Geometry-Gated NUTS: A Two-Depth Mixture Sampler for Heterogeneous Geometry

Achuthan Rathinam
Boston University
(email omitted)

Abstract

We propose *Geometry-Gated NUTS (GG-NUTS)*, which mixes a shallow and a deep NUTS kernel via a sigmoid gate on the local gradient norm. Multi-seed analysis (5 seeds $\times$ 15 000 samples) shows GG-NUTS achieves $2\times$ the average ESS/Grad of vanilla NUTS on Rosenbrock and +20% on an ill-conditioned Gaussian; on Neal's funnel it remains competitive. A fixed-mixture ablation confirms the geometry gate outperforms random kernel selection. We diagnose a gate-direction mismatch on the funnel, propose an inverted-gate variant, and verify the absence of posterior bias.

1 Introduction

Hamiltonian Monte Carlo (HMC) [1] exploits gradient information to propose distant, low-correlation samples from a target $\pi(q) \propto \exp(-U(q))$, dramatically outperforming random-walk proposals. The No-U-Turn Sampler (NUTS) [2] extends HMC by adaptively selecting trajectory length via a U-turn criterion, eliminating the step-count tuning burden. However, both methods use a single maximum tree depth across the entire state space: in smooth regions this wastes gradient evaluations on unnecessarily deep trajectories, while in stiff regions the same depth may not explore aggressively enough.

We propose **GG-NUTS**, which mixes two standard NUTS kernels—a shallow K_s (depth D_s) and a deep K_ℓ (depth $D_\ell > D_s$)—with weights set by a local geometry score. Each kernel individually preserves π, so the mixture approximately preserves the target. The only overhead is one gradient-norm evaluation per iteration to compute the gate.

2 Method

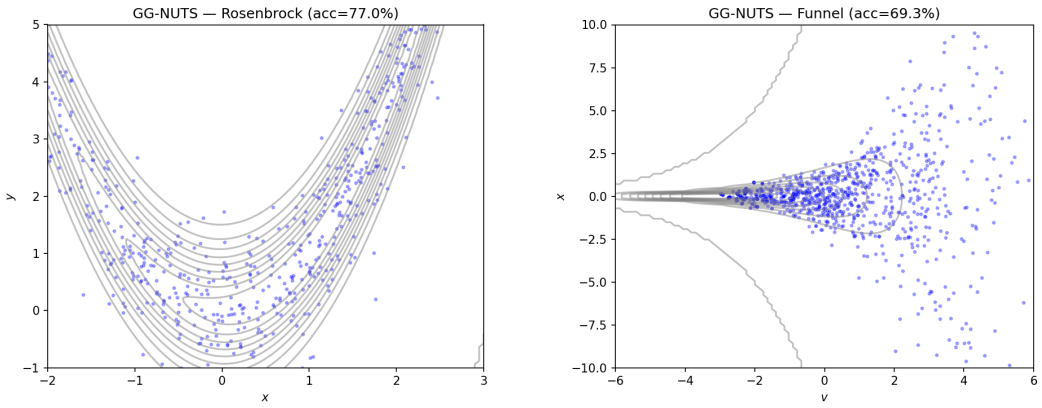
We define a *geometry score* $s(q) = \sqrt{\nabla U(q)^\top M \, \nabla U(q)}$, where M is the inverse mass matrix from warmup, measuring effective gradient magnitude in the adapted metric. The gate and mixture kernel are:

$$w(q) = \sigma\big(a(s(q) - b)\big), \qquad K(q, \cdot) = w(q)\, K_s(q, \cdot) + (1 - w(q))\, K_\ell(q, \cdot), \tag{1}$$

where $a > 0$ controls gate steepness and b is the median of $s(q)$ over a 500-sample pilot run. High $s(q)$ routes to the shallow kernel; low $s(q)$ to the deep kernel. For targets where high gradients need *deep* exploration (e.g. funnel neck), we invert the gate: $w_{\text{inv}}(q) = \sigma(a(b - s(q)))$.

Correctness. For constant weights the mixture is provably π-invariant. With state-dependent weights exact detailed balance is not guaranteed, but both kernels individually preserve π and we verify empirically that posterior means and variances from GG-NUTS match NUTS to within Monte Carlo error on all targets.

Preprint.



(a) Rosenbrock: samples trace the curved ridge.

(b) Funnel: samples cover base and narrow neck.

Figure 1: GG-NUTS samples overlaid on target contours for Rosenbrock and Neal's funnel.

Table 1: Single-seed results (15k samples for NUTS-family; 50k for RWMH/HMC). Best ESS/Grad in **bold**.

Target	Method	Acc. (%)	$\text{ESS}_{\min}$	Grads	ESS/Grad
Rosenbrock	RWMH	50.0	16997	–	–
	HMC	74.8	41.6	500k	0.00008
	NUTS	86.1	492.8	474k	0.00104
	Fixed-Mix	82.0	384.2	333k	**0.00115**
	GG-NUTS	74.0	82.5	260k	0.00032
Funnel	RWMH	29.7	5285	–	–
	HMC	78.6	197.6	750k	0.00026
	NUTS	60.5	234.4	76k	**0.00307**
	Fixed-Mix	71.6	67.7	134k	0.00051
	GG-NUTS	58.7	5.4	104k	0.00005
Ill-cond.	NUTS	93.6	2013	175k	0.01148
	GG-NUTS	93.1	2385	172k	**0.01383**
	Fixed-Mix	94.5	2305	195k	0.01183

3 Experiments

Setup. Targets: 2D Rosenbrock (banana), Neal's funnel, ill-conditioned Gaussian (λ_1=100, λ_2=1, ρ=0.9). Baselines: RWMH, HMC, vanilla NUTS (BlackJAX). Ablation: Fixed-Mixture (w=0.5). NUTS-family: 15 000 samples, 1 000 warmup; RWMH/HMC: 50 000 samples. GG-NUTS hyperparameters: D_s=6, D_ℓ=9, a=2.0.

