SIGMOID CONTACT SMOOTHING RECOVERS A LOCAL BOUNDARY TERM IN FIRST-ORDER PATHWISE GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

First-order pathwise gradients through rigid-contact simulation can be biased: the return of a rollout may jump when a contact event switches on or off, and differentiating a fixed rollout never sees that jump. We analyze one deterministic remedy used in recent differentiable locomotion pipelines, in which hard contact-mode selection is replaced by a sigmoid gate on the contact impulse, and identify precisely which part of the true gradient the gate supplies. Near a single regular switching surface in the space of rollout randomness, the gradient of the expected return equals the expectation of the pathwise gradient plus a boundary integral over that surface, weighted by the jump of the return. This decomposition is classical in form, a transport (Reynolds) identity; we give a rigorous local proof of it under mild hypotheses, from which it follows that hard pathwise differentiation misses exactly the boundary term. Our main result concerns the gate: whenever the smoothed return depends on the contact activation through a differentiable interpolation whose endpoints reproduce the two hard branches, the gate part of the smoothed pathwise gradient converges in expectation to the boundary term as the gate hardens. In this limit, analytic gate smoothing supplies exactly the term that hard differentiation discards. We complement this with a finite-sharpness theory: the second moment of the gate term grows at most linearly in the sharpness, and under one extra smoothness hypothesis its bias and second moment have explicit expansions whose orders are set by the readout and the noise density at the surface, giving a bias–variance scaling law in the batch size; we illustrate the results numerically on a single-contact model, including a two-dimensional-noise configuration.

1 INTRODUCTION

Differentiable simulators promise first-order policy gradients of lower variance than zeroth-order estimates. Contact is where the promise breaks. A rigid-contact rollout switches between contact modes, and its return can jump when a small change of the parameters makes a contact event fire or not fire; automatic differentiation through a fixed rollout never sees that jump, so the average of the pathwise derivatives is not the derivative of the average return. Suh et al. (2022b) document this bias of first-order gradients under contact and interpolate toward a zeroth-order estimate when the first-order one falls outside a confidence interval around the zeroth-order one, a test Onoda et al. (2026) replace by a lighter one. In simulation optimization this is the classical failure of infinitesimal perturbation analysis at a discontinuity (Glasserman, 2004; Ho & Cao, 1983), repaired by conditioning in smoothed perturbation analysis (Fu & Hu, 1997) or by estimators that evaluate or eliminate the jump term exactly (Lakshtanov, 2026; Peng et al., 2018; Ren et al., 2025).

Practitioners respond by smoothing. Randomized smoothing (Le Lidec et al., 2022; Pang et al., 2023; Suh et al., 2022a) acts on the objective through the law of injected noise: the expected return becomes differentiable, but each rollout stays discontinuous and averaging pathwise derivatives remains biased. Analytic smoothing acts on each rollout through the simulator itself: Tassa & Todorov (2010) replace the complementarity problem of contact by its stochastic counterpart and obtain continuously differentiable dynamics; Schwarke et al. (2025) scale the contact impulse of a Moreau time-stepper by a sigmoid of the penetration depth with a sharpness parameter κ, and the locomo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

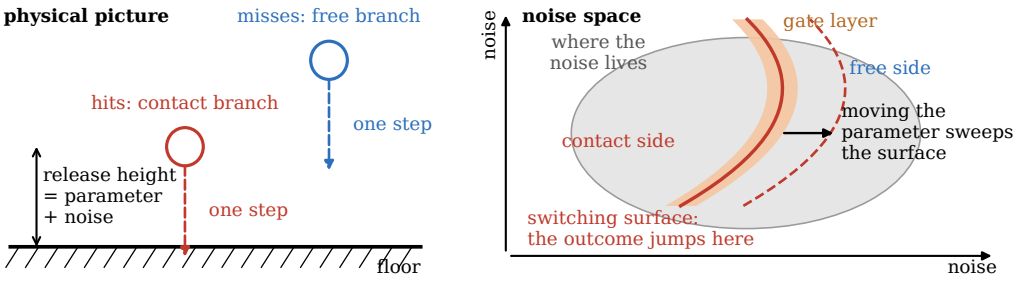

Figure 1: Left: a ball released above a floor; whether it hits the floor within the horizon, and hence the outcome, depends on the release height, the parameter plus random noise. Right: the same situation in noise space. The set where the outcome switches between the free and the contact branch is the switching surface, across which the outcome jumps. Moving the parameter sweeps this surface through the noise, and the true gradient picks up a term from the sweep, proportional to the jump, that differentiating any single rollout never sees. A sigmoid gate replaces the hard switch by a thin layer around the surface; the gradient through that layer recovers the missing term, in expectation, as the layer thins.

tion controllers learned this way transfer to hardware; Paulus et al. (2026) replace the hard case distinctions of collision detection by sigmoid-weighted interpolations between the cases; Song et al. (2024) avoid switching by differentiating a simplified continuous model. The impulse gate makes the contact impulse a differentiable function of the gap and gives informative gradients in practice, but it also changes what the gradient computes, and no existing analysis covers a gate inside the dynamics. This paper gives that analysis for one isolated surface; well-separated surfaces add up, and intersecting events are left open (Section 6).

The answer has two parts (Figure 1). Near a single regular switching surface in the space of rollout randomness, the gradient of the expected return equals the expected pathwise gradient plus a boundary integral over the surface, weighted by the jump of the return across it. The decomposition is classical in form: it is the boundary correction to reparameterization gradients (Lee et al., 2018), the edge term of differentiable rendering (Bangaru et al., 2020; Li et al., 2018; Loubet et al., 2019), and the pathwise-plus-boundary sensitivity of a discontinuous payoff (Liu & Hong, 2011). We prove a local version for the contact setting under C^1 hypotheses, and it shows that hard pathwise differentiation misses exactly the boundary term. The second part concerns the gate. When the smoothed return depends on the contact activation $a = \sigma(-\kappa g)$ through a differentiable function whose endpoints $a = 0$ and $a = 1$ reproduce the two hard branches, the chain rule splits the smoothed pathwise gradient into a frozen-activation term and a gate term, and our main result is that the expectation of the gate term converges to the boundary term as $\kappa \to \infty$. In expectation the gate term is a kernel smoother of the boundary term with bandwidth $1/\kappa$, the kernel estimator of Liu & Hong (2011) pushed inside the dynamics; the kernel is shaped by how the return depends on the activation and is in general neither symmetric nor positive.

Contributions. *Gate consistency.* For a general gated return, with the sigmoid inside the dynamics rather than in a payoff, the gate part of the smoothed pathwise gradient converges in expectation to the missing boundary term (Theorem 1). For a return affine in the activation the gate term coincides with the sigmoid-family IPA estimator of Yang & Zhang (2026); the general case covers returns nonlinear in the activation and a switching hypersurface in compactly supported noise with a bounded density, with a co-area density whose regularity we derive. *Finite-κ theory.* The second moment of the gate term grows at most linearly in κ (Proposition 2); under one additional smoothness hypothesis on the co-area densities the bias expands with leading order κ^{-r}, $r \in \{1, 2\}$, set by the readout and the noise density at the surface, and balancing bias against variance over a batch of N rollouts gives the scaling law $\kappa^*(N) \sim N^{1/(2r+1)}$. *Numerical validation.* On a single-contact Moreau stepper where the gated form is exact, the gate term matches the closed-form boundary term and a finite difference of the objective, and the measured bias, second-moment, and $\kappa^*(N)$ exponents reproduce the predictions for $r = 1$ and $r = 2$, also with two-dimensional noise, where both terms of the decomposition are nonzero.

2 BACKGROUND

2.1 SETUP

Let $\theta \in \mathbb{R}^d$ collect the parameters to be optimized and let $\xi \in \mathbb{R}^m$ be the randomness of one rollout, with density p. The objective is $F(\theta) = \mathbb{E}_\xi[L(\theta, \xi)] = \int_{\mathbb{R}^m} L(\theta, \xi)\, p(\xi)\, d\xi$, where $L(\theta, \xi)$ is the return of the rollout. If L were continuously differentiable in θ, the pathwise identity $\nabla_\theta F = \mathbb{E}_\xi[\nabla_\theta L]$ would hold and one backpropagation per rollout would give an unbiased gradient estimate; contact breaks this because the return is only piecewise smooth. We analyze the gradient at a fixed θ_0 near one switching surface in noise space, $\Sigma = \{\xi \in U : \ g(\theta_0, \xi) = 0\}$, defined by a scalar switching function g on a neighborhood U of the noise support, across which the return selects one of two branches, $L = L^+ \mathbf{1}\{g > 0\} + L^- \mathbf{1}\{g < 0\}$. In a contact simulator g is the signed gap, positive when the bodies are separated and equal to minus the penetration depth when they overlap; L^- is the return with the contact engaged ($g < 0$) and L^+ the return with the bodies separated ($g > 0$). The jump of the return across the surface is $[\![L]\!](\theta_0, \xi) = L^+(\theta_0, \xi) - L^-(\theta_0, \xi)$ for $\xi \in \Sigma$; if $[\![L]\!] \equiv 0$ the surface is a kink (continuous return, discontinuous derivative) rather than a jump. Below, $\mathcal{H}^{m-1}$ is the surface measure on $\mathbb{R}^m$ and $\|\cdot\|$ the Euclidean norm.

Assumption 1 (One regular switching surface)**.** There are a compact $K \supset \operatorname{supp} p$, an open neighborhood Θ of θ_0, and an open bounded neighborhood U of K such that $g, L^+, L^- \in C^1(\Theta \times U)$, $\nabla_\xi g(\theta_0, \xi) \neq 0$ for every $\xi \in K \cap \Sigma$, and $p \in L^\infty(\mathbb{R}^m)$ is continuous on a neighborhood of $K \cap \Sigma$.

In words: each branch is C^1, the surface is regular wherever the noise lives, and the noise is bounded with a density continuous near the surface; the itemized form (S1)–(S5) cited in the proofs is in Appendix A. The pathwise gradient of the hard return, $\nabla_\theta L = \nabla_\theta L^+ \mathbf{1}\{g > 0\} + \nabla_\theta L^- \mathbf{1}\{g < 0\}$, is defined p-almost everywhere and is what automatic differentiation through a fixed rollout computes.

2.2 THE TRANSPORT IDENTITY

Differentiating an integral whose integrand jumps across a moving surface produces a bulk term and a surface term, the Leibniz rule for moving domains (Flanders, 1973). We prove the local version we need under Assumption 1 alone (Appendix B).

Proposition 1 (One-surface transport identity)**.** *Under Assumption 1, the directional derivative* $D_u F(\theta_0) = \frac{d}{d\eta}\big|_{\eta=0} F(\theta_0 + \eta u)$ *exists for every* $u \in \mathbb{R}^d$ *and is linear in* u*, with*

$$\nabla_\theta F(\theta_0) = \int_{\mathbb{R}^m} \nabla_\theta L(\theta_0, \xi)\, p(\xi)\, d\xi + B, \qquad B := \int_\Sigma [\![L]\!]\, \frac{\nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1} \in \mathbb{R}^d, \tag{1}$$

all integrands evaluated at (θ_0, ξ)*.*

Corollary 1 (Bias of hard pathwise differentiation)**.** $\mathbb{E}_\xi[\nabla_\theta L(\theta_0, \xi)] = \nabla_\theta F(\theta_0) - B$*: the hard pathwise estimator is biased by exactly* $-B$*. If* $[\![L]\!] \equiv 0$ *on* Σ *(a kink), then* $B = 0$ *and it is unbiased.*

The weight $\nabla_\theta g / \|\nabla_\xi g\|$ is the normal speed at which θ moves the surface through the noise.

2.3 SIGMOID CONTACT GATING

The impulse gate of Schwarke et al. (2025) scales the impulse of a contact with gap g by the activation $a_\kappa(g)$, where

$$\sigma(t) = \frac{1}{1 + e^{-t}}, \qquad a_\kappa(s) = \sigma(-\kappa s), \qquad a'_\kappa(s) = -\kappa\, \sigma'(-\kappa s), \qquad \sigma' = \sigma(1 - \sigma) \in \left(0, \tfrac{1}{4}\right], \tag{2}$$

so that $a_\kappa \to 1$ on the engaged side $g < 0$ and $a_\kappa \to 0$ on the separated side $g > 0$ as $\kappa \to \infty$. The gate derivative has height $O(\kappa)$ on a layer of width $O(1/\kappa)$ around the surface and decays like $e^{-\kappa|s|}$ outside it (Appendix A).

3 GATE CONSISTENCY

3.1 THE GATED FORM AND THE GATE TERM

Assumption 2 (Local gated form)**.** There is an open neighborhood of K contained in U such that, for every $\theta \in \Theta$ and every ξ in that neighborhood, the smoothed return has the form

$$L_\kappa(\theta, \xi) \;=\; \Lambda\big(a_\kappa(g(\theta, \xi)), \theta, \xi\big), \tag{3}$$

where $\Lambda : [0, 1] \times \Theta \times U \to \mathbb{R}$ is continuous, continuously differentiable in its first argument a, with $\partial_a \Lambda$ jointly continuous in (a, θ, ξ). Moreover the two activation endpoints reproduce the two hard branches on the surface at θ_0:

$$\Lambda(0, \theta_0, \xi) = L^+(\theta_0, \xi), \qquad \Lambda(1, \theta_0, \xi) = L^-(\theta_0, \xi), \qquad \xi \in \Sigma. \tag{4}$$

Finally, $\partial_a \Lambda(\cdot, \theta_0, \cdot)$ is bounded on $[0, 1] \times K$ and $\nabla_\theta g(\theta_0, \cdot)$ is bounded on K.

Assumption 2 is the abstraction of an impulse gate. The map $a \mapsto \Lambda(a, \theta, \xi)$ is the rollout return as a function of the contact-activation level a, with the explicit dependence on θ and the realized randomness ξ held fixed; the endpoints $a \in \{0, 1\}$ recover the two hard counterfactuals. The substantive content is that Λ be C^1 in a: interpolating the activation produces a differentiable family of returns. The boundedness conditions are automatic from continuity on the compact set.

