
Model-Based First-Order Policy Gradient for Contact Dynamics

Shenao Zhang¹ Wanxin Jin² Zhaoran Wang³

Abstract

In model-based reinforcement learning (RL), the learned models are typically smooth approximators of the environment dynamics. This is problematic in robotic systems that experience hard contact and have non-smooth or even discontinuous local behaviors. Despite the large amount of data needed to fit these behaviors, the inaccurate model gradient can lead to poor performance when applying First-Order Policy Gradient (FOPG). Therefore, we study the physics-guided complementarity-based model, which, unfortunately, is shown to be stiff. Our theory further implies that the stiffness can cause the gradient variance to explode, leading to slow convergence and optimization difficulties. For this reason, we present the softened complementarity model that corresponds to a barrier-smoothed objective. We propose the Analytic Barrier-Smoothed FOPG that avoids the large variance resulting from the impact contact. Besides, we control the gradient bias using a contact-aware adaptive centering parameter, with which the analytic smoothing is the best linear approximation of the exact unsmoothed system and the gradient bias can be upper bounded. Experimental results also support our theory and method.

1. Introduction

Model-Based Reinforcement Learning (MBRL) has achieved great success in sequential decision-making applications including board and video games (Schrittwieser et al., 2020; Kaiser et al., 2019). First-Order Policy Gradient (FOPG) that backpropagates through the path of cumulative reward computation, is the most straightforward approach in differentiable simulation (Xu et al., 2022; Freeman et al., 2021) and shows its potential in general control tasks by learning a model (Clavera et al., 2020; Li et al., 2021; Amos

et al., 2021). However, physics-based systems, e.g. robotic locomotion and manipulation, have stiff dynamics with extreme curvatures (Parmar et al., 2021; Anitescu & Potra, 2002) due to geometrical constraints and contact events. Most modern MBRL algorithms that fit the dynamical systems 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, the black-box models typically require a large amount of data to learn the contact behaviors while still suffering from large first-order gradient bias in long-horizon problems (Hochlehnert et al., 2021).

In this work, we study the physics-guided model built upon the complementarity problem that underlies hard-contact (differentiable) simulators (Geilinger et al., 2020; Howell et al., 2022; Werling et al., 2021). By solving this complementarity problem using Interior Point Method (IPM), the impact and friction contact force ensure non-penetration and maximum dissipation. Although the contact dynamics can be well approximated by fitting the physics parameter, we show that the stiffness of complementarity-based models leads to optimization difficulties when performing FOPG. Specifically, we first establish the convergence of model-based FOPG algorithms that depends on the gradient variance and bias. Moreover, the upper bound of the gradient variance has polynomial dependencies on the Lipschitz continuity of the model, where the degrees are linear in the task horizon. When the model is stiff, long chains of chaotic (Boltz, 2000) mappings lead to exploding variance and slow convergence, a phenomenon that is also observed in experiments (Parmas et al., 2018; Metz et al., 2021).

To alleviate this issue, we define a class of μ -softened Linear Complementarity Systems (LCS) that corresponds to the solution of the IPM solver with centering parameter μ . We prove that the Lipschitz upper bound of the μ -softened LCS scales inversely with μ . Since the softened LCS can be shown to be the optimality condition of a barrier-smoothed objective, we present Analytic Barrier Smoothing that prevents μ from reaching 0 by setting a positive threshold to terminate the IPM iterations. However, simply applying analytic smoothing everywhere can lead to large gradient bias.

To best trade-off between the gradient variance and bias, we

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

propose to use a contact-aware adaptive centering parameter that decreases with the minimum distance-to-obstacle of the inactive impact contacts. Intuitively, since the root of stiffness is the sudden change of the impact contact force, we only need local smoothing near the impact contact while maintaining globally accurate gradients. Built upon the equivalence between the proposed method and the randomized smoothing (Suh et al., 2022b;a), we prove that analytic smoothing is the best linear approximation of the exact unsmoothed LCS. Furthermore, we also provide the bias upper bound of our method.

