Do Differentiable Simulators Give Better Policy Gradients?

H.J. Terry Suh [1] **Max Simchowitz** [1] **Kaiqing Zhang** [1] **Russ Tedrake** [1]

Abstract

Differentiable simulators promise faster computation time for reinforcement learning by replacing zeroth-order gradient estimates of a stochastic objective with an estimate based on first-order gradients. However, it is yet unclear what factors decide the performance of the two estimators on complex landscapes that involve long-horizon planning and control on physical systems, despite the crucial relevance of this question for the utility of differentiable simulators. We show that characteristics of certain physical systems, such as stiffness or discontinuities, may compromise the efficacy of the first-order estimator, and analyze this phenomenon through the lens of bias and variance. We additionally propose an α-order gradient estimator, with $\alpha \in [0, 1]$, which correctly utilizes exact gradients to combine the efficiency of first-order estimates with the robustness of zero-order methods. We demonstrate the pitfalls of traditional estimators and the advantages of the α-order estimator on some numerical examples.

1. Introduction

Consider the problem of minimizing a *stochastic objective*,

$$\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{w}} f(\boldsymbol{\theta}, \mathbf{w}).$$

At the heart of many algorithms for reinforcement learning (RL) lies *zeroth-order* estimation of the gradient ∇F (Sutton et al., 2000; Schulman et al., 2017). Yet, in domains that deal with structured systems, such as linear control, physical simulation, or robotics, it is possible to obtain *exact* gradients of f, which can also be used to construct a *first-order* estimate of ∇F. The availability of both options begs the question: given access to exact gradients of f, which estimator should we prefer?

In stochastic optimization, the theoretical benefits of using first-order estimates of ∇F over zeroth-order ones have mainly been understood through the lens of variance and convergence rates (Ghadimi & Lan, 2013; Mahamed et al., 2020): the first-order estimator often (*not always*) results in much less variance compared to the zeroth-order one, which leads to faster convergence rates to a local minima of general nonconvex smooth objective functions.

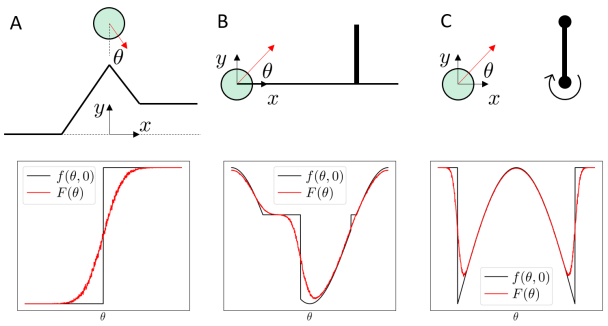

Figure 1. Examples of simple optimization problems on physical systems. Goal is to: A. maximize y position of the ball after dropping. B. maximize distance thrown, with a wall that results in inelastic impact. C. maximize transferred angular momentum to the pivoting bar through collision. Second row: the original objective and the stochastic objective after randomized smoothing.

However, the landscape of RL objectives that involve long-horizon sequential decision making (e.g. policy optimization) is challenging to analyze, and convergence properties in these landscapes are relatively poorly understood, except for structured settings such as finite-state MDPs (Agarwal et al., 2020; Zhang et al., 2020) or linear control (Fazel et al., 2019; Bhandari & Russo, 2020). In particular, physical systems with contact, as we show in Figure 1, can display complex characteristics including nonlinearities, non-smoothness, and discontinuities (van der Schaft & Schumacher, 2000; Mason, 2001; Suh et al., 2021).

Nevertheless, lessons from convergence rate analysis tell us that there may be benefits to using the exact gradients even for these complex physical systems. Such ideas have been championed through the term "differentiable simulation", where forward simulation of physics is programmed in a manner that is consistent with automatic differentiation (Freeman et al., 2021; Hu et al., 2020; Tedrake, 2022; Werling et al., 2021; Geilinger et al., 2020), or computation of analytic derivatives (Carpentier et al., 2019). These methods have shown promising results in decreasing computation

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: H.J.Terry Suh <hjsuh@mit.edu>.

time compared to zeroth-order methods (Huang et al., 2021; Freeman et al., 2021; Gradu et al., 2021; Du et al., 2020; de Avila Belbute-Peres et al., 2018; Mora et al., 2021).

Existing literature in differentiable simulation mainly focuses on the use of exact gradients for *deterministic* optimization. However, (Suh et al., 2021; Le Lidec et al., 2021) show that using exact gradients for a deterministic objective can lead to suboptimal behavior of certain systems due to their landscapes. In these systems, stochasticity can be used to *regularize* the landscapes with randomized smoothing (Duchi et al., 2015). We illustrate how the landscapes change upon injecting noise (Figure 1), and list some benefits of considering a *surrogate* stochastic objective.

- **Stochasticity smooths local minima**. As noted in (Suh et al., 2021; Metz et al., 2021), stochasticity can alleviate some of the high-frequency local minima that deterministic gradients will be stuck on. For instance, the small discontinuity on the right side of Figure 1.B is filtered by Gaussian smoothing.
- **Stochasticity alleviates flat regions**. In systems of Figure 1, the gradients in some of the regions can be completely flat. This stalls progress of gradient descent. The stochastic objective, however, still has non-zero gradient as some samples escape the flat regions and provide an informative direction of improvement.
- **Stochasticity encodes robustness.** In Figure 1.C, following the gradient to increase the transferred momentum causes the ball to miss the pivot and land in a high-cost region. In contrast, the stochastic objective has a local minimum within the safe region, as the samples provide information about missing the pivot.

Thus, our work attempts to compare two versions of gradient estimators in the stochastic setting: the first-order estimator and the zeroth-order one. This setting rules out the case that zeroth-order estimates perform better simply because of stochasticity, and sets equal footing for the two methods to be compared against.

When f is continuous, these quantities both converge to the same quantity (∇F) in expectation. We first show that even with continuous f, the first-order gradient estimate *can* result in more variance than the zeroth-order one due to the *stiffness* of the dynamics or due to compounding of gradients in chaotic systems (Metz et al., 2021).

In addition, we show that the assumption of continuous f can be violated in many relevant physical systems that are nearly/strictly *discontinuous* in the underlying landscape. These discontinuities are commonly caused by contact and geometrical constraints. We provide minimal examples to highlight specific challenges in Figure 1. These are not mere pathologies, but abstractions of more complicated examples that are rich with contact, such as robotic manipulation.

We show that the presence of such discontinuities causes the first-order gradient estimator to be *biased*, while the zeroth-order one still remains unbiased under discontinuities. Furthermore, we show that stiff continuous approximations of discontinuities, even if asymptotically unbiased, can still suffer from what we call *empirical bias* under finite-sample settings. This results in a bias-variance tradeoff between the biased first-order estimator and the *often* high-variance, yet unbiased zeroth-order estimator. Intriguingly, we find that the bias-variance tradeoff in this setting manifests itself not through convergence rates, but through different local minima. This shows that the two estimators may fundamentally operate on different landscapes implicitly.

The presence of discontinuities need not indicate that we need to commit ourselves to uniformly using one of the estimators. Many physical systems are *hybrid* by nature (van der Schaft & Schumacher, 2000); they consist of smooth regions that are separated by manifolds of non-smoothness or discontinuities. This suggests that we may be able to utilize the first-order estimates far away from these manifolds to obtain benefits of convergence rates, while switching to zeroth-order ones in the vicinity of discontinuities to obtain unbiased estimates.

For this purpose, we further attempt to answer the question: how can we then correctly utilize exact gradients of f for variance reduction when we know the objective is nearly discontinuous? Previous works show that the two estimators can be combined by interpolating based on empirical variance (Parmas et al., 2018; Metz et al., 2019; Mahamed et al., 2020). However, we show that in the presence of near-discontinuities, selecting based on empirical variance alone can lead to highly inaccurate estimates of ∇F, and propose a robustness constraint on the accuracy of the interpolated estimate to remedy this effect.

Contributions. In summary, we 1) shed light on some of the inherent problems of RL using differentiable simulators, and answer which gradient estimator can be more useful under different characteristics of underlying systems such as discontinuities, stiffness, and chaos; and 2) present the α-order gradient estimator, a robust interpolation strategy between the two gradient estimators that utilizes exact gradients of the physical system without falling into the identified pitfalls of the previous methods.

We hope both contributions inspire new algorithms for policy optimization using differentiable simulators, as well as design guidelines for simulators to accelerate reinforcement learning for physical systems.

2. Preliminaries

Notation. We denote the expectation of a random vector $\mathbf{z}$ as $\mathbb{E}[\mathbf{z}]$, and its variance as $\mathbf{Var}[\mathbf{z}] := \mathbb{E}[\|\mathbf{z} - \mathbb{E}[\mathbf{z}]\|^2]$. Expectations are defined in almost-sure sense, so that the law of large numbers holds (see Appendix A.1 for details).

Setting. We study a discrete-time, finite-horizon, continuous-state control problem with states $\mathbf{x} \in \mathbb{R}^n$, inputs $\mathbf{u} \in \mathbb{R}^m$, transition function $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, and horizon $H \in \mathbb{N}$. Given a sequence of costs $c_h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, a family of policies $\pi_h(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$, and a sequence of injected noise terms $\mathbf{w}_{1:H} \in (\mathbb{R}^m)^H$, we define the cost-to-go functions

$$V_h(\mathbf{x}_h, \mathbf{w}_{h:H}, \boldsymbol{\theta}) = \textstyle\sum_{h'=h}^{H} c_h(\mathbf{x}_{h'}, \mathbf{u}_{h'}),$$
$$\text{s.t. } \mathbf{x}_{h'+1} = \phi(\mathbf{x}_{h'}, \mathbf{u}_{h'}),\ \ \mathbf{u}_{h'} = \pi(\mathbf{x}_{h'}, \boldsymbol{\theta}) + \mathbf{w}_{h'},\ \ h' \geq h.$$

Our aim is to minimize the policy optimization objective

$$F(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_h \overset{\text{i.i.d.}}{\sim} p} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}), \qquad (1)$$

where ρ is a distribution over initial states $\mathbf{x}_1$, and $\mathbf{w}_1, \ldots, \mathbf{w}_H$ are independent and identically distributed according to some distribution p. In the main text, we make the following assumption on the distributions ρ and p:

Assumption 2.1. We assume that ρ has finite moments, and that $p = \mathcal{N}(0, \sigma^2 I_n)$ for some $\sigma > 0$.

Our rationale for Gaussian p is that we view $\mathbf{w}_{1:H}$ as *smoothing* to regularize the optimization landscape (Duchi et al., 2011; Berahas et al., 2019). To simplify the main text, we take $\mathbf{x}_1$ to be deterministic (ρ is a dirac-delta), with general ρ being addressed in the appendix. Setting $\bar{\mathbf{w}} = \mathbf{w}_{1:H}$, $\bar{p} = \mathcal{N}(0, \sigma^2 I_{nH})$, and $f(\boldsymbol{\theta}, \bar{\mathbf{w}}) = V_1(\mathbf{x}_1, \bar{\mathbf{w}}, \boldsymbol{\theta})$, we can express $F(\boldsymbol{\theta})$ as a *stochastic optimization problem*,

$$F(\boldsymbol{\theta}) := \mathbb{E}_{\bar{\mathbf{w}} \sim \bar{p}} f(\boldsymbol{\theta}, \bar{\mathbf{w}}).$$

Trajectory optimization. Our parametrization also includes open-loop trajectory optimization. Letting the policy parameters be an open-loop sequence of inputs $\boldsymbol{\theta} = \{\boldsymbol{\theta}_h\}_{h=1}^H$ and having no feedback $\pi(\mathbf{x}_h, \boldsymbol{\theta}) = \boldsymbol{\theta}_h$, we optimize over sequence of inputs to be applied to the system.

One-step optimization. We illustrate some key ideas in the open-loop case where $H = 1$: $\pi : \mathbb{R}^d \to \mathbb{R}^m$ is the identity function with $\bar{\mathbf{w}} = \mathbf{w} \in \mathbb{R}^m$, $d = m$ and $c : \mathbb{R}^m \to \mathbb{R}$,

$$F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w} \sim p} f(\boldsymbol{\theta}, \mathbf{w}), \quad f(\boldsymbol{\theta}, \mathbf{w}) = c(\boldsymbol{\theta} + \mathbf{w}). \qquad (2)$$

2.1. Gradient Estimators

In order to minimize $F(\boldsymbol{\theta})$, we consider iterative optimization using stochastic estimators of its gradient $\nabla F(\boldsymbol{\theta})$. We say a function $\psi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ has *polynomial growth* if there exist constants a, b such that, for all $\mathbf{z} \in \mathbb{R}^{d_1}$, $\|\psi(\mathbf{z})\| \leq a(1 + \|\mathbf{z}\|^b)$. The following assumption ensures these gradients are well-defined.

Assumption 2.2. We assume that the policy $\pi(\cdot, \cdot)$ is continuously differentiable everywhere, and the dynamics $\phi(\cdot, \cdot)$, as well as the cost $c_h(\cdot, \cdot)$ have polynomial growth.

Even when the costs or dynamics are *not* differentiable, the expected cost $F(\boldsymbol{\theta})$ is differentiable due to the smoothing $\bar{\mathbf{w}}$. $\nabla F(\boldsymbol{\theta})$ is referred to as the *policy gradient*.

Zeroth-order estimator. The policy gradient can be estimated only using samples of the function values.

Definition 2.3. Given a single zeroth-order estimate of the policy gradient $\hat{\nabla}^{[0]} F_i(\boldsymbol{\theta})$, we define the zeroth-order batched gradient (ZoBG) $\bar{\nabla}^{[0]} F(\boldsymbol{\theta})$ as the sample mean,

$$\hat{\nabla}^{[0]} F_i(\boldsymbol{\theta}) \coloneqq \frac{1}{\sigma^2} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}^i, \boldsymbol{\theta}) \left[\sum_{h=1}^{H} \mathrm{D}_{\boldsymbol{\theta}} \pi(\mathbf{x}_h^i, \boldsymbol{\theta})^\intercal \mathbf{w}_h^i \right]$$
$$\bar{\nabla}^{[0]} F(\boldsymbol{\theta}) \coloneqq \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \hat{\nabla}^{[0]} F_i(\boldsymbol{\theta}),$$

where $\mathbf{x}_h^i$ is the state at time h of a trajectory induced by the noise $\mathbf{w}_{1:H}^i$, i is the index of the sample trajectory, and $\mathrm{D}_{\boldsymbol{\theta}} \pi$ is the Jacobian matrix $\partial \pi / \partial \boldsymbol{\theta} \in \mathbb{R}^{m \times d}$.

The hat notation denotes a per-sample Monte-Carlo estimate, and bar-notation a sample mean. The ZoBG is also referred to as the REINFORCE (Williams, 1992), score function, or the likelihood-ratio gradient.

Baseline. In practice, a baseline term b is subtracted from $V_1(\mathbf{x}_1, \mathbf{w}_{1:H}^i, \boldsymbol{\theta})$ for variance reduction. We use the zero-noise rollout as the baseline $b = V_1(\mathbf{x}_1, \mathbf{0}_{1:H}, \boldsymbol{\theta})$:

$$\frac{1}{\sigma^2} \left[V_1(\mathbf{x}_1, \mathbf{w}_{1:H}^i, \boldsymbol{\theta}) - b \right] \left[\sum_{h=1}^{H} \mathrm{D}_{\boldsymbol{\theta}} \pi(\mathbf{x}_h^i, \boldsymbol{\theta})^\intercal \mathbf{w}_h^i \right].$$

First-order estimator. In differentiable simulators, the gradients of the dynamics ϕ and costs c_h are available *almost surely* (i.e., with probability one). Hence, one may compute the exact gradient $\nabla_{\boldsymbol{\theta}} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$ by automatic differentiation and average them to estimate $\nabla F(\boldsymbol{\theta})$.

Definition 2.4. Given a single first-order gradient estimate $\hat{\nabla}^{[1]} F_i(\boldsymbol{\theta})$, we define the first-order batched gradient (FoBG) as the sample mean:

$$\hat{\nabla}^{[1]} F_i(\boldsymbol{\theta}) \coloneqq \nabla_{\boldsymbol{\theta}} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}^i, \boldsymbol{\theta})$$
$$\bar{\nabla}^{[1]} F(\boldsymbol{\theta}) \coloneqq \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \hat{\nabla}^{[1]} F_i(\boldsymbol{\theta}).$$

The FoBG is also referred to as the reparametrization gradient (Kingma et al., 2015), or the pathwise derivative (Schulman et al., 2015). Finally, we define the empirical variance.