Wherever Λ is in addition differentiable in θ, the chain rule applied to (3) splits the pathwise gradient of L_κ into two pieces,

$$\nabla_\theta L_\kappa(\theta, \xi) = \underbrace{\partial_\theta \Lambda\big(a_\kappa(g), \theta, \xi\big)}_{\text{frozen-activation bulk}} \;+\; \underbrace{\partial_a \Lambda\big(a_\kappa(g), \theta, \xi\big)\, a'_\kappa(g)\, \nabla_\theta g}_{\text{gate derivative}}, \tag{5}$$

and we name the gate derivative at θ_0:

$$Z_\kappa(\xi) \;:=\; \partial_a \Lambda\big(a_\kappa(g(\theta_0, \xi)), \theta_0, \xi\big)\; a'_\kappa\big(g(\theta_0, \xi)\big)\; \nabla_\theta g(\theta_0, \xi) \;\in\; \mathbb{R}^d. \tag{6}$$

The gate derivative Z_κ in (6) is well-defined from $\partial_a \Lambda$ alone, with no derivative of Λ in θ; consequently the headline theorem below, which concerns only $\lim_{\kappa\to\infty} \mathbb{E}[Z_\kappa]$, needs only the C^1-in-a regularity of Assumption 2, while the full decomposition (5) (and its frozen-activation bulk term) enters only in the full-gradient corollary. Automatic differentiation through the smoothed simulator returns the whole of (5), both terms; the gate derivative Z_κ is precisely the term that has no analogue in hard branch selection, which differentiates only within the frozen contact mode.

The limit is computed on the level sets of the switching function. By Assumption 1 there are an open neighbourhood V of $K \cap \Sigma$ on which $\|\nabla_\xi g(\theta_0, \cdot)\| \geq c_0 > 0$ and a $\delta_0 > 0$ such that every $\xi \in K$ with $|g(\theta_0, \xi)| < \delta_0$ lies in V (Lemma 2, Appendix A). For $|s| < \delta_0$ and $a \in (0, 1)$ define the co-area density of the gate weight,

$$n(s, a) \;:=\; \int_{\{g(\theta_0, \cdot) = s\} \cap V} \partial_a \Lambda(a, \theta_0, \xi)\, \nabla_\theta g(\theta_0, \xi)\, \frac{p(\xi)}{\|\nabla_\xi g(\theta_0, \xi)\|}\, d\mathcal{H}^{m-1}(\xi) \;\in\; \mathbb{R}^d. \tag{7}$$

It is bounded and, for fixed a, continuous at $s = 0$; unlike the gated form of Assumption 2 this is not an independent hypothesis but follows from the standing assumptions (Proposition 4, Appendix C).

3.2 THE GATE TERM CONVERGES TO THE BOUNDARY TERM

Theorem 1 (The sigmoid gate term converges to the boundary term)**.** *Under Assumptions 1 and 2,*

$$\lim_{\kappa\to\infty} \mathbb{E}_\xi[Z_\kappa] \;=\; \int_\Sigma [\![L]\!](\theta_0, \xi)\, \frac{\nabla_\theta g(\theta_0, \xi)}{\|\nabla_\xi g(\theta_0, \xi)\|}\, p(\xi)\, d\mathcal{H}^{m-1}(\xi) \;=\; B.$$

Proof idea and kernel view. Split $\mathbb{E}[Z_\kappa]$ at $|g| = \delta_0$. Outside, $|a'_\kappa| \leq \kappa e^{-\kappa\delta_0}$ makes the contribution exponentially small; inside, the co-area formula along the level sets of g and the substitution $a = \sigma(-\kappa s)$ give

$$\mathbb{E}[Z_\kappa] = -\int_{\sigma(-\kappa\delta_0)}^{\sigma(\kappa\delta_0)} n\Big(-\tfrac{\sigma^{-1}(a)}{\kappa}, a\Big)\, da + O(\kappa e^{-\kappa\delta_0}), \qquad \sigma^{-1}(a) = \log \tfrac{a}{1-a}. \tag{8}$$

As $\kappa \to \infty$ the window fills $(0,1)$ and, by the continuity of n at $s=0$ and dominated convergence, the integral tends to $-\int_0^1 n(0,a)\,da$, which equals B because $\int_0^1 \partial_a \Lambda\,da = \Lambda(1)-\Lambda(0) = -[\![L]\!]$ on Σ by the endpoint condition (4) (Appendix D). Identity (8) is the kernel view: the gate term reads the co-area density on the level set $g=-\sigma^{-1}(a)/\kappa$, a smoother of bandwidth $1/\kappa$ whose weights, $\partial_a\Lambda$ along the gate, integrate to $-[\![L]\!]$. In one dimension, where $n(s,a)=\partial_a\Lambda(a)\,p(\xi_\star+s)$, it is a convolution, $\mathbb{E}[Z_\kappa] = [\![L]\!]\,(k_\kappa * p)(\xi_\star)$ with $k_\kappa(s) = -\partial_a\Lambda(\sigma(\kappa s))\,\kappa\sigma'(\kappa s)/[\![L]\!]$ and $\int k_\kappa = 1$: a kernel estimator of the boundary term in the sense of Liu & Hong (2011), with a kernel shaped by the readout, symmetric when $\partial_a\Lambda$ is constant and in general neither symmetric nor positive.

Corollary 2 (Consistency of the full smoothed gradient)**.** *Suppose, in addition to Assumptions 1 and 2, that Λ is jointly continuously differentiable in (a,θ), with $\partial_\theta\Lambda$ jointly continuous in (a,θ,ξ) (so that, with the gated form holding on a neighbourhood of K, the decomposition (5) is valid throughout $\operatorname{supp} p$), and that the branch-matching (4) extends off the surface, i.e. $\Lambda(0,\theta,\xi)=L^+(\theta,\xi)$ and $\Lambda(1,\theta,\xi)=L^-(\theta,\xi)$ for (θ,ξ) in a neighbourhood of $\{\theta_0\}\times K$. Then*

$$\mathbb{E}_\xi[\nabla_\theta L_\kappa(\theta_0,\xi)] \xrightarrow[\kappa\to\infty]{} \nabla_\theta F(\theta_0).$$

For Λ affine in a, Z_κ coincides with the sigmoid-family IPA estimator of Yang & Zhang (2026), a logistic sigmoid of sharpness κ (in our notation) in place of the indicator of a discontinuous payoff; under Gaussian noise they prove asymptotic unbiasedness, an $O(\kappa^{-2})$ bound on the boundary part of the bias, and an $O(N^{-4/5})$ mean-squared error in the sample size N at a plug-in κ built from kernel-density estimates at the surface. Theorem 1, the case $C_1=0$ of Remark 1, and the case $r=2$ of Corollary 3 are the counterparts under compactly supported noise; a nonlinear Λ is what makes the kernel and the order r readout-dependent, and we give a scaling law rather than a plug-in rule.

4 Finite-κ theory

Theorem 1 identifies the limit of the gate term; this section quantifies the cost of approaching it. An upper bound on the second moment needs nothing beyond the standing assumptions. Exact expansions of bias and second moment, and the tradeoff they imply, need one more order of smoothness of the co-area densities: alongside n of (7) define, on the same domain, the scalar co-area density of the squared gate weight,

$$m(s,a) \;:=\; \int_{\{g(\theta_0,\cdot)=s\}\cap V} \partial_a\Lambda(a,\theta_0,\xi)^2\,\|\nabla_\theta g(\theta_0,\xi)\|^2\,\frac{p(\xi)}{\|\nabla_\xi g(\theta_0,\xi)\|}\,d\mathcal{H}^{m-1}(\xi), \tag{9}$$

bounded on $(-\delta_0,\delta_0)\times(0,1)$ by the same argument as n (Appendix C).

Proposition 2 (Gate-gradient variance bound)**.** *Under Assumptions 1 and 2, $\mathbb{E}\|Z_\kappa\|^2 = O(\kappa)$ as $\kappa\to\infty$; consequently, for the empirical mean $\widehat{Z}_{\kappa,N} = \frac{1}{N}\sum_{i=1}^N Z_\kappa(\xi_i)$ of N independent draws $\xi_i\sim p$, $\operatorname{tr}\operatorname{Cov}(\widehat{Z}_{\kappa,N}) = O(\kappa/N)$.*

Assumption 3 (Smooth co-area densities)**.** For every $a\in(0,1)$ the maps $s\mapsto n(s,a)$ and $s\mapsto m(s,a)$ are C^2 on $(-\delta_0,\delta_0)$, and $\partial_s n, \partial_s^2 n, \partial_s m, \partial_s^2 m$ are bounded on $(-\delta_0,\delta_0)\times(0,1)$.

Unlike Assumptions 1 and 2, this is a hypothesis on derived objects, stated on the densities because that is what the proof uses; in both configurations of Section 5 the densities are explicit and C^∞ in s.

Proposition 3 (Bias and second-moment expansions)**.** *Under Assumptions 1, 2, and 3, as $\kappa\to\infty$,*

$$\mathbb{E}[Z_\kappa] - B = \frac{C_1}{\kappa} + O(\kappa^{-2}), \qquad \mathbb{E}\|Z_\kappa\|^2 = V_1\,\kappa + V_0 + O(\kappa^{-1}), \tag{10}$$

with

$$\begin{aligned} C_1 &= \int_0^1 \sigma^{-1}(a)\,\partial_s n(0,a)\,da, \\ V_1 &= \int_0^1 a(1-a)\,m(0,a)\,da, \qquad V_0 = -\int_0^1 a(1-a)\,\sigma^{-1}(a)\,\partial_s m(0,a)\,da, \end{aligned} \tag{11}$$

where the integrals converge because $\sigma^{-1}\in L^1(0,1)$ and the derivatives are bounded. If moreover $s\mapsto n(s,a)$ is C^3 with $\partial_s^3 n$ bounded on $(-\delta_0,\delta_0)\times(0,1)$, then $\mathbb{E}[Z_\kappa]-B = C_1/\kappa + C_2/\kappa^2 + O(\kappa^{-3})$ with $C_2 = -\frac{1}{2}\int_0^1 \sigma^{-1}(a)^2\,\partial_s^2 n(0,a)\,da$.

The proof (Appendix E) Taylor expands n in (8), and m in the analogous identity $\mathbb{E}\|Z_\kappa\|^2 = \kappa \int a(1-a)\, m(-\sigma^{-1}(a)/\kappa, a)\, da + O(\kappa^2 e^{-2\kappa\delta_0})$, in their first argument; the moments $\int_0^1 |\sigma^{-1}(a)|^k\, da$ of the logistic law are finite, which makes every remainder of the stated order.

Corollary 3 (Bias–variance scaling law)**.** *Under the hypotheses of Proposition 3, suppose $V_1 > 0$ and let $b_r \kappa^{-r}$ be the leading term of the bias: $r = 1$ and $b_1 = C_1$ if $C_1 \neq 0$, or, under the C^3 clause, $r = 2$ and $b_2 = C_2$ if $C_1 = 0 \neq C_2$. Then the mean-squared error $\mathrm{MSE}(\kappa, N) := \mathbb{E}\|\widehat{Z}_{\kappa,N} - B\|^2$ of the N-sample mean satisfies*

$$\mathrm{MSE}(\kappa, N) = \|b_r\|^2 \kappa^{-2r}\big(1 + O(\kappa^{-1})\big) + \frac{V_1\kappa + O(1)}{N}, \tag{12}$$

its two leading terms have minimal sum at $\kappa_N = (2r\|b_r\|^2 N/V_1)^{1/(2r+1)} \propto N^{1/(2r+1)}$, where both are $\Theta(N^{-2r/(2r+1)})$, as is $\mathrm{MSE}(\kappa_N, N)$.

Remark 1 (What the scaling law is and is not). Corollary 3 balances the two leading terms of (12); it is a scaling law in the batch size, not a rule for choosing κ, whose constants b_r and V_1 are unknown in practice. The order r is set jointly by the readout and the noise. If the return is affine in the activation, $\partial_a \Lambda$ does not depend on a, so neither do n and m, and $C_1 = V_0 = 0$ for every noise density, because σ^{-1} is antisymmetric about $a = \frac{1}{2}$ while $a(1-a)$ is symmetric; the bias is then $O(\kappa^{-2})$, the symmetric kernel of (8).

5 Experiments

We illustrate the theorems on the smallest model that realizes Assumption 2 literally. The rollout integrates a ball moving in the vertical plane under gravity, released at height $\theta \in \mathbb{R}$ above a fixed plane and undergoing a single inelastic, fully sticking ("all-stick") impact resolved by a Moreau time-stepper, with the contact impulse gated by $a_\kappa(g_n) = \sigma(-\kappa g_n)$ of the instantaneous normal gap g_n; two scalar readouts of the post-impact state are recorded, the horizontal velocity and the kinetic energy. In the *one-step* instance (a single integration step, released from rest in the normal direction) the rollout randomness is a scalar perturbation $\xi \in \mathbb{R}$ of the release height, drawn from a compactly supported C^∞ mollifier bump density p of radius $R = 0.02$ centered at ξ_c, and the base parameter is $\theta_0 = 0$. The gap is exactly $g = \theta + \xi$, the switching "surface" is the single point $\Sigma = \{\xi_\star\}$ with $\xi_\star = -\theta_0 = 0$, and the smoothed return is *exactly* the gated form $L_\kappa = \Lambda(a_\kappa(g), \theta, \xi)$ of (3) with a κ-independent Λ: Assumption 2 holds with equality, not as an approximation. Because the noise is one-dimensional, the boundary term (1) reduces to the point value $B = [\![L]\!](\theta_0, \xi_\star)\, p(\xi_\star)$, and since θ and ξ enter the gap only through $\theta + \xi$, $\nabla_\theta g = \nabla_\xi g$ and the geometric factor $\nabla_\theta g / \|\nabla_\xi g\|$ equals 1 exactly. Appendix F reports the same measurements on a 20-step unroll of the stepper. *Predicted rates.* In the one-step instance the return depends on the randomness only through the activation, $\Lambda = \Lambda(a)$, with

$$\Lambda_{\mathrm{vel}}(a) = (1-a)\, w, \qquad \Lambda_{\mathrm{KE}}(a) = \tfrac{1}{2}(1-a)^2\big(w^2 + u_N^2\big),$$

w the horizontal launch speed and u_N the normal impact speed; both reproduce the hard branches, $\Lambda(0) = L^+$ and $\Lambda(1) = 0 = L^-$ (the all-stick impact zeroes the post-impact state), with jumps $[\![L]\!] = w$ and $[\![L]\!] = \frac{1}{2}(w^2 + u_N^2)$. Since $\{g(\theta_0, \cdot) = s\} = \{s\}$, the co-area densities are $n(s,a) = \partial_a \Lambda(a)\, p(s)$ and $m(s,a) = \partial_a \Lambda(a)^2\, p(s)$, so Assumption 3 and the C^3 clause hold for the C^∞ bump, and the constants of Proposition 3 factor into a density part and a readout part,

$$C_1 = p'(\xi_\star)\, J, \qquad J := \int_0^1 \partial_a \Lambda(a)\, \sigma^{-1}(a)\, da, \qquad C_2 = -\tfrac{1}{2}\, p''(\xi_\star) \int_0^1 \partial_a \Lambda(a)\, \sigma^{-1}(a)^2\, da,$$

and $V_1 = p(\xi_\star) \int_0^1 \partial_a \Lambda(a)^2\, a(1-a)\, da > 0$, so the $O(\kappa)$ of Proposition 2 is attained. Because $\int_0^1 \sigma^{-1}(a)\, da = 0$ and $\int_0^1 a\, \sigma^{-1}(a)\, da = \frac{1}{2}$, a constant part of $\partial_a \Lambda$ contributes nothing to J and a linear part $c_1 a$ contributes $\frac{1}{2} c_1$. For velocity $\partial_a \Lambda = -w$ is constant, so $C_1 = 0$ for every bump while $C_2 = \frac{\pi^2}{6}\, w\, p''(\xi_\star)$ is nonzero for the bump used: $r = 2$. For kinetic energy $\partial_a \Lambda = -(1-a)(w^2 + u_N^2)$ has linear part $c_1 = w^2 + u_N^2$, so $C_1 = \frac{1}{2}(w^2 + u_N^2)\, p'(\xi_\star)$, nonzero for the off-centre bump $\xi_c = \xi_\star - R/2$: $r = 1$. The factorization $C_1 = p'(\xi_\star) J$ has two independent zeros, hence two routes to $r = 2$, the readout's ($J = 0$) and the bump's ($p'(\xi_\star) = 0$, the centred bump), and we exhibit both.