2. Background

2.1. Reinforcement Learning

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

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

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

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

2.2. Stochastic Gradient Estimation

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

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

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

First-Order (or 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:

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

Derived from the *law of the unconscious statistician* (LO-TUS) (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:

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

2.3. Rigid-Body Dynamics

We consider a standard approach to modeling robotic systems — the framework of rigid-body systems with contacts. The continuous-time equation of motion is

$$M(q)dv = (n(q, v) + u)dt + J(q)^\top \lambda,$$

where we let q denote the generalized coordinates, v the generalized velocities, $u \in \mathbb{R}^{n_u}$ the applied control force, $M(q)$ 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 $\lambda := (\gamma^{(1)}, \beta^{(1)}, \dots, \gamma^{(c)}, \beta^{(c)}) \in \mathbb{R}^{n_\lambda}$ as the (unknown) contact space force, where γ and β are the normal *impact* forces and *friction* forces, respectively, and c denotes the number of contact points. The state x usually contains q and v .

Using Euler approximation and multiplying by M_t^{-1} , the discrete-time dynamics can be modeled in contact space by

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

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

The friction and impacts are constrained by the system's configuration and the applied contact impulses. The impact problem is encoded with the following constraints:

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

where \circ is the element-wise (Hadamard) product, $\phi(q_{t+1}) = \phi(x_t, u_t)$ is the signed distance from the contact points to obstacles, $\vec{0}$ is the zero vector, and the equality, inequality are also element-wise. The intuition behind (2.4) is that the magnitude of the normal impact forces must be non-negative and can only be non-zero to maintain non-negative gaps (non-penetration) when there is a contact.

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

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

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

3. Complementarity-Based Contact Models

3.1. Softened Linear Complementarity Systems

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

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

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

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

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 S_μ as the solver of the μ -complementarity problem, which returns the solution $\lambda_t = S_\mu(Dx_t + Eu_t + d) \in \mathbb{R}^{d_\lambda}$.

In simulation, $\mu = 0$ corresponds to the exact Linear Complementarity Problem (LCP) where the LCS $f_{\mu=0}$ resembles the reality. Obviously, solving for the contact space impulse λ_t is our main problem, as x_{t+1} is readily obtained from the dynamics. Next, we introduce the assumption and method for solving the exact LCP.

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

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

3.2. Smoothed Objective with Barrier Function

To efficiently and accurately solve the convex constrained optimization problem (3.1), we adopt the *Interior-Point Method* (IPM) (Wright et al., 1999) that leads to a sequence

of relaxed problems with decreasing $\mu > 0$ to reliably converge to a solution of the exact LCS $f_{\mu=0}$.

We show that the softened LCS is the optimality condition of a barrier-smoothed objective with the following lemma. We defer all the proofs 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

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

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 indicates that the softened LCS is in fact the perturbed Karush–Kuhn–Tucker (KKT) conditions of (3.2), where the perturbation corresponds to smoothing the objective with barrier functions. By replacing the hard contact constraints in the LCS $f_{\mu=0}$, the logarithmic barrier functions in (3.2) discourages the solution to reach the boundary of the polytope constructed by the hard constraints. Therefore, μ is a centering parameter as it restrains the solution within the analytic center of the constraint polytope.

The barrier terms can be thought of as the potential of a force field whose strength is inversely proportional to the distance to the constraint boundary (Boyd et al., 2004; Pang et al., 2022). When applying IPM with a sequence of centering parameters, the intermediate problems with $\mu > 0$ achieve a smoothing effect similar to the “force-at-a-distance” relaxation of the complementarity constraints. In other words, μ controls both the *stiffness* and the *accuracy* of the softened LCS model f_μ . In the following section, we show that both of them are determining factors for the quality of first-order gradient estimation and the convergence of the policy gradient algorithm.

4. Model-Based First-Order Policy Gradient

In this section, we first provide a general framework of model-based First-Order Policy Gradient (FOPG). Then we establish the convergence of model-based FOPG and study the relationship between its convergence rate and the gradient bias, variance. Based on our analysis on the gradient variance and the stiffness of the complementarity-based models, we find that non-smooth local behaviors at contact points can lead to optimization difficulties, which motivates us to analytically smooth the system.

4.1. Framework

The pseudocode of model-based FOPG is presented in Algorithm 1, where two update procedures are performed iteratively. Namely, the model and policy are updated in every iteration $n \in [N]$, which give us sequences of $\{f_{\psi_n}\}_{n \in [N]}$ and $\{\pi_{\theta_n}\}_{n \in [N]}$, respectively.