Definition 2.5 (Empirical variance)**.** For $k \in \{0, 1\}$, we define the empirical variance by

$$\hat{\sigma}_k^2 = \tfrac{1}{N-1} \textstyle\sum_{i=1}^{N} \|\hat{\nabla}^{[k]} F_i(\boldsymbol{\theta}) - \bar{\nabla}^{[k]} F(\boldsymbol{\theta})\|^2.$$

3. Pitfalls of First-order Estimates

What are the cases for which we would prefer to use the ZoBG over the FoBG in policy optimization using differentiable simulators? Throughout this section, we analyze the performance of the two estimators through their bias and variance properties, and find pathologies where using the first-order estimator blindly results in worse performance.

3.1. Bias under discontinuities

Under standard regularity conditions, it is well-known that both estimators are unbiased estimators of the true gradient $\nabla F(\boldsymbol{\theta})$. However, care must be taken to define these conditions precisely. Fortunately, the ZoBG is still unbiased under mild assumptions.

Lemma 3.1. *Under Assumption 2.1 and Assumption 2.2, the* ZoBG *is an unbiased estimator of the stochastic objective.*

$$\mathbb{E}[\bar{\nabla}^{[0]} F(\boldsymbol{\theta})] = \nabla F(\boldsymbol{\theta}).$$

In contrast, the FoBG requires strong continuity conditions in order to satisfy the requirement for unbiasedness. However, under Lipschitz continuity, it is indeed unbiased.

Lemma 3.2. *Under Assumption 2.1 and Assumption 2.2, and if $\phi(\cdot,\cdot)$ is locally Lipschitz and $c_h(\cdot,\cdot)$ is continuously differentiable, then $\bar{\nabla}^{[1]} F(\boldsymbol{\theta})$ is defined almost surely, and*

$$\mathbb{E}[\bar{\nabla}^{[1]} F(\boldsymbol{\theta})] = \nabla F(\boldsymbol{\theta}).$$

The proofs and more rigorous statements of both lemmas are provided in Appendix A. Notice that Lemma 3.1 permits V_h to have discontinuities (via discontinuities of c_h and ϕ), whereas Lemma 3.2 does not.

Bias of FoBG **under discontinuities.** The FoBG can fail when applied to discontinuous landscapes. We illustrate a simple case of biasedness through a counterexample.

Example 3.3 (**Heaviside**). (Bangaru et al., 2021; Suh et al., 2021) Consider the Heaviside function,

$$f(\boldsymbol{\theta}, \mathbf{w}) = H(\boldsymbol{\theta} + \mathbf{w}), \quad H(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases},$$

whose stochastic objective becomes the error function

$$F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w}}[H(\boldsymbol{\theta} + \mathbf{w})] = \mathrm{erf}(-\boldsymbol{\theta}; \sigma^2),$$

where $\mathrm{erf}(t; \sigma^2) := \int_t^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/\sigma^2} \mathrm{d}x$ is the Gaussian tail integral. Defining the gradient of the Monte-Carlo objective $H(\boldsymbol{\theta} + \mathbf{w})$ requires subtlety. It is common in physics to define $\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta} + \mathbf{w}) = \delta(\boldsymbol{\theta} + \mathbf{w})$ as a dirac-delta function, where integration is interpreted so that the fundamental theorem of calculus holds. This is *irreconcilable* with using *expectation* to define the integral, which presupposes that the law of large numbers hold. Indeed, since $\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta} + \mathbf{w}) = 0$ for all $\boldsymbol{\theta} \neq -\mathbf{w}$, we have $\mathbb{E}_{\mathbf{w}_i} \delta(\boldsymbol{\theta} + \mathbf{w}_i) = 0$. Hence, the FoBG is biased, because the gradient of the stochastic objective at any $\boldsymbol{\theta}$ is non-zero: $\nabla_{\boldsymbol{\theta}} \mathrm{erf}(-\boldsymbol{\theta}; \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(\boldsymbol{\theta} - \mathbf{w})/2\sigma^2) \neq 0$.

It is worth noting that the empirical variance of the FoBG estimator in this example is zero, since all the samples are identically zero. On the other hand, the ZoBG escapes this problem and provides an unbiased estimate, since it always takes finite intervals that include the integral of the delta.

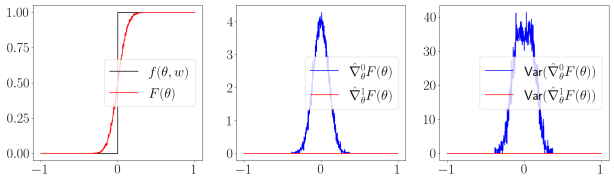

Figure 2. From left: heaviside objective $f(\boldsymbol{\theta}, \mathbf{w})$ and stochastic objective $F(\boldsymbol{\theta})$, empirical values of the gradient estimates, and their empirical variance.

3.2. The "Empirical bias" phenomenon

One might argue that *strict* discontinuity is simply an artifact of modeling choice in simulators; indeed, many simulators approximate discontinuous dynamics as a limit of continuous ones with growing Lipschitz constant (Geilinger et al., 2020; Elandt et al., 2019). In this section, we explain how this can lead to a phenomenon we call *empirical bias*, where the FoBG appears to have low empirical variance, but is still highly inaccurate; i.e. it "looks" biased when a finite number of samples are used. Through this phenomenon, we claim that performance degradation of first-order gradient estimates do not require strict discontinuity, but is also present in continuous, yet *stiff* approximations of discontinuities.

Definition 3.4 (Empirical bias)**.** Let $\mathbf{z}$ be a vector-valued random variable with $\mathbb{E}[\|\mathbf{z}\|] < \infty$. We say $\mathbf{z}$ has (β, Δ, S)-empirical bias if there is a random event $\mathcal{E}$ such that $\Pr[\mathcal{E}] \geq 1 - \beta$, and $\|\mathbb{E}[\mathbf{z} \mid \mathcal{E}] - \mathbb{E}[\mathbf{z}]\| \geq \Delta$, but $\|\mathbf{z} - \mathbb{E}[\mathbf{z} \mid \mathcal{E}]\| \leq S$ almost surely on $\mathcal{E}$.

A paradigmatic example of empirical bias is a random scalar $\mathbf{z}$ which takes the value 0 with probability $1 - \beta$, and $\frac{1}{\beta}$ with probability β. Setting $\mathcal{E} = \{\mathbf{z} = 0\}$, we see $\mathbb{E}[\mathbf{z}] = 1$, $\mathbb{E}[\mathbf{z} \mid \mathcal{E}] = 0$, and so $\mathbf{z}$ satisfies $(\beta, 1, 0)$-empirical bias. Note that $\mathbf{Var}[\mathbf{z}] = 1/\beta - 1$; in fact, small-$\beta$ empirical bias implies large variance more generally.

Lemma 3.5. *Suppose* $\mathbf{z}$ *has* (β, Δ, S)*-empirical bias. Then* $\mathbf{Var}[\mathbf{z}] \geq \frac{\Delta_0^2}{\beta}$*, where* $\Delta_0 := \max\{0, (1-\beta)\Delta - \beta\|\mathbb{E}[\mathbf{z}]\|\}$.

Empirical bias naturally arises for discontinuities or stiff continuous approximations. We give two examples of common discontinuities that arise in differentiable simulation, that permit continuous approximations.

Example 3.6 (**Coulomb friction**). The Coulomb model of friction is discontinuous in the relative tangential velocity between two bodies. In many simulators (Geilinger et al., 2020; Castro et al., 2020), it is common to consider a continuous approximation instead. We idealize such approximations through a piecewise linear relaxation of the Heaviside that is continuous, parametrized by the width of the middle linear region ν (which corresponds to *slip tolerance*).

$$\bar{H}_\nu(t) = \begin{cases} 2t/\nu & \text{if } |t| \leq \nu/2 \\ H(t) & \text{else} \end{cases}.$$

In practice, lower values of ν lead to more realistic behavior in simulation (Tedrake, 2022), but this has adverse effects for empirical bias. Considering $f_\nu(\boldsymbol{\theta}, \mathbf{w}) = \bar{H}_\nu(\boldsymbol{\theta} + \mathbf{w})$, we have $F_\nu(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w}}[\bar{H}_\nu(\boldsymbol{\theta} + \mathbf{w})] := \mathrm{erf}(\nu/2 - \theta; \sigma^2)$. In particular, setting $c_\sigma := \frac{1}{\sqrt{2\pi}\sigma}$, then at $\boldsymbol{\theta} = \nu/2$, $\nabla F_\nu(\boldsymbol{\theta}) = c_\sigma$, whereas, with probability at least $c_\sigma\nu$, $\nabla f_\nu(\boldsymbol{\theta}, \mathbf{w}) = 0$. Hence, the FoBG has $(c_\sigma\nu, c_\sigma, 0)$ empirical bias, and its variance scales with $1/\nu$ as $\nu \to 0$. The limiting $\nu = 0$ case, corresponding to the Coulomb model, is the Heaviside from Example 3.3, where the limit of high empirical bias, as well as variance, becomes biased in expectation (but, surprisingly, zero variance!). We empirically illustrate this effect in Figure 3. We also note that more complicated models of friction (e.g. that incorporates the Stribeck effect (Stribeck, 1903)) would suffer similar problems.

Example 3.7. (**Discontinuity in geometry**). Another source of discontinuity in simulators comes from the discontinuity of surface normals. We show this in Figure 4, where balls that collide with a rectangular geometry create discontinuities. It is possible to make a continuous relaxation (Elandt et al., 2019) by considering a smoother geometry, depicted by the addition of the dome in Figure 4. While this makes FoBG no longer biased asymptotically, the stiffness of the relaxation still results in high empirical bias.

3.3. High variance first-order estimates

Even in the absence of empirical bias, we present other cases in which FoBG suffers simply due to high variance.

Scenario 1: Persistent stiffness. When the dynamics are *stiff*, such as contact models with stiff spring approximations (Hunt & Crossley, 1975), the high norm of the gradient can contribute to high variance of the FoBG.

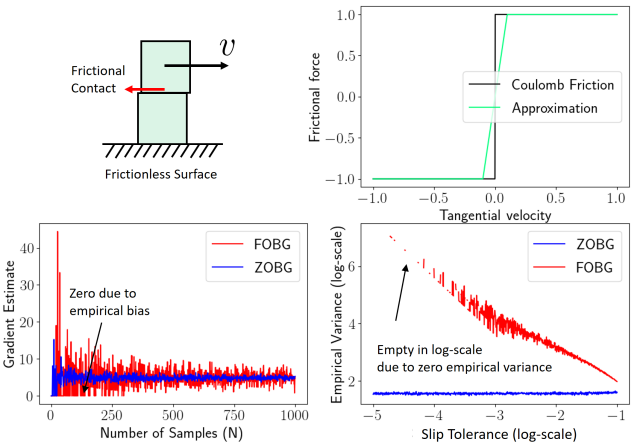

Figure 3. Top column: illustration of the physical system and the relaxation of Coulomb friction. Bottom column: the values of estimators and their empirical variances depending on number of samples and slip tolerance. Values of FoBG are zero in low-sample regimes due to empirical bias. As $\nu \to 0$, the empirical variance of FoBG goes to zero, which shows as empty in the log-scale. Expected variance, however, blows up as it scales with $1/\nu$.

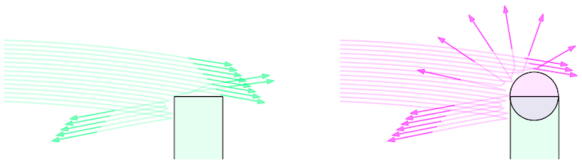

Figure 4. Left: More detailed example of ball hitting the wall in Figure 1.B. Left: The green trajectories hit a rectangular wall, displaying discontinuities. Right: the pink trajectories collide with the dome on top, and show continuous but stiff behavior.

Example 3.8. (**Pushing with stiff contact**). We demonstrate this phenomenon through a simple 1D pushing example in Figure 5, where the ZoBG has lower variance than the FoBG as stiffness increases, until numerical semi-implicit integration becomes unstable under a fixed timestep.

In practice, lowering the timestep can alleviate the issue at the cost of more computation time. Less stiff formulations of contact dynamics (Stewart & Trinkle, 2000; Mirtich, 1996) also addresses this problem effectively.

Scenario 2: Chaos. As noted in (Metz et al., 2021), even if the gradient of the dynamics is small at every h, their compounding product can cause $\|\nabla_{\boldsymbol{\theta}} V_1\|$ to be large if the system is chaotic. Yet, in expectation, the gradient of the stochastic objective $\nabla F = \nabla \mathbb{E}[V_1]$ can be benign and well-behaved (Lasota & Mackey, 1996).

Example 3.9. (**Chaos of double pendulum**). We demonstrate this in Figure 6 for a classic chaotic system of the double pendulum. As the horizon of the trajectory increases, the variance of FoBG becomes higher than that of ZoBG.

Comparison to ZoBG. Compared to the pitfalls of FoBG, the ZoBG variance can be bounded as follows.

Lemma 3.10. *If for all* $\mathbf{x}$ *and* $\bar{\mathbf{w}}$, $|V_1(\mathbf{x}, \bar{\mathbf{w}}, \boldsymbol{\theta})| \leq B_V$ *and* $\|\mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}, \boldsymbol{\theta})\|_{\mathrm{op}} \leq B_\pi$, *then*

$$\mathbf{Var}[\bar{\nabla}^{[0]}F(\boldsymbol{\theta})] = \frac{1}{N}\mathbf{Var}[\hat{\nabla}^{[0]}F_i(\boldsymbol{\theta})] \leq \frac{B_V^2 B_\pi^2}{N} \cdot \frac{Hn}{\sigma^2}.$$

We refer to Appendix B.2 for proof. Lemma 3.10 is intended to provide a qualitative understanding of the zeroth-order variance: it scales with the horizon-dimension product Hn, but *not* the scale of the derivatives. On the other hand, the variance of FoBG does; when $\frac{Hn}{\sigma^2} \gg \mathbf{Var}[\hat{\nabla}^{[1]}F(\boldsymbol{\theta})] = \mathbf{Var}[\nabla_{\boldsymbol{\theta}} V(\mathbf{x}_1, \bar{\mathbf{w}}, \boldsymbol{\theta})]$, the ZoBG has higher variance.

4. α-order Gradient Estimator

Previous examples give us insight on which landscapes are better fit for first-order estimates of policy gradient, and which are better fit for zeroth-order ones. As shown in Figure 7, even on a single policy optimization objective, it is best to adaptively switch between the first and zeroth-order estimators depending on the local characteristics of the landscape. In this section, we propose a strategy to achieve this adaptively, interpolating between the two estimators to reap the benefits of both approaches simultaneously.

Definition 4.1. Given $\alpha \in [0, 1]$, we define the alpha-order batched gradient (AoBG) as:

$$\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta}) = \alpha\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) + (1-\alpha)\bar{\nabla}^{[0]}F(\boldsymbol{\theta}).$$

When interpolating, we use independent trajectories to generate $\bar{\nabla}^{[1]}F(\boldsymbol{\theta})$ and $\bar{\nabla}^{[0]}F(\boldsymbol{\theta})$ (see Appendix C.1). We consider strategies for selecting α in a *local fashion*, as a function of the observed sample, as detailed below.

4.1. A robust interpolation protocol

A potential approach might be to select α based on achieving minimum variance (Parmas et al., 2018; Metz et al., 2019),

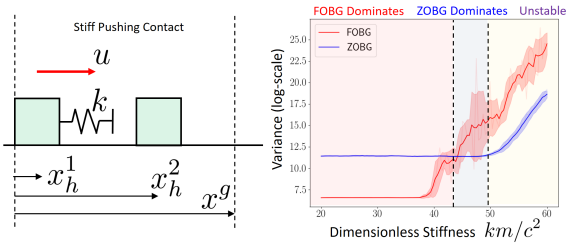

Figure 5. The variance of the gradient of V_1, with running cost $c_h = \|\mathbf{x}_h^2 - \mathbf{x}^g\|^2$, with respect to input trajectory as spring constant k increases. Mass m and damping coefficient c are fixed.

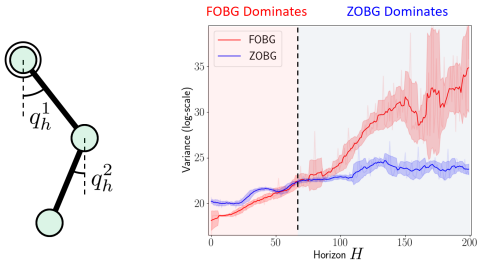

Figure 6. Variance of the gradient of the terminal cost $\|q_H - q^g\|^2$ with respect to the initial position q_1. As horizon grows through a chaotic system, the ZoBG dominates the FoBG.

considering empirical variance as an estimate. However, in light of the *empirical bias* phenomenon detailed in Section 3 (or even actual bias in the presence of discontinuities), we see that the empirical variance is unreliable, and can lead to inaccurate estimates for our setting. For this reason, we consider an additional criterion of *uniform accuracy*:

Definition 4.2 (Accuracy)**.** α is (γ, δ)-accurate if the bound on the *error* of AoBG is satisfied with probability δ:

$$\|\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \leq \gamma. \tag{3}$$

To remedy the limitations of considering empirical variance in isolation, we propose an interpolation protocol that can satisfy an accuracy guarantee, while still attempting to minimize the variance.

$$\begin{aligned} \min_{\alpha\in[0,1]} \quad & \alpha^2\hat{\sigma}_1^2 + (1-\alpha)^2\hat{\sigma}_0^2 \\ \text{s.t.} \quad & \epsilon + \alpha \underbrace{\|\bar{\nabla}^{[1]}F - \bar{\nabla}^{[0]}F\|}_{B} \leq \gamma. \end{aligned} \tag{4}$$

We explain the terms in Eq (4) below in detail.