For the second moment, $V_0 = 0$ for velocity (Remark 1) while $V_0 \neq 0$ for kinetic energy with the off-centre bump, so $\mathbb{E}\|Z_\kappa\|^2/(V_1\kappa) - 1$ decays like κ^{-2} and κ^{-1} respectively. Corollary 3 then predicts the $\kappa^*(N)$ exponents $\frac{1}{5}$ and $\frac{1}{3}$ and the $\mathrm{MSE}^*(N)$ exponents $-\frac{4}{5}$ and $-\frac{2}{3}$ of Table 1; the constants themselves are checked against the data in Appendix F. On the one-step instance, where the gated form is exact, three checks pin down the identifications of Section 3. First, the gate term computed by forward-mode automatic differentiation matches the closed-form Z_κ of (6) to 1.4×10^{-14} (maximum over a grid in (ξ, κ), both readouts): the estimator the simulator differentiates is exactly the object the theorem analyzes. Second, the frozen-pathwise gradient (the same rollout differentiated with the gate argument held fixed, i.e. the first-order pathwise estimator of Suh et al. (2022b)) is *identically zero* for both readouts; in the all-stick regime the frozen-activation bulk term of (5) vanishes, so the entire true gradient is the boundary term and hard pathwise differentiation misses all of it, the sharpest instance of Corollary 1. Third, the closed-form B of (1) is matched from two independent directions: the large-κ value of $\mathbb{E}[Z_\kappa]$ (at $\kappa = 5\times10^5$) approaches it, which is Theorem 1, and a finite difference of $F(\theta) = \mathbb{E}[L]$ recovers it, which is the transport identity with the bulk term vanishing. For velocity $B = 5.93696$, with $|\mathbb{E}[Z_\kappa] - B| = 5.0\times10^{-7}$ and $|\partial_\theta F - B| = 1.3\times10^{-4}$; for kinetic energy $B = 0.630801$, with $|\mathbb{E}[Z_\kappa] - B| = 1.1\times10^{-4}$ and $|\partial_\theta F - B| = 1.4\times10^{-5}$. The jump that drives B is positive in both cases, $[\![L]\!] = +0.2$ for velocity and $+0.02125$ for kinetic energy, so $B > 0$: the sign supplied by the decreasing-sigmoid orientation (Remark 4) is the one observed. *Measured rates.* These predictions are borne out in Table 1 and Figure 2, with the proviso that the expansions are asymptotic in κR: the leading forms hold once the gate layer (width $\sim 1/\kappa$) is narrow against the bump (width $\sim R$), so each slope in Table 1 is a least-squares fit from the lower edge at which the sliding fit has settled to within 0.04 of its plateau (Appendix F). The collapse predicted by $C_1 = p'(\xi_\star)\,J$ is exhibited directly: moving the bump from $\xi_c = \xi_\star - R/2$ (where $p'(\xi_\star) \neq 0$) to $\xi_c = \xi_\star$ (where $p'(\xi_\star) = 0$) changes the measured kinetic-energy slope from -1.042 to -1.966, matching the centred-bump velocity slope -1.966, the two bias curves being proportional for the symmetric bump. The measured second-moment slopes sit slightly above 1 for two different reasons: for kinetic energy $V_0 < 0$, and the mean of $-V_0/(V_1\kappa)$ over the fit window, $+0.039$, is the measured excess; for velocity $V_0 = 0$ and the excess is pre-asymptotic, the window reaching down to $\kappa R \approx 1$ (Appendix F). *Both terms over a genuine surface.* The scalar-noise instance leaves two elements of the theory untested, the surface integral with a varying co-area weight and the split (1) with both terms nonzero; a second configuration removes both while changing only the initial conditions of the same one-step stepper. The noise is two-dimensional, $\xi = (\xi_1, \xi_2)$, with independent mollifier bumps p_1 (radius $R_1 = 0.02$, centered at -0.01 as before) and p_2 (radius $R_2 = 0.007$, centered at 0): the ball is released at height $z_0 = \theta + \xi_1 - h(\xi_2)$ with horizontal speed $w(\xi_2) = w_0 + \beta_w\,\xi_2$ ($w_0 = 0.2$, $\beta_w = 5$), where $h(\xi_2) = h_1\xi_2 + \frac{1}{2}h_2\xi_2^2$ ($h_1 = 0.8$, $h_2 = 40$) is a smooth nonlinear channel of the second noise coordinate. With γ the gravitational acceleration and Δt the step size ($\gamma = 1$, $\Delta t = 0.05$), the normal impact speed of the single step is $u_N = \gamma\Delta t = 0.05$. The gap is exactly $g = \theta + \xi_1 - h(\xi_2)$, so the switching set is the curve $\Sigma = \{\xi_1 = \varphi(\xi_2)\}$ with $\varphi = h - \theta_0$, along which the geometric weight $\nabla_\theta g/\|\nabla_\xi g\| = 1/\sqrt{1 + h'(\xi_2)^2}$ ranges from 0.68 to 0.89. The readout is the mixed return $L_\lambda = L_{\mathrm{KE}} + \lambda\gamma\,z_1$ with $\lambda = 1$, which adds to the kinetic energy the potential energy of the terminal height $z_1 = \theta + \xi_1 - h(\xi_2) - \frac{1}{2}(1-a)\,u_N\Delta t$, whose endpoints reproduce the hard branches and carry the discretization jump $[\![z_1]\!] = -\frac{1}{2}\gamma\Delta t^2$ (Appendix F). The frozen-activation derivative is the constant $\partial_\theta\Lambda_\lambda = \lambda\gamma$, so the bulk term no longer vanishes, and the jump along the surface is $[\![L_\lambda]\!](\xi_2) = \frac{1}{2}\big(w(\xi_2)^2 + u_N^2\big) - \frac{1}{2}\lambda\gamma\,u_N\Delta t > 0$. Because Σ is the graph of φ and $\partial_{\xi_1} g \equiv 1$, the arc-length element $\sqrt{1+\varphi'^2}\,d\xi_2$ is the exact reciprocal of the weight and the two cancel: the boundary term is the one-dimensional quadrature $B = \int [\![L_\lambda]\!](\xi_2)\,p_1(\varphi(\xi_2))\,p_2(\xi_2)\,d\xi_2 = 0.551479$, whereas the estimator runs in raw ξ-space and involves neither the surface nor its parametrization, so the comparison is between two independent computations; dropping the weight would change B by 28%, over 10^3 times the residual $|\mathbb{E}[Z_\kappa] - B|$ at $\kappa = 5\times10^5$. The same cancellation gives the co-area densities in closed form, $n(s,a) = \int \partial_a\Lambda_\lambda(a;\xi_2)\,p_1(s+\varphi(\xi_2))\,p_2(\xi_2)\,d\xi_2$ and m likewise with $\partial_a\Lambda_\lambda^2$, so Assumption 3 and its C^3 clause hold and Proposition 3 gives $C_1 = \int p_1'(\varphi(\xi_2))\,J(\xi_2)\,p_2(\xi_2)\,d\xi_2$ with $J(\xi_2) = \frac{1}{2}\big(w(\xi_2)^2 + u_N^2\big)$, the height term $\lambda\gamma z_1$ contributing only the constant $\frac{1}{2}\lambda\gamma u_N\Delta t$ to $\partial_a\Lambda_\lambda$ and hence nothing to J; numerically $C_1 = -61.0755$, so $r = 1$, while for the velocity readout $J \equiv 0$ and $C_1 = 0$ as before.

Measured, with both terms alive. The frozen-pathwise gradient now returns $\lambda\gamma = 1$ identically, so hard pathwise differentiation carries a nonzero baseline and is biased by exactly $-B$ (Corollary 1

with both terms present). At $\kappa = 5 \times 10^5$, $\mathbb{E}[Z_\kappa]$ lies within 1.2×10^{-4} of B, equal to the predicted leading term $|C_1|/\kappa$; a finite difference of F gives $\partial_\theta F = 1.551467$ against $\lambda\gamma + B = 1.551479$ (within 1.2×10^{-5}), which is the transport identity (1) verified with both terms simultaneously nonzero; and the full smoothed gradient equals $\lambda\gamma + \mathbb{E}[Z_\kappa]$ to machine precision, as Corollary 2 asserts. The rates persist in the surface-integral geometry: the bias slope is -0.966 on its window from $\kappa \approx 91$ (plateau -1.003), with $\kappa\,|\mathbb{E}[Z_\kappa] - B| = 61.19$ at $\kappa = 10^5$ against the predicted $|C_1| = 61.08$, and the velocity readout in the same geometry retains $r = 2$ (Table 1). What the experiments establish is therefore the boundary term (its value, sign, and surface-integral form, and its recovery by the gate as $\kappa \to \infty$), the split with both terms nonzero, the $O(\kappa)$ cost, and the predicted finite-κ orders and leading bias constant. Left to future numerical work are a parameter-dependent numerator ($\nabla_\theta g$ varying along Σ), several simultaneously active or intersecting surfaces (Section 6), and readouts evaluated after an earlier event has fired, where the branch-continuation gap loses smoothness.

readout	instance	bias slope	2nd-mom. slope	$\kappa^*(N)$	$\mathrm{MSE}^*(N)$
velocity (r=2)	one-step	-1.9609 (-2)	$+1.0386$ ($+1$)	0.2097 ($\frac{1}{5}$)	-0.7732 ($-\frac{4}{5}$)
	two-dim.	-1.9614 (-2)	$+1.0389$ ($+1$)	0.2191 ($\frac{1}{5}$)	-0.7647 ($-\frac{4}{5}$)
kinetic energy (r=1)	one-step	-1.0422 (-1)	$+1.0390$ ($+1$)	0.3175 ($\frac{1}{3}$)	-0.6663 ($-\frac{2}{3}$)
mixed energy, λ=1 (r=1)	two-dim.	-0.9662 (-1)	$+1.0401$ ($+1$)	0.3236 ($\frac{1}{3}$)	-0.6626 ($-\frac{2}{3}$)

Table 1: Measured power-law exponents of the gate estimator, with the values predicted by Proposition 3 and Corollary 3 in parentheses. Entries are log–log least-squares slopes: bias slope $= \mathrm{d}\log|\mathbb{E}[Z_\kappa] - B|/\mathrm{d}\log\kappa$, second-moment slope $= \mathrm{d}\log\mathbb{E}\|Z_\kappa\|^2/\mathrm{d}\log\kappa$, and the $\kappa^*(N), \mathrm{MSE}^*(N)$ exponents from log–log fits over $N \in [10^2, 10^8]$. Every fit has $R^2 \geq 0.992$; fit windows are described in Appendix F.

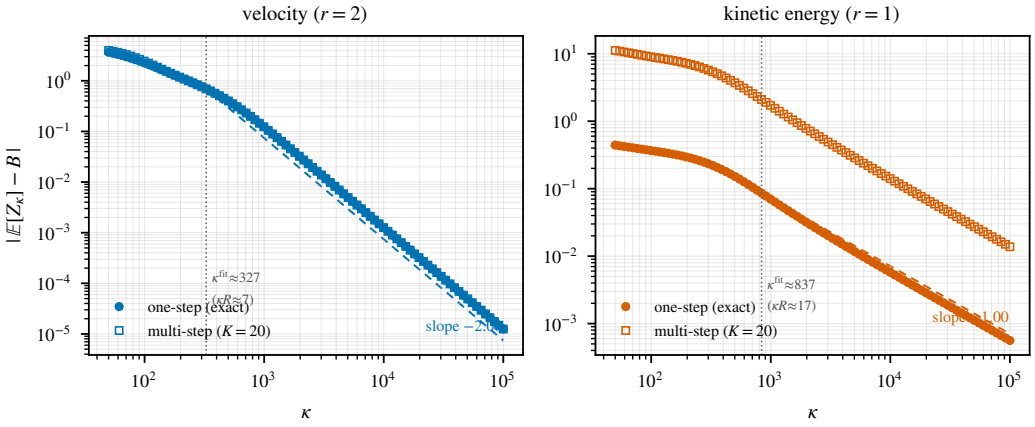

Figure 2: Bias $|\mathbb{E}[Z_\kappa] - B|$ of the gate estimator against gate sharpness κ (log–log) for the velocity (left, $r = 2$) and kinetic-energy (right, $r = 1$) readouts. Filled markers: one-step instance; open markers: the 20-step unroll of Appendix F, a distinct configuration whose absolute bias level differs for kinetic energy, but whose asymptotic slope is shared. Dashed lines: reference slopes -2 and -1; dotted vertical lines: observed fit-window onsets ($\kappa \approx 327$, $\kappa R \approx 7$ for velocity; $\kappa \approx 837$, $\kappa R \approx 17$ for kinetic energy).

6 Limitations

(L1) *Locality.* Everything is proved for one regular surface. Finitely many surfaces whose tubes are pairwise disjoint add up: the one-surface arguments run independently in each tube and sum to $\nabla_\theta F = \mathbb{E}[\nabla_\theta L] + \sum_i B_i$ and $\sum_i \mathbb{E}[Z_\kappa^{(i)}] \to \sum_i B_i$. Intersecting or simultaneous events are not covered: away from the intersection Theorem 1 still applies locally, and a transverse intersection has codimension two and no boundary term of its own, but the

behaviour of the gate where two activations vary at once needs stratified regularity of the arrangement and of Λ in several activations, which we do not develop.