Algorithm 1 Model-Based First-Order Policy Gradient

Input: Number of iterations N , transition data set $\mathcal{D} = \emptyset$
 1: **for** iteration $n \in [N]$ **do**
 2: Update the model parameter ψ_n by minimizing (4.1)
 3: Update the policy parameter θ_n by (4.3), where $f(x, u)$ is returned by the IPM solver (App. Alg. 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 model is learned from real-world data $\mathcal{D} = \{(x_t^*, u_t^*, x_{t+1}^*)\}_{t=1}^T$. For rigid-body systems that experience hard contact, we learn a physically grounded model $x_{t+1} = f(x_t, u_t; \psi)$ where the state $x_t \in \mathbb{R}^{d_x}$ is the system's configuration (including velocity v_t , coordinate q_t , etc.), and f returns the solution of (2.3) constrained by (2.4), (2.5). Instead of parameterized by a black-box neural network, the ψ contains all *estimated* physics parameters in (2.3), (2.4), (2.5), e.g., the inertia matrix M , Jacobian matrix J , friction coefficient α , and signed-distance function ϕ . The model training loss is given by

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

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

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

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

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

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

4.2. Convergence of Model-Based FOPG

To begin with, we impose a common regularity condition on the policy functions following previous works (Xu et al.,

2019; Pirotta et al., 2015; Zhang et al., 2020; Agarwal et al., 2021). The assumption below essentially ensures the smoothness of the objective $\mathcal{J}(\pi_\theta)$, which is required by most existing analyses of policy gradient methods (Wang et al., 2019; Bastani, 2020; Agarwal et al., 2020).

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

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

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

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

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

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

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

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

4.3. Gradient Variance and LCS Stiffness

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

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

$$v_n \leq O\left(H^4 \tilde{L}_f^{4H} \tilde{L}_\pi^{4H} / M\right). \quad (4.4)$$

We observe that the variance upper bound has polynomial dependence on the Lipschitz of the model and policy, where the degrees are linear in the effective horizon. This makes intuitive sense as the system is chaotic: The stochasticity during training can lead to diverging trajectories and stochastic gradient directions, causing large gradient variance. The

optimization difficulties imposed by non-smooth models, e.g. the hard contact models, result in slow convergence or training failure even in simple toy tasks (Parmas et al., 2018; Suh et al., 2022a).

The above analysis holds for general model-based FOPG. When adopting the complementarity-based contact model f , studying its stiffness, i.e. the Lipschitz L_f , is especially important since they are inherently non-smooth or discontinuous at local mode-switching points. We characterize the stiffness of the softened LCS with the following theorem.

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

$$L_f \leq (\|A\|_F + \|B\|_F) + d_\lambda^2 \|C\|_F (\|D\|_F + \|E\|_F) \cdot l(\mu),$$

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

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

Theorem 4.5 indicates that the model stiffness is largely determined by the centering parameter μ : The upper bound of L_f (and thus of the variance (4.4)) is at least inversely proportional to μ . This is problematic when performing first-order policy gradient based on the softened LCS model f_μ : The accurate dynamics is obtained when solving the exact LCP ($\mu \rightarrow 0$), which, however, causes the gradient variance to explode since $l(\mu) \rightarrow \infty$. The optimization challenges, e.g. chaotic and non-smooth landscapes, are posed even when contact occurs occasionally in a full model unroll.

5. Contact-Aware Analytic Barrier Smoothing

5.1. Method

A natural idea to alleviate the exploding FOPG variance issue is to prevent μ from reaching 0 by setting a positive threshold to terminate the IPM iterations. The solutions correspond to trajectories that do *not* obey the physics laws. According to Lemma 3.3, maintaining a positive μ is equivalent to analytically smoothing the complementarity problem with log-barrier functions. For this reason, we call this vanilla method *analytic barrier smoothing*.

Unfortunately, analytic smoothing with a constant μ can lead to large gradient bias. Therefore, to achieve a good convergence in Thm. 4.2, additional care must be taken to trade-off between the variance and bias.

In this work, we propose the *contact-aware* analytic barrier-smoothed FOPG: When calculating the gradient in (4.2), we differentiate through $f_{\mu(x_t, u_t)}$ — the intermediate solution of the IPM that corresponds to an *adaptive* $\mu(x_t, u_t) > 0$,

whose value scales *inversely* with the minimum distance-to-obstacle $|\phi(x_t, u_t)|$ of the *inactive impact contacts*.

This contact-aware design is based on the observation that the stiffness of complementarity-based models mainly comes from the sudden change of the impact force when the penetration first arises. For example, in systems of Fig. 1(a) and 1(b), the velocity is continuous everywhere except at the hard contact point $z = 0$. For system states that contain the velocity information, the dynamics are stiff around $z = 0$ (see Sec. 7.1 for details). Therefore, when performing FOPG, we only need *local* smoothing near the impact contact to avoid large variance, while maintaining *globally* accurate gradients for a small overall bias.