Objective. Since we interpolate the FoBG and ZoBG using independent samples, $\alpha^2\hat{\sigma}_1^2 + (1-\alpha)^2\hat{\sigma}_0^2$ is an unbiased estimate of $N \cdot \mathbf{Var}[\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta})]$. Thus, our objective is to choose α to minimize this variance.

Constraint. Our constraint serves to enforce accuracy. Since the FoBG is potentially biased, we use ZoBG as a surrogate of $\nabla F(\boldsymbol{\theta})$. For this purpose, we use $\epsilon > 0$ as a confidence bound on $\|\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\|$ from the obtained samples. When ϵ is a valid confidence bound that holds with probability δ, we prove that our constraint in Eq (4) guarantees accuracy in Eq (3).

Lemma 4.3 (Robustness)**.** *Suppose that* $\epsilon + \alpha B \leq \gamma$ *with probability* δ. *Then,* α *is* (γ, δ)*-accurate.*

Proof. By repeated applications of the triangle inequality. See Appendix C.3 for a detailed proof. □

Specifying the confidence $\epsilon > 0$**.** We select $\epsilon > 0$ based on a Bernstein vector concentration bound (Appendix C.4),

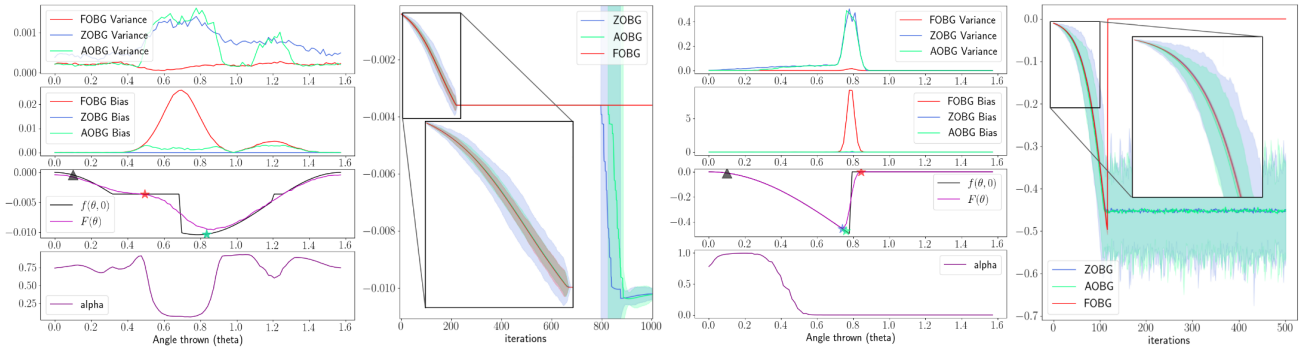

Figure 7. First Column: **Ball with wall** example. In the third row, the triangle is the initial point, and red/blue/green stars are the optimum achieved by FoBG, ZoBG, and AoBG respectively (blue and green stars overlap). Second column: Iteration vs. Cost plot of different gradients. Right columns: Same plot repeated for the **Momentum Transfer** example. Standard deviation plotted 10 fold for visualization.

which only requires a prior upper bound on the magnitude of the value function $V_1(\cdot)$ and gradients $\mathrm{D}_{\boldsymbol{\theta}}\pi(\cdot, \boldsymbol{\theta})$.

Asymptotic feasibility. Eq (4) is not feasible if $\epsilon > \gamma$, which would indicate that we simply do not have enough samples to guarantee (γ, δ)-accuracy. In this case, we choose to side on conservatism and fully use the ZoBG by setting $\alpha = 0$. Asymptotically, as the number of samples $N \to \infty$, the confidence interval $\varepsilon \to 0$, which implies that Eq (4) will always be feasible.

Finally, we note that Eq (4) has a closed form solution, whose proof is provided in Appendix C.2.

Lemma 4.4. *With $\gamma = \infty$, the optimal α is $\alpha_\infty := \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}$. For finite $\gamma \geq \epsilon$, Eq (4) is*

$$\alpha_\gamma := \begin{cases} \alpha_\infty & \text{if} \quad \alpha_\infty B \leq \gamma - \varepsilon \\ \frac{\gamma - \varepsilon}{B} & \text{otherwise} . \end{cases} \tag{5}$$

We give some qualitative characteristics of the solution:

- If we are within constraint and $\hat{\sigma}_0^2 \gg \hat{\sigma}_1^2$, as we can expect from benign smooth systems, then $\alpha \approx 1$, and we rely more on the FoBG.
- In pathological cases where we are unbiased yet $\hat{\sigma}_1^2 \gg \hat{\sigma}_0^2$ (e.g. stiffness and chaos), then $\alpha \approx 0$.
- If there is a large difference between the ZoBG and the FoBG such that $B \gg 0$, we expect strict/empirical bias from discontinuities and tend towards using ZoBG.

5. Landscape Analysis & Case Studies

5.1. Landscape analysis on examples

Though we have characterized the bias-variance characteristics of different gradients, their convergence properties in landscapes of physical systems remain to be investigated. We visualize the performance of fixed-step gradient descent with the FoBG, ZoBG, and AoBG on examples of Figure 1.

Ball with wall. On the system of Figure 1.B, the FoBG fails to make progress at the region of flatness, while the ZoBG and AoBG successfully find the minima of the landscape (Figure 7). In addition, the interpolation scheme switches to prioritizing ZoBG near discontinuities, while using more information from FoBG far from discontinuities; as a result, the variance of AoBG is lower than that of ZoBG.

Angular momentum transfer. Next, we show results for the momentum transfer system of Figure 1.C in Figure 7. Running gradient descent results in both the ZoBG and AoBG converging to the robust local minima of the solution. However, the bias of FoBG forces it off the cliff and the optimizer is unable to recover. Again, our interpolation scheme smoothly switches to prioritizing the ZoBG near the discontinuity, enabling it to stay within the safe region while maximizing the transferred momentum.

Bias-variance leads to different minima. Through these examples with discontinuities, we claim that the bias-variance characteristics of gradients in these landscapes not only lead to different convergence rates, but convergence to different minima. The same argument holds for *nearly discontinuous* landscapes that display high empirical bias. Both estimators are unbiased in expectation, and the high variance of FoBG should manifest itself in worse convergence rates. Yet, the high *empirical bias* in the finite-sample regime leads to low empirical variance and different minima, leading to performance that is indistinguishable from when the underlying landscape is truly discontinuous.

Combined with the benefits of stochasticity in Section 1, we believe that this might explain why zero-order methods in RL are solving problems for physical systems where deterministic (even *stochastic*) first order methods have struggled.

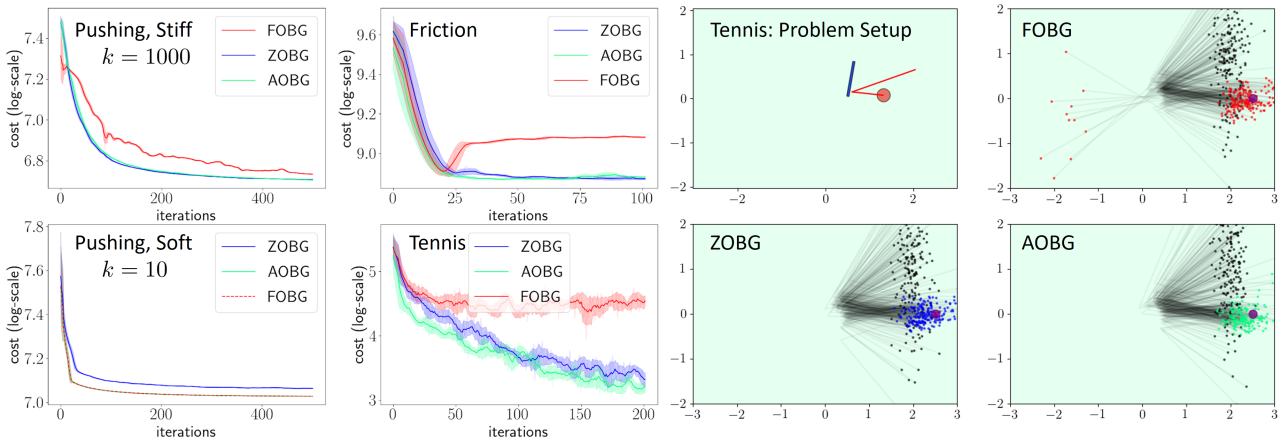

Figure 8. 1st column: trajectory optimization on pushing example with different contact models. AoBG and FoBG overlaps in soft pushing example. 2nd column: trajectory optimization on friction contact, and policy optimization on the tennis example. 3rd / 4th column: Visualization of policy performance for tennis. Black dots correspond to initial positions and colored dots correspond to final position.

5.2. Policy optimization case studies

To validate our results on policy optimization problems with differentiable simulators, we compare the performance of different gradients on time-stepping simulations written in `torch` (Paszke et al., 2019). For all of our examples, we validate the correctness of the analytic gradients by comparing the values of FoBG and ZoBG on a one-step cost.

Pushing & friction: Trajectory optimization. We describe performance of gradients on the pushing (Figure 5) and friction (Figure 3) environments, where contact is modeled using the *penalty method* (i.e. stiff spring) with additional viscous damping on the velocities of the system. We use horizon of $H = 200$ to find the optimal force sequence of the first block to minimize distance between the second block and the goal position. Our results in Figure 8 show that for soft springs ($k = 10$), the FoBG outperforms the ZoBG, but stiffer springs ($k = 1000$) results in the ZoBG outperforming the FoBG. Our interpolated gradient AoBG is able to automatically choose the one that performs better.

In addition, although the FoBG initially converges faster in the friction environment, it is unaware of the discontinuity that occurs when it slides off the box. On the other hand, the AoBG and ZoBG successfully optimize the trajectory, with AoBG showing slightly faster convergence.

Tennis: Policy optimization. Next, we describe the performance of different gradients on a tennis environment (similar to breakout), where the paddle needs to bounce the ball to some desired target location. We use a linear feedback policy with $d = 21$ parameters, and horizon of $H = 200$. In order to correctly obtain analytic gradients, we use continuous event detection with the time of impact formulation (Hu et al., 2020). The results of running policy optimization is presented in Figure 8. While the ZoBG and the AoBG are successful in finding a policy that bounces the balls through different initial conditions, the FoBG suffers from the discontinuities of geometry, and still misses many of the balls. Furthermore, the AoBG still converges slightly faster than the ZoBG by utilizing first-order information.

Implicit time-stepping. A large class of simulators rely on optimization-based implicit time-stepping (Todorov et al., 2012; Coumans & Bai, 2016–2021; Macklin et al., 2014; Pang, 2021), which can be made differentiable by sensitivity analysis (Boyd & Vandenberghe, 2004). While these simulators suffer less from stiffness, we expect that geometrical discontinuities will remain problematic. We leave detailed empirical study using these simulators to future work.

6. Conclusion

Do differentiable simulators give better policy gradients? We have shown that the answer depends intricately on the underlying characteristics of the physical systems. While Lipschitz continuous systems with reasonably bounded gradients may enjoy fast convergence given by the low variance of first-order estimators, using the gradients of differentiable simulators may *hurt* for problems that involve nearly/strictly discontinuous landscapes, stiff dynamics, or chaotic systems. Moreover, due to the empirical bias phenomenon, bias of first-order estimators in nearly/strictly discontinuous landscapes cannot be diagnosed from empirical variance alone. We believe that many challenging tasks that both RL and differentiable simulators try to address necessarily involve dealing with physical systems with such characteristics, such as those that are rich with contact.

These limitations of using differentiable simulators for planning and control need to be addressed from both the design of simulator and algorithms: from the simulator side, we have shown that certain modeling decisions such as stiffness of contact dynamics can have significant underlying consequences in the performance of policy optimization that uses

gradients from these simulators. From the algorithm side, we have shown we can automate the procedure of deciding which one to use online via interpolation.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.

Bangaru, S. P., Michel, J., Mu, K., Bernstein, G., Li, T.-M., and Ragan-Kelley, J. Systematically differentiating parametric discontinuities. *ACM Trans. Graph.*, 40(4), July 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459775.

Berahas, A. S., Cao, L., Choromanski, K., and Scheinberg, K. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *arXiv: Optimization and Control*, 2019.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods, 2020.

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

Carpentier, J., Saurel, G., Buondonno, G., Mirabel, J., Lamiraux, F., Stasse, O., and Mansard, N. The pinocchio c++ library : A fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pp. 614–619, 2019. doi: 10.1109/SII.2019.8700380.

Castro, A. M., Qu, A., Kuppuswamy, N., Alspach, A., and Sherman, M. A transition-aware method for the simulation of compliant contact with regularized friction. *IEEE Robotics and Automation Letters*, 5(2):1859–1866, Apr 2020. ISSN 2377-3774. doi: 10.1109/lra.2020.2969933. URL http://dx.doi.org/10.1109/LRA.2020.2969933.

Çinlar, E. *Probability and stochastics*, volume 261. Springer, 2011.

Coumans, E. and Bai, Y. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

de Avila Belbute-Peres, F., Smith, K., Allen, K., Tenenbaum, J., and Kolter, J. Z. End-to-end differentiable physics for learning and control. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/842424a1d0595b76ec4fa03c46e8d755-Paper.pdf.

Du, T., Li, Y., Xu, J., Spielberg, A., Wu, K., Rus, D., and Matusik, W. D3{pg}: Deep differentiable deterministic policy gradients, 2020. URL https://openreview.net/forum?id=rkxZCJrtwS.

Duchi, J., Bartlett, P., and Wainwright, M. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22, 03 2011. doi: 10.1137/110831659.

Duchi, J., Jordan, M., Wainwright, M., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61, 12 2015. doi: 10.1109/TIT.2015.2409256.

Elandt, R., Drumwright, E., Sherman, M., and Ruina, A. A pressure field model for fast, robust approximation of net contact force and moment between nominally rigid objects. *IROS*, pp. 8238–8245, 2019.

Ern, A. and Guermond, J.-L. *Theory and practice of finite elements*, volume 159. Springer Science & Business Media, 2013.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator, 2019.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=VdvDlnnjzIN.

Geilinger, M., Hahn, D., Zehnder, J., Bächer, M., Thomaszewski, B., and Coros, S. Add: Analytically differentiable dynamics for multi-body systems with frictional contact, 2020.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://doi.org/10.1137/120880811.

Gradu, P., Hallman, J., Suo, D., Yu, A., Agarwal, N., Ghai, U., Singh, K., Zhang, C., Majumdar, A., and Hazan, E. Deluca – a differentiable control library: Environments, methods, and benchmarking, 2021.

Hu, Y., Anderson, L., Li, T.-M., Sun, Q., Carr, N., Ragan-Kelley, J., and Durand, F. Difftaichi: Differentiable programming for physical simulation. *ICLR*, 2020.

Huang, Z., Hu, Y., Du, T., Zhou, S., Su, H., Tenenbaum, J. B., and Gan, C. Plasticinelab: A soft-body manipulation benchmark with differentiable physics. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xCcdBRQEDW.

Hunt, K. H. and Crossley, F. R. E. Coefficient of Restitution Interpreted as Damping in Vibroimpact. *Journal of Applied Mechanics*, 42(2):440–445, 06 1975. ISSN 0021-8936. doi: 10.1115/1.3423596. URL https://doi.org/10.1115/1.3423596.

Kakade, S. M. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Lasota, A. and Mackey, M. C. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*. Cambridge university press, 1996.

Le Lidec, Q., Montaut, L., Schmid, C., Laptev, I., and Carpentier, J. Leveraging Randomized Smoothing for Optimal Control of Nonsmooth Dynamical Systems. working paper or preprint, December 2021. URL https://hal.archives-ouvertes.fr/hal-03480419.

Macklin, M., Müller, M., Chentanez, N., and Kim, T.-Y. Unified particle physics for real-time applications. *ACM Trans. Graph.*, 33(4), jul 2014. ISSN 0730-0301. doi: 10.1145/2601097.2601152. URL https://doi-org.libproxy.mit.edu/10.1145/2601097.2601152.

Mahamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte carlo gradient estimation in machine learning. In Dy, J. and Krause, A. (eds.), *Journal of Machine Learning Research*, volume 21, pp. 1–63, 4 2020.

Mason, M. T. *Mechanics of Robotic Manipulation*. The MIT Press, 06 2001. ISBN 9780262256629. doi: 10.7551/mitpress/4527.001.0001. URL https://doi.org/10.7551/mitpress/4527.001.0001.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Metz, L., Maheswaranathan, N., Nixon, J., Freeman, D., and Sohl-Dickstein, J. Understanding and correcting pathologies in the training of learned optimizers. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4556–4565. PMLR, 09–15 Jun 2019.

Metz, L., Freeman, C. D., Schoenholz, S. S., and Kachman, T. Gradients are not all you need, 2021.

Mirtich, B. V. *Impulse-Based Dynamic Simulation of Rigid Body Systems*. PhD thesis, 1996. AAI9723116.

Mora, M. A. Z., Peychev, M., Ha, S., Vechev, M., and Coros, S. Pods: Policy optimization via differentiable simulation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7805–7817. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mora21a.html.

Pang, T. A convex quasistatic time-stepping scheme for rigid multibody systems with contact and friction. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6614–6620, 2021.

Parmas, P., Rasmussen, C. E., Peters, J., and Doya, K. PIPPS: Flexible model-based policy search robust to the curse of chaos. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4065–4074. PMLR, 10–15 Jul 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Schulman, J., Heess, N., Weber, T., and Abbeel, P. Gradient estimation using stochastic computation graphs. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

Stein, E. M. and Shakarchi, R. *Real analysis*. Princeton University Press, 2009.

Stewart, D. and Trinkle, J. J. An implicit time-stepping scheme for rigid body dynamics with coulomb friction. volume 1, pp. 162–169, 01 2000. doi: 10.1109/ROBOT.2000.844054.

Stribeck, R. *Die wesentlichen Eigenschaften der Gleit- und Rollenlager*. Mitteilungen über Forschungsarbeiten auf dem Gebiete des Ingenieurwesens, insbesondere aus den Laboratorien der technischen Hochschulen. Julius Springer, 1903.

Suh, H. J. T., Pang, T., and Tedrake, R. Bundled gradients through contact via randomized smoothing. *arXiv preprint*, 2021.

Sutton, R., Mcallester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst*, 12, 02 2000.

Tedrake, R. Drake: A planning, control, and analysis toolbox for nonlinear dynamical systems, 2022. URL `http://drake.mit.edu`.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *IROS*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL `http://dx.doi.org/10.1561/2200000048`.

van der Schaft, A. and Schumacher, H. *An Introduction to Hybrid Dynamical Systems*. Springer Publishing Company, Incorporated, 1st edition, 2000. ISBN 978-1-4471-3916-4.

Wasserman, L. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Werling, K., Omens, D., Lee, J., Exarchos, I., and Liu, C. K. Fast and feature-complete differentiable physics for articulated rigid bodies with contact, 2021.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 3, 05 1992.

Zhang, K., Koppel, A., Zhu, H., and Başar, T. Global convergence of policy gradient methods to (almost) locally optimal policies, 2020.

Supplementary Materials for "Do Differentiable Simulators give Better Gradients for Policy Optimization?"

A. Formal Expected Gradient Computations

This section establishes rigorous unbiasedness guarantees for the ZoBG (under general conditions) and of the FoBG (under more restrictive conditions). Specifically, Corollary A.12 provides a rigorous version of the ZoBG guarantee, Lemma 3.1, which is a special case of Proposition A.11 which holds for general, possibly non-Gaussian noise distributions. The FoBG estimator is addressed in Proposition A.15, which provides the rigorous statement of Lemma 3.2. We present a lengthy preliminaries section, Appendix A.1, to formalize the results that follow. We then follow with formal statements of the results, Appendix A.2, and defer proofs to Appendix A.3. The preliminaries below are requisites only for the results and proofs within this section, and are not needed in future appendices.

A.1. Preliminaries

Throughout, $\|\cdot\|$ denotes the Euclidean norm of vectors. We begin by specifying our sense of expectations and derivatives, and then turn to other, less-standard preliminaries. To rigorously describe expectations of non-continuous functions and of derivatives of non-smooth functions, we start with some preliminaries from measure theory.

Lebesgue measurability. For a background on measure theory, we direct the reader to (Stein & Shakarchi, 2009). Here, we recall a few definitions. We define the set of Lebesgue measurable sets $\mathscr{L}(\mathbb{R}^D)$ as the collection of subset $\mathcal{Z} \subset \mathbb{R}^D$ for which the Lebesgue measure is well-defined. We let $\mathscr{B}(\mathbb{R}^{D'}) \subset \mathscr{L}(\mathbb{R}^D)$ be the collection of Borel measurable sets on $\mathbb{R}^{D'}$. We say a mapping $\Phi : \mathbb{R}^D \to \mathbb{R}^{D'}$ is *Lebesgue measurable* if for all $\mathcal{Z}' \in \mathscr{B}(\mathbb{R}^{D'})$, $\Phi^{-1}(\mathcal{Z}') \in \mathscr{L}(\mathbb{R}^D)$. We say it is *Borel measurable* if, more strongly, it holds that $\Phi^{-1}(\mathcal{Z}') \in \mathscr{B}(\mathbb{R}^D)$. The composition of Borel measurable functions are Borel measurable, but the same is not true more generally for Lebesgue measurable functions. Throughout, all functions are assumed Borel measurable unless otherwise specified, so their compositions are also Borel measurable.

More generally, given a Lebesgue measurable set $\mathcal{Z} \subset \mathbb{R}^D$, we define $\mathscr{L}(\mathcal{Z})$ as the set $\{\mathcal{Z} \cap \tilde{\mathcal{Z}} : \tilde{\mathcal{Z}} \in \mathscr{L}(\mathbb{R}^{R^d})\}$, and say a function $\Phi : \mathcal{Z} \to \mathbb{R}^{D'}$ is Lebesgue mearuable on its domain if for all $\mathcal{Z}' \in \mathscr{B}(\mathbb{R}^{D'})$, $\Phi^{-1}(\mathcal{Z}') \in \mathscr{L}(\mathbb{R}^D)$.

Lebesgue complete distribution. We consider probability distributions $\mathcal{D}$ on $\mathbb{R}^D$ which assign probability to *all Lebesgue measurable sets* $\mathcal{Z} \subset \mathbb{R}^D$: i.e., $\Pr_{\mathbf{z}\sim\mathcal{D}}[\mathbf{z} \in \mathcal{Z}]$ is well defined. Note that these distribution do not need to have density with respect to the Lebesgue measure: indeed, continuous, discrete, and mixture of continuous and discrete distributions all can be defined to assign probabilities to all Lebesgue-measurable sets.

We say $\mathcal{Z} \subset \mathbb{R}^D$ is $\mathcal{D}$*-null* if $\Pr_{\mathbf{z}\sim\mathcal{D}}[\mathbf{z} \in \mathcal{Z}] = 0$. We assume without loss of generality that $\mathcal{D}$ is *complete*, so that given a $\mathcal{D}$-null Lebesgue measurable set $\mathcal{Z}$, $\Pr_{\mathbf{z}\sim\mathcal{D}}[\mathbf{z} \in \mathcal{Z}']$ is well defined and equal to zero for all $\mathcal{Z}' \subset \mathcal{Z}$. We call distributions which are complete and assign probability to all Lebesgue sets *Lebesgue complete*. We shall assume without comment that all distributions are Lebesgue complete.

Almost-everywhere functions and expectation. Given a Lebesgue complete distribution $\mathcal{D}$ on $\mathbb{R}^D$, we define expectation of a Lebesgue measurable $\Phi : \mathbb{R}^D \to \mathbb{R}^{D'}$ in the standard way. We say a function Φ is defined $\mathcal{D}$-*almost-surely* if there exists a Lebesgue-measurable set $\mathcal{Z} \subset \mathbb{R}^D$ such that Φ is a Lebesgue measurable as mapping $\mathcal{Z} \to \mathbb{R}^{D'}$, and $\mathcal{Z}^c = \mathbb{R}^D \setminus \mathcal{Z}$ is $\mathcal{D}$-null. Given such a function Φ, we define its expectation

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\Phi(\mathbf{z})] := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\tilde{\Phi}(\mathbf{z})], \quad \text{where } \tilde{\Phi}(\mathbf{z}) = \begin{cases} \Phi(\mathbf{z}) & \mathbf{z} \in \mathcal{Z} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

One can verify that $\tilde{\Phi}(\mathbf{z})$ is Lebesgue measurable. Note that this definition is independent of the choice of $\mathcal{Z}$: if $\mathcal{Z}'$ is another set witnessing the almost-sure definition of Φ, then the induced map $\tilde{\Phi}'$ defined by applying Eq (6) with $\mathcal{Z}'$ is also Lebesgue measurable, $\tilde{\Phi}' = \tilde{\Phi}$ $\mathcal{D}$-almost surely, so that the integrals coincide.

Example A.1 (Heaviside, revisited). With definition in Eq (6), we see that the derivative of the example in Example 3.3 is 0 almost surely under $\mathbf{w} \sim p$; that is, the event on which the derivative of the Heaviside is both undefined has probability zero when $\mathbf{w} \sim p$, and outside this event, its derivative is identically zero.

Stated simply, we ignore values of Φ defined outside the probability-one set $\mathcal{Z}$. This definition has numerous advantages. For one, it satisfies the law of large numbers. That is,

- If $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\|\tilde{\Phi}(\mathbf{z})\| < \infty$, then for $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)} \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, $\frac{1}{N}\sum_{i=1}^{N} \Phi(\mathbf{z}^{(i)})$ converges to $\mathbb{E}[\Phi(\mathbf{z})]$ in probability.
- If $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\|\tilde{\Phi}(\mathbf{z})\|^2 < \infty$, this convergence holds almost surely.

For further discussion, we direct the readers to a standard reference on probability theory (e.g. (Çinlar, 2011)).

Multivariable derivative. We provide conditions under which the multivariable function $F(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Formally, we say that a function $\Phi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$ is *differentiable* at a point $\mathbf{z} \in \mathbb{R}^d$ if there exists a linear map $\mathrm{D}\Phi(\mathbf{z}) \in \mathbb{R}^{d_2 \times d_1}$ such that

$$\lim_{\|\mathbf{h}\|\to 0} \left\| \frac{\Phi(\mathbf{h}+\mathbf{z}) - \Phi(\mathbf{z})}{\|\mathbf{h}\|} - \mathrm{D}\Phi(\mathbf{z}) \cdot \mathbf{h} \right\| = 0.$$

The limit is defined in the sense of $\lim_{\|\mathbf{h}\|\to 0}(\cdot) = \lim_{t\to 0} \sup_{\|\mathbf{h}\|\le t}(\cdot)$. Existence of a multivariable derivative slightly stronger that $\Phi(\cdot)$ having directional derivatives, and in particular, stronger than the existence of a gradient (see (Rudin et al., 1964, Chapter 9) for reference).

Finite moments and polynomial growth. To ensure all expectations are defined, we consider distributions for which all moments are finite.

Definition A.2. We say that a (Lebesgue complete) distribution ρ over a random variable $\mathbf{z}$ *has finite moments* if $\mathbb{E}_{\mathbf{z}\sim\rho}\|\mathbf{z}\|^a < \infty$ for all $a > 0$.

The class of function which have finite expectations under distributions with finite moments are functions which have polynomial growth, in the following sense.

Definition A.3. We say that a function $\psi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ has *polynomial growth* if there exists constants $a, b > 0$ such that $\|\psi(\mathbf{x})\| \le a(1 + \|\mathbf{z}\|^b)$ for all $\mathbf{z} \in \mathbb{R}^d$. We say that a matrix (or tensor) valued function has polynomial growth if the vector-valued function corresponding to flattening its entries into a vector has polynomial growth (for matrices, this means $\|\psi(\mathbf{z})\|_{\mathrm{F}} \le a(1 + \|\mathbf{z}\|^b)$).

The following lemma is clear.

Lemma A.4. *Suppose ρ is a distribution over variables $\mathbf{x}$ which has finite moments, and suppose $g(\mathbf{x})$ has polynomial growth. Then $\mathbb{E}[g(\mathbf{x})]$ is well defined.*

A second useful (and straightforward to check) fact is that polynomial growth is preserved under marginalization.

Lemma A.5. *Suppose ρ is a distribution over variables $\mathbf{x}$ which has finite moments, and suppose $g(\mathbf{z}, \mathbf{x})$ has polynomial growth in its argument $(\mathbf{z}, \mathbf{x})$. Then $\mathbf{z} \mapsto \mathbb{E}[g(\mathbf{z}, \mathbf{x})]$ is well defined and has polynomial growth in $\mathbf{z}$.*

Lipschitz functions. To establish the unbiasedness of the FoBG for non-smooth functions, we invoke the Lipschitz continuity assumption. We say a function $\Phi : \mathbb{R}^D \to \mathbb{R}^{D'}$ is locally-Lipschitz if, for every $\mathbf{z} \in \mathbb{R}^D$, there is a neighborhood a neighborhood $\mathcal{U}$ of $\mathbf{z}$ such that there exists an $L > 0$ such that for all $\mathbf{z}', \mathbf{z}'' \in \mathcal{U}$, $\|\Phi(\mathbf{z}') - \Phi(\mathbf{z}'')\| \leq L\|\mathbf{z}' - \mathbf{z}''\|$. Locally Lipschitz functions are continuous, and thus Borel measurable.

Lemma A.6 (Rademacher's Theorem)**.** *Every locally Lipschitz function $\psi : \mathbb{R}^D \to \mathbb{R}$ is differentiable on a set of $\mathcal{Z} \subset \mathbb{R}^D$ such that $\mathcal{Z}^c = \mathbb{R}^D \setminus \mathcal{Z}$ has Lebesgue measure zero.*

The above result is standard (see, e.g. Ern & Guermond (2013, Chapter 2)).

To ensure convergence of integrals, we consider functions where the Lipschitz constant grows polynomially in the radius of the domain.

Definition A.7 (Polynomially Lipschitz)**.** We say that

- A function $\psi(\mathbf{z}) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is *polynomially-Lipschitz* if there are constants $a, b > 0$ such that for all radii $R \geq 1$ and all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{d_1}$ such that $\|\mathbf{z}\|, \|\mathbf{z}'\| \leq R$, $\|\psi(\mathbf{z}) - \psi(\mathbf{z}')\| \leq aR^b$.
- We say a function $\psi(\mathbf{z}; \mathbf{x}) : \mathbb{R}^{d_1} \times \mathbb{R}^n \to \mathbb{R}^{d_2}$ is *parametrized-polynomially-Lipschitz* if for all radii $R \geq 1$ and all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{d_1}$ and $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{z}\|, \|\mathbf{z}'\|, \|\mathbf{x}\| \leq R$, $\|\psi(\mathbf{z}; \mathbf{x}) - \psi(\mathbf{z}'; \mathbf{x})\| \leq aR^b$.

One can check that polynomially Lipschitz functions are locally Lipschitz.

A.2. Formal results

We now state our formal results. Throughout, our smoothing noise w has distribution p which has the following form.

Assumption A.8. The distribution p admits a density $p(\mathbf{w}) = e^{\alpha - \psi(\mathbf{w})}$, where

(a) $\psi(\mathbf{w}) \geq a\|\mathbf{w}\| - b$ for some constants $a > 0$ and $b \in \mathbb{R}$.

(b) ψ is twice differentiable everywhere, and $\nabla^2 \psi(\mathbf{w})$ has polynomial growth.

Example A.9 (Gaussian distribution). The cannonical example is the Gaussian distribution $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_n)$, where $p(\mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\|\mathbf{w}\|^2}{2\sigma^2})$. Here, $\psi(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2\sigma^2}$, which has polynomial growth and, being quadratic, is twice differentiable. In addition,

$$\nabla\psi(\mathbf{w}) = \frac{\mathbf{w}}{\sigma^2}, \quad \mathbb{E}[\nabla\psi(\mathbf{w})] = 0. \tag{7}$$

Zeroth-order unbiasedness. We now stipulate highly general conditions under which the zeroth-order estimator is unbiased. In the interest of generality, we allow time-varying policies and costs.

Definition A.10. We say that a tuple $(\rho, p, \phi, c_{1:H}; \pi_{1:H})$ is a *benign planning problem* if (a) ρ has finite moments (b) p satisfies Assumption A.8, (c) the dynamics $\phi(\cdot, \cdot)$ and costs $c_h(\cdot, \cdot)$ have polynomial growth (for all $h \in H$), and (d), for each $\mathbf{x} \in \mathbb{R}^n$ and $h \in [H]$, $\mathbf{u} \mapsto \pi_h(\mathbf{x}, \mathbf{u})$ is twice-differentiable in $\mathbf{u}$ and its second-order derivative has polynomial growth in $\mathbf{x}$. In addition, we assume $\phi, c_{1:H}, \pi_{1:H}$ are all Borel measurable.