(L2) *Bias, not optimization.* Theorem 1 is a statement about $\lim_{\kappa\to\infty}\mathbb{E}[Z_\kappa]$, and Propositions 2–3 with Corollary 3 about the moments of Z_κ at finite κ. Nothing here concerns the convergence of an optimization procedure that uses these gradients.

(L3) *The gated form is an assumption.* Assumption 2 holds exactly for a single gated Moreau contact step (Section 5) and is not derived from the equations of any general solver; whether a particular simulator's smoothing has the form $\Lambda(a_\kappa(g),\theta,\xi)$ with endpoints reproducing the hard branches is a modelling fact about that simulator, to be checked case by case.

(L4) *Assumption 3 is assumed, not derived.* The continuity of the co-area density that Theorem 1 uses follows from the standing assumptions (Proposition 4); the C^2 smoothness that Proposition 3 uses does not, and we verify it only for the configurations of Section 5.

(L5) *Bulk consistency needs more.* Convergence of the full smoothed gradient (Corollary 2) requires joint C^1 regularity of Λ in (a,θ) and branch-matching off the surface; Theorem 1 itself is only about the gate term.

(L6) *Only jumps matter for the bias.* At a kink ($[\![L]\!]\equiv 0$) $B=0$ and hard pathwise differentiation is already exactly unbiased (Corollary 1); gating replaces that exactness by the asymptotic unbiasedness of Corollary 2, at the sampling cost of Proposition 2.

(L7) *Compactly supported noise.* Assumption 1 requires $\operatorname{supp} p$ compact. This is what keeps the argument local (the tube of Lemma 2, the bounds (S5), finite surface integrals), but it excludes unbounded noise laws such as Gaussian initial-state or domain-randomization distributions; non-compact support would need tail and p-weighted integrability conditions on g, L^+, L^-, $\nabla_\theta g$ and the surface integral, which we do not pursue.

7 Conclusion

Near one regular switching surface, the gradient of the expected return of a contact rollout is the expected pathwise gradient plus a boundary integral over the surface weighted by the jump of the return, and hard pathwise differentiation misses exactly the boundary term. When the simulator gates the contact impulse by a sigmoid and the return depends on the activation through a differentiable interpolation between the hard branches, the gate part of the smoothed pathwise gradient is a kernel smoother of the co-area density at the surface, with a kernel shaped by the readout, and its expectation converges to the boundary term as the gate hardens. At finite sharpness the second moment of the gate term grows at most linearly; under one extra smoothness hypothesis its bias and second moment expand explicitly, with an order set by the readout and the noise density, and balance at a batch-size scaling law. A single-contact stepper for which the gated form is exact reproduces the boundary term, the predicted orders, and the leading bias constant, also on a genuine surface with both terms of the decomposition nonzero. Separated events add surface by surface; intersecting events and the multi-step solvers in which the gate is deployed remain open.

AI use statement

In this work, we used generative AI tools for writing, for coding, and for identifying related literature. We have reviewed all AI-assisted work: the code was verified against automatic differentiation and against independent quadrature and Monte Carlo estimates, and every citation was checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

All proofs are in Appendices A–E; the standing assumptions are stated in Section 2 and the hypotheses of each result in the result itself. The numerical study is specified in Section 5 and Appendix F (model constants, quadrature rules, grids, the fit-window rule, seeds). The supplementary material contains the code: a Python package with the stepper, the forward-mode gate-term estimator, the quadrature and Monte Carlo routines, the experiment scripts that produce Table 1, Figure 2 and the values of Section 5 and Appendix F, the cached outputs with a checksum manifest, and the test suite.

REFERENCES

Sai Praveen Bangaru, Tzu-Mao Li, and Frédo Durand. Unbiased warped-area sampling for differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, revised edition, 2015.

Herbert Federer. *Geometric Measure Theory*, volume 153 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1969.

Harley Flanders. Differentiation under the integral sign. *The American Mathematical Monthly*, 80 (6):615–627, 1973.

Michael C. Fu and Jian-Qiang Hu. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer, 1997.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2004.

Yu-Chi Ho and Xiren Cao. Perturbation analysis and optimization of queueing networks. *Journal of Optimization Theory and Applications*, 40(4):559–582, 1983.

Evgeny Lakshtanov. Unbiased Monte Carlo Greeks for discontinuous payoffs, 2026. arXiv:2609.06137.

Quentin Le Lidec, Louis Montaut, Cordelia Schmid, Ivan Laptev, and Justin Carpentier. Leveraging randomized smoothing for optimal control of nonsmooth dynamical systems, 2022. arXiv:2203.03986.

Wonyeol Lee, Hangyeol Yu, and Hongseok Yang. Reparameterization gradient for non-differentiable models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable Monte Carlo ray tracing through edge sampling. *ACM Transactions on Graphics*, 37(6), 2018.

Guangwu Liu and L. Jeff Hong. Kernel estimation of the Greeks for options with discontinuous payoffs. *Operations Research*, 59(1):96–108, 2011.

Guillaume Loubet, Nicolas Holzschuch, and Wenzel Jakob. Reparameterizing discontinuous integrands for differentiable rendering. *ACM Transactions on Graphics*, 38(6), 2019.

Ku Onoda, Paavo Parmas, Manato Yaguchi, and Yutaka Matsuo. Does "Do differentiable simulators give better policy gradients?" give better policy gradients? In *International Conference on Learning Representations (ICLR)*, 2026.

Tao Pang, H. J. Terry Suh, Lujie Yang, and Russ Tedrake. Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models. *IEEE Transactions on Robotics*, 39 (6):4691–4711, 2023.

Anselm Paulus, A. René Geist, Pierre Schumacher, Vít Musil, Simon Rappenecker, and Georg Martius. Differentiable simulation of hard contacts with soft gradients for learning and control. In *International Conference on Learning Representations (ICLR)*, 2026.

Yijie Peng, Michael C. Fu, Jian-Qiang Hu, and Bernd Heidergott. A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Operations Research*, 66(2):487–499, 2018.

Xingyu Ren, Michael C. Fu, and Pierre L'Ecuyer. Stochastic derivative estimation for discontinuous sample performances: A Leibniz integration perspective, 2025. arXiv:2511.00006.

Clemens Schwarke, Victor Klemm, Joshua Bagajo, Jean-Pierre Sleiman, Ignat Georgiev, Jesus Tordesillas, and Marco Hutter. Learning deployable locomotion control via differentiable simulation. In *Proceedings of the 9th Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3665–3684. PMLR, 2025.

Yunlong Song, Sangbae Kim, and Davide Scaramuzza. Learning quadruped locomotion using differentiable simulation. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 258–271. PMLR, 2024.

H. J. Terry Suh, Tao Pang, and Russ Tedrake. Bundled gradients through contact via randomized smoothing. *IEEE Robotics and Automation Letters*, 7(2):4000–4007, 2022a.

H. J. Terry Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do differentiable simulators give better policy gradients? In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022b.

Yuval Tassa and Emanuel Todorov. Stochastic complementarity for local control of discontinuous dynamics. In *Robotics: Science and Systems*, 2010.

Senyun Yang and Zhimin Zhang. Smoothing discontinuities in option Greeks estimation: A sigmoid-family infinitesimal perturbation analysis approach. *Mathematics and Computers in Simulation*, 248:668–685, 2026.

A TECHNICAL PRELIMINARIES

Notation. Gradients $\nabla_\theta(\cdot) \in \mathbb{R}^d$ and $\nabla_\xi(\cdot) \in \mathbb{R}^m$ are column vectors. The directional derivative of F at θ_0 in direction $u \in \mathbb{R}^d$ is $D_u F(\theta_0) = \frac{d}{d\eta}\big|_{\eta=0} F(\theta_0 + \eta u)$ when it exists. $\mathcal{H}^{m-1}$ is the $(m-1)$-dimensional Hausdorff (surface) measure on $\mathbb{R}^m$ and $d\xi$ is Lebesgue measure. The support $\operatorname{supp} f$ is the closure of $\{f \neq 0\}$, and $A \Subset B$ means $\overline{A}$ is compact with $\overline{A} \subset B$.

Assumption 1, itemized. With K, Θ, and U as in Assumption 1, the proofs cite the following items:

- (S1) $g \in C^1(\Theta \times U)$;
- (S2) $\nabla_\xi g(\theta_0, \xi) \neq 0$ for every $\xi \in K \cap \Sigma$ *(the surface is regular on the support of p)*;
- (S3) $p \in L^\infty(\mathbb{R}^m)$, $\operatorname{supp} p \subset K$, and p is continuous on a neighborhood of $K \cap \Sigma$;
- (S4) $L^+, L^- \in C^1(\Theta \times U)$;
- (S5) $L^+, L^-, \nabla_\theta L^+, \nabla_\theta L^-$ and $\nabla_\theta g$ are bounded on $\Theta_1 \times K'$ for some $\Theta_1 \Subset \Theta$ containing θ_0 and some K' with $K \subset K' \Subset U$.

Items (S1)–(S4) restate Assumption 1. Item (S5) is a consequence of (S1) and (S4): once $\Theta_1 \Subset \Theta$ and $K' \Subset U$ are fixed, $\overline{\Theta_1} \times \overline{K'}$ is compact and contained in the open set $\Theta \times U$, so the continuous functions $L^+, L^-, \nabla_\theta L^+, \nabla_\theta L^-, \nabla_\theta g$ are bounded there; it is listed separately because it is the property the proofs invoke. All items are local: they constrain g, L^+, L^-, p only on a neighborhood of the noise support K, and the switching surface only where K meets it. Under (S1)–(S2), $K \cap \Sigma$ is a compact C^1 hypersurface and hence Lebesgue-null, so the value of L on Σ is irrelevant for F and the branchwise formula of Section 2.1 determines $\nabla_\theta L$ p-a.e.

Sigmoid tail bound. For σ as in (2), $\sigma'(t) = e^{-|t|}/(1+e^{-|t|})^2 \le e^{-|t|}$ for all t; hence $|a'_\kappa(s)| \le \kappa e^{-\kappa|s|}$ by (2).

We use the co-area formula in the following standard form; see Evans & Gariepy (2015) or Federer (1969).

Lemma 1 (Co-area formula)**.** *Let $w : \mathbb{R}^m \to \mathbb{R}$ be Lipschitz and $f : \mathbb{R}^m \to \mathbb{R}$ Lebesgue-integrable. Then*

$$\int_{\mathbb{R}^m} f(\xi)\, \|\nabla w(\xi)\|\, d\xi \;=\; \int_{\mathbb{R}} \Big(\int_{\{w=s\}} f \, d\mathcal{H}^{m-1} \Big)\, ds.$$

In particular, on any open set where $\|\nabla w\| \ge c_0 > 0$, replacing f by $f/\|\nabla w\|$ gives

$$\int f(\xi)\, d\xi \;=\; \int_{\mathbb{R}} \Big(\int_{\{w=s\}} \frac{f}{\|\nabla w\|}\, d\mathcal{H}^{m-1} \Big)\, ds \tag{13}$$

for every f supported in that set.

The formula localizes: if f has compact support and w is merely C^1 on an open set containing $\operatorname{supp} f$, one may replace w by ζw, with ζ a C^1 compactly supported cutoff equal to 1 on a neighborhood of $\operatorname{supp} f$ inside that set. The product is Lipschitz on all of $\mathbb{R}^m$ and has the same gradient and the same level sets as w near $\operatorname{supp} f$, so neither side of (13) changes under the replacement. Every later application of (13) is of this kind. The two geometric facts we use repeatedly follow immediately.

Lemma 2 (Tube and gradient lower bound)**.** *Under Assumption 1 there exist an open neighborhood V of $K \cap \Sigma$ and constants $c_0 > 0$, $\delta_0 > 0$, such that*

$$\|\nabla_\xi g(\theta_0, \xi)\| \;\geq\; c_0 \qquad \textit{for all } \xi \in V, \tag{14}$$

and such that the tube

$$T_\delta \;:=\; \{\xi \in V :\; |g(\theta_0, \xi)| < \delta\} \tag{15}$$

satisfies, for every $0 < \delta \leq \delta_0$,

$$\{\xi \in K :\; |g(\theta_0, \xi)| < \delta\} \;\subset\; T_\delta. \tag{16}$$

Proof. By (S2), $\nabla_\xi g(\theta_0, \cdot)$ is nonzero on the compact set $K \cap \Sigma$ and, by (S1), continuous; hence there are an open neighborhood V of $K \cap \Sigma$ and $c_0 > 0$ with (14) on V. It remains to produce δ_0 with (16). If no such δ_0 existed, then for each $k \in \mathbb{N}$ there would be $\xi_k \in K$ with $|g(\theta_0, \xi_k)| < 1/k$ but $\xi_k \notin V$. By compactness of K a subsequence converges, $\xi_{k_j} \to \xi_\star \in K$, and by continuity $g(\theta_0, \xi_\star) = 0$, so $\xi_\star \in K \cap \Sigma \subset V$. As V is open and $\xi_{k_j} \to \xi_\star$, we get $\xi_{k_j} \in V$ for large j, contradicting $\xi_{k_j} \notin V$. □

Lemma 2 is the only place locality becomes quantitative: the part of the noise support within g-distance δ of Σ lies inside a tube on which g is a genuine submersion, with $\|\nabla_\xi g\| \geq c_0$. Outside $\{|g| < \delta\}$ we require no regular foliation of the level sets of g, and no continuity of p along such level sets; only the global boundedness and support assumptions on p remain in force. The smoothing terms of Section 3 are exponentially small away from Σ.

B Proof of Proposition 1

Proof. Fix $u \in \mathbb{R}^d$ and write $\theta_\eta = \theta_0 + \eta u$ for $|\eta|$ small enough that $\theta_\eta \in \Theta_1$. Decompose the return by the sign of g,

$$F(\theta_\eta) = I^+(\eta) + I^-(\eta), \qquad I^\pm(\eta) = \int_{\{\pm g(\theta_\eta, \cdot) > 0\}} L^\pm(\theta_\eta, \xi)\, p\, d\xi.$$

Each integrand is smooth, but the domain moves with η, so the derivative has a bulk part (from the integrand) and a boundary part (from the moving domain). We make this rigorous locally and patch.