The choice of $\mu(x_t, u_t)$ can be problem-dependent, as long as it decreases with $|\phi(x_t, u_t)|$ (e.g., we use (??) in our experiments). In the following section, we show that analytic barrier smoothing has close relationship with randomized smoothing and, when the contact-aware $\mu(x_t, u_t)$ takes certain forms, enjoys small gradient bias.

5.2. Analysis

Since we are interested in controlling the system stiffness and large gradient variance brought by the impact contact, we study frictionless systems with a single contact point in this section. This simplifies our analysis by reducing d_λ to 1. Although the results might generalize to broader settings, their forms are beyond the scope of this paper.

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

Proposition 5.1 (Equivalence with Randomized Smoothing). Denote $z_t := Dx_t + Eu_t + d \in \mathbb{R}$. Recall that the solution of the exact LCP is $S_{\mu=0}(z_t)$ and the analytically smoothed LCP solution is $S_{\mu(z_t)}(z_t)$ (c.f. Defn. 3.1). For any centering function $\mu(z_t)$, analytic 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 the analytic barrier smoothing inherently smooths the contact impulse λ_t (with respect to z_t), and thus smoothing the dynamics $x_{t+1} = f_\mu(x_t, u_t)$ since x_t, u_t are prefixed. More importantly, by choosing a proper (adaptive) centering parameter $\mu(z_t)$, the proposed method can cover any randomized smoothing method while avoiding its drawbacks when calculating first-order gradients, which we will discuss in more detail.

Benefited from Proposition 5.1, we can work directly on the randomization-smoothed model when studying the bias of analytic smoothing. This gives us the following results.

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

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

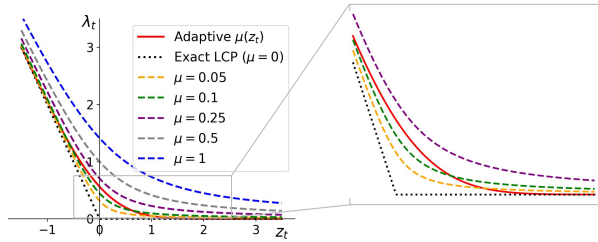
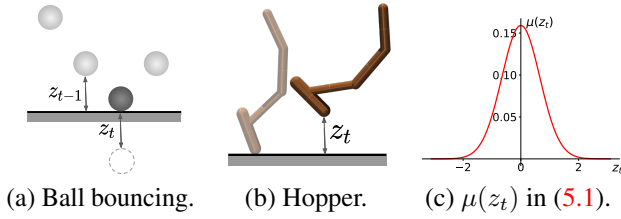
and $c_1, c_2 \in \mathbb{R}$ are tunable constants. Consider the problem of regressing the exact LCP solution $S_{\mu=0}$ with parameters (K, W) such that the residual around z_t distributed according to Gaussian is minimized, formally:

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

The solution K^*, W^* that achieves the minimum is the analytically smoothed surrogate and its gradient:

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

The above proposition shows that analytic smoothing is the best linear approximation of the LCP solution around z_t . Therefore, with a small approximation error, we can conclude the model gradient bias of analytic barrier smoothing.



(d) Contact impulses λ_t of the exact and the softened LCP, including constant μ and the 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 proposed $\mu(z_t)$ is contact-aware and has better balance of stiffness smoothing and bias reduction: Around the contact point $z = 0$, adaptive $\mu(z_t)$ gives smoother dynamics (compared to $\mu \leq 0.1$) and meanwhile best approximates the exact LCP at contactless regions $z > 0$.

We provide the example systems and the plots of $\mu(z_t), \lambda_t$ in Figure 1. We observe that μ that adapts with z_t is contact-aware: μ is positive only when around the contact point around 0. This contact-aware design supports our intuition:

When the body is away from contact, we can safely solve the exact LCP and get accurate simulation; When experiencing contact, the proposed method smooths the LCP to obtain non-stiff local dynamics and small gradient variance.

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

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

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

In Theorem 5.3, we bound the bias of gradient when analytically smoothing the model with an adaptive centering parameter $\mu(z_t)$. Therefore, when the model parameters ψ are accurately fitted with supervised learning, i.e. $f_{\mu=0} \approx f^*$, then the softened LCS $f_{\mu(z_t)}$ and its gradient achieves the best linearization error and has small gradient bias b_n .