We consider the resulting stochastic optimization objective.

$$F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H} \left[V_1(\mathbf{x}_h, \mathbf{w}_{1:H}, \boldsymbol{\theta}) \right] \\ \text{s.t. } \mathbf{x}_{h+1} = \phi(\mathbf{x}_h, \mathbf{u}_h), \quad \mathbf{u}_h = \pi(\mathbf{x}_h, \boldsymbol{\theta}) + \mathbf{w}_h.$$

Note that we define the expectation jointly over $\mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H}$, so as not to assume Fubini's theorem holds (even though, under our assumptions, it does). Our first result is a rigorous statement of the unbiasedness of the zeroth-order estimator.

Proposition A.11. *Suppose that $(\rho, p, \phi, c_{1:H}; \pi_{1:H})$ is a benign planning problem. Then, the objective $F(\boldsymbol{\theta})$ defined in Eq* (1) *is differentiable, and*

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H} \left[\sum_{h=1}^{H} (\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta}))^\intercal \psi(\mathbf{w}_h) V_h(\mathbf{x}_h, \mathbf{w}_{h:H}, \boldsymbol{\theta}) \right].$$

If, in addition $\mathbb{E}_{\mathbf{w} \sim p}[\nabla \psi(\mathbf{w})] = 0$, we also have

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H} \left[V_1(\mathbf{x}_h, \mathbf{w}_{1:H}, \boldsymbol{\theta}) \sum_{h=1}^{H} (\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta}))^\intercal \psi(\mathbf{w}_h) \right].$$

Eq (7) yields the following corollary for Gaussian distributions, which recovers Lemma 3.1 in the main text.

Corollary A.12. *In the special case where $p = \mathcal{N}(0, \sigma^2 I)$, we have*

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H} \left[V_1(\mathbf{x}_h, \mathbf{w}_{1:H}, \boldsymbol{\theta}) \sum_{h=1}^{H} (\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta}))^\intercal \mathbf{w}_h \right].$$

First-order unbiasedness under Lipschitzness. Next, we turn to the formal result under Lipschitzness. We consider objectives which have the following additional assumptions:

Definition A.13. We say that a tuple $(\rho, p, \phi, c_{1:H}; \pi_{1:H})$ is a *benign Lipschitz planning problem* if it is a benign planning problem, and in addition, (a) c_h and π_h are everywhere-differentiable and their derivatives have polynomial growth, and (b) ϕ is polynomially Lipschitz.

In addition, we require one more technical condition which ensures measurability of the set on which the analytic gradients are defined.

Definition A.14. We say that the distribution ρ is *decomposable* if there exists a Lebesgue-measurable function $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and a countable set of atoms $\mathbf{a}_1, \mathbf{a}_2, \ldots$, with weights $\nu_1, \nu_2, \ldots$ such that, for any $\mathcal{X} \subset \mathbb{R}^n$, $\Pr_{\mathbf{x}_1 \sim \rho}[\mathbf{x}_1 \in \mathcal{X}] = \int_{\mathcal{X}} \mu(\mathbf{x}_1) \mathrm{d}\mathbf{x}_1 + \sum_{i \geq 1} \mathbf{a}_i \nu_i$.

More general conditions can be established, but we adopt the above for simplicity. We assume that the distribution over initial state $\mathbf{x}_1 \sim \rho$ satisfies decomposability, which in particular encompasses the deterministic distribution over initial states considered in the body of the paper. The following lemma formalizes Lemma 3.2.

Proposition A.15. *Suppose that $(\rho, p, \phi, c_{1:H}; \pi_{1:H})$ is a benign Lipschitz planning problem. If ρ is decomposable, then*

(a) For each $\boldsymbol{\theta}$, there exists a set Lebesgue-measurable set $\mathcal{Z} \subset \mathbb{R}^{n+mH}$ such that $\Pr_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H}[(\mathbf{x}_1, \mathbf{w}_{1:H}) \in \mathcal{Z}] = 1$ and $\boldsymbol{\theta} \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$ is differentiable for all $(\mathbf{x}_1, \mathbf{w}_{1:H}) \in \mathcal{Z}$.

(b) $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H}[\nabla V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})]$, where expectations are taken in the sense of Eq (6).

If ρ is not necessarilly decomposable, but for given $\boldsymbol{\theta} \in \mathbb{R}^D$, the set $\{(\mathbf{x}_1, \mathbf{w}_{1:H}) : V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$ is differentiable$\}$ is Lebesgue measureable, then points (a) and (b) still hold.

Example A.16 (Piecewise linear). As an example, piecewise linear, or piecewise-polynomial dynamics statisfy the conditions of the above proposition.

A.2.1. Separable functions

A key step in establishing the unbiasedness of the zeroth-order estimator for policy optimization is the special case of separable functions. We begin by stating guarantees for simple functions which the noise enters in the following *separable* fashion.

Definition A.17 (Benign separability)**.** We say that a function $f(\boldsymbol{\theta}, \mathbf{w})$ has *benign separability* if there exists an everywhere differentiable function $g_{\text{in}}(\boldsymbol{\theta})$ and a Lebesgue measurable function $g_{\text{out}}(\cdot)$ with polynomial growth such that

$$f(\boldsymbol{\theta}, \mathbf{w}) = g_{\text{out}}(g_{\text{in}}(\boldsymbol{\theta}) + \mathbf{w}).$$

A slightly more general version of the above definition is as follows.

Definition A.18. We say that a $\mathbf{x}$-parameterized function function $f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x})$ has *parametrized benign separability* if there exists Lebesgue-measure functions $g_{\text{out}}(\cdot;\cdot)$ and $g_{\text{in}}(\cdot;\cdot)$ such that $g_{\text{in}}(\cdot;\cdot)$ is differentiable for all $\mathbf{x}$, and

$$f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x}) = g_{\text{out}}(g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x}) + \mathbf{w}; \mathbf{x}),$$

where (a) $(\mathbf{z}, \mathbf{x}) \mapsto g_{\text{out}}(\mathbf{z}; \mathbf{x})$ has polynomial growth, (b) for each $\boldsymbol{\theta}$, the mapping $\mathbf{x} \mapsto \mathrm{D}_{\boldsymbol{\theta}} g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x})$ has polynomial growth, $(\mathbf{z}, \mathbf{x}) \mapsto g_{\text{out}}(\mathbf{z}; \mathbf{x})$ has polynomial growth, and (c) for some $\epsilon_0 > 0$, there is a function $\tilde{g}(\mathbf{x})$ with polynomial growth such that such that for all $\Delta : \|\Delta\| \le \epsilon$,

$$\|g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x}) - g_{\text{in}}(\boldsymbol{\theta} + \Delta; \mathbf{x}) - \mathrm{D} g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x}) \cdot \Delta\| \le \|\Delta\|^2 \tilde{g}(\mathbf{x}). \tag{8}$$

We note that Eq (8) is satisfied when $g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x})$ has a second derivative by having polynomial growth.

The following gives an expression for the derivative of separable functions. Note that we do not require $g_{\text{out}}(\cdot)$ to be differentiable, and depend only on the derivatives of $\psi(\mathbf{w}) = \log p(\mathbf{w}) +$ const. from the density p, as well as the derivative of $g_{\text{in}}(\boldsymbol{\theta})$. The following statement establishes the well-known (Williams, 1992) computation at our level of generality.

Proposition A.19. *Suppose that p satisfies Assumption A.8. Then, if $f(\boldsymbol{\theta}, \mathbf{w})$ is benign separable, then the expectation $F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w} \sim p}[f(\boldsymbol{\theta}, \mathbf{w})]$ is well defined, differentiable, and has*

$$\nabla F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w} \sim p}[\mathrm{D} g_{\text{in}}(\boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}) \cdot f(\boldsymbol{\theta}, \mathbf{w})].$$

More generally, if ρ has finite moments and $f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x})$ has benign parametrized separability, then $F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{w} \sim p}[f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x})]$ satisfies

$$\nabla F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{w} \sim p}[\mathrm{D} g_{\text{in}}(\boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}) \cdot f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x})].$$

A.3. Proofs

A.3.1. Proof of Proposition A.19

Lemma A.20. *Let p be the distribution of $\mathbf{w}$ satisfying Assumption A.8 (and, by abuse of notation, its density with respect to the Lebesgue measure). Then, the following statements are true*

(a) The distribution p has finite moments. In particular, for any function $g(\cdot) : \mathbb{R}^m \to \mathbb{R}^d$ with polynomial growth, $\mathbb{E}_{\mathbf{w}\sim p}[\|g(\mathbf{w})\|] < \infty$. This only requires Assumption A.8 part (a).

(b) $\nabla\psi(\mathbf{w})$ has polynomial growth, and $\mathbb{E}[\nabla\psi(\mathbf{w})] = 0$.

(c) For any $B > 0$, there exists a function with polynomial growth such that $\tilde{g}(\cdot)$, for all $\Delta : \|\Delta\| \leq B$,

$$|p(\mathbf{w}) - p(\mathbf{w}+\Delta) - p(\mathbf{w})\langle -\nabla\psi(\mathbf{w}), \Delta\rangle| \leq \|\Delta\|^2 p(\mathbf{w}) \cdot \tilde{g}(\mathbf{w}).$$

(d) Let $g(\cdot,\cdot) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ have polynomial growth. Then $\mathbf{x} \mapsto \mathbb{E}_{\mathbf{w}} \nabla\psi(\mathbf{w}) \cdot g(\mathbf{w}, \mathbf{x})$ is well defined and polynomial growth in $\mathbf{x}$.

Proof. Since $p(\mathbf{w})$ decays exponentially in $\mathbf{w}$, p has finite moments. Thus, part (a) follows from Lemma A.4. Part (b) follows from the fundamental theorem of calculus:

$$\|\nabla\psi(\mathbf{w})\| = \left\| \int_0^1 \nabla^2 \psi(t\mathbf{w}) \cdot \mathbf{w} \mathrm{d}t \right\| \leq \|\nabla\psi(0)\| + \|\mathbf{w}\| \max_{t\in[0,1]} \|\nabla^2 \psi(t\mathbf{w})\|_{\mathrm{op}},$$

the upper bound on which has polynomial growth since $\|\nabla^2 \psi(t\mathbf{w})\|_{\mathrm{op}}$ does.

To prove part (c), we have that since $p(\mathbf{w}) = e^{\alpha - \psi(\mathbf{w})}$ for ψ differentiable

$$\nabla p(\mathbf{w}) = -p(\mathbf{w}) \cdot \nabla\psi(\mathbf{w}), \quad \nabla^2 p(\mathbf{w}) = p(\mathbf{w}) \cdot \underbrace{\left(\nabla\psi(\mathbf{w})\nabla\psi(\mathbf{w})^\intercal - \nabla^2 \psi(\mathbf{w})\right)}_{:= M(\mathbf{w})}.$$

Note that, by part (b) and the map that $\nabla^2 \psi(\mathbf{w})$ has polynomial growth, $M(\mathbf{w})$ has polynomial growth. Therefore, for any bound $B > 0$, the function $\tilde{g}(\mathbf{w}) := \sup_{\|\Delta\|\leq B} M(\mathbf{w}+\Delta)$ has polynomial growth. Finally, by the intermediate value theorem and for any $\Delta : \|\Delta\| \leq B$,

$$\begin{aligned} |p(\mathbf{w}) - p(\mathbf{w}+\Delta) - p(\mathbf{w})\langle\nabla\psi(\mathbf{w}), \Delta\rangle| &= |p(\mathbf{w}) - p(\mathbf{w}+\Delta) - \langle\nabla p(\mathbf{w}), \Delta\rangle| \\ &\leq \|\Delta\|^2 p(\mathbf{w}) M(\mathbf{w} + t\Delta), \quad \text{for some } t \in [0,1] \\ &\leq \|\Delta\|^2 p(\mathbf{w}) \tilde{g}(\mathbf{w}), \end{aligned}$$

as needed.

Part (d) is a consequence of part (b) and Lemma A.5. □

Proof of Proposition A.19. Consider the non-parametric case. Since $g_{\mathrm{out}}(\cdot)$ has polynomial growth, one can verify that $\mathbf{w} \mapsto f(\boldsymbol{\theta}, \mathbf{w})$ has polynomial growth. Hence the expectation $F(\boldsymbol{\theta})$ is well-defined by Lemma A.20,

part (a). We now prove that $F(\boldsymbol{\theta})$ is differentiable. Fix a $\boldsymbol{\theta}$, and let $\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \epsilon$.

$$\begin{aligned}
F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}') &= \int \big(f(\boldsymbol{\theta}, \mathbf{w}) - f(\boldsymbol{\theta}', \mathbf{w})\big)\, p(\mathbf{w}) \mathrm{d}\mathbf{w} \\
&= \int \big(g_{\text{out}}(g_{\text{in}}(\boldsymbol{\theta}) + \mathbf{w}) - g_{\text{out}}(g_{\text{in}}(\boldsymbol{\theta}') + \mathbf{w})\big)\, p(\mathbf{w}) \mathrm{d}\mathbf{w} \\
&= \int g_{\text{out}}(\underbrace{g_{\text{in}}(\boldsymbol{\theta}) + \mathbf{w}}_{\mathbf{w}_1}) p(\mathbf{w}) \mathrm{d}\mathbf{w} - \int g_{\text{out}}(\underbrace{g_{\text{in}}(\boldsymbol{\theta}') + \mathbf{w}}_{\mathbf{w}_2}) p(\mathbf{w}) \mathrm{d}\mathbf{w} \\
&= \int g_{\text{out}}(\mathbf{w}_1) p(\mathbf{w}_1 - g_{\text{in}}(\boldsymbol{\theta})) \mathrm{d}\mathbf{w}_1 - \int g_{\text{out}}(\mathbf{w}_2) p(\mathbf{w}_2 - g_{\text{in}}(\boldsymbol{\theta}')) \mathrm{d}\mathbf{w}_2 \\
&= \int \big(p(\mathbf{w} - g_{\text{in}}(\boldsymbol{\theta})) - p(\mathbf{w} - g_{\text{in}}(\boldsymbol{\theta}'))\big) \cdot g_{\text{out}}(\mathbf{w}) \mathrm{d}\mathbf{w} \\
&= \int \big(p(\mathbf{w}) - p(\mathbf{w} + g_{\text{in}}(\boldsymbol{\theta}) - g_{\text{in}}(\boldsymbol{\theta}'))\big) \cdot g_{\text{out}}(\mathbf{w} + g_{\text{in}}(\boldsymbol{\theta})) \mathrm{d}\mathbf{w} \\
&= \int \big(p(\mathbf{w}) - p(\mathbf{w} + g_{\text{in}}(\boldsymbol{\theta}) - g_{\text{in}}(\boldsymbol{\theta}'))\big) \cdot f(\boldsymbol{\theta}, \mathbf{w}) \mathrm{d}\mathbf{w} \\
&= \mathbb{E}_{\mathbf{w} \sim p} \left[\left(\frac{p(\mathbf{w}) - p(\mathbf{w} + g_{\text{in}}(\boldsymbol{\theta}) - g_{\text{in}}(\boldsymbol{\theta}'))}{p(\mathbf{w})} \right) \cdot f(\boldsymbol{\theta}, \mathbf{w}) \right].
\end{aligned}$$

Setting $\Delta = g_{\text{in}}(\boldsymbol{\theta}) - g_{\text{in}}(\boldsymbol{\theta}')$, Lemma A.20 implies that the remainder term enjoyes the following property.

$$R(\mathbf{w}) := p(\mathbf{w}) - p(\mathbf{w} + \Delta) - p(\mathbf{w}) \langle -\nabla \psi(\mathbf{w}), \Delta \rangle \text{ satisifies } |R(\mathbf{w})| \leq \|\Delta\|^2 \tilde{g}(\mathbf{w}) p(\mathbf{w}),$$

where $\tilde{g}(\mathbf{w})$ has polynomial growth, and thus $\tilde{g}(\mathbf{w}) \cdot f(\boldsymbol{\theta}, \mathbf{w})$ integrable under p. Thus, there exists a constant $C_w > 0$ such that

$$\big| \big(F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}')\big) - \mathbb{E}_{\mathbf{w} \sim p} \left[\langle -\nabla \psi(\mathbf{w}), \Delta \rangle f(\boldsymbol{\theta}, \mathbf{w}) \right] \big| \leq C_w \|\Delta\|^2,$$

where the integral on the right hand side exists because $\psi(\mathbf{w})$ and $f(\boldsymbol{\theta}, \mathbf{w})$ have polynomial growth. Simplying and dividing by $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ and substituting again$\Delta = g_{\text{in}}(\boldsymbol{\theta}) - g_{\text{in}}(\boldsymbol{\theta}')$,

$$\left| \frac{F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}') - \left\langle g_{\text{in}}(\boldsymbol{\theta}') - g_{\text{in}}(\boldsymbol{\theta}), \mathbb{E}_{\mathbf{w}} \left[\nabla \psi(\mathbf{w}) \cdot f(\boldsymbol{\theta}, \mathbf{w}) \right] \right\rangle}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} \right| \leq C_w \frac{\|g_{\text{in}}(\boldsymbol{\theta}') - g_{\text{in}}(\boldsymbol{\theta})\|^2}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|}. \tag{9}$$

The result now follows from taking $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \to 0$ and using differentiability of $g_{\text{out}}(\cdot)$ concludes.