Step 1: localization. By Lemma 2, $\nabla_\xi g(\theta_0, \cdot) \neq 0$ on the open set $V \supset K \cap \Sigma$. We cover $K \cap \Sigma$ by coordinate cylinders carrying a straightening chart. Fix $\xi_\star \in K \cap \Sigma$; since $\nabla_\xi g(\theta_0, \xi_\star) \neq 0$, after relabelling the noise coordinates as $\xi = (\xi_1, \xi') \in \mathbb{R} \times \mathbb{R}^{m-1}$ we may assume $\partial_{\xi_1} g(\theta_0, \xi_\star) \neq 0$, so the map $\Phi(\xi) := (g(\theta_0, \xi), \xi')$ has block upper-triangular Jacobian with $\det D\Phi(\xi_\star) = \partial_{\xi_1} g(\theta_0, \xi_\star) \neq 0$. By the inverse function theorem there is an open ball $B_{\xi_\star} \subset V$ about $\xi_\star$, shrunk if necessary so that p is continuous on it (possible by (S3) since $\xi_\star \in K \cap \Sigma$), on which Φ is a C^1 diffeomorphism onto the open set $O_{\xi_\star} := \Phi(B_{\xi_\star})$, with $\Phi(\xi_\star) = (0, y_\star) \in O_{\xi_\star}$ (as $g(\theta_0, \xi_\star) = 0$). Since $O_{\xi_\star}$ is open, choose a bounded open product box $I_{\xi_\star} \times Y_{\xi_\star} \ni (0, y_\star)$ with $\overline{I_{\xi_\star}} \times \overline{Y_{\xi_\star}} \subset O_{\xi_\star}$, and set the coordinate cylinder $C_{\xi_\star} := \Phi^{-1}(I_{\xi_\star} \times Y_{\xi_\star}) \subset B_{\xi_\star}$, an open neighbourhood of $\xi_\star$. The cylinders $\{C_{\xi_\star}\}$ cover the compact set $K \cap \Sigma$; take a finite subcover $C_1, \ldots, C_N$, keeping for each its chart Φ_j, inverse $\Psi_j := \Phi_j^{-1}$, box $I_j \times Y_j$, and image $O_j = \Phi_j(B_j) \supset \overline{I_j} \times \overline{Y_j}$. Cover $K \setminus \bigcup_j C_j$ by an open set W with $\overline{W} \cap \Sigma = \emptyset$ (possible since $K \cap \Sigma \subset \bigcup_j C_j$ and K is compact). Take a C^1 partition of unity $\{\chi_0, \chi_1, \ldots, \chi_N\}$ subordinate to $\{W, C_1, \ldots, C_N\}$ with $\sum_{j \geq 0} \chi_j \equiv 1$ on a neighborhood of K and $\operatorname{supp} \chi_j \Subset C_j$ for $j \geq 1$. It suffices to differentiate $\int \chi_j L(\theta_\eta, \cdot) p\, d\xi$ for each j and sum, because $\sum_j \chi_j L = L$ on $\operatorname{supp} p$.

On the patch W, only $\operatorname{supp} \chi_0 \cap K$ contributes to $\int \chi_0 L(\theta_\eta, \cdot) p\, d\xi$; this set is compact (a closed subset of K) and disjoint from Σ (as $\overline{W} \cap \Sigma = \emptyset$), so $|g(\theta_0, \cdot)| \geq \delta > 0$ on it. By continuity of g

in θ, the perturbed zero set $\{g(\theta_\eta, \cdot) = 0\}$ also avoids $\operatorname{supp}\chi_0 \cap K$ for $|\eta|$ small; there $\chi_0 L(\theta_\eta, \cdot)$ equals $\chi_0 L^+(\theta_\eta, \cdot)$ or $\chi_0 L^-(\theta_\eta, \cdot)$, a C^1 function of θ. Ordinary differentiation under the integral sign, justified by (S5) and dominated convergence of difference quotients, gives

$$\frac{d}{d\eta}\int \chi_0 L(\theta_\eta, \cdot)\, p\, d\xi\Big|_0 = \int \chi_0\, u^\top \nabla_\theta L(\theta_0, \cdot)\, p\, d\xi,$$

with no boundary contribution.

Step 2: a single regular patch. Fix a surface patch $j \geq 1$ and write $\chi = \chi_j$, $\Phi = \Phi_j$, $\Psi = \Psi_j$, $B = B_j$, $O = O_j = \Phi(B)$, and $I \times Y = I_j \times Y_j$ for the data attached to C_j in Step 1. The argument has three parts: we straighten $g(\theta_0, \cdot)$ into a coordinate, represent the perturbed surface as a graph over that coordinate, and differentiate the graph in η.

(2a) Straightening $g(\theta_0, \cdot)$ into a coordinate. Write $\xi = \Psi(s, y)$ for $(s, y) \in O$; by Step 1 the chart $\Phi = (g(\theta_0, \cdot), \xi')$ is a C^1 diffeomorphism of the ball B onto O, so $\Psi = \Phi^{-1}$ is C^1 on O. The box $I \times Y$ from Step 1 has $\overline{I} \times \overline{Y} \subset O$ and $0 \in I$ (it contains the centre's image $(0, y_\star)$), so $(0, y) \in I \times Y \subset O$ for every $y \in Y$; and $\operatorname{supp}\chi \Subset C_j = \Psi(I \times Y)$ gives $\Phi(\operatorname{supp}\chi) \Subset I \times Y$. Every uniform bound below is taken over the compact set $\overline{I} \times \overline{Y} \subset O$, on which the C^1 data are bounded. Equating first coordinates in $\Phi(\Psi(s, y)) = (s, y)$ yields the *straightening identity*

$$g\big(\theta_0, \Psi(s, y)\big) = s \qquad \text{for all } (s, y) \in O : \tag{17}$$

in these coordinates the switching function $g(\theta_0, \cdot)$ *is* the first coordinate. Two consequences follow from (17) alone. First, since Ψ is a bijection onto B, a point $\xi = \Psi(s, y) \in B$ obeys $\xi \in \Sigma \Leftrightarrow g(\theta_0, \xi) = 0 \Leftrightarrow s = 0$, so

$$\Sigma \cap B = \Psi\big(\{(s, y) \in O :\ s = 0\}\big). \tag{18}$$

Second, differentiating (17) by the chain rule (the argument θ_0 is held fixed) gives, for every $(s, y) \in O$,

$$\nabla_\xi g(\theta_0, \Psi)^\top\, \partial_s \Psi = \partial_s s = 1, \qquad \nabla_\xi g(\theta_0, \Psi)^\top\, \partial_{y_k} \Psi = \partial_{y_k} s = 0 \quad (1 \leq k \leq m-1), \tag{19}$$

so $\partial_s \Psi$ is transverse to the level sets of $g(\theta_0, \cdot)$ and the $\partial_{y_k}\Psi$ are tangent to them. The first identity in (19) is the relation $\partial_s\big(g(\theta_0, \Psi(s, y))\big) = 1$ used repeatedly below.

(2b) The perturbed surface as a graph. For $|\eta| < \eta_0$, so that $\theta_\eta = \theta_0 + \eta u \in \Theta$, define

$$H(\eta, s, y) := g\big(\theta_\eta, \Psi(s, y)\big), \qquad (\eta, s, y) \in (-\eta_0, \eta_0) \times O,$$

which is C^1 as a composition of the C^1 maps g, the affine map $\eta \mapsto \theta_\eta$, and Ψ. By (17), $H(0, s, y) = s$, hence

$$H(0, 0, y) = 0, \qquad \partial_s H(0, s, y) \equiv 1. \tag{20}$$

Since $\partial_s H$ is continuous with $\partial_s H(0, \cdot, \cdot) \equiv 1$ by (20), and $\overline{I} \times \overline{Y}$ is compact, we may shrink η_0 so that

$$\partial_s H(\eta, s, y) \geq \tfrac{1}{2} \qquad \text{on } [-\eta_0, \eta_0] \times \overline{I} \times \overline{Y}. \tag{21}$$

Put $M := \sup_{[-\eta_0, \eta_0] \times \overline{I} \times \overline{Y}} |\partial_\eta H| < \infty$ (finite, as $H \in C^1$ on a compact set) and fix $\rho > 0$ with $[-\rho, \rho] \subset I$; shrinking η_0 once more we assume $\eta_0 \leq \rho/(2M)$. We show that for every $(\eta, y) \in (-\eta_0, \eta_0) \times Y$ the equation $H(\eta, \cdot, y) = 0$ has a unique root $\varphi(\eta, y) \in (-\rho, \rho)$, and that φ is C^1.

Uniqueness. By (21), $s \mapsto H(\eta, s, y)$ is strictly increasing on I, so it equals 0 at most once.

Existence. Since $H(0, 0, y) = 0$, $|H(\eta, 0, y)| = \big|\int_0^\eta \partial_\eta H(t, 0, y)\, dt\big| \leq M|\eta|$. Integrating (21) in s outward from 0,

$$H(\eta, \rho, y) \geq H(\eta, 0, y) + \tfrac{1}{2}\rho \geq \tfrac{1}{2}\rho - M|\eta|, \qquad H(\eta, -\rho, y) \leq H(\eta, 0, y) - \tfrac{1}{2}\rho \leq M|\eta| - \tfrac{1}{2}\rho.$$

For $|\eta| < \eta_0 \leq \rho/(2M)$ the first quantity is positive and the second negative, so by the intermediate value theorem $H(\eta, \cdot, y)$ has a root in $(-\rho, \rho)$; by uniqueness it is the only one. (Integrating from 0 to that root gives $|\varphi(\eta, y)| \leq 2M|\eta|$, whence $\varphi(0, y) = 0$.)

Regularity. At a root $\big(\eta_1, \varphi(\eta_1, y_1), y_1\big)$ one has $H = 0$ and $\partial_s H \geq \frac{1}{2} \neq 0$; the implicit function theorem (unknown s, parameters (η, y)) then gives a C^1 map solving $H = 0$ near (η_1, y_1), which

coincides with φ by the uniqueness above. Since (η_1, y_1) was arbitrary, $\varphi \in C^1\big((-\eta_0, \eta_0) \times Y\big)$, with $\nabla_{(\eta,y)}\varphi = -(\partial_s H)^{-1}\,\nabla_{(\eta,y)} H$ at $s = \varphi$. Thus

$$g\big(\theta_\eta, \Psi(\varphi(\eta, y), y)\big) = H\big(\eta, \varphi(\eta, y), y\big) = 0, \tag{22}$$

and by (18) the graph $\{s = \varphi(\eta, y) : y \in Y\}$ is exactly $\{g(\theta_\eta, \cdot) = 0\} \cap C_j$ in the coordinates (s, y), where $C_j = \Psi(I \times Y)$ is the cylinder carrying $\operatorname{supp}\chi$; this is the only region the localized integral uses. At $\eta = 0$, $H(0, s, y) = s$ vanishes only at $s = 0$, so uniqueness forces

$$\varphi(0, y) \equiv 0. \tag{23}$$

On the cylinder C_j (i.e. for $y \in Y$, $s \in I$), strict monotonicity in s also identifies the two sides of the surface:

$$g(\theta_\eta, \Psi(s, y)) > 0 \iff s > \varphi(\eta, y), \qquad g(\theta_\eta, \Psi(s, y)) < 0 \iff s < \varphi(\eta, y). \tag{24}$$

(2c) Normal speed of the surface. Differentiate the identity (22) in η at fixed y. Both $\eta \mapsto (\eta, \varphi(\eta, y), y)$ and H are C^1, so the chain rule gives

$$0 = \frac{d}{d\eta} H\big(\eta, \varphi(\eta, y), y\big) = \partial_\eta H\big(\eta, \varphi, y\big) + \partial_s H\big(\eta, \varphi, y\big)\, \partial_\eta \varphi(\eta, y).$$

Evaluate at $\eta = 0$: there $\varphi(0, y) = 0$ by (23), $\partial_s H(0, 0, y) = 1$ by (20), and, again by the chain rule with $\frac{d}{d\eta}\theta_\eta = u$,

$$\partial_\eta H(0, 0, y) = \nabla_\theta g\big(\theta_0, \Psi(0, y)\big)^\top u.$$

Solving for $\partial_\eta \varphi$ (this is the implicit-function gradient formula $\partial_\eta \varphi = -\,\partial_\eta H / \partial_s H$, here with denominator 1) gives

$$\partial_\eta \varphi(0, y) = -\,u^\top \nabla_\theta g\big(\theta_0, \Psi(0, y)\big). \tag{25}$$

Geometrically, $\partial_\eta \varphi(0, y)$ is the speed, in the straightened coordinate $s = g(\theta_0, \cdot)$, at which the switching surface moves as θ advances along u; it is this normal speed that drives the boundary term below.

Consider the positive-side piece $J^+(\eta) = \int_{\{g(\theta_\eta, \cdot) > 0\}} \chi\, L^+(\theta_\eta, \cdot)\, p\, d\xi$. Its integrand is supported in $\operatorname{supp}\chi \Subset C_j = \Psi(I \times Y)$, so it pushes to (s, y)-coordinates over $I \times Y$; there $\{g(\theta_\eta, \cdot) > 0\} \cap C_j = \{s > \varphi(\eta, y)\}$ by (24), and $d\xi = \mathcal{J}(s, y)\, ds\, dy$ with $\mathcal{J} = |\det D\Psi| > 0$. Writing $F^+(s, y) := \chi(\Psi(s, y))\, L^+(\theta_\eta, \Psi(s, y))\, p(\Psi(s, y))\, \mathcal{J}(s, y)$, which vanishes for (s, y) outside a compact subset of $I \times Y$,

$$J^+(\eta) = \int_Y \int_{\varphi(\eta, y)}^{\infty} F^+(s, y)\, ds\, dy,$$

the upper limit being immaterial since $F^+(\cdot, y)$ is compactly supported in I. Differentiating in η by the Leibniz rule in the inner s-integral (legitimate because only $L^+(\theta_\eta, \cdot)$ depends on η, with $\partial_\eta L^+ = u^\top \nabla_\theta L^+$ bounded since $\nabla_\theta L^+$ is continuous on the compact $\overline{\Theta_1} \times \overline{\operatorname{supp}\chi} \subset \Theta \times U$ (S4) and the remaining factors $\chi, p, \mathcal{J}$ bounded on the compact $\Phi(\operatorname{supp}\chi) \Subset I \times Y$, while the boundary trace at $s = \varphi$ is taken near $K \cap \Sigma$, where p is continuous by (S3)) gives, at $\eta = 0$,

$$\frac{d}{d\eta} J^+\Big|_0 = \int_Y \int_0^{\infty} \chi\, u^\top \nabla_\theta L^+\, p\, \mathcal{J}\, ds\, dy \;-\; \int_Y \big(\chi\, L^+\, p\big)\big|_{s=0}\, \partial_\eta \varphi(0, y)\, \mathcal{J}(0, y)\, dy.$$

The first term equals $\int_{\{g(\theta_0, \cdot) > 0\}} \chi\, u^\top \nabla_\theta L^+\, p\, d\xi$. For the second, insert (25) so that $-\partial_\eta \varphi(0, y) = u^\top \nabla_\theta g$. To convert the resulting y-integral to a surface integral, use the area factor of the slice map $y \mapsto \Psi(s, y)$, $y \in Y$, which for each fixed $s \in I$ parametrizes the slice $\{g(\theta_0, \cdot) = s\} \cap C_j$ injectively; by the area formula for this C^1 parametrization, $d\mathcal{H}^{m-1} = \big(\det(\partial_y \Psi^\top \partial_y \Psi)\big)^{1/2} dy$ on the slice. The identities (19) turn this area element into the Jacobian $\mathcal{J} = |\det D\Psi|$: with $\nu := \nabla_\xi g(\theta_0, \Psi) / \|\nabla_\xi g(\theta_0, \Psi)\|$ the unit normal, the columns $\partial_{y_k}\Psi$, independent because $D\Psi$ is invertible, span the tangent space $\nu^\perp$ of the slice, and the remaining column decomposes as $\partial_s \Psi = \|\nabla_\xi g(\theta_0, \Psi)\|^{-1} \nu + \tau$ with τ a linear combination of the $\partial_{y_k}\Psi$; subtracting τ from $\partial_s \Psi$ leaves the determinant unchanged, and $|\det[\,\nu \;\; \partial_y \Psi\,]| = \big(\det(\partial_y \Psi^\top \partial_y \Psi)\big)^{1/2}$ because ν is a unit vector orthogonal to the columns of $\partial_y \Psi$, so

$$\mathcal{J} = |\det D\Psi| = \frac{\big(\det(\partial_y \Psi^\top \partial_y \Psi)\big)^{1/2}}{\|\nabla_\xi g(\theta_0, \Psi)\|} \qquad \text{at every } (s, y) \in O.$$

Hence the surface element on $\{s = 0\}$ satisfies $d\mathcal{H}^{m-1} = \mathcal{J}(0, y)\, \|\nabla_\xi g(\theta_0, \Psi(0, y))\|\, dy$, i.e.

$$\mathcal{J}(0, y)\, dy \;=\; \frac{d\mathcal{H}^{m-1}}{\|\nabla_\xi g(\theta_0, \cdot)\|} \qquad \text{on } \Sigma \cap C_j. \tag{26}$$

Therefore the boundary part of $\frac{d}{d\eta} J^+|_0$ is $\int_{\Sigma \cap C_j} \chi\, L^+ \dfrac{u^\top \nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1}$.