Discussion on Randomized Smoothing. Although equivalence can be established between the analytically smoothed system $f_{\mu>0}(x)$ and the randomized smoothed surrogate $\mathbb{E}_{w \sim \rho}[f_{\mu=0}(x + w)]$, applying randomized smoothing in FOPG suffers from the empirical bias (Suh et al., 2022b;a) and the noisy gradients (Howell et al., 2022).

Randomized smoothing estimates the FOPG by averaging the stochastic gradients of multiple noise-induced samples, each of which is calculated by differentiating through $f_{\mu=0}$. The empirical bias phenomenon happens under discontinuities or stiffness. Consider the Heaviside step dynamics when experiencing contact (e.g. pushing an object under friction), differentiating through which will give zero gradient. Therefore, the randomized smoothed FOPG is also zero, causing large bias. Even if the system is non-stiff, sampling and averaging the stochastic gradients is noisy and computationally expensive. In contrast, analytic smoothing by softening the complementarity constraints and solving the program, directly differentiates through the smoothed system $f_{\mu>0}$ and prevents the above issues.

6. Related Work

Differentiable Simulation. The physics-guided (Jiang et al., 2018; Pizzuto & Mistry, 2021) complementarity-based model that is used in this work is adopted in various differentiable hard-contact engines, such as Dojo (Howell et al., 2022), DART (Werling et al., 2021), and Bullet (Heiden et al., 2021). The (sub)gradients of simulation outcomes

w.r.t. control actions are readily available in these differentiable simulators. However, the extreme curvatures of contact events prevent the (sub-)gradients to be effective when performing FOPG, and our method serves as a potential solution. On the other hand, simulators such as MuJoCo (Todorov et al., 2012) and PhysX, implement soft contacts and can generate physics-violated behaviors (Howell et al., 2022). Besides, their non-differentiable nature requires expensive finite-difference to obtain the gradients.

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

7. Experiments

7.1. Contact Behaviors and System Stiffness

We first provide the insight behind the proposed analytic smoothing method. In Figure 2, we plot the dynamics and derivatives of the velocity w.r.t. the coordinate in the simple ball bouncing example, where the ball is thrown with an initial horizontal velocity, subjected to the gravity and impact force (when contacting the ground). We note that similar contact behaviors can happen in almost all robotics tasks.

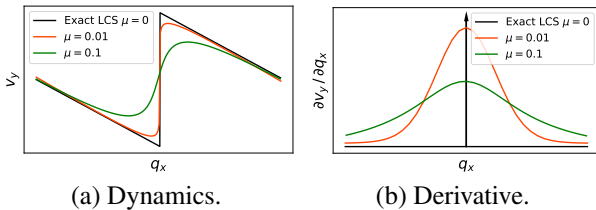


Figure 2. Contact behaviors in the Fig. 1(a) ball bouncing example. 2(a): The vertical velocity v_y w.r.t. the x -coordinate q_x in the exact LCS and in the μ -smoothed system. 2(b): Derivative of v_y w.r.t. q_x . The black arrow represents the impulse function, i.e. $\partial v_y / \partial q_x = \infty$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

At the contact point, the velocity is discontinuous due to the sudden change of contact impulse γ from 0 to a positive value. This results in the stiffness of the system $x' = f_{\mu=0}(x)$ where the state $x = (q_x, v_y)$. By analytically smoothing the complementarity system, the dynamics becomes less stiff with larger centering parameter μ .

7.2. Dependencies of the Gradient Variance

We 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. We plot the mean gradient variance during the FOPG training in two locomotion tasks in Figure 3.

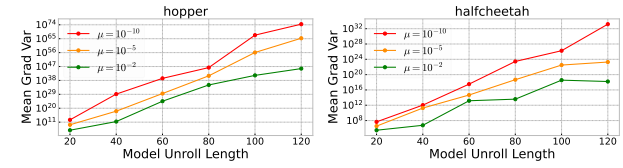


Figure 3. The mean gradient variance with different model unroll lengths when changing the μ value.

We observe that the gradient variance of FOPG can explode with exponential order w.r.t. the horizon or model unroll length. When μ increases, which indicates larger Lipschitz constant of the complementarity-based model, the variance decreases. This supports our result in Theorem 4.4.