Parametrized case. Now consider the parametrized case, and define $F(\boldsymbol{\theta}; \mathbf{x}) = \mathbb{E}_{\mathbf{w} \sim p} f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x})$. Then the analogue of Eq (10) holds pointwise for each $\mathbf{x}$:

$$\left| \frac{F(\boldsymbol{\theta}; \mathbf{x}) - F(\boldsymbol{\theta}'; \mathbf{x}) - \left\langle g_{\text{in}}(\boldsymbol{\theta}'; \mathbf{x}) - g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x}), \mathbb{E}_{\mathbf{w}} \left[\nabla \psi(\mathbf{w}; \mathbf{x}) \cdot f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x}) \right] \right\rangle}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} \right| \leq C_w \frac{\|g_{\text{in}}(\boldsymbol{\theta}'; \mathbf{x}) - g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x})\|^2}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|}. \tag{10}$$

Using Eq (8), the triangle inequality and Cauchy Schwartz, we obtain, for some integrable function $\tilde{g}$ with polynomial growth,

$$\begin{aligned}
&\left| \frac{F(\boldsymbol{\theta}; \mathbf{x}) - F(\boldsymbol{\theta}'; \mathbf{x}) - \left\langle \mathrm{D} g_{\text{in}}(\boldsymbol{\theta}; \mathbf{x})(\boldsymbol{\theta}' - \boldsymbol{\theta}), \mathbb{E}_{\mathbf{w}} \left[\nabla \psi(\mathbf{w}; \mathbf{x}) \cdot f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x}) \right] \right\rangle}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} \right| \\
&\qquad \leq \tilde{g}(\mathbf{x}) \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \cdot \mathbb{E}_{\mathbf{w}} \left\| \nabla \psi(\mathbf{w}) \cdot f(\boldsymbol{\theta}, \mathbf{w}; \mathbf{x}) \right\| + C_w \frac{\|g_{\text{out}}(\boldsymbol{\theta}'; \mathbf{x}) - g_{\text{out}}(\boldsymbol{\theta}; \mathbf{x})\|^2}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|}.
\end{aligned}$$

Applying Eq (8) again, we can bound

$$\begin{aligned}\|g_{\text{out}}(\boldsymbol{\theta}';\mathbf{x}) - g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\|^2 &= \left\|\left(g_{\text{out}}(\boldsymbol{\theta}';\mathbf{x}) - g_{\text{out}}(\boldsymbol{\theta};\mathbf{x}) - \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\right) + \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\right\|^2 \\ &= 2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\|^2 + 2\left\|g_{\text{out}}(\boldsymbol{\theta}';\mathbf{x}) - g_{\text{out}}(\boldsymbol{\theta};\mathbf{x}) - \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\right\|^2 \\ &\le 2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\|^2 + 2\tilde{g}(\mathbf{x})^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^4.\end{aligned}$$

Thus,

$$\begin{aligned}&\left|\frac{F(\boldsymbol{\theta};\mathbf{x}) - F(\boldsymbol{\theta}';\mathbf{x}) - \left\langle \mathrm{D}g_{\text{in}}(\boldsymbol{\theta};\mathbf{x})(\boldsymbol{\theta}'-\boldsymbol{\theta}), \mathbb{E}_{\mathbf{w}}\left[\nabla\psi(\mathbf{w};\mathbf{x})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\right]\right\rangle}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|}\right| \\ &\quad\le \tilde{g}(\mathbf{x})\cdot\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\cdot\|\mathbb{E}_{\mathbf{w}\sim p})\nabla\psi(\mathbf{w})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\| + 2C_w\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\|^2 + 2C_w\tilde{g}(\mathbf{x})^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^3.\end{aligned}$$

To conclude, observe that, by assumption, $\tilde{g}(\mathbf{x})$, $\mathbf{x}\mapsto\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})$, and (by Lemma A.20 part (d)) $\mathbf{x}\mapsto\mathbb{E}_{\mathbf{w}\sim p}\left[\nabla_{\mathbf{w}}\psi(\mathbf{w})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\right]$ all have polynomial growth. Thus, the fact that ρ has finite moments ensures that all terms have expectations under $\mathbf{x}\sim\rho$, and thus,

$$\begin{aligned}&\left|\frac{\mathbb{E}_{\mathbf{x}\sim\rho}[F(\boldsymbol{\theta};\mathbf{x}) - F(\boldsymbol{\theta}';\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\rho}[\left\langle \mathrm{D}g_{\text{in}}(\boldsymbol{\theta};\mathbf{x})(\boldsymbol{\theta}'-\boldsymbol{\theta}), \mathbb{E}_{\mathbf{w}}\left[\nabla\psi(\mathbf{w};\mathbf{x})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\right]\right\rangle]}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|}\right| \\ &\quad\le \mathbb{E}_{\mathbf{x}\sim\rho}\left|\frac{F(\boldsymbol{\theta};\mathbf{x}) - F(\boldsymbol{\theta}';\mathbf{x}) - \left\langle \mathrm{D}g_{\text{in}}(\boldsymbol{\theta};\mathbf{x})(\boldsymbol{\theta}'-\boldsymbol{\theta}), \mathbb{E}_{\mathbf{w}}\left[\nabla\psi(\mathbf{w};\mathbf{x})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\right]\right\rangle}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|}\right| \\ &\quad\le \mathbb{E}_{\mathbf{x}\sim\rho}\left[\tilde{g}(\mathbf{x})\cdot\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\cdot\|\mathbb{E}_{\mathbf{w}\sim p}\nabla\psi(\mathbf{w})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\| + 2C_w\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\|\nabla_{\boldsymbol{\theta}} g_{\text{out}}(\boldsymbol{\theta};\mathbf{x})\|^2 + 2C_w\tilde{g}(\mathbf{x})^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^3\right].\end{aligned}$$

Again, since all expectations are finite, the all terms on the last line above tend to zero as $\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\to 0$, so that

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}}\left(\mathbb{E}_{\mathbf{x}\sim\rho}[F(\boldsymbol{\theta};\mathbf{x})]\right) \\ &= \mathbb{E}_{\mathbf{x}\sim\rho}\left[\mathrm{D}g_{\text{in}}(\boldsymbol{\theta};\mathbf{x})^\intercal\mathbb{E}_{\mathbf{w}\sim p}\left[\nabla_{\mathbf{w}}\psi(\mathbf{w})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\right]\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{w}\sim p}\left[\nabla_{\boldsymbol{\theta}}\mathrm{D}g_{\text{in}}(\boldsymbol{\theta};\mathbf{x})^\intercal\nabla_{\mathbf{w}}\psi(\mathbf{w})\cdot f(\boldsymbol{\theta},\mathbf{w};\mathbf{x})\right]\end{aligned}$$

where in the last step, measurability and polynomial-growth conditions allow the application of Fubini's theorem. □

A.3.2. Proof of Proposition A.11

We prove a slightly different proof from that of the standard REINFORCE lemma to accommodate the fact that the state space is continuous, but the distribution over states may not have a density with respect to the Lebesgue measure. Instead, we adopt an approach based on the performance difference lemma (Kakade, 2003, Lemma 5.2.1).

To begin, define the expected cost to go function and expected costs

$$\bar{V}_h(\mathbf{x}_h,\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w}_{h:H}\overset{\text{i.i.d.}}{\sim}p} V(\mathbf{x}_h,\mathbf{w}_{h:H},\boldsymbol{\theta}) \tag{a}$$

$$V^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}',\mathbf{w}_h;\mathbf{x}_h) = \bar{V}_{h+1}(\phi(\mathbf{x}_h,\pi_h(\mathbf{x}_h,\boldsymbol{\theta}')+\mathbf{w}_t),\boldsymbol{\theta}) \tag{b.1}$$

$$\bar{V}^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}';\mathbf{x}_h) = \mathbb{E}_{\mathbf{w}_h\sim p} V^+_h(\boldsymbol{\theta}',\mathbf{w}_h;\mathbf{x}_h,\boldsymbol{\theta}). \tag{b.2}$$

$$c_h(\boldsymbol{\theta},\mathbf{w}_h;\mathbf{x}_h) = c_h(\mathbf{x}_h,\pi_h(\mathbf{x}_h,\boldsymbol{\theta})+\mathbf{w}_t), \tag{c.1}$$

$$\bar{c}_h(\boldsymbol{\theta};\mathbf{x}_h) = \mathbb{E}_{\mathbf{w}_h\sim p} c_h(\boldsymbol{\theta},\mathbf{w}_h;\mathbf{x}_h) \tag{c.2}$$

which describe (a) the expected cost-to-go under $\mathbf{x}_h, \boldsymbol{\theta}$, and (b) the expected cost-to-go from the next stage h after starting in state $\mathbf{x}_h$, acting according to $\boldsymbol{\theta}'$ in stage h, and subsequently acting according to $\boldsymbol{\theta}$, and (c) expected cost in state $\mathbf{x}_h$ under policy θ and. By the well known performance-difference lemma, we have

$$
\begin{aligned}
&F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}') \qquad (11)\\
&= \mathbb{E}_{\mathbf{x}_1 \sim \rho}\left[\bar{V}_1(\mathbf{x}_1, \boldsymbol{\theta}) - \bar{V}_1(\mathbf{x}_1, \boldsymbol{\theta}')\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\theta};h}[\left(\bar{c}_h(\boldsymbol{\theta}; \mathbf{x}_h) - \bar{c}_h(\boldsymbol{\theta}'; \mathbf{x}_h)\right) + \left(\bar{V}^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}; \mathbf{x}_h) - \bar{V}^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}'; \mathbf{x}_h)\right)] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\theta};h}\mathbb{E}_{\mathbf{w}_h \sim p}[c_h(\boldsymbol{\theta}, \mathbf{w}_h; \mathbf{x}_h) - c_h(\boldsymbol{\theta}', \mathbf{w}_h h \mathbf{x}_h)] + \mathbb{E}_{\mathbf{x}_h \sim \boldsymbol{\theta};h}\mathbb{E}_{\mathbf{w}_h \sim p}[V^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{w}_h; \mathbf{x}_h) - V^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}', \mathbf{w}_h; \mathbf{x}_h)] \\
&= \sum_{h=1}^{H} (F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}) - F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}')) + (F_{h,\boldsymbol{\theta};V}(\boldsymbol{\theta}) - F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}')) \qquad (12)
\end{aligned}
$$

where $\mathbb{E}_{\boldsymbol{\theta},h}$ denotes expectations over $\mathbf{x}_h$ under the dynamics

$$\mathbf{x}_1 \sim \rho, \quad \mathbf{x}_{t+1} = \phi(\mathbf{x}_t, \mathbf{u}_t), \quad \mathbf{u}_t = \pi(\mathbf{x}_t, \boldsymbol{\theta}) + \mathbf{w}_t,$$

and where we define

$$F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}') := \mathbb{E}_{\mathbf{x}_h \sim \boldsymbol{\theta};h}\mathbb{E}_{\mathbf{w}_h \sim p} c_h(\boldsymbol{\theta}', \mathbf{w}_h; \mathbf{x}_h)], \quad F_{h,\boldsymbol{\theta};V}(\boldsymbol{\theta}') := \mathbb{E}_{\mathbf{x}_h \sim \boldsymbol{\theta};h}\mathbb{E}_{\mathbf{w}_h \sim p} V^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}', \mathbf{w}_h \mid \mathbf{x}_h)].$$

Hence, if the functions $F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}')$ and $F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}')$ are differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$ for $h = 1, 2, \ldots, H$, Eq (12) implies

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \sum_{h=1}^{H} \left(\nabla_{\boldsymbol{\theta}'} F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}') + \nabla_{\boldsymbol{\theta}'} F_{h,\boldsymbol{\theta};V}(\boldsymbol{\theta}')\right)\big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}. \qquad (13)$$

We establish differentiability and compute the derivatives by appealing to Proposition A.19. First, we establish a couple of useful claims.

Claim A.21. *The marginal distribution over $\mathbf{x}_h$ under $\mathbb{E}_{h;\boldsymbol{\theta}}$ has all moments.*

Proof. Observe that the polynomial growth conditions on the dynamics map $\phi(\cdot, \cdot)$ imply that as a function, $\mathbf{x}_h = \mathbf{x}_h(\mathbf{x}_1, \mathbf{w}_{1:h-1})$, $\mathbf{x}_h$ has polynomial growth in $\mathbf{x}_h(\mathbf{x}_1, \mathbf{w}_{1:h-1})$. Thus, since the distributions over $\mathbf{x}_1$ and $\mathbf{w}_{1:h-1}$ have all moments, so does the distribution over $\mathbf{x}_h$. □

Claim A.22. *The function $\boldsymbol{\theta} \mapsto c_h(\boldsymbol{\theta}, \mathbf{w}_h; \mathbf{x}_h) = c_h(\mathbf{x}_h, \pi(\mathbf{x}_h; \boldsymbol{\theta}) + \mathbf{w}_t)$ satisfies benign parametrized separability.*

Proof. Take $g_{\mathrm{out}}(\cdot; \mathbf{x}_h) = c_h(\mathbf{x}_h, \cdot)$ and $g_{\mathrm{in}} = \pi(\mathbf{x}_h, \boldsymbol{\theta})$. Since $c_h(\cdot, \cdot)$ has polynomial growth, the requisite growth condition on $g_{\mathrm{out}}(\cdot, \cdot)$ holds. The polynomial growth in $\mathbf{x}$ of the second-order differentials of $\boldsymbol{\theta} \mapsto \pi_h(\mathbf{x}, \boldsymbol{\theta})$ implies that the first order differential of $\boldsymbol{\theta} \mapsto \pi_h(\mathbf{x}, \boldsymbol{\theta})$ has polynomial growth in $\mathbf{x}$, and that g_{in} also satisfies Eq (8) by Taylor's theorem. Hence, $g_{\mathrm{out}}, g_{\mathrm{in}}$ satisfy the requisite conditions. □

Claim A.23. *The function $\boldsymbol{\theta} \mapsto V^+_{h,\boldsymbol{\theta}_0}(\boldsymbol{\theta}, \mathbf{w}_h; \mathbf{x}_h) = \bar{V}_{h+1}(\phi(\mathbf{x}_h, \pi_h(\mathbf{x}_h, \boldsymbol{\theta}') + \mathbf{w}_t), \boldsymbol{\theta}_0)$ satisfies benign parametrized separability.*

Proof. Take $g_{\mathrm{out}}(\mathbf{u}; \mathbf{x}_h) = \bar{V}_{h+1}(\phi(\mathbf{x}_h, \mathbf{u}), \boldsymbol{\theta}_0)$ and $g_{\mathrm{in}} = \pi_h(\mathbf{x}_h, \boldsymbol{\theta})$. As shown in Claim A.22, g_{in} satisifes the requisite conditions for benign parametrized separability. To conclude, it suffices to show that $(\mathbf{u}, \mathbf{x}_h) \mapsto \bar{V}_{h+1}(\phi(\mathbf{x}_h, \mathbf{u}), \boldsymbol{\theta}_0)$ has polynomial growth. By Lemma A.5, it suffices to show that

$$(\mathbf{u}, \mathbf{x}_h, \mathbf{w}_{h+1:H}) \mapsto V_{h+1}(\phi(\mathbf{x}_h, \mathbf{u}), \mathbf{w}_{h+1:H}\boldsymbol{\theta}_0)$$

has polynomial growth. This holds since we have

$$V_{h+1}(\phi(\mathbf{x}_h, \mathbf{u}), \mathbf{w}_{h+1:H}, \boldsymbol{\theta}_0) = \sum_{i=h+1}^{H} c_i(\mathbf{x}_i, \pi_h(\mathbf{x}_i, \boldsymbol{\theta}_0) + \mathbf{w}_i, \quad \text{s.t. } \mathbf{x}_{i+1} = \phi(\mathbf{x}_i, \pi_h(\mathbf{x}_i, \boldsymbol{\theta}_0)) + \mathbf{w}_i.$$

Just as in the proof of Claim A.21, $\mathbf{x}_i, i > 1$ have polynomial growth when viewed as functions of $\mathbf{w}_{h+1:H}$, $\mathbf{x}_h$ and $\mathbf{u}$ (since the dynamics ϕ) have polynomial growth. Since c_i also have polynomial growth, we conclude $V_{h+1}(\phi(\mathbf{x}_h, \mathbf{u}), \mathbf{w}_{h+1:H}, \boldsymbol{\theta}_0)$ must as well. □