The negative-side piece $J^-(\eta) = \int_{\{g(\theta_\eta, \cdot) < 0\}} \chi\, L^-(\theta_\eta, \cdot)\, p\, d\xi$ is identical (with the same η-differentiability and continuous-trace justification) except that the domain is $\{s < \varphi(\eta, y)\}$, so the moving limit is now the *upper* limit and the boundary term carries the opposite sign:

$$\frac{d}{d\eta} J^- \Big|_0 = \int_{\{g(\theta_0, \cdot) < 0\}} \chi\, u^\top \nabla_\theta L^-\, p\, d\xi \;-\; \int_{\Sigma \cap C_j} \chi\, L^- \frac{u^\top \nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1}.$$

Adding, the bulk parts combine into $\int \chi\, u^\top \nabla_\theta L(\theta_0, \cdot)\, p\, d\xi$ by the definition of $\nabla_\theta L$ in Section 2.1, and the boundary parts combine into

$$\int_{\Sigma \cap C_j} \chi\, (L^+ - L^-) \frac{u^\top \nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1} = \int_{\Sigma \cap C_j} \chi\, [\![L]\!] \frac{u^\top \nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1}.$$

Step 3: summation. Summing over $j = 1, \dots, N$ and adding the boundary-free patch W, and using $\sum_j \chi_j \equiv 1$ on $\operatorname{supp} p$ and on $\Sigma \cap \operatorname{supp} p$,

$$D_u F(\theta_0) = \int_{\mathbb{R}^m} u^\top \nabla_\theta L(\theta_0, \xi)\, p\, d\xi + \int_\Sigma [\![L]\!] \frac{u^\top \nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1},$$

the surface integral over $\Sigma \setminus \bigcup_j C_j$ vanishing because $\operatorname{supp} p \cap \Sigma \subset \bigcup_j C_j$. Both terms are finite (the bulk term by (S5) and $p \in L^1$; the boundary term by (S5), (14), and finiteness of $\mathcal{H}^{m-1}(\Sigma \cap K)$, since $\Sigma \cap K$ is compact and covered by finitely many regular C^1 hypersurface patches) and linear in u. Hence $D_u F(\theta_0) = \langle \nabla_\theta F(\theta_0), u \rangle$ with $\nabla_\theta F(\theta_0)$ as in (1). □

Remark 2 (Sense of differentiability). Proposition 1 asserts that the two-sided directional derivative exists in every direction and is linear in u; this is Gâteaux differentiability at θ_0, with gradient (1). If, in addition, Assumption 1 holds uniformly for θ in a neighborhood of θ_0, so that $\Sigma(\theta)$ varies C^1 and the right-hand side of (1) is continuous in the base point, then $F \in C^1$ near θ_0. We neither need nor assume the stronger statement.

C Regularity of the co-area density

Proposition 4 (Regularity of the co-area density)**.** *Let δ_0, V be as in Lemma 2 and n the co-area density (7). Under Assumptions 1 and 2, the integral (7) is well-defined, n is bounded on $(-\delta_0, \delta_0) \times (0, 1)$, and for every fixed $a \in (0, 1)$,*

$$n(s, a) \longrightarrow n(0, a) \qquad \text{as } s \to 0. \tag{27}$$

Remark 3. Proposition 4 controls n only on the tube T_{δ_0}; nothing about n away from Σ is used. Its proof rests on three facts furnished by the standing assumptions on the tube: the level sets $\{g(\theta_0, \cdot) = s\}$ form a regular C^1 foliation near Σ (Lemma 2), p is continuous there by (S3), and $\partial_a \Lambda$ is jointly continuous by Assumption 2. These combine with the boundedness of $\partial_a \Lambda$ and $\nabla_\theta g$ on K, $p \in L^\infty$, $\|\nabla_\xi g\| \geq c_0$, and a uniform bound on $\mathcal{H}^{m-1}(\{g = s\} \cap K)$ for $|s| < \delta_0$.

Proof. We reuse the localization constructed in the proof of Proposition 1 (Appendix B, Step 1): coordinate cylinders $C_1, \dots, C_N$ covering the compact set $K \cap \Sigma$, each with a C^1 straightening chart $\Phi_j = (g(\theta_0, \cdot), \xi')$ and C^1 inverse $\Psi_j = \Phi_j^{-1}$ on an open set $O_j \supset \overline{I_j} \times \overline{Y_j}$, satisfying

$$g\big(\theta_0, \Psi_j(s, y)\big) = s \;\text{ on } O_j, \qquad C_j = \Psi_j(I_j \times Y_j), \qquad 0 \in I_j; \tag{28}$$

a patch $W \supset K \setminus \bigcup_j C_j$ with $\overline{W} \cap \Sigma = \emptyset$; and a C^1 partition of unity $\chi_0, \chi_1, \dots, \chi_N$ subordinate to $\{W, C_1, \dots, C_N\}$ with $\sum_{j \geq 0} \chi_j \equiv 1$ on a neighborhood of K and $\operatorname{supp} \chi_j \Subset C_j$ for $j \geq 1$. By

the shrinking performed there using (S3), p is continuous on each cylinder C_j; and each C_j lies in the set V of Lemma 2, so $\|\nabla_\xi g(\theta_0,\cdot)\| \geq c_0$ on $\bigcup_j C_j$.

Choice of δ_0. Since the open cylinders $C_1,\dots,C_N$ cover the compact set $K\cap\Sigma$, the compactness argument that proves (16) in Lemma 2 yields a radius $\delta_1 \in (0,\delta_0]$ with $\{\xi \in K : |g(\theta_0,\xi)| < \delta_1\} \subset \bigcup_{j\geq 1} C_j$; shrinking δ_1 once more (possible since $0\in I_j$ and each I_j is open) we also arrange $(-\delta_1,\delta_1)\subset I_j$ for every j. The radius furnished by Lemma 2 may be taken equal to this δ_1: doing so leaves (14) unchanged and (16) valid, and every statement in the paper invokes the tube only through the defining properties (14)–(16). We write δ_0 for this radius and assume henceforth

$$\{\xi \in K : |g(\theta_0,\xi)| < \delta_0\} \subset \bigcup_{j=1}^{N} C_j, \qquad (-\delta_0,\delta_0) \subset \bigcap_{j=1}^{N} I_j. \tag{29}$$

The slice area element. Fix $j\geq 1$ and $s\in I_j$. By (28) the slice $\{g(\theta_0,\cdot)=s\}\cap C_j$ is the image of the injective C^1 map $y\mapsto \Psi_j(s,y)$, $y\in Y_j$. The computation behind (26) in Appendix B (the area formula for this injective C^1 parametrization, combined with the column decomposition of $\partial_s\Psi_j$ recorded there) is pointwise in (s,y), hence valid at every level $s\in I_j$ and not only at $s=0$; with $\mathcal{J}_j := |\det D\Psi_j| > 0$ it yields

$$\frac{1}{\|\nabla_\xi g(\theta_0,\cdot)\|}\, d\mathcal{H}^{m-1} \;=\; \mathcal{J}_j(s,y)\,dy \qquad \text{on } \{g(\theta_0,\cdot)=s\}\cap C_j. \tag{30}$$

Both $\mathcal{J}_j$ and $(s,y)\mapsto\|\nabla_\xi g(\theta_0,\Psi_j(s,y))\|$ are continuous on the compact set $\overline{I_j}\times\overline{Y_j}$.

Well-definedness. Fix $|s|<\delta_0$. The integrand of (7) is supported in $\operatorname{supp} p\subset K$; by (29), $\{g(\theta_0,\cdot)=s\}\cap\operatorname{supp} p\subset\bigcup_j C_j$, a set on which p is continuous, while on the open complement of $\operatorname{supp} p$ the integrand vanishes. Hence the integrand has a continuous trace along the slice $\{g(\theta_0,\cdot)=s\}\cap V$; being bounded with support of finite surface measure (by the estimate (31) below), it is $\mathcal{H}^{m-1}$-integrable, so $n(s,a)$ is a well-defined finite vector.

Boundedness. Put $\Lambda' := \sup_{[0,1]\times K}|\partial_a\Lambda(\,\cdot\,,\theta_0,\cdot)|$ and $G' := \sup_K\|\nabla_\theta g(\theta_0,\cdot)\|$, both finite by Assumption 2. The integrand of (7) is supported in $\operatorname{supp} p\subset K$, where $|\partial_a\Lambda|\leq\Lambda'$ and $\|\nabla_\theta g\|\leq G'$; moreover $|p|\leq\|p\|_\infty$ and $\|\nabla_\xi g(\theta_0,\cdot)\|\geq c_0$ on V. Therefore, for $|s|<\delta_0$ and $a\in(0,1)$, using the triangle inequality for the vector integral, (29), subadditivity of $\mathcal{H}^{m-1}$, and (30),

$$\begin{aligned}\|n(s,a)\| \;&\leq\; \frac{\Lambda'\,G'\,\|p\|_\infty}{c_0}\,\mathcal{H}^{m-1}\big(\{g(\theta_0,\cdot)=s\}\cap K\big)\\ &\leq\; \frac{\Lambda'\,G'\,\|p\|_\infty}{c_0}\sum_{j=1}^{N}\int_{Y_j}\mathcal{J}_j(s,y)\,\|\nabla_\xi g\big(\theta_0,\Psi_j(s,y)\big)\|\,dy.\end{aligned} \tag{31}$$

Each summand is at most $|Y_j|\cdot\sup_{\overline{I_j}\times\overline{Y_j}}\big(\mathcal{J}_j\,\|\nabla_\xi g(\theta_0,\Psi_j)\|\big)<\infty$, a bound independent of s and a. Hence $\sup_{(-\delta_0,\delta_0)\times(0,1)}\|n\|<\infty$.

Continuity at $s=0$. Fix $a\in(0,1)$. Since $\sum_{j\geq 0}\chi_j\equiv 1$ on a neighborhood of K, hence on $\{g(\theta_0,\cdot)=s\}\cap\operatorname{supp} p$ for $|s|<\delta_0$, we may write $n(s,a)=\sum_{j=0}^{N} n_j(s,a)$, where n_j is the integral (7) with the extra factor χ_j inserted under the integral sign.

Patch $j=0$. The set $\operatorname{supp}\chi_0\cap K$ is compact and disjoint from Σ (as $\overline{W}\cap\Sigma=\emptyset$), so $|g(\theta_0,\cdot)|\geq\delta_W$ on it for some $\delta_W>0$. For $|s|<\delta_W$ the slice $\{g(\theta_0,\cdot)=s\}$ misses $\operatorname{supp}\chi_0\cap\operatorname{supp} p$ (because $\operatorname{supp} p\subset K$), so the n_0-integrand vanishes on it; thus $n_0(s,a)=0$, and in particular $n_0(s,a)\to 0=n_0(0,a)$ as $s\to 0$.

Patches $j\geq 1$. Fix $j\geq 1$. For $|s|<\delta_0$ we have $s\in I_j$ by (29), and $\operatorname{supp}\chi_j\Subset C_j=\Psi_j(I_j\times Y_j)$, so the n_j-integrand is supported in C_j; pushing it to the chart by (30),

$$n_j(s,a)=\int_{Y_j}G_j(s,y,a)\,dy, \qquad G_j(s,y,a):=\Big(\chi_j\,\partial_a\Lambda(a,\theta_0,\cdot)\,\nabla_\theta g(\theta_0,\cdot)\,p\,\mathcal{J}_j\Big)\big(\Psi_j(s,y)\big).$$

For fixed a and y, the map $s\mapsto\Psi_j(s,y)$ is continuous into C_j, on which p is continuous; $\partial_a\Lambda(a,\theta_0,\cdot)$ is continuous by Assumption 2, $\nabla_\theta g(\theta_0,\cdot)$ by (S1), and $\chi_j,\mathcal{J}_j$ are continuous. Hence

$G_j(s,y,a) \to G_j(0,y,a)$ as $s \to 0$. Moreover $\|G_j(s,y,a)\| \le \sup_{\overline{I_j}\times\overline{Y_j}\times[0,1]}\|G_j\| =: M_j < \infty$, a constant integrable over the bounded set Y_j. By dominated convergence, $n_j(s,a) \to n_j(0,a)$ as $s \to 0$.