References

- Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.
- Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- Amos, B., Stanton, S., Yarats, D., and Wilson, A. G. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, pp. 6–20. PMLR, 2021.
- Anitescu, M. and Potra, F. A. A time-stepping method for stiff multibody dynamics with contact and friction. *International journal for numerical methods in engineering*, 55(7):753–784, 2002.
- Aydinoglu, A., Preciado, V. M., and Posa, M. Contact-aware controller design for complementarity systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1525–1531. IEEE, 2020.

- Aydinoglu, A., Sieg, P., Preciado, V. M., and Posa, M. Stabilization of complementarity systems via contact-aware controllers. *IEEE Transactions on Robotics*, 2021.
- Bastani, O. Sample complexity of estimating the policy gradient for nearly deterministic dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 3858–3869. PMLR, 2020.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.
- Boltt, 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.
- Degrís, T., White, M., and Sutton, R. S. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.
- Drumwright, E. and Shell, D. A. Extensive analysis of linear complementarity problem (lcp) solver performance on randomly generated rigid body contact problems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5034–5039. IEEE, 2012.
- Freeman, C. D., Frey, E., Raichuk, A., Girgin, S., Mordatch, I., and Bachem, O. Brax—a differentiable physics engine for large scale rigid body simulation. *arXiv preprint arXiv:2106.13281*, 2021.
- Geilinger, M., Hahn, D., Zehnder, J., Bäcker, 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. Neursim: 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.
- 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.
- Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Koza-kowski, P., Levine, S., et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.
- Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.
- Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.
- Li, C., Wang, Y., Chen, W., Liu, Y., Ma, Z.-M., and Liu, T.-Y. Gradient information matters in policy optimization by back-propagating through model. In *International Conference on Learning Representations*, 2021.
- Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.
- 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. Pippis: 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.
- 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., Simchowit, 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.
- Werling, K., Omens, D., Lee, J., Exarchos, I., and Liu, C. K. Fast and feature-complete differentiable physics engine for articulated rigid bodies with contact constraints. In *Robotics: Science and Systems*, 2021.
- Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.
- Wright, S., Nocedal, J., et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.
- Xu, J., Makoviychuk, V., Narang, Y., Ramos, F., Matusik, W., Garg, A., and Macklin, M. Accelerated policy learning with parallel differentiable simulation. *arXiv preprint arXiv:2204.07137*, 2022.
- Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.
- Zhang, K., Koppel, A., Zhu, H., and Basar, T. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

A. Proofs

A.1. Proof of Lemma 3.3

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

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

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

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

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

$$Dx_t + Eu_t + F\lambda_t + d = \epsilon_t, \quad (\text{A.3})$$

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

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

A.2. Proof of Theorem 4.2

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

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

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

$$\nabla_\theta \mathcal{J}(\pi_{\theta_n}) = \left(\nabla_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}[\hat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})] \right) - \left(\hat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) - \mathbb{E}[\hat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})] \right) + \hat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}).$$

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

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

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

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

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

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

By rearranging terms,

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

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

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

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

$$\|y + z\|_2^2 \leq \|y\|_2^2 + \|z\|_2^2 + 2\|y\|_2 \cdot \|z\|_2 \leq 2\|y\|_2^2 + 2\|z\|_2^2. \quad (\text{A.8})$$

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

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

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

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.7), we further have

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

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

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

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

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

A.3. Proof of Theorem 4.4

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

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

We start with the case when the sample size $M = 1$, which can naturally generalize to $N > 1$. Specifically, consider an *arbitrary* trajectory obtained by unrolling the model under policy π_{θ_n} . Denote the pathwise gradient $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n})$ of this trajectory as g' . Then we have

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

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

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

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

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

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

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

$$\frac{du_t}{d\theta} = \frac{\partial u_t}{\partial x_t} \cdot \frac{dx_t}{d\theta} + \frac{\partial u_t}{\partial \theta}, \quad (\text{A.11})$$

$$\frac{dx_t}{d\theta} = \frac{\partial x_t}{\partial x_{t-1}} \cdot \frac{dx_{t-1}}{d\theta} + \frac{\partial x_t}{\partial u_{t-1}} \cdot \frac{du_{t-1}}{d\theta}. \quad (\text{A.12})$$

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

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

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

$$\left\| \frac{dx_t}{d\theta} \right\|_2 \leq L_f \tilde{L}_\pi \cdot \left\| \frac{dx_{t-1}}{d\theta} \right\|_2 + L_f L_\theta.$$