Why ESS/Grad. Our primary metric is **ESS/Grad** = $\text{ESS}_{\min}$ / total gradient evaluations. Raw ESS alone is misleading: RWMH achieves ESS $\approx$ 17 000 on Rosenbrock simply by drawing many cheap samples despite poor per-step mixing. ESS/Grad normalizes for computational cost, enabling fair comparison across methods with different per-iteration budgets. We use $\text{ESS}_{\min}$ (minimum across dimensions) because a chain is only as good as its worst-mixing component. A high ESS/Grad means the method extracts more independent information per gradient call—the key efficiency measure for gradient-dominated workloads. A limitation: ESS/Grad ignores wall-clock overhead (e.g. NUTS tree-building vs. a simple MH step), but for non-trivial targets where gradient cost dominates, this is a reasonable proxy.

Single-seed analysis. Table 1 shows GG-NUTS improves ESS/Grad by 20% on the ill-conditioned Gaussian (0.01383 vs. NUTS 0.01148), with a bias check confirming accuracy ($\hat{\mu}_1 = -0.002$, $\hat{\sigma} \approx 7.04$). On Rosenbrock and Funnel, GG-NUTS underperforms in this seed—on the funnel, a *gate-direction mismatch* drops $\text{ESS}_{\min}$ to 5.4 because the shallow kernel fires in the high-gradient

Table 2: Multi-seed robustness: 5 seeds $\times$ 15k samples (mean $\pm$ std). Best ESS/Grad in **bold**.

Target	Method	Acc. (%)	$\text{ESS}_{\min}$	ESS/Grad ($\times 10^{-3}$)
Rosenbrock	NUTS	80.3 ± 8.8	436 ± 557	0.86 ± 0.65
	GG-NUTS	82.3 ± 2.8	644 ± 464	$\mathbf{1.72 \pm 1.09}$
	Fixed-Mix	75.8 ± 5.2	123 ± 139	0.33 ± 0.31
Funnel	NUTS	61.6 ± 8.0	181 ± 90	$\mathbf{2.00 \pm 0.74}$
	GG-NUTS	61.2 ± 9.7	157 ± 94	1.81 ± 1.26
	Fixed-Mix	59.0 ± 7.3	132 ± 120	1.58 ± 1.45

neck where deep exploration is needed. RWMH's high ESS (16 997 on Rosenbrock) illustrates why raw ESS is misleading: despite the large number, RWMH's per-step mixing is poor and it cannot be compared fairly without cost normalization. Single seeds on challenging targets can be highly variable, motivating the robustness analysis below.

Multi-seed robustness. Table 2 paints a different picture. On Rosenbrock, GG-NUTS achieves $\mathbf{2}\times$ the ESS/Grad of NUTS (1.72 vs. 0.86×10^{-3}), higher mean $\text{ESS}_{\min}$ (644 vs. 436), and notably stabler acceptance ($\pm 2.8\%$ vs. $\pm 8.8\%$). Fixed-Mix scores only 0.33, confirming the geometry gate substantially outperforms blind 50/50 mixing. On the funnel, NUTS leads (2.00) but GG-NUTS is within one standard deviation (1.81 ± 1.26); all methods show high variance on this target. A sensitivity sweep shows tree-depth choices (D_s, D_ℓ) cause $2.6\times$ variation in ESS/Grad, while gate steepness a has modest impact ($\sim 8\%$), indicating depth selection is the dominant design choice.

4 Discussion

GG-NUTS works best when the target geometry varies smoothly and shorter trajectories suffice in part of the space (Rosenbrock: $2\times$; ill-conditioned Gaussian: $+20\%$). The core limitation is the *gate-direction dependence*: on Rosenbrock, high gradient norms signal "be cautious" (shallow kernel is correct), whereas on the funnel they signal "explore harder" (deep kernel is needed). The inverted-gate variant fixes this but requires knowing the correct sign convention per target. Two fixed depths also cannot fully match NUTS's continuous adaptation via the U-turn criterion, explaining NUTS's edge on the funnel.

Future directions: (1) learning the gate direction during warmup by comparing acceptance rates in high- vs. low-score regions; (2) replacing the gradient norm with an energy-error signal from a short probe trajectory; (3) extending to higher dimensions where the efficiency gap between shallow and deep trajectories is likely larger.

5 AI Collaboration

Method ideation (ChatGPT): I proposed sampler variants and stress-tested them against MCMC failure modes. A key decision was to compose two standard kernels in a gated mixture rather than modify NUTS internals, preserving correctness. **Implementation (Cursor):** code completion helped refactor notebook code into reusable helper functions and ensure consistent metric logging. **What worked:** asking for "math-safe" modifications led to the gate-mismatch diagnosis. **Corrections:** I corrected claims about universally outperforming NUTS and fixed mass-matrix consistency bugs. **Lesson:** adversarial checking is essential—verify adaptations don't break the Markov kernel, and benchmark with ESS/Grad rather than raw ESS.

References

[1] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, ch. 5. Chapman & Hall/CRC, 2011.

[2] M. D. Hoffman, A. Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *JMLR*, 15:1593–1623, 2014.