Summing the finitely many patch limits gives $n(s,a) \to n(0,a)$, which is (27). □

D Proof of Theorem 1 and Corollary 2

Proof. Fix any $\delta \in (0, \delta_0]$ and split the expectation according to the tube (15):

$$\mathbb{E}[Z_\kappa] = \underbrace{\int_{\{|g(\theta_0,\cdot)|<\delta\}} Z_\kappa\, p\, d\xi}_{=:A_\kappa^{\text{in}}} + \underbrace{\int_{\{|g(\theta_0,\cdot)|\ge\delta\}} Z_\kappa\, p\, d\xi}_{=:A_\kappa^{\text{out}}}.$$

Step 1: the tail is exponentially small. On $\{|g(\theta_0,\cdot)| \ge \delta\}$, by (2) and $\sigma'(t) \le e^{-|t|}$,

$$|a'_\kappa(g)| = \kappa\,\sigma'(-\kappa g) \le \kappa\, e^{-\kappa|g|} \le \kappa\, e^{-\kappa\delta}.$$

With $C_{\text{out}} := \sup|\partial_a\Lambda| \cdot \sup\|\nabla_\theta g\| < \infty$ (Assumption 2) and $\int p = 1$,

$$\|A_\kappa^{\text{out}}\| \le \int_{\{|g|\ge\delta\}} \|Z_\kappa\|\, p\, d\xi \le C_{\text{out}}\,\kappa\, e^{-\kappa\delta} \xrightarrow[\kappa\to\infty]{} 0.$$

Step 2: co-area on the tube. By (16), $\{\xi \in \operatorname{supp} p : |g(\theta_0,\xi)| < \delta\} \subset T_\delta \subset V$, where $\|\nabla_\xi g(\theta_0,\cdot)\| \ge c_0 > 0$. The integrand $Z_\kappa\, p$ is supported in this set, so Lemma 1 in the form (13) applies with $w = g(\theta_0,\cdot)$. Both factors $a_\kappa(g) = \sigma(-\kappa s)$ and $a'_\kappa(g) = -\kappa\sigma'(-\kappa s)$ depend on ξ only through $s = g(\theta_0,\xi)$. Hence, with n as in (7),

$$A_\kappa^{\text{in}} = \int_{-\delta}^{\delta} \big(-\kappa\,\sigma'(-\kappa s)\big)\, n\big(s, \sigma(-\kappa s)\big)\, ds.$$

Step 3: change of variables $a = \sigma(-\kappa s)$. On $s \in (-\delta,\delta)$ the map $s \mapsto a = \sigma(-\kappa s)$ is a strictly decreasing C^1 bijection onto $\big(\sigma(-\kappa\delta),\, 1-\sigma(-\kappa\delta)\big)$, with

$$da = -\kappa\,\sigma'(-\kappa s)\, ds, \qquad s = -\frac{\sigma^{-1}(a)}{\kappa}, \qquad \sigma^{-1}(a) = \log\frac{a}{1-a}.$$

Because the map is decreasing, $s = -\delta \mapsto a = 1-\sigma(-\kappa\delta)$ and $s = \delta \mapsto a = \sigma(-\kappa\delta)$; ordering the limits ascending introduces a single minus sign:

$$A_\kappa^{\text{in}} = \int_{a=1-\sigma(-\kappa\delta)}^{a=\sigma(-\kappa\delta)} n\Big(-\tfrac{\sigma^{-1}(a)}{\kappa}, a\Big)\, da = -\int_{\sigma(-\kappa\delta)}^{1-\sigma(-\kappa\delta)} n\Big(-\tfrac{\sigma^{-1}(a)}{\kappa}, a\Big)\, da.$$

Step 4: dominated convergence. Fix $a \in (0,1)$. For κ large enough, a lies in the integration window, and $-\sigma^{-1}(a)/\kappa \to 0$, so by (27) the integrand tends to $n(0,a)$. It is dominated, uniformly in κ, by the constant $M := \sup_{(-\delta_0,\delta_0)\times(0,1)}\|n\| < \infty$ (Proposition 4), which is integrable on $(0,1)$. Extending the integrand by 0 outside the window and applying dominated convergence,

$$\lim_{\kappa\to\infty} A_\kappa^{\text{in}} = -\int_0^1 n(0,a)\, da.$$

Step 5: telescoping and identification with B. Expand $n(0,a)$ via (7); the level set is $\{g(\theta_0,\cdot) = 0\} \cap V = \Sigma \cap V \supset \Sigma \cap \operatorname{supp} p$. By Fubini (legitimate since the integrand is bounded and $\Sigma \cap K$ has finite surface measure),

$$-\int_0^1 n(0,a)\, da = -\int_\Sigma \Big(\int_0^1 \partial_a\Lambda(a,\theta_0,\xi)\, da\Big)\nabla_\theta g(\theta_0,\xi)\, \frac{p(\xi)}{\|\nabla_\xi g(\theta_0,\xi)\|}\, d\mathcal{H}^{m-1}(\xi).$$

By the fundamental theorem of calculus and the endpoint condition (4),

$$\int_0^1 \partial_a \Lambda(a, \theta_0, \xi)\, da = \Lambda(1, \theta_0, \xi) - \Lambda(0, \theta_0, \xi) = L^-(\theta_0, \xi) - L^+(\theta_0, \xi) = -[\![L]\!](\theta_0, \xi).$$

Substituting, the two minus signs cancel:

$$-\int_0^1 n(0, a)\, da = \int_\Sigma [\![L]\!] \frac{\nabla_\theta g}{\|\nabla_\xi g\|}\, p\, d\mathcal{H}^{m-1} = B.$$

Combining with Step 1, $\mathbb{E}[Z_\kappa] = A_\kappa^{\text{in}} + A_\kappa^{\text{out}} \to B + 0 = B$. □

Remark 4 (Where the sign lives). Two minus signs enter and compose to give $+B$. The first is in Step 3: because $a = \sigma(-\kappa s)$ is decreasing (equivalently $a'_\kappa = -\kappa\sigma' < 0$), ordering the limits of integration ascending introduces a single minus, so $A_\kappa^{\text{in}} \to -\int_0^1 n(0, a)\, da$. The second is the endpoint orientation (4), fixed as $\Lambda(0, \theta_0, \cdot) = L^+$, $\Lambda(1, \theta_0, \cdot) = L^-$ rather than the reverse: it makes $\int_0^1 \partial_a \Lambda\, da = L^- - L^+ = -[\![L]\!]$. The two minus signs cancel, yielding $+B$. They are not independent degrees of freedom: choosing the decreasing gate $a_\kappa(s) = \sigma(-\kappa s)$ both fixes the change-of-variables orientation (Step 3) and forces the endpoint convention (4), since only $\Lambda(0, \theta_0, \cdot) = L^+$, $\Lambda(1, \theta_0, \cdot) = L^-$ makes the activation endpoints reproduce the correct branches. These are the two bookkeeping points where a sign enters, and they cancel by construction.

Proof of Corollary 2. By (5), $\mathbb{E}[\nabla_\theta L_\kappa] = \mathbb{E}[\partial_\theta \Lambda(a_\kappa(g), \theta_0, \cdot)] + \mathbb{E}[Z_\kappa]$. We show the first expectation tends to $\mathbb{E}[\nabla_\theta L]$ by dominated convergence. *Pointwise:* off Σ, $a_\kappa(g(\theta_0, \xi)) \to 0$ on $\{g > 0\}$ and $\to 1$ on $\{g < 0\}$; differentiating the extended branch-matching in θ at θ_0 gives $\partial_\theta \Lambda(0, \theta_0, \cdot) = \nabla_\theta L^+$ and $\partial_\theta \Lambda(1, \theta_0, \cdot) = \nabla_\theta L^-$, so joint continuity of $\partial_\theta \Lambda$ yields $\partial_\theta \Lambda(a_\kappa(g), \theta_0, \xi) \to \nabla_\theta L(\theta_0, \xi)$ for p-a.e. ξ (recall $\Sigma \cap \operatorname{supp} p$ is Lebesgue-null). *Domination:* $\partial_\theta \Lambda$ is jointly continuous, hence bounded on the compact set $[0, 1] \times \{\theta_0\} \times K$ by some C'; since $a_\kappa(g) \in [0, 1]$ and $\operatorname{supp} p \subset K$, the integrands are bounded uniformly in κ by the p-integrable constant C'. Dominated convergence then gives $\mathbb{E}[\partial_\theta \Lambda(a_\kappa(g), \theta_0, \cdot)] \to \mathbb{E}[\nabla_\theta L]$. The second expectation tends to B by Theorem 1. Adding and using Proposition 1, $\mathbb{E}[\nabla_\theta L_\kappa] \to \mathbb{E}[\nabla_\theta L] + B = \nabla_\theta F(\theta_0)$. □

E Proofs of Propositions 2 and 3 and Corollary 3

Proof of Proposition 2. From (6) and (2), with $g = g(\theta_0, \xi)$ and C a common bound on $|\partial_a \Lambda|$ and $\|\nabla_\theta g\|$ over $[0, 1] \times K$ (Assumption 2),

$$\|Z_\kappa(\xi)\|^2 = \big(\partial_a \Lambda\big)^2\, \kappa^2\, \sigma'(-\kappa g)^2\, \|\nabla_\theta g\|^2 \le C^4\, \kappa^2\, \sigma'(-\kappa g)^2.$$

Split the expectation at $|g| = \delta_0$. On the tail $\{|g| \ge \delta_0\}$, $\sigma'(t) \le e^{-|t|}$ gives $\kappa^2 \sigma'(-\kappa g)^2 \le \kappa^2 e^{-2\kappa\delta_0}$, so

$$\int_{\{|g| \ge \delta_0\}} \|Z_\kappa\|^2 p\, d\xi \le C^4 \kappa^2 e^{-2\kappa\delta_0} \to 0.$$

On the tube $\{|g| < \delta_0\} \cap \operatorname{supp} p \subset V$, Lemma 1 applies and $\sigma'(-\kappa g)^2$ depends on ξ only through $s = g(\theta_0, \xi)$; with $\ell(s) := \int_{\{g(\theta_0, \cdot) = s\} \cap V} p \,/\, \|\nabla_\xi g(\theta_0, \cdot)\|\, d\mathcal{H}^{m-1}$,

$$\int_{\{|g| < \delta_0\}} \|Z_\kappa\|^2 p\, d\xi \le C^4 \int_{-\delta_0}^{\delta_0} \kappa^2 \sigma'(-\kappa s)^2\, \ell(s)\, ds \le C^4 H\, \kappa^2 \int_{\mathbb{R}} \sigma'(-\kappa s)^2\, ds,$$

where $\ell \le H < \infty$ on $(-\delta_0, \delta_0)$ because its integrand is at most $\|p\|_\infty / c_0$ and vanishes off $\operatorname{supp} p \subset K$, while $\sup_{|s| < \delta_0} \mathcal{H}^{m-1}(\{g(\theta_0, \cdot) = s\} \cap K) < \infty$ by (31). With $t = \kappa s$, $\kappa^2 \int_{\mathbb{R}} \sigma'(-\kappa s)^2\, ds = \kappa \int_{\mathbb{R}} \sigma'(-t)^2\, dt$, and $\int_{\mathbb{R}} \sigma'(-t)^2\, dt < \infty$ because σ' decays exponentially. Hence this contribution is $O(\kappa)$, and $\mathbb{E}\|Z_\kappa\|^2 = O(\kappa)$. For independent samples,

$$\operatorname{tr} \operatorname{Cov}\big(\widehat{Z}_{\kappa, N}\big) = \frac{1}{N} \operatorname{tr} \operatorname{Cov}(Z_\kappa) \le \frac{1}{N} \mathbb{E}\|Z_\kappa\|^2 = O\Big(\frac{\kappa}{N}\Big).$$ □

Proof of Proposition 3. Write $\alpha_\kappa := \sigma(-\kappa\delta_0)$, so that $\alpha_\kappa \le e^{-\kappa\delta_0}$ and $1-\alpha_\kappa = \sigma(\kappa\delta_0)$, and let $M_j < \infty$ bound $\|\partial_s^j n\|$ and $|\partial_s^j m|$ on $(-\delta_0,\delta_0)\times(0,1)$ for every order j available under Assumption 3, with $M_0 < \infty$ by (31); the maps $a \mapsto \partial_s^j n(0,a)$ and $a \mapsto \partial_s^j m(0,a)$ are measurable, being pointwise limits of difference quotients of $a \mapsto n(s,a)$ and $a \mapsto m(s,a)$, which are continuous by dominated convergence since $\partial_a\Lambda$ is jointly continuous and bounded. The logistic law has finite moments of every order, $\mu_k := \int_0^1 |\sigma^{-1}(a)|^k\, da < \infty$, with $\mu_2 = \pi^2/3$; and for $\alpha \le \frac12$, $\int_0^\alpha |\sigma^{-1}(a)|\, da \le \int_0^\alpha \log\frac{1}{a}\, da = \alpha(1-\log\alpha)$, so every window end below is exponentially small in κ.

Bias. Steps 1–3 of the proof of Theorem 1, with $\delta = \delta_0$, give

$$\mathbb{E}[Z_\kappa] = -\int_{\alpha_\kappa}^{1-\alpha_\kappa} n(s_a,a)\, da + A_\kappa^{\mathrm{out}}, \qquad s_a := -\frac{\sigma^{-1}(a)}{\kappa}, \qquad \|A_\kappa^{\mathrm{out}}\| \le C_{\mathrm{out}}\,\kappa\, e^{-\kappa\delta_0}, \quad (32)$$

with $s_a \in (-\delta_0,\delta_0)$ for a in the window. Taylor's formula with integral remainder along $s \mapsto n(s,a)$ gives, for a in the window,

$$n(s_a,a) = n(0,a) + s_a\,\partial_s n(0,a) + R(a), \qquad R(a) = \int_0^{s_a} (s_a - t)\,\partial_s^2 n(t,a)\, dt,$$

so that $\|R(a)\| \le \frac12 M_2\, s_a^2 = M_2\,\sigma^{-1}(a)^2/(2\kappa^2)$. Inserting this into (32) and extending the integrals from the window to $(0,1)$,

$$\mathbb{E}[Z_\kappa] = -\int_0^1 n(0,a)\, da + \frac{1}{\kappa}\int_0^1 \sigma^{-1}(a)\,\partial_s n(0,a)\, da + E_\kappa,$$

where E_κ collects the remainder, of norm at most $M_2\mu_2/(2\kappa^2)$, the tail A_κ^{out}, and the two window ends $(0,\alpha_\kappa)$ and $(1-\alpha_\kappa,1)$, on which the integrands are bounded by M_0 and by $M_1|\sigma^{-1}|/\kappa$ and which therefore contribute $O(e^{-\kappa\delta_0})$ and $O(\alpha_\kappa(1-\log\alpha_\kappa)/\kappa)$. Hence $\|E_\kappa\| = O(\kappa^{-2})$, and $-\int_0^1 n(0,a)\, da = B$ by Step 5 of the same proof, which is the first expansion in (10) with C_1 as in (11). Under the C^3 clause, the same argument with one more Taylor term, $\frac12 s_a^2\,\partial_s^2 n(0,a)$, and remainder at most $M_3|s_a|^3/6$ gives $C_2 = -\frac12\int_0^1 \sigma^{-1}(a)^2\,\partial_s^2 n(0,a)\, da$ with error $O(\kappa^{-3})$.