Applying the above recursion gives us

$$\left\| \frac{dx_t}{d\theta} \right\|_2 \leq L_f L_\theta \cdot \sum_{j=0}^{t-1} L_f^j \tilde{L}_\pi^j \leq i \cdot L_\theta L_f^{t+1} \tilde{L}_\pi^t, \quad (\text{A.14})$$

where the first inequality follows from the induction

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

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

$$\left\| \frac{du_t}{d\theta} \right\|_2 \leq L_\pi \cdot \left\| \frac{dx_t}{d\theta} \right\|_2 + L_\theta \leq t \cdot L_\theta L_f^{t+1} \tilde{L}_\pi^{t+1} + L_\theta. \quad (\text{A.16})$$

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

$$\left\| \frac{dy_t}{d\theta} \right\|_2 = \left\| \frac{dx_t}{d\theta} \right\|_2 + \left\| \frac{du_t}{d\theta} \right\|_2 \leq K(t) := 2t \cdot L_\theta L_f^{t+1} \tilde{L}_\pi^{t+1} + L_\theta, \quad (\text{A.17})$$

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

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

$$\begin{aligned} & \mathbb{E}_{\bar{y}_t} \left[\left\| \nabla_\theta r(y_t) - \nabla_\theta r(\bar{y}_t) \right\|_2 \right] \\ &= \mathbb{E}_{\bar{y}_t} \left[\left\| \nabla r(y_t) \nabla_\theta y_t - \nabla r(\bar{y}_t) \nabla_\theta \bar{y}_t \right\|_2 \right] \\ &\leq \mathbb{E}_{\bar{y}_t} \left[\left\| \nabla r(y_t) \nabla_\theta y_t - \nabla r(y_t) \nabla_\theta \bar{y}_t \right\|_2 \right] + \mathbb{E} \left[\left\| \nabla r(y_t) \nabla_\theta \bar{y}_t - \nabla r(\bar{y}_t) \nabla_\theta \bar{y}_t \right\|_2 \right] \\ &\leq L_r \cdot \left(\mathbb{E}_{\bar{x}_n} \left[\left\| \frac{dx_t}{d\theta} - \frac{d\bar{x}_t}{d\theta} \right\|_2 \right] + \mathbb{E}_{\bar{u}_n} \left[\left\| \frac{du_t}{d\theta} - \frac{d\bar{u}_t}{d\theta} \right\|_2 \right] \right) + 2L_r \cdot K(t), \end{aligned} \quad (\text{A.18})$$

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

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

$$\begin{aligned} v_n &\leq L_r \cdot \left(\sum_{t=0}^{H-1} \left(\mathbb{E}_{\bar{x}_t} \left[\left\| \frac{dx_t}{d\theta} - \frac{d\bar{x}_t}{d\theta} \right\|_2 \right] + \mathbb{E}_{\bar{u}_t} \left[\left\| \frac{du_t}{d\theta} - \frac{d\bar{u}_t}{d\theta} \right\|_2 \right] + 2K(t) \right) \right)^2 \\ &\leq O \left(\left(\sum_{t=0}^{H-1} t^2 \tilde{L}_f^{2t} \tilde{L}_\pi^{2t} \right)^2 \right) = O \left(H^4 \tilde{L}_f^{4H} \tilde{L}_\pi^{4H} \right), \end{aligned} \quad (\text{A.19})$$

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

Lemma A.1. Denote $e := \sup \mathbb{E}_{\bar{x}_0} [\|dx_0/d\theta - d\bar{x}_0/d\theta\|_2]$, which is a constant that only depends on the initial state distribution¹. For any timestep $t \geq 1$ and the corresponding state x_t , control input u_t , we have the following inequality results:

