle is large, and $\mu \approx 10^{-2}$ when this distance approaches zero. To accomplish this, the adaptive $\mu(x_t, u_t)$ is designed as

$$\mu(x_t,u_t) = 10^{-2}\big(100d^2+1\big)^{-4} = 10^{-2}\big(100\min_{i\in\mathcal{I}'}|\phi(x_t,u_t)^{(i)}|^2+1\big)^{-4}. \tag{C.2}$$

Here, the set $\mathcal{I}$ of the inactive impact contact points is approximated by $\mathcal{I}' = \{1 \le i \le c \,|\, \gamma_{t,\mu}^{(i)} \le 1\}$, where $\gamma_{t,\mu}$ is solved from the μ-softened complementarity problem.

We also visualize the three tasks that are used in our experiments for performance evaluation, namely, the ball-throwing task and the hopper, half-cheetah locomotion tasks.

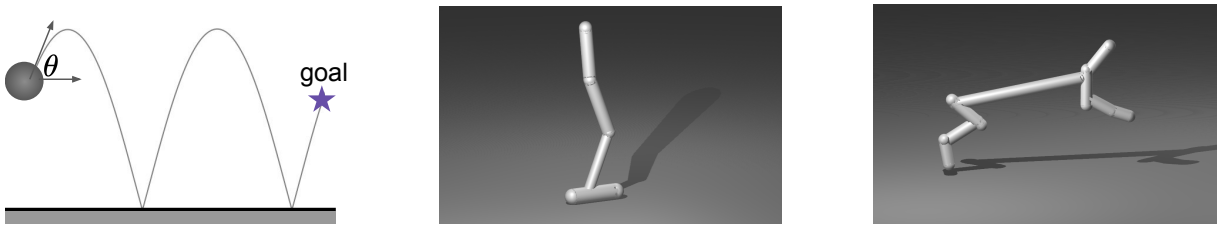


Figure 7. 7(a): Throw a ball and optimize θ to reach the goal at a certain velocity. The return curve in Section 8.3 is the negative mean squared error at the final timestep. 7(b),7(c): Screenshots of the locomotion tasks using the Dojo (Howell et al., 2022) simulator.