Second moment. The proof of Proposition 2 splits $\mathbb{E}\|Z_\kappa\|^2$ at $|g| = \delta_0$ with a tail at most $C^4\kappa^2 e^{-2\kappa\delta_0}$, and on the tube (13) applies with all ξ-dependence of $\kappa^2\sigma'(-\kappa g)^2$ through $s = g(\theta_0,\xi)$ and that of $\partial_a\Lambda$ through $a = \sigma(-\kappa s)$, so the tube contributes $\int_{-\delta_0}^{\delta_0} \kappa^2\sigma'(-\kappa s)^2\, m\big(s,\sigma(-\kappa s)\big)\, ds$. The substitution $a = \sigma(-\kappa s)$ of Step 3, with $\sigma'(-\kappa s) = a(1-a)$ and $\kappa^2\sigma'(-\kappa s)^2\, ds = -\kappa\, a(1-a)\, da$, turns this into

$$\kappa\int_{\alpha_\kappa}^{1-\alpha_\kappa} a(1-a)\, m(s_a,a)\, da.$$

Taylor expanding $m(s_a,a) = m(0,a) + s_a\,\partial_s m(0,a) + O(M_2 s_a^2)$ and extending the window as before (the ends now contribute $O(\kappa\alpha_\kappa^2)$ and $O(\alpha_\kappa)$, since $a(1-a) \le a$ and $a(1-a)|\sigma^{-1}(a)| \le 1$ there),

$$\mathbb{E}\|Z_\kappa\|^2 = \kappa\int_0^1 a(1-a)\, m(0,a)\, da - \int_0^1 a(1-a)\,\sigma^{-1}(a)\,\partial_s m(0,a)\, da + O(\kappa^{-1}),$$

which is the second expansion in (10) with V_1, V_0 as in (11). □

Proof of Corollary 3. For the mean of N independent draws, $\mathbb{E}\|\widehat{Z}_{\kappa,N} - B\|^2 = \|\mathbb{E}[Z_\kappa] - B\|^2 + \operatorname{tr}\operatorname{Cov}(Z_\kappa)/N$, and $\operatorname{tr}\operatorname{Cov}(Z_\kappa) = \mathbb{E}\|Z_\kappa\|^2 - \|\mathbb{E}[Z_\kappa]\|^2 = V_1\kappa + O(1)$ by Proposition 3, since $\mathbb{E}[Z_\kappa] = B + O(\kappa^{-1})$; the bias expansion gives $\|\mathbb{E}[Z_\kappa] - B\|^2 = \|b_r\|^2\kappa^{-2r}\big(1 + O(\kappa^{-1})\big)$, which is (12). The function $\kappa \mapsto \|b_r\|^2\kappa^{-2r} + V_1\kappa/N$ is minimized at $\kappa_N = (2r\|b_r\|^2 N/V_1)^{1/(2r+1)}$, where both terms are $\Theta(N^{-2r/(2r+1)})$, so $\mathrm{MSE}(\kappa_N, N)$ is too: at $\kappa = \kappa_N$ the corrections in (12), of orders κ_N^{-2r-1} and N^{-1}, are $O(N^{-1}) = o(N^{-2r/(2r+1)})$. □

F EXPERIMENTAL DETAILS

Configurations. One-step instance: $\gamma = 1$, $\Delta t = T = 0.05$ (one step, $u_N = \gamma\Delta t = 0.05$), $w = 0.2$, $\theta_0 = 0$, $\xi_\star = 0$, bump radius $R = 0.02$ centred at $\xi_\star - R/2 = -0.01$; the centred variant has the bump at $\xi_\star$. The bump is $p(\xi) = (RZ_f)^{-1}\exp\big(-1/(1-u^2)\big)\mathbf{1}\{|u| < 1\}$ with $u = (\xi - \xi_c)/R$ and $Z_f = 0.443994$, so that at $\xi_\star$ ($u = \frac{1}{2}$) $p = 29.6848$, $p' = -2638.65$, $p'' = -381138$, while for the centred bump $p = 41.4284$, $p' = 0$, $p'' = -207142$. Multi-step instance: the same ball released from $T = 1$, i.e. 20 steps of $\Delta t = 0.05$ with the contact at step 19, $u_N = \gamma T = 1$, friction coefficient 0.3 (all-stick, since $0.3\,u_N \geq w$), $\xi_\star = 0.475$ (root-found on the hard stepper), bump centred at $\xi_\star - R/2 = 0.465$; the smallest non-contact gap on the support is $g_{\min} = 0.0175$. Two-dimensional configuration: p_1 as in the one-step instance, p_2 of radius 0.007 centred at 0, $h(\xi_2) = 0.8\,\xi_2 + 20\,\xi_2^2$, so that $h' \in [0.52, 1.08]$ and the weight $1/\sqrt{1+h'^2} \in [0.679, 0.887]$ on $\operatorname{supp} p_2$, $w(\xi_2) = 0.2 + 5\xi_2 \in [0.165, 0.235]$, mixed readout with $\lambda = 1$, jump along Σ in $[0.0136, 0.0276]$; the surface stays inside the support of p_1, at a distance between $0.27R$ and $0.83R$ from its centre. The terminal height $z_1 = \theta + \xi_1 - h(\xi_2) - \frac{1}{2}(1-a)\,u_N\Delta t$ is the Moreau position half-step taken with the post-impulse velocity; its free-side endpoint is the exact constant-gravity drop and its stuck-side endpoint freezes the height at the midpoint configuration, so the height carries the discretization jump $[\![z_1]\!] = -\frac{1}{2}\gamma\Delta t^2$: in continuous time position is continuous across an impact, but the stepper's branches differ by half a step. At $\lambda = 1$ the readout is the terminal mechanical energy with the floor as datum. Before being combined, the mixed readout was run at scalar noise and the velocity and kinetic-energy readouts in the two-dimensional geometry, each validated separately; their constants appear in Table 2.

Estimator and moments. $Z_\kappa(\xi)$ is the difference of two forward-mode tangent rollouts of the same stepper, one propagating the sensitivity of the gate argument and one freezing it; in the one-step instance it agrees with the closed form $\partial_a\Lambda\, a'_\kappa(g)$ to 1.4×10^{-14} over 257 values of ξ and 30 values of $\kappa \in [30, 2\times10^4]$. $\mathbb{E}[Z_\kappa]$ and $\mathbb{E}\|Z_\kappa\|^2$ are computed by a 2000-node Gauss–Legendre rule in $t = \kappa(\xi - \xi_\star)$ on $|t| \leq 45$ clipped to the support, cross-checked against adaptive quadrature with a break at $\xi_\star$ (agreement 10^{-13} up to $\kappa = 5\times10^4$) and against Monte Carlo (20 000 samples, 64 repetitions). In two dimensions the same rule is applied on each of 400 Gauss–Legendre nodes in ξ_2 weighted by p_2; doubling the outer rule changes B and $\mathbb{E}[Z_\kappa]$ by less than 3×10^{-14}. In two dimensions the tangent rollout matches $\partial_a\Lambda(a;\xi_2)\,a'_\kappa(g)$ to 2.8×10^{-14} over 21 slices in ξ_2, 401 values of ξ_1 and 25 values of κ; setting $h \equiv 0$ and w constant reproduces the scalar-noise values to 5×10^{-15}, and $\lambda = 0$ reproduces the kinetic-energy readout in the same geometry exactly. B is the closed form $[\![L]\!]\,p(\xi_\star)$ with the branches read off the hard stepper on either side of $\xi_\star$, and in two dimensions the outer-rule sum of the slice values; keeping the arc-length element without the co-area weight would give 0.704 instead of 0.551 for the mixed readout, 1.2×10^3 times the residual $|\mathbb{E}[Z_\kappa] - B|$ at $\kappa = 5\times10^5$. $\partial_\theta F$ is a central difference with step 10^{-4} of $F(\theta) = \mathbb{E}[L]$, integrated adaptively with a break at the moving jump.

Grids, fit windows, tradeoff. Every sweep in Table 1 uses 90 log-spaced values of $\kappa \in [50, 10^5]$ (κR from 1 to 2000); the multi-step slopes quoted below are fitted on a 30-point grid in $[50, 5000]$, on $\kappa \geq 1000$ (11 points); the multi-step curves of Figure 2 are drawn on the same 90-point grid as the one-step instance. Slopes are ordinary least-squares fits of $\log|\cdot|$ on $\log\kappa$. Windows are determined, not chosen: the lower edge is slid upward with the upper edge fixed at 10^5 and at least six points, the plateau slope is the median over the 17 most asymptotic windows, and the reported window starts at the smallest lower edge beyond which every window lies within 0.04 of the plateau slope. Lower edges and plateau slopes: one-step velocity, bias 327 ($\kappa R = 6.5$), plateau -2.000, second moment 59; one-step kinetic energy, bias 837 ($\kappa R = 16.7$), plateau -1.004, second moment 327; centred-bump kinetic energy, bias 77, plateau -2.000; two-dimensional velocity, bias 502, plateau -2.000, second moment 50; mixed energy, bias 91, plateau -1.003, second moment 388. The $\kappa^*(N)$ and $\mathrm{MSE}^*(N)$ exponents are built from the quadrature bias and variance, $\mathrm{MSE}(\kappa, N) = \big(\mathbb{E}[Z_\kappa] - B\big)^2 + \mathrm{Var}(Z_\kappa)/N$ on the κ grid, with $\kappa^*(N)$ the grid minimizer refined by a parabola through its neighbours in $\log\kappa$ and $\mathrm{MSE}^*(N)$ the grid minimum, for 18 log-spaced $N \in [10^2, 10^8]$; the minimizers lie in $[403, 7967]$ for velocity and $[782, 61\,947]$ for kinetic energy.

Multi-step instance. Its bias and second-moment slopes are -1.988 and $+1.004$ (velocity) and -1.104 and $+1.043$ (kinetic energy) against the predicted $-2, +1$ and $-1, +1$; its bias curves are

the open markers of Figure 2. The 19 inert gates perturb the single-gate form $\Lambda(a_\kappa(g_{19}))$, with g_{19} the contact-step gap of the hard rollout, by at most 8.8% of $\max|L_\kappa|$ over the support at $\kappa = 50$ (velocity; 7.1% for kinetic energy), 0.08% at $\kappa = 200$, 10^{-5} at $\kappa = 300$, and by nothing beyond double precision for $\kappa \geq 750$; on the fitted range $\kappa \geq 1000$ the unrolled return is the single-gate form to machine precision. The measured decay is $e^{-0.0475\kappa}$, where $0.0475 = \min_{n\neq 19} g_n(\xi_\star)$ is the nearest inert gap at the switching point; the uniform bound $e^{-\kappa g_{\min}}$ with $g_{\min} = 0.0175$, the minimum of that gap over the support, holds but is not sharp, because where the inert gap is smallest the contact gate is engaged and multiplies the perturbation by $1 - a_\kappa \approx e^{-\kappa(\xi_\star - \xi)}$. We prove neither statement. On the fitted range the multi-step gate term is therefore the one-step gate term with $u_N = 1$ in place of 0.05: the ratios of bias and variance to the one-step values are 1 and 1 for velocity and $24.47 = 1.04/0.0425$ and 24.47^2 for kinetic energy at every $\kappa \geq 1000$, and its $\kappa^*(N)$ and $\mathrm{MSE}^*(N)$ exponents, 0.2097 and -0.7732 (velocity) and 0.3175 and -0.6663 (kinetic energy), coincide with the one-step ones to six figures because both are invariant under a rescaling of Z_κ. The instance checks that the inert gates leave the plateau intact; it is not independent evidence of the rates.

Constants. Table 2 compares the constants of Proposition 3, evaluated from the closed forms of Section 5 (for kinetic energy $V_1 = p(\xi_\star)S^2/20$ and $V_0 = p'(\xi_\star)S^2/24$ with $S = w^2 + u_N^2$, since $\int_0^1 a(1-a)^3\sigma^{-1}(a)\,da = -\frac{1}{24}$; for velocity $V_1 = p(\xi_\star)w^2/6$) and the bump derivatives above, with the data at the top of the grid, $\kappa = 10^5$: $\kappa|\mathbb{E}[Z_\kappa] - B| \to |C_1|$, $\kappa^2|\mathbb{E}[Z_\kappa] - B| \to |C_2|$ when $C_1 = 0$, $\mathbb{E}\|Z_\kappa\|^2/\kappa \to V_1$, and $\mathbb{E}\|Z_\kappa\|^2 - V_1\kappa \to V_0$. For kinetic energy $V_0 < 0$ ($p'(\xi_\star) < 0$), and the mean of $-V_0/(V_1\kappa)$ over its second-moment fit window, $+0.039$, equals the measured excess of the slope over 1; for velocity $V_0 = 0$ and the same excess is pre-asymptotic, the window reaching $\kappa R \approx 1$, with slope 1.001 on $\kappa \geq 1000$.

instance, readout	$\lvert C_1\rvert$ pred. / data	$\lvert C_2\rvert$ pred. / data	V_1 pred. / data	V_0 pred. / data
one-step, velocity	0 / 0	125389 / 125389	0.19790 / 0.19790	0 / −0.016
one-step, kinetic energy	56.07 / 56.20	13323 / 13334	0.002681 / 0.002679	−0.1986 / −0.1989
centred, kinetic energy	0 / 0	7240.6 / 7240.6	—	—
one-step, mixed λ=1	56.07 / 56.20	—	0.002426 / 0.002424	−0.1869 / −0.1872
two-dim., velocity	0 / 0	—	0.18383 / 0.18383	0 / −0.016
two-dim., kinetic energy	61.08 / 61.20	—	0.002471 / 0.002469	−0.2293 / −0.2296
two-dim., mixed λ=1	61.08 / 61.19	—	0.002234 / 0.002232	−0.2166 / −0.2169

Table 2: Constants of Proposition 3 predicted from the closed forms against their finite-κ estimates at $\kappa = 10^5$. For $C_1 \neq 0$ the data entry for $|C_2|$ is $\kappa^2|\mathbb{E}[Z_\kappa] - B - C_1/\kappa|$; the residuals in the $|C_1|$ column (0.13 for kinetic energy) are $|C_2|/\kappa$, and those in the V_1 column are V_0/κ. Data entries of 0 are quantities that decay to 0 ($\kappa|\mathbb{E}[Z_\kappa] - B| = 1.25$ at $\kappa = 10^5$ for one-step velocity, against $C_2/\kappa = 1.25$).