The above three claims allow us to invoke Proposition A.19, so that

$$\nabla_{\boldsymbol{\theta}'} F_{h,\boldsymbol{\theta};c}(\boldsymbol{\theta}') = \mathbb{E}_{\mathbf{x}_h \sim \boldsymbol{\theta};h} \mathbb{E}_{\mathbf{w}_h \sim p} \left[\mathrm{D}_{\boldsymbol{\theta}'} \pi_h(\mathbf{x}_h, \boldsymbol{\theta}')^\intercal \nabla \psi(\mathbf{w}_h) c_h(\boldsymbol{\theta}', \mathbf{w}_h; \mathbf{x}_h)] \right]$$
$$\nabla_{\boldsymbol{\theta}'} F_{h,\boldsymbol{\theta};V}(\boldsymbol{\theta}') = \mathbb{E}_{\mathbf{x}_h \sim \boldsymbol{\theta};h} \mathbb{E}_{\mathbf{w}_h \sim p} \left[\mathrm{D}_{\boldsymbol{\theta}'} \pi_h(\mathbf{x}_h, \boldsymbol{\theta}')^\intercal \nabla \psi(\mathbf{w}_h) \cdot V^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}', \mathbf{w}_h \mid \mathbf{x}_h) \right].$$

Therefore, from Eq (13), we conclude

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \sum_{h=1}^{H} \mathbb{E}_{\mathbf{x}_h \sim \boldsymbol{\theta};h} \mathbb{E}_{\mathbf{w}_h \sim p} \left[\mathrm{D}_{\boldsymbol{\theta}'} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) \left(c_h(\boldsymbol{\theta}', \mathbf{w}_h; \mathbf{x}_h) + V^+_{h,\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{w}_h \mid \mathbf{x}_h) \right)] \right].$$

Thus, the various polynomial growth conditions imply we can use Fubini's theorem (and the definition of $V^+_{h,\boldsymbol{\theta}}$) , so that the above is equal to

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) &= \sum_{h=1}^{H} \mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:H} \sim p^H} \left[\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) \left(c_h(\mathbf{x}_h, \mathbf{u}_h) + V_{h+1}(\mathbf{x}_h, \mathbf{w}_{h+1:H}, \boldsymbol{\theta}) \right)] \right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:H} \sim p^H} \left[\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) \cdot V_h(\mathbf{x}_h, \mathbf{w}_{h:H}, \boldsymbol{\theta})] \right] \\
&= \mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:H} \sim p^H} \left[\sum_{h=1}^{H} \mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) \cdot V_h(\mathbf{x}_h, \mathbf{w}_{h:H}, \boldsymbol{\theta}) \right].
\end{aligned}$$

This completes the first part of the proof. Next, we simplify in the special case where $\mathbb{E}_{\mathbf{w} \sim p}[\nabla \psi(\mathbf{w})] = 0$. Observe that the last line of the above display is equal to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:H} \sim p^H} \left[\sum_{h=1}^{H} \mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) \cdot V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}) \right] \\
&\quad \underbrace{- \mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:H} \sim p^H} \left[\sum_{h=1}^{H} \mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) \cdot \left(\sum_{i=1}^{h-1} c_i(\mathbf{x}_i, \mathbf{u}_i) \right) \right]}_{(b)},
\end{aligned}$$

where in the last line, we use that $V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}) = \sum_{i=1}^{h-1} c_i(\mathbf{x}_i, \mathbf{u}_i) = V_h(\mathbf{x}_h, \mathbf{w}_{1:H}, \boldsymbol{\theta})$. It suffices to show term (b) is zero. This follows since, for each $i < h$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:H} \sim p^H} \left[\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal \nabla \psi(\mathbf{w}_h) c_i(\mathbf{x}_i, \mathbf{u}_i) \right] \\
&\quad = \mathbb{E}_{\mathbf{x}_1 \sim \rho} \mathbb{E}_{\mathbf{w}_{1:h-1} \sim p^{h-1}} [\mathrm{D}_{\boldsymbol{\theta}} \pi_h(\mathbf{x}_h, \boldsymbol{\theta})^\intercal c_i(\mathbf{x}_i, \mathbf{u}_i)] \cdot \mathbb{E}_{\mathbf{w}_h \sim p}[\nabla \psi(\mathbf{w}_h)] = 0.
\end{aligned}$$

Here, we used that $\mathbf{w}_h$ is independent of $\mathbf{x}_1, \mathbf{w}_{1:h-1}$, and the assumption that $\mathbb{E}_{\mathbf{w}_h \sim p}[\nabla \psi(\mathbf{w}_h)] = 0$.

A.3.3. Proof of Proposition A.15

First, we establish almost-everywhere differentiability. Let $\tilde{\phi}_h$ denote the transitions under noise $\mathbf{w}$ and policy $\boldsymbol{\theta}$, defined as

$$\tilde{\phi}_h(\mathbf{x}, \mathbf{w}, \boldsymbol{\theta}) = \phi(\mathbf{x}, \pi_h(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{w}).$$

Set Φ_h to be their composition

$$\Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta}) = \tilde{\phi}_{h-1}(\cdot, \mathbf{w}_h, \boldsymbol{\theta}) \circ \tilde{\phi}_{h-2}(\cdot, \mathbf{w}_{h-1}, \boldsymbol{\theta}) \circ \cdots \circ \tilde{\phi}_1(\mathbf{x}_1, \mathbf{w}_1, \boldsymbol{\theta}).$$

Notice that $\mathbf{x}_h = \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta})$, where $\mathbf{x}_h$ is generated according to the dynamics $\mathbf{x}_{i+1} = \phi(\mathbf{x}_i, \pi_i(\mathbf{x}_i, \boldsymbol{\theta}) + \mathbf{w}_i)$. We now establish two key claims.

Claim A.24. *Fix $(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$. If $\boldsymbol{\theta}' \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta}')$ for all is differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$ for all $h \in [H]$, then $\boldsymbol{\theta}' \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}')$ is differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$.*

Proof. Defining $\tilde{c}_h(\mathbf{x}, \boldsymbol{\theta}; \mathbf{w}) = c_h(\mathbf{x}, \pi_h(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{w})$, we have

$$V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}) = \sum_{h=1}^{H} \tilde{c}_h(\Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta}), \boldsymbol{\theta}; \mathbf{w}_h).$$

Since both $c_h(\cdot, \cdot)$ and $\pi_h(\cdot, \cdot)$ are everywhere differentiable (jointly in their arguments), $(\mathbf{x}, \boldsymbol{\theta}) \mapsto \tilde{c}_h(\mathbf{x}, \boldsymbol{\theta}; \mathbf{w})$ is everywhere differentiable. Thus, under the assumptions of the claim, the composition $\boldsymbol{\theta}' \mapsto \tilde{c}_h(\Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta}'), \boldsymbol{\theta}'; \mathbf{w}_h)$ is differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$. □

The next claim provides a sufficient condition for Claim A.24 to hold.

Claim A.25. *Fix $(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$. Then if $\mathbf{w}'_{1:h-1} \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta})$ is differentiable at $\mathbf{w}'_{1:h-1} = \mathbf{w}_{1:h-1}$ for all $h \in [H]$, then $\boldsymbol{\theta}' \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta}')$ is differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$ for all $h \in [H]$.*

Proof. Fix $\mathbf{x}_1, \mathbf{w}_{1:h-1}$. Let $\boldsymbol{\delta} \in \mathbb{R}^d$ denote perturbations of $\boldsymbol{\theta}$. It suffices to show that, for each $h = 1, 2, \ldots, H$, the mapping $\Psi_h(\boldsymbol{\delta}) := \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta} + \boldsymbol{\delta})$ is differenitable if $\boldsymbol{\delta}$. By induction, it is straightforward to verify the identity

$$\Psi_h(\boldsymbol{\delta}) := \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta} + \boldsymbol{\delta}) = \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1} + \tilde{\mathbf{w}}_{1:h}(\boldsymbol{\delta})), \boldsymbol{\theta}) \tag{14}$$

where we have defined the noise term $\tilde{\mathbf{w}}_{1:h-1}(\boldsymbol{\delta})$ so as to transition from policy $\boldsymbol{\theta}$ to policy $\boldsymbol{\theta} + \boldsymbol{\delta}$:

$$\tilde{\mathbf{w}}_i(\boldsymbol{\delta}) = \pi_i(\Psi_i(\boldsymbol{\delta}), \boldsymbol{\theta} + \boldsymbol{\delta}) - \pi_i(\Psi_i(\boldsymbol{\delta}), \boldsymbol{\theta}). \tag{15}$$

We now argue by induction on little h that if $\mathbf{w}'_{1:i-1} \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:i-1}, \boldsymbol{\theta})$ for all $i \le H$, then $\Psi_i(\boldsymbol{\delta})$ is differentiable at $\boldsymbol{\delta} = 0$ for all $i \le h$.

For $h = 1$, both maps are the constant map $\Phi_1(\cdot, \cdot, \mathbf{x}_1) = \mathbf{x}_1$, so the result holds trivially. Now suppose the inductive hypothesis holds at some $h \ge 1$. Then, since each $\pi_i(\cdot, \cdot)$ is everywhere differentiable in its arguments, and since $\Psi_i(\boldsymbol{\delta})$ is differentiable at $\boldsymbol{\delta} = 0$ for all $i \le h$ by inductive hypothesis, $\tilde{\mathbf{w}}_i(\boldsymbol{\delta})$ defined in Eq (15) is differentiable at $\boldsymbol{\delta} = 0$ for each $i \le h$. Hence, $\tilde{\mathbf{w}}_{1:h}(\boldsymbol{\delta})$ is differentiable at $\boldsymbol{\delta} = 0$. Now, by assumption $\mathbf{w}'_{1:h} \mapsto \Phi_{h+1}(\mathbf{x}_1, \mathbf{w}'_{1:h}, \boldsymbol{\theta})$ is differentiable at $\mathbf{w}'_{1:h} = \mathbf{w}_{1:h}$. Therefore, at $\boldsymbol{\delta} = 0$, $\Psi_h(\boldsymbol{\delta})$ is given by the composition of two maps which are differentiable, and hence is differentiable. □

Define the set $\mathcal{W}_h(\mathbf{x}_1)$ as the set of $\mathbf{w}_{1:H} \in \mathbb{R}^{mH}$ such that the map $\mathbf{w}_{1:H} \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta})$ is differentiable (for simplicity, we augmented the map to be a function of all noises $\mathbf{w}_{1:H}$). Synthesizing Claims A.24 and A.25, we see that if $\mathbf{w}_{1:H} \in \bar{\mathcal{W}}(\mathbf{x}_1) := \bigcap_{h=1}^{H} \mathcal{W}_h(\mathbf{x}_1)$, then $\boldsymbol{\theta}' \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}')$ is differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$.

Furthermore, define $\mathcal{Z}_h$ as the set of $(\mathbf{x}_1, \mathbf{w}_{1:H}) \in \mathbb{R}^{d+mH}$ such that $(\mathbf{x}_1, \mathbf{w}_{1:H}) \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta})$ is differentiable. Here, we've just added $\mathbf{x}_1$ as a nuissance variable, so Claims A.24 and A.25 also imply that, on $\bar{\mathcal{Z}} := \bigcap_{h=1}^{H} \mathcal{Z}_h$, $\boldsymbol{\theta}' \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}')$ is differentiable at $\boldsymbol{\theta}' = \boldsymbol{\theta}$.

We invoke Rademacher's theorem. Since $\mathbf{w}_{1:H} \mapsto \Phi_h(\mathbf{x}_1, \mathbf{w}_{1:h-1}, \boldsymbol{\theta})$ is given by the composition of locally Lipschitz maps (note that differentiable maps are locally Lipschitz), Lemma A.6 implies that $\mathbb{R}^{mH} \setminus \mathcal{W}_h(\mathbf{x}_1)$ has Lebesgue measure zero for each h, so that $\mathbb{R}^{mH} \setminus \bar{\mathcal{W}}(\mathbf{x}_1)$ has Lebesgue measure zero by a union bound for each fixed $\mathbf{x}_1$. Similarly, $\mathbb{R}^{d+mH} \setminus \bar{\mathcal{Z}}$ has measure zero.

Proof under decomposability. Assume ρ is decomposable with atoms $\mathbf{a}_1, \mathbf{a}_2, \ldots$. Define the set

$$\mathcal{Z} := \bar{\mathcal{Z}} \cap \bigcap_{i \geq 1} \{\mathbf{a}_i\} \times \bar{\mathcal{W}}(\mathbf{a}_i).$$

The set $\mathcal{Z}$ is Lebesgue measurable because it is the intersection of Lebesgue measurable sets. Moroever, by the above discussion, $\boldsymbol{\theta} \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$ is differentiable everywhere on $\mathcal{Z}$. Lastly, one can verify by decomposability and the fact that $\bar{\mathcal{Z}}$ and $\mathcal{W}$ are the complement of Lebesgue measure-zero sets that $\Pr_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim \rho^H}[(\mathbf{x}_1, \mathbf{w}_{1:H}) \in \mathcal{Z}] = 1$. This proves part (a).

To prove part (b), one can use the polynomial Lipschitz conditions to verify that $\nabla_{\boldsymbol{\theta}} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})$, has polynomial growth wherever defined. Hence, its expectation (in the sense of Eq (6)) is well-defined. To prove part (b), one can verify that, via polynomial-Lipschitzness of the dynamics, policies and costs that the quotients satisfy

$$\frac{V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}) - V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta} + \boldsymbol{\delta})}{\|\boldsymbol{\delta}\|} \leq \mathrm{poly}(\|\mathbf{x}_1\|, \|\mathbf{w}_{1:H}\|).$$

Hence, the quotients are uniformly integrable, and one can apply the dominate convergence theorem to show that, for any sequence $\boldsymbol{\delta}_n \to 0$

$$\lim_{n \to \infty} \frac{F(\boldsymbol{\theta} + \boldsymbol{\delta}_n) - F(\boldsymbol{\theta})}{\|\boldsymbol{\delta}_n\|} = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H} \left[\lim_{n \to \infty} \frac{V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}) - V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta} + \boldsymbol{\delta}_n)}{\|\boldsymbol{\delta}_n\|} \right].$$

By considering $\boldsymbol{\delta}_n = t_n \mathbf{v}$ for a direction $\mathbf{v} \in \mathbb{R}^d$ and a sequence $t_n \to 0$, one can equate directional derivatives

$$\langle \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}), \mathbf{v} \rangle = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H}[\langle \nabla_{\boldsymbol{\theta}} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}), \mathbf{v} \rangle].$$

This proves that[1]

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_1 \sim \rho, \mathbf{w}_{1:H} \sim p^H}[\nabla_{\boldsymbol{\theta}} V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta})].$$

Proof under measurability assumption. Consider the set

$$\mathcal{Z}_0 := \{(\mathbf{x}_1, \mathbf{w}_{1:H}) \text{ s.t. } \boldsymbol{\theta}' \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}') \text{ is differentiable at } \boldsymbol{\theta}\}$$

and define its slices

$$\mathcal{Z}_0(\mathbf{x}_1) := \{\mathbf{w}_{1:H} \quad \text{s.t. } \boldsymbol{\theta}' \mapsto V_1(\mathbf{x}_1, \mathbf{w}_{1:H}, \boldsymbol{\theta}') \text{ is differentiable at } \boldsymbol{\theta}\}.$$

If we assume that $\mathcal{Z}_0$ is Lebesgue measurable, then by Fubini's theorem,

$$\Pr_{\mathbf{x} \sim \rho, \mathbf{w}_{1:H} \sim p^H}[(\mathbf{x}_1, \mathbf{w}_{1:H}) \in \mathcal{Z}_0] = \mathbb{E}_{\mathbf{x} \sim \rho} \Pr_{\mathbf{w}_{1:H} \sim p^H}[\mathbf{w}_{1:H} \in \mathcal{Z}_0(\mathbf{x}_1)].$$

[1] Note that , $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$ is differentiable by Proposition A.11, so we do not make the mistake of using existence of partial derivatives to imply differentiability.

