Model-Based First-Order Policy Gradient for Contact Dynamics

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

1. Background

1.1. 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 $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$, respectively. When taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, the agent receives reward $r(s, a)$ and the MDP transitions to a new state according to probability $s' \sim f^*(\cdot \,|\, s, a)$.

We are interested in controlling the system by finding a policy π_θ that maximizes the expected cumulative reward. For $j \in \{0, \cdots, H-1\}$, denote the state and state-action value function associated with π by $V_j^\pi : \mathcal{S} \to \mathbb{R}$ and $Q_j^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, respectively, which are defined as

$$V_j^\pi(s) = \mathbb{E}_{\pi, f^*}\bigg[\sum_{i=j}^{H-1} r(s_i, a_i) \,\bigg|\, s_j = s\bigg],$$

$$Q_j^\pi(s, a) = \mathbb{E}_{\pi, f^*}\bigg[\sum_{i=j}^{H-1} r(s_i, a_i) \,\bigg|\, s_j = s, a_j = a\bigg],$$

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

We denote by ζ the initial state distribution. The objective is

$$\mathcal{J}(\pi) = \mathbb{E}_{s_0 \sim \zeta}\big[V_0^\pi(s_0)\big] = \mathbb{E}_{p_\pi(\alpha)}\big[\sum_{i=0}^{H-1} r(s, a)\big], \quad (1.1)$$

where $p_\pi(\alpha)$ is the distribution over rollouts $\alpha := ((s_0, a_0), \cdots, (s_{H-1}, a_{H-1}))$ when executing π, formally, $s_0 \sim \zeta(\cdot)$, $a_i \sim \pi(\cdot \,|\, s_i)$, and $s_{i+1} \sim f^*(\cdot \,|\, s_i, a_i)$.

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 $\nabla_\theta \mathbb{E}_{p(x;\theta)}[y(x)]$. In RL, we set $p(x; \theta)$ as the trajectory distribution conditioned on policy parameter θ, and $y(x)$ as the cumulative reward.

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

In the sequel, we introduce two commonly used gradient estimators in RL.

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

$$\begin{aligned}\nabla_\theta \mathbb{E}_{p(x;\theta)}\big[y(x)\big] &= \int y(x) \nabla_\theta p(x; \theta) \mathrm{d}x \\ &= \mathbb{E}_{p(x;\theta)}\big[y(x) \nabla_\theta \log p(x; \theta)\big]. \quad (1.2)\end{aligned}$$

ReParameterization (RP) Gradient: RP gradient benefits from the structural characteristics of the objective, i.e., how the overall objective is affected by the operations applied to the sources of randomness as they pass through the measure and into the cost function (Mohamed et al., 2020). From the simulation property of continuous distribution, we have the following equivalence between direct and indirect ways of drawing samples:

$$\widehat{x} \sim p(x; \theta) \;\equiv\; \widehat{x} = g(\epsilon; \theta), \quad \epsilon \sim p. \quad (1.3)$$

Derived from the *law of the unconscious statistician* (LOTUS) (Grimmett & Stirzaker, 2020), i.e., $\mathbb{E}_{p(x;\theta)}[y(x)] = \mathbb{E}_{p(\epsilon)}[y(g(\epsilon; \theta))]$, the RP gradient can be expressed as

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

1.2. 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

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

where we let q denote the generalized coordinates, v the generalized velocities, $u \in \mathbb{R}^{n_u}$ the applied control force, $M(q)$ 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 impulse, where γ and β are the normal impact forces and friction forces, respectively, and c denotes the number of contact points.

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

$$v_{t+1} = v_t + M_t^{-1}(n_t + u_t)h + M_t^{-1}J_t^\top \lambda_t,$$
$$q_{t+1} = q_t + hv_{t+1} \tag{1.4}$$

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:

$$\gamma_{t+1} \circ \phi(q_{t+1}) = \vec{0}, \qquad \gamma_{t+1}, \phi(q_{t+1}) \geq \vec{0}, \tag{1.5}$$

where $\circ$ is the element-wise (Hadamard) product, ϕ is the signed-distance function, and $\vec{0}$ is the zero vector. The intuition behind (1.5) is that the magnitude of the normal forces must be non-negative and can only be non-zero if there is a contact to maintain non-negative gaps (non-penetration).