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

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

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

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

According to the triangle inequality, we continue with

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

Similarly, we have from (A.11) that

$$\begin{aligned}
 &\mathbb{E}_{\bar{u}_n} \left[\left\| \frac{du_t}{d\theta} - \frac{d\bar{u}_t}{d\theta} \right\|_2 \right] \\
 &= \mathbb{E} \left[\left\| \frac{\partial u_t}{\partial x_t} \cdot \frac{dx_t}{d\theta} + \frac{\partial u_t}{\partial \theta} - \frac{\partial \bar{u}_t}{\partial \bar{x}_t} \cdot \frac{d\bar{x}_t}{d\theta} - \frac{\partial \bar{u}_t}{\partial \theta} \right\|_2 \right] \\
 &\leq \mathbb{E} \left[\left\| \frac{\partial u_t}{\partial x_t} \cdot \frac{dx_t}{d\theta} - \frac{\partial \bar{u}_t}{\partial \bar{x}_t} \cdot \frac{d\bar{x}_t}{d\theta} \right\|_2 \right] + \mathbb{E} \left[\left\| \frac{\partial \bar{u}_t}{\partial \bar{x}_t} \cdot \frac{dx_t}{d\theta} - \frac{\partial \bar{u}_t}{\partial \bar{x}_t} \cdot \frac{d\bar{x}_t}{d\theta} \right\|_2 \right] + \mathbb{E} \left[\left\| \frac{\partial u_t}{\partial \theta} - \frac{\partial \bar{u}_t}{\partial \theta} \right\|_2 \right] \\
 &\leq 2L_\pi \cdot \mathbb{E} \left[\left\| \frac{dx_t}{d\theta} \right\|_2 \right] + L_\pi \cdot \mathbb{E} \left[\left\| \frac{dx_t}{d\theta} - \frac{d\bar{x}_t}{d\theta} \right\|_2 \right] + 2L_\theta. \tag{A.21}
 \end{aligned}$$

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

$$\begin{aligned}
 &\mathbb{E}_{\bar{x}_t} \left[\left\| \frac{dx_t}{d\theta} - \frac{d\bar{x}_t}{d\theta} \right\|_2 \right] \\
 &\lesssim 4L_f \tilde{L}_\pi \cdot \left(\left\| \frac{dx_{t-1}}{d\theta} \right\|_2 + \left\| \frac{du_{t-1}}{d\theta} \right\|_2 \right) + L_f \tilde{L}_\pi \cdot \mathbb{E}_{\bar{x}_{t-1}} \left[\left\| \frac{dx_{t-1}}{d\theta} - \frac{d\bar{x}_{t-1}}{d\theta} \right\|_2 \right] + 2L_f L_\theta \\
 &\leq 4L_f \tilde{L}_\pi \cdot K(t-1) + L_f \tilde{L}_\pi \cdot \mathbb{E}_{\bar{x}_{t-1}} \left[\left\| \frac{dx_{t-1}}{d\theta} - \frac{d\bar{x}_{t-1}}{d\theta} \right\|_2 \right] + 2L_f L_\theta,
 \end{aligned}$$

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

Applying this recursion gives us

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

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

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

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

This concludes the proof. \square

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.2 (Implicit Function Theorem). An implicit function $g : \mathbb{R}^{d_s} \times \mathbb{R}^{d_w} \rightarrow \mathbb{R}^{d_s}$ 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

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

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

$$\frac{dg}{dw} = \frac{\partial g}{\partial w} + \frac{\partial g}{\partial s} \frac{\partial s}{\partial w}.$$

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

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

□

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

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

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

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

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

By the Implicit Function Theorem A.2, we have

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

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

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

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

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

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

By the definition of Frobenius norm, we know that

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

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

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

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

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

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

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

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

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

By the Cauchy-Schwartz inequality we have

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

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

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

Solving (A.29) gives

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

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

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

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

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

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

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

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

A.5. Proof of Proposition 5.1

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

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

By rewriting the above function 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:

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

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

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

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

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

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

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

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

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

This concludes the proof. \square

A.6. Proof of Proposition 5.2

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

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

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

The solution is the linearization of the smoothed surrogate:

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

Proof. The proof is originally provided in (Pang et al., 2022). We adapt it here for completeness.

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

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

Therefore, we obtain the solution

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

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

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

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

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

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

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

The solution of the above equation is given by

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

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

A.7. Proof of Theorem 5.3

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

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

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

We define the linearization residual at point $z_t + w$ as

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

Then we have from (A.37) that

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

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

$$|\nabla_z S_{\mu(z_t)}(z_t) - \nabla_z S_{\mu=0}(z_t)| \leq \frac{F^2|w|}{2} + \frac{\nu(w) + \varsigma}{|w|}. \quad (\text{A.38})$$

From Proposition 5.2, we know that

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

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

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

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

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

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

□

B. IPM Solver

TODO Algorithm 2.

C. Details of Experiments

C.1. Dynamics in the Ball Bouncing Example

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

Without loss of generality, we assume that the discretization timestep size, the mass of the ball, and its initial velocity v_0 are all 1. Denote the initial vertical coordinate of the ball is q_0 . Then the distance of the ball to the ground is given by $q_0 - \int \int (g - \gamma_t) dt dt$, 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

$$\gamma_t \left(q_0 - \int \int (g - \gamma_t) dt dt \right) = \mu.$$

This can be rewrite as

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

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

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

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