Notice that, for any given $\mathbf{x}_1$, the above proof under decomposability shows that $\mathcal{Z}_0(\mathbf{x}_1) \supseteq \bar{\mathcal{W}}(\mathbf{x}_1)$, and thus the complement of $\mathcal{Z}_0(\mathbf{x}_1)$ in $\mathbb{R}^{mH}$ has Lebesgue measure zero. Hence $\Pr_{\mathbf{w}_{1:H}\sim p^H}[\mathbf{w}_{1:H} \in \mathcal{Z}_0(\mathbf{x}_1)] = 1$, so that $\Pr_{\mathbf{x}\sim\rho,\mathbf{w}_{1:H}\sim p^H}[(\mathbf{x}_1, \mathbf{w}_{1:H}) \in \mathcal{Z}_0] = 0$. This proves part (a). Part (b) follows by the same dominated convergence argument. □

B. Additional Proofs from Section 3

B.1. Proof of Lemma 3.5

Recall that empirical bias means there exists an event $\mathcal{E}$ such that $\|\mathbb{E}[\mathbf{z} \mid \mathcal{E}] - \mathbb{E}[\mathbf{z}]\| \geq \Delta$, and $\Pr[\mathcal{E}] \geq 1 - \beta$. Since the target lower bound increases as β decreases, we may assume that $\Pr[\mathcal{E}] = 1 - \beta$ with equality (since choosing a small β so that equality holds gives a larger variance lower bound). We begin

$$\begin{aligned}
\Delta &\leq \|\mathbb{E}[\mathbf{z} \mid \mathcal{E}] - \mathbb{E}[\mathbf{z}]\| \\
&= \|(1-\beta)^{-1}\mathbb{E}[\mathbf{z}\mathbb{I}\{\mathcal{E}\}] - \mathbb{E}[\mathbf{z}]\| \\
&\leq \|(1-\beta)^{-1}\mathbb{E}[\mathbf{z}\mathbb{I}\{\mathcal{E}\}] + (1-\beta)^{-1}\mathbb{E}[\mathbf{z}]\| - \|\mathbb{E}[\mathbf{z}]\| \cdot |1 - (1-\beta)^{-1}| \\
&\leq (1-\beta)^{-1}\|\mathbb{E}[\mathbf{z}\mathbb{I}\{\mathcal{E}^c\}]\| - \|\mathbb{E}[\mathbf{z}]\| \cdot |1 - (1-\beta)^{-1}|.
\end{aligned}$$

Rearranging, we have

$$\|\mathbb{E}[\mathbf{z}\mathbb{I}\{\mathcal{E}^c\}]\| \geq \Delta_0 := \max\{0, (1-\beta)\Delta - \beta\|\mathbb{E}[\mathbf{z}]\|\}.$$

And thus, since $\Pr[\mathcal{E}^c] = \beta$,

$$\|\mathbb{E}[\mathbf{z} \mid \mathcal{E}^c]\| \geq \frac{\Delta_0}{\beta}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{z}\|^2] &\geq \mathbb{E}[\|\mathbf{z}\|^2\mathbb{I}\{\mathcal{E}^c\}] \\
&= \Pr[\mathcal{E}^c] \cdot \mathbb{E}[\|\mathbf{z}\|^2 \mid \mathcal{E}^c] \\
&\geq \Pr[\mathcal{E}^c] \cdot \|\mathbb{E}[\mathbf{z} \mid \mathcal{E}^c]\|^2 \\
&\geq \beta \cdot \frac{\Delta_0^2}{\beta^2} = \frac{\Delta_0^2}{\beta}.
\end{aligned}$$

□

B.2. Proof of Lemma 3.10

Let's consider that $\hat{\nabla}^{[0]}$ estimator with a single sample, and drop the superscript i. We accommodate the general case with $\mathbf{x}_1 \sim \rho$. Since $\mathbf{Var}[\mathbf{z}] \leq \mathbb{E}[\|\mathbf{z}\|^2]$ for any random vector $\mathbf{z}$, we have

$$\begin{aligned}
\mathbf{Var}\left[\frac{1}{\sigma^2}V_1(\mathbf{x}_1, \bar{\mathbf{w}}^i, \boldsymbol{\theta})\left[\sum_{h=1}^{H}\nabla_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})^\top \mathbf{w}_h^i\right]\right] &\leq \mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left\|\frac{1}{\sigma^2}V_1(\mathbf{x}_1, \bar{\mathbf{w}}^i, \boldsymbol{\theta})\cdot\sum_{h=1}^{H}\mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_h^i, \boldsymbol{\theta})^\top \mathbf{w}_h^i\right\|_2^2 \\
&\leq \frac{B_V^2}{\sigma^4}\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left\|\sum_{h=1}^{H}\mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})^\top \mathbf{w}_h^i\right\|_2^2 \\
&= \frac{B_V^2}{\sigma^4}\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left[\sum_{h_1=1}^{H}\sum_{h_1=1}^{H}\left\langle \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_{h_1}, \boldsymbol{\theta})^\top \mathbf{w}_{h_1}^i, \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_{h_2}, \boldsymbol{\theta})^\top \mathbf{w}_{h_2}^i\right\rangle\right] \\
&= \frac{B_V^2}{\sigma^4}\sum_{h_1=1}^{H}\sum_{h_1=1}^{H}\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left[\left\langle \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_{h_1}, \boldsymbol{\theta})^\top \mathbf{w}_{h_1}^i, \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_{h_2}, \boldsymbol{\theta})^\top \mathbf{w}_{h_2}^i\right\rangle\right].
\end{aligned}$$

We claim that $\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left[\left\langle \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_{h_1}, \boldsymbol{\theta})^\top \mathbf{w}_{h_1}^i, \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_{h_2}, \boldsymbol{\theta})^\top \mathbf{w}_{h_2}^i\right\rangle\right] = 0$ unless $h_1 = h_2$. Suppose $h_1 \neq h_2$. Since inner products are symmetric, we may assume without loss of genearlity that $h_1 < h_2$. Then, $\mathbf{x}_{h_2}$, $\mathbf{x}_{h_1}$ and $\mathbf{w}_{h_1}$are all functions of $\mathbf{x}_1$ and $\mathbf{w}_{1:h_2-1}$,whereas $\mathbf{w}_2$ is independent of these. Hence, since $\mathbb{E}[\mathbf{w}_2] = 0$, the cross term vanishes. Thus, we are left with

$$\begin{aligned}
\mathbf{Var}\left[\frac{1}{\sigma^4}V_1(\mathbf{x}_1, \bar{\mathbf{w}}^i, \boldsymbol{\theta})\left[\sum_{h=1}^{H}\nabla_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})^\top \mathbf{w}_h^i\right]\right] &\leq \frac{B_V^2}{\sigma^4}\sum_{h=1}^{H}\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left[\left\langle \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})^\top \mathbf{w}_h^i, \mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})^\top \mathbf{w}_h^i\right\rangle\right] \\
&\leq \frac{B_V^2}{\sigma^4}\sum_{h=1}^{H}\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left[\|\mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})\|_{\mathrm{op}}\|\mathbf{w}_h^i\|^2\right] \\
&\leq \frac{B_V^2 B_{\pi^2}}{\sigma^4}\sum_{h=1}^{H}\mathbb{E}_{\bar{\mathbf{w}}^i, \mathbf{x}_1}\left[\ \|\mathbf{w}_h^i\|^2\right] = \frac{B_V^2 B_{\pi^2}}{\sigma^4}\cdot Hn\sigma^2 = \frac{HnB_V^2 B_\pi^2}{\sigma^2},
\end{aligned}$$

as needed. □

C. Interpolation

C.1. Bias and variance of the interpolated estimator

Here we describe the bias and variance of the interpolated estimator. The first is a straightforward consequence of linearity of expectation and the expectation computations in Eq (4).

Lemma C.1 (Interpolated bias)**.** *Assuming the costs and dynamics satisfies the conditions of Lemma 3.1 (formally, Corollary A.12), then for all $\alpha \in [0, 1]$,*

$$\mathbb{E}[\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta})] - \nabla F(\boldsymbol{\theta}) = \alpha\left(\mathbb{E}[\bar{\nabla}^{[1]}F(\boldsymbol{\theta})] - \nabla F(\boldsymbol{\theta})\right).$$

If in addition, the costs and dynamics satisfy the conditions of Lemma 3.2 (formally, Proposition A.15), then $\mathbb{E}[\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta})] = \nabla F(\boldsymbol{\theta})$.

Lemma C.2 (Interpolated variance)**.** *Assume that $\bar{\nabla}^{[1]}F(\boldsymbol{\theta})$ and $\bar{\nabla}^{[0]}F(\boldsymbol{\theta})$ are constructed using two independent sets of N trajectories. Then We have that*

$$\begin{aligned}\mathbf{Var}[\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta})] &= \alpha^2\mathbf{Var}[\bar{\nabla}^{[1]}F(\boldsymbol{\theta})] + (1-\alpha)^2\mathbf{Var}[\bar{\nabla}^{[0]}F(\boldsymbol{\theta})] \\ &= \frac{\alpha^2}{N}\mathbf{Var}[\hat{\nabla}^{[0]}F_i(\boldsymbol{\theta})] + \frac{(1-\alpha)^2}{N}\mathbf{Var}[\hat{\nabla}^{[1]}F_i(\boldsymbol{\theta})].\end{aligned}$$

Proof. Let $X = \bar{\nabla}^{[1]}F(\boldsymbol{\theta})$ and $Y = \bar{\nabla}^{[1]}F(\boldsymbol{\theta})$. Since the ZoBG and FoBG are assumed to use independent trajectories, X and Y are independent, and thus

$$\begin{aligned}\mathbf{Var}[\alpha X + (1-\alpha)Y] &= \mathbb{E}[\|\alpha(X - \mathbb{E}[X]) + (1-\alpha)(Y - \mathbb{E}[Y])\|^2] \\ &= \alpha^2\mathbb{E}\|X - \mathbb{E}[X]\|^2 + (1-\alpha)^2\mathbb{E}\|Y - \mathbb{E}[Y]\|^2 + \underbrace{\alpha\,\mathbb{E}[\langle X - \mathbb{E}[X], Y - \mathbb{E}[Y]\rangle]}_{=0} \\ &= \alpha^2\mathbf{Var}[X] + (1-\alpha)^2\mathbf{Var}[Y],\end{aligned}$$

which establishes the first equality. The second equality follows from decompsing each of $X = \bar{\nabla}^{[1]}F(\boldsymbol{\theta})$ and $Y = \bar{\nabla}^{[1]}F(\boldsymbol{\theta})$ as the empirical mean of N i.i.d random variables. □

The following lemma justifies using $\left(\alpha^2\hat{\sigma}_1^2 + (1-\alpha)^2\hat{\sigma}_2^2\right)$ as a proxy for the variance:

Lemma C.3 (Empirical variance)**.** *For $k = 0, 1$, we have*

$$\frac{1}{N}\mathbb{E}[\hat{\sigma}_k^2] = \mathbf{Var}[\bar{\nabla}^{[k]}].$$

Thus,

$$\mathbb{E}[\left(\alpha^2\hat{\sigma}_1^2 + (1-\alpha)^2\hat{\sigma}_2^2\right)] = N \cdot \mathbf{Var}[\bar{\nabla}^{[\alpha]}].$$

Proof. The first part of the lemma follows from a standard unbiasedness computation for a sample variance (see, e.g. Wasserman (2004, Theorem 3.17) for the scalar case). The second part of the lemma follows from Lemma C.2. □

C.2. Closed-form for interpolation

Recall Lemma 4.4: With $\gamma = \infty$, the optimal α is $\alpha_\infty := \frac{\sigma_0^2}{\sigma_1^2+\sigma_0^2}$. For finite $\gamma \geq \epsilon$, Eq (4) is

$$\alpha_\gamma := \begin{cases} \alpha_\infty & \text{if} \quad \alpha_\infty B \leq \gamma - \varepsilon \\ \frac{\gamma-\varepsilon}{B} & \text{otherwise .} \end{cases} \tag{16}$$

Proof. Intuitively, the objective is convex with a linear constraint, so meets its optimality either at the unconstrained minimum or at the constraint surface. This is implied by complementary slackness of the KKT conditions, since an optimal α^* satisfies:

$$\begin{aligned}2\alpha^*\hat{\sigma}_1^2 + 2(1-\alpha^*)\hat{\sigma}_0^2 + \lambda B &= 0 \\ \lambda(\epsilon - \gamma + \alpha^* B) &= 0,\end{aligned}$$

where the first line is stationarity of the Lagrangian and the second line is complementary slackness. Clearly, either $\lambda = 0$ and the minimum is met at the inverse-weighted solution of the variances, or the constraint is zero and we have $\alpha^* = (\gamma - \epsilon)/B$. □

C.3. Proof of Lemma 4.3

We give a more detailed proof of Lemma 4.3 here.

$$
\begin{aligned}
&\|\bar{\nabla}^{[\alpha]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \\
&= \|\alpha\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) + (1-\alpha)\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \\
&= \|\alpha\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) + (1-\alpha)\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \alpha\nabla F(\boldsymbol{\theta}) - (1-\alpha)\nabla F(\boldsymbol{\theta})\| \\
&\le (1-\alpha)\|\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| + \alpha\|\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \\
&\le (1-\alpha)\|\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| + \alpha\left(\|\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) - \bar{\nabla}^{[0]}F(\boldsymbol{\theta})\| + \|\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\|\right) \\
&\le \|\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| + \alpha\|\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) - \bar{\nabla}^{[0]}F(\boldsymbol{\theta})\| \\
&\le \epsilon + \alpha\|\bar{\nabla}^{[1]}F(\boldsymbol{\theta}) - \bar{\nabla}^{[0]}F(\boldsymbol{\theta})\| \\
&\le \gamma.
\end{aligned}
$$

C.4. Empirical Bernstein confidence

Here describe our confidence estimate based on the ZoBG. Recall that that ZoBG is

$$
\bar{\nabla}^{[0]}F(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}^{[0]}F_i(\boldsymbol{\theta}), \quad \text{where} \quad \hat{\nabla}^{[0]}F_i(\boldsymbol{\theta})\sum_{i=1}^{N}V_1(\mathbf{x}_1^i, \mathbf{w}_{1:H}^i, \boldsymbol{\theta}) \cdot \left[\sum_{i=1}^{H}\mathrm{D}_{\boldsymbol{\theta}}\pi(\mathbf{x}_h, \boldsymbol{\theta})^\top \mathbf{w}_h\right].
$$

Our estimate is based on the matrix Bernstein inequality due (see, e.g. (Tropp, 2015)) specified below.

Lemma C.4 (Matrix Bernstein inequality)**.** *Let $X_1, \ldots, X_N$ be N i.i.d random d-dimensional random* vectors *with $\|X_1 - \mathbb{E}[X_1]\| \le R$ almost surely, and $\mathbb{E}[\|X_1\|^2] \le \sigma^2$. Then,*

$$
\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N}X_i - \mathbb{E}[X]\right\| \ge t\right] \le (d+1)\exp\left(\frac{-Nt^2/2}{\sigma^2 + Rt/3}\right)
$$

Hence, with probability, for any $\delta > 0$,

$$
\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N}X_i - \mathbb{E}[X]\right\| \ge \sqrt{\frac{2\sigma^2\log\frac{d+1}{\delta}}{N}} + \frac{2R}{3N}\log\frac{d+1}{\delta}\right] \le 1-\delta.
$$

As stated, Lemma C.4 does not apply to our setting because (a) the variance of each $X_i := \hat{\nabla}^{[0]}F_i(\boldsymbol{\theta})$ is unknown, and (b) X_i are not uniformly bounded (due to the Gaussian noise $\mathbf{w}_h^i$ being unbounded.) We address point (a) by replacing $\mathbf{Var}[X_i]$ with the following empirical upper bound

$$
\bar{\sigma}_0^2 := \sum_i \|\hat{\nabla}^{[0]}F_i(\boldsymbol{\theta})\|^2 \ge \hat{\sigma}_0^2.
$$

To address point (b), we take R to be some educated guess on the problem using the gradient samples from the system (e.g. $R = \max_i \|\hat{\nabla}^{[0]}F_i(\boldsymbol{\theta}) - \bar{\nabla}^{[0]}F(\boldsymbol{\theta})\|$). In practice, since the confidence bound ϵ directly scales with R, and the user needs to set some threshold term γ on $\epsilon + \alpha B$, a guess on the scale of R is already decided by the user threshold γ. Thus, rather than viewing R as a rigorous absolute bound on the max deviation that we have to compute, we interpret it as a hyperparameter balancing how much we should be cautious against an extreme deviation outside the events covered by the variance term. We find that this approach, while not entirely rigorous, performs well in simulation. The following remark sketches how a rigorous confidence interval could be derived.

Remark C.5. For a statistically rigorous confidence interval, one would have to (a) control the error introduced by using an empirical estimate of the variance, and (b) control the non-boundedness of the X_i vectors. The first point could be addressed by generalizing the empirical Bernstein inequality (Maurer & Pontil, 2009) (which slightly inflates the confidence intervals to accomodate fluctuations in empirical variance) to vector-valued random variables. Point (b) can be handled by a truncation argument, leveraging the light-tails of Gaussian vectors. Nevertheless, we find that our naive approach which substitutes in the empirical variance for the true variance and our choice of R has good performance in simulation, so we do not pursue more complicated machinery. In fact, we conjecture that a more rigorous concentration bound may be overly conservative and worse in experiments.