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

$$\beta_{t+1} \circ \eta_{t+1} = \vec{0}, \qquad \beta_{t+1}, \eta_{t+1} \geq \vec{0},$$
$$B(q_{t+1})v_{t+1} + \omega_{t+1}\vec{1} - \eta_{t+1} = \vec{0},$$
$$\omega_{t+1} \cdot (\alpha\gamma_{t+1} - \beta_{t+1}) = \vec{0}, \tag{1.6}$$

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

2. Complementarity-Based Contact Models

In MBRL, 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 (1.4) constrained by (1.5), (1.6). Instead of parameterized by a black-box neural network, the ψ contains all *estimated* physics parameters in (1.4), (1.5), (1.6), e.g., the inertia matrix M, Jacobian matrix J, friction coefficient α, and signed-distance function ϕ. The model training loss is given by

$$L(\psi; \mathcal{D}) = \sum_{t=1}^{T} \frac{1}{2}\left\| f(x_t^*, u_t^*; \psi) - x_{t+1}^*\right\|_2^2. \tag{2.1}$$

2.1. Linear Complementarity Systems

The dynamic (1.4) describes a hybrid system where different modes are controlled by the contact force λ under the complementarity constraints (1.5), (1.6). To simplify our analysis, we study the more abstracted Linear Complementarity Systems (LCS), which effectively capture the local behaviors of the transition and are widespread in the robotics community (Aydinoglu et al., 2021; Tassa & Todorov, 2010; Drumwright & Shell, 2012).

We define the LCS model f_μ as a generalization of the original LCS $f_{\mu=0}$, which is a special case when $\mu = 0$.

Definition 2.1 (LCS Model)**.** A model $x_{t+1} = f_\mu(x_t, u_t)$ is an LCS model if the evolution of state $x \in \mathbb{R}^{d_x}$ is governed by a linear dynamics and a softened LC problem (LCP):

$$x_{t+1} = Ax_t + Bu_t + C\lambda_t + c,$$
$$\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu\vec{1},$$
$$\lambda_t \geq \vec{0}, \quad Dx_t + Eu_t + F\lambda_t + d \geq \vec{0}, \tag{2.2}$$

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 the scalar $\mu \geq 0$. Denote $\lambda_t \in \mathbb{R}^{d_\lambda}$ returned by the solver S_μ of the softened LCP (the last two lines of (2.2)) as $\lambda_t = S_\mu(Dx_t + Eu_t + d)$.

In simulation, $\mu = 0$ corresponds to the exact LCP where the system $f_{\mu=0}$ resembles the reality. Obviously, solving for the contact space impulse λ_t is the 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 2.2 (P-Matrix)**.** Assume F in the LCS (2.2) is a P-matrix, defined as a matrix whose principal minors are all positive, i.e., the determinants of its principal submatrices $\det(F_{\alpha\alpha}) > 0, \forall \alpha \subseteq \{1, \cdots, d_\lambda\}$.

Assumption 2.2 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).

2.2. Interior-Point Method

To efficiently and accurately solve the convex constrained optimization problem (2.2), we adopt the interior-point method (IPM) that leads to a sequence of relaxed problems by choosing a positive $\mu > 0$ that converges to zero to reliably converge to a solution of the original LCS ($\mu = 0$).

Lemma 2.3 (Primal Problem with Log-Barrier Function)**.** The softened LCS model (2.2) with $\mu \geq 0$ is the first-order optimality condition of the following program

$$\min_{\lambda_t, \epsilon_t} \quad \lambda_t^\top \epsilon_t - \mu \sum_{i=1}^{d_\lambda} \left(\log \lambda_t^{(i)} + \log \epsilon_t^{(i)}\right)$$
$$\text{s.t.} \quad Dx_t + Eu_t + F\lambda_t + d = \epsilon_t,$$
$$Ax_t + Bu_t + C\lambda_t + c = x_{t+1}, \tag{2.3}$$

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

Lemma 2.3 indicates that the system (2.2) is in fact the solution of the smoothed objective function (2.3). Instead of the hard contact constraints in LCS, the logarithmic barrier function in (2.3) discourages the contact impulse to get closer to the boundary. The intermediate solution at $\mu > 0$ corresponds to a force field whose strength is inversely proportional to the distance to the constraint boundary. In other words, μ controls the *stiffness* of the dynamical system, which is a major determining factor that affects the quality of the first-order gradient estimation and the policy gradient algorithm convergence. We now make the formal analysis.

3. First-Order Policy Gradient

We present our main theoretical results in this section, with the proofs deferred to §A. Specifically, we establish the convergence of model-based First-Order Policy Gradient (FOPG) methods. Moreover, we study the relationship between the convergence rate, gradient bias, variance, and the model stiffness, approximation error.

3.1. Convergence of 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 3.1 (Lipschitz Score Function and Boundedness)**.** Assume that the score function of policy π_θ is Lipschitz continuous and has bounded norm for $(s, a) \in \mathcal{S} \times \mathcal{A}$:

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

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

Theorem 3.2 (Convergence to Stationary Points)**.** Define the gradient bias b_n and variance v_n as

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

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

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

Theorem 3.2 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 3.3 (Lipschitz Continuity)**.** Assume the policy, model, and reward are L_π, L_f, L_r Lipschitz continuous.

3.2. Gradient Variance and LCS Stiffness

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

Proposition 3.4 (Gradient Variance)**.** Under Assumption 3.3, at any iteration $n \in [N]$, the gradient variance of FOPG is bounded by

$$v_n \leq O\Big(H^4 \widetilde{L}_f^{4H} \widetilde{L}_\pi^{4H} / M\Big). \tag{3.1}$$

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).

It's worth noting that the above analysis holds for general model-based FOPG. Studying the model stiffness is especially important when adopting complementarity-based contact models since they are inherently non-smooth or discontinuous at local mode-switching points. We characterize the stiffness of the LCS models with the following proposition.

Proposition 3.5 (LCS Stiffness)**.** 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 2.2, the Lipschitz L_f of the LCS model f_μ defined in (2.2) 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

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

Proposition 3.5 indicates that the stiffness of the LCS model is largely determined by the centering parameter μ: The upper bound of L_f (and thus of the variance (3.1)) is at least inversely proportional to μ. This raises the optimization issue when performing first-order policy gradient: The exact LCP solution λ_t, x_{t+1} is obtained when $\mu \to 0$, which, however, causes the gradient variance exploding ($l(\mu) \to \infty$) even when contact occurs occasionally in a full model unroll.

4. Contact-Aware Analytic Barrier Smoothing

4.1. Method

A natural idea to alleviate the exploding FOPG variance issue is to prevent μ from reaching 0. The solutions correspond to trajectories that do *not* well obey the physics laws. According to Lemma 2.3, setting μ to positive values is equivalent to solving a smoothed constrained optimization problem with log-barrier functions. For this reason, we call this vanilla approach *analytic barrier smoothed* FOPG.

Unfortunately, simply softening the complementarity system with a constant μ can lead to large model simulation error and gradient bias. Therefore, to achieve a good convergence in Thm. 3.2, additional care must be taken to trade-off between the variance and bias.

In this work, we propose a *contact-aware* analytic barrier smoothing method. It is based on the observation that the existence of contact is the fundamental reason for the stiffness of complementarity-based models. Therefore, we only need to smooth the local dynamics at contact points to avoid large variance, while preferring globally accurate simulation for a small overall bias. This translates to setting the centering parameter μ as a function $\mu(x_t, u_t)$ such that it is aware of the position of and distance to the contact points. By doing so, the sequence of μ in the IPM converges to 0 at any timestep that is far from contact, and the μ sequence stops at a positive value when the contact is nearby.

We don't fix the function $\mu(\cdot,\cdot)$ since it should be problem-dependent. However, we will show that our analytic smoothing methods *do* have interesting properties and, when $\mu(\cdot,\cdot)$ takes certain forms, enjoy small gradient bias.

4.2. Analysis

Studying the bias of analytic barrier smoothing requires more fine-grained analysis. In this section, we focus on the frictionless setting, which reduces d_λ to 1 corresponding to the contact impulses. Although the results might generalize to broader settings, their forms are prohibitive for analysis.

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 (but analytically smoothing the complementarity constraints has its own advantage in practice, which we will discuss).

Proposition 4.1 (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. 2.1). 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 contact-aware centering parameter $\mu(z_t)$, the proposed method can cover any randomized smoothing method while avoiding differentiating through the samples, which is known to give wrong gradients under discontinuities or stiff approximations (Suh et al., 2022a).

As a benefit of Proposition 4.1, we can work directly on the randomization-smoothed model when studying the bias of analytic smoothing. This gives us the following results.

Proposition 4.2 (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 contact-aware centering parameter as $\mu(z_t) = \kappa\cdot(z_t + F\kappa)$. Here,

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

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

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

The 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.

Proposition 4.3 (Model Gradient Bias)**.** With the same definition of $\mu(z_t)$ in Proposition 4.2, the gradient of the LCS model $f_{\mu(z_t)}$ approximately matches the gradient of LCS $f_{\mu=0}$, with the bias upper bounded by

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

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

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.

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.

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

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.

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

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.

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.

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.

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

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

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

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.

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

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.

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 2.3

Proof. Corresponding to the constrained optimization problem (2.3) 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 (2.3) since x_{t+1} can be directly calculated when λ_t is obtained.

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

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

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

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

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

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

A.2. Proof of Theorem 3.2

As a preparation before proving Theorem 3.2, we first present the following lemma stating that the objective in (1.1) is Lipschitz smooth under Assumption 3.1.

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

$$L := H^2 r_{\mathrm{m}} \cdot L_1 + 2H^3 r_{\mathrm{m}} \cdot B_\theta^2.$$

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

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. From the policy update rule, we know that $\widehat{\nabla}_\theta \mathcal{J}(\pi_{\theta_n}) = (\theta_{n+1} - \theta_n)/\eta$. By the Lipschitz Assumption 3.3, we have

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

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

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

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

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

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

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

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

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

By rearranging terms,

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

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

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

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

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

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

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

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

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

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

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

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

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

A.3. Proof of Proposition 3.4

In what follows, we interchangeably write $\nabla_a b$ and $\mathrm{d}b/\mathrm{d}a$ as the gradient, and use the notation $\partial b/\partial a$ to denote the partial derivative. With slight abuse of notation, for vectors s and 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

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

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

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

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

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

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

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

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

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

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

By the Cauchy-Schwarz inequality and the Lipschitz Assumption 3.3, we have

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

Applying the above recursion gives us

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

where the first inequality follows from the induction

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

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

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

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

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

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

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

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

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

Plugging (A.18) into (A.10) and (A.9) gives us

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

where the second inequality follows from the results from Lemma A.2 and by plugging the definition of K in (A.17). Since the analysis above considers batch size $M = 1$, the bound of gradient variance v_n is established by dividing M, which concludes the proof. □

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

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

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

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

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

According to the triangle inequality, we continue with

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

Similarly, we have from (A.11) that

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

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

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

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

Applying this recursion gives us

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

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

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

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

This concludes the proof. $\square$

A.4. Proof of Proposition 3.5

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 Proposition 3.5. To begin with, we first study the Jacobian $\partial x_{t+1}/\partial x_t$, and the Jacobian $\partial x_{t+1}/\partial u_t$ can be analyzed using similar techniques.

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

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

We rewrite the contact equation $\lambda_t \circ (Dx_t + Eu_t + F\lambda_t + d) = \mu \vec{1}$ in (2.2) as

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

By the Implicit Function Theorem A.3, we have

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

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

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

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

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

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

By the definition of Frobenius norm, we know that

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

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

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

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

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

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

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.26). The third inequality follows from (A.27) 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 (2.2) as

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

By the Cauchy-Schwartz inequality we have

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

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

$$\|F\|_F \cdot \|\Lambda_t\|_F^2 + e\cdot\|\Lambda_t\|_F - \mu \ge 0. \qquad \text{(A.29)}$$

Solving (A.29) gives

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

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

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

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

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

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

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

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

A.5. Proof of Proposition 4.1

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

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

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 unsmoothed LCP as follows:

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

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

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

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

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

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

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

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

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

This concludes the proof. □

A.6. Proof of TODO

Lemma A.4 (Randomized Smoothing as Linearization Minimizer (Pang et al., 2022))**.** Let $\rho(w)=\mathcal{N}(w;0,\Sigma)$ be a zero-mean Gaussian. Consider the problem of regressing the unsmoothed LCS $f_{\mu=0}$ with parameters (K,W) such that the residual around $\overline{x}$ distributed according to ρ is minimized:

$$K^*,W^*=\operatorname*{argmin}_{K,W}\mathbb{E}_{w\sim\rho(w)}\Big[\big\|f_{\mu=0}(\overline{x}+w)-Ww-K\big\|_2^2\Big].$$

The solution is the linearization of the smoothed surrogate:

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

The above lemma shows that randomized smoothing is the best linear approximation of f around any nominal point $\overline{x}$ distributed according to Gaussian. Therefore,

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

$$\left|\frac{S_{\mu=0}(z_r+w)-S_{\mu=0}(z_t)}{w}-\nabla_z S_{\mu=0}(z_t)\right|\le |w|\cdot\sup\frac{|\nabla_z^2 S_{\mu=0}(z_t)|}{2}=\frac{F^2|w|}{2},\tag{A.33}$$

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

We define the linearization residual at point z_t+w as

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

Then we have from (A.33) that

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

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

$$\big|\nabla_z S_{\mu(z_t)}(z_t)-\nabla_z S_{\mu=0}(z_t)\big|\le\frac{F^2|w|}{2}+\frac{\nu(w)+\varsigma}{|w|}.\tag{A.34}$$

From Proposition 4.2, we know that

$$\mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}[\nu(w)] = \delta. \tag{A.35}$$

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

This can be proved by contradiction: Suppose $\forall w \in [\sigma\mathcal{Q}(2/3), \sigma\mathcal{Q}(3/4)]$, $\nu(w) > 12\delta$. Then the expectation $\mathbb{E}_{w\sim\mathcal{N}(0,\sigma)}[\nu(w)] > (3/4-2/3)\cdot 12\delta = \delta$. This contradicts with (A.35). Therefore, this claim is correct.

Using the above claim, we have from (A.34) that

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

We conclude the proof by applying chain rule in the LCS model (2.2):

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

$\square$