# Assignment 1: [Hybrid MALA + Global Jump]

**Jingwei Li**
Boston University
li0518@bu.edu

## Abstract

We propose a **Hybrid MALA + Global Jump** sampler, a mixture-of-kernels Markov chain Monte Carlo method that alternates between gradient-guided Metropolis-Adjusted Langevin Algorithm (MALA) local moves and large isotropic global jump proposals. At each step, with probability $1 - q$ the chain executes a MALA step with Metropolis correction; with probability $q$ it proposes a global jump from $\mathcal{N}(\theta, \sigma^2_{\text{global}} I)$. We evaluate the sampler on two benchmark distributions—the Rosenbrock (banana) distribution and Neal's Funnel—and compare it to Random Walk Metropolis–Hastings (RWMH) and Hamiltonian Monte Carlo (HMC) baselines. On the Rosenbrock distribution, the Hybrid achieves comparable coverage to HMC with a slightly higher effective sample size ($\text{ESS}_x = 55$ vs. 42). On Neal's Funnel, the Hybrid most closely recovers the true marginal variance ($\hat{\sigma}_v = 3.51$ vs. true $\sigma_v = 3$) and achieves more than twice the ESS of HMC for the hierarchy variable $v$ ($\text{ESS}_v = 68$ vs. 32), while RWMH severely undersamples the funnel ($\hat{\sigma}_v = 1.26$). An ablation study on the Rosenbrock confirms that a small global-jump fraction ($q \approx 0.10$) yields the best trade-off between local gradient efficiency and global exploration.

## 1 Introduction

In this assignment, we focus on a common problem in MCMC: it is hard for one sampler to be both **fast locally** and **good at exploring globally**. Random Walk Metropolis–Hastings (RWMH) is easy to code, but it depends a lot on step size. On curved targets like the Rosenbrock "banana," small steps give high acceptance but move very slowly, while large steps often get rejected. Gradient-based methods like MALA can move more efficiently by using $\nabla \log \pi(\theta)$, but with one fixed step size they can still struggle on targets with very different scales, such as Neal's Funnel.

To address this, we use a **mixture-kernel sampler** with two types of moves. At each step, with probability $1 -$ `jump_prob` we take a **MALA** local move (good for following narrow ridges). With probability `jump_prob` we take a **global Gaussian jump** with scale $\sigma_{\text{global}}$ (good for making bigger moves and escaping slow regions). Both moves are Metropolis-corrected, so the sampler still targets the correct distribution. The method is easy to tune with three parameters: the MALA step size $\epsilon$, the global jump scale $\sigma_{\text{global}}$, and the jump probability `jump_prob`.

## 2 Method

Let $\pi(\theta) \propto \exp(\ell(\theta))$ be the target distribution. We use a simple **mixture-kernel Metropolis–Hastings** sampler with two move types: a **MALA** local move and a **global Gaussian jump**. At each iteration we draw $b \sim \text{Bernoulli}(q)$ with $q =$ `jump_prob`. If $b = 0$ (probability $1 - q$), we take a MALA step: compute $g = \nabla \ell(\theta)$ and propose $\theta' = \theta + \frac{\varepsilon^2}{2} g + \varepsilon \eta$ with $\eta \sim \mathcal{N}(0, I)$, where $\varepsilon =$ `step_size`. Since the MALA proposal is asymmetric, we accept $\theta'$ using the standard MALA MH ratio $\alpha = \min\Big(1, \frac{\pi(\theta')\, q_{\text{MALA}}(\theta|\theta')}{\pi(\theta)\, q_{\text{MALA}}(\theta'|\theta)}\Big)$.

If $b = 1$ (probability $q$), we take a global jump by proposing $\theta' \sim \mathcal{N}(\theta, \sigma_{\text{global}}^2 I)$, where $\sigma_{\text{global}} =$ `sigma_global`. This proposal is symmetric, so the acceptance probability simplifies to $\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right)$. We implement the full sampler in JAX using `jax.lax.scan` for fast compiled loops and `jax.lax.cond` to switch between the two kernels each step, with gradients computed by `jax.value_and_grad`. The method has three easy-to-read hyperparameters: $\varepsilon$ (local step size), $\sigma_{\text{global}}$ (jump scale), and $q$ (jump frequency); in our experiments we set $q$ around 0.10–0.15 and tune $\varepsilon$ for a stable MALA acceptance rate.

## 3 Experiments

We draw $N = 50{,}000$ samples per method on each benchmark, starting from $\theta_0 = (0, 0)$. All diagnostics are computed with ArviZ. Since we run a single chain, $\hat{R}$ is not available; we report $\text{ESS}_{\text{bulk}}$ as the main efficiency metric.

### 3.1 Benchmark 1: Rosenbrock (Banana)

The Rosenbrock target has log density $\log p(x, y) = -0.05(1 - x)^2 - (y - x^2)^2$. It features a thin, curved ridge and thus tests whether a sampler can move efficiently along a strongly correlated geometry. Hyperparameters: RWMH uses $\sigma = 1.0$; HMC uses $\epsilon = 0.2$, $L = 10$; Hybrid uses $\epsilon = 0.3$, $\sigma_{\text{global}} = 3.0$, $q = 0.15$.

Table 1: Rosenbrock benchmark ($N = 50{,}000$).

| Method | Acc. rate | $\hat{\sigma}_x$ | $\hat{\sigma}_y$ | $\text{ESS}_x$ | $\text{ESS}_y$ |
|---|---|---|---|---|---|
| RWMH | 50.0% | 0.637 | 0.659 | 16,997 | 18,094 |
| HMC | 74.8% | 2.202 | 5.012 | 42 | 93 |
| Hybrid (ours) | 56.7% | 2.090 | 6.316 | 26 | 34 |

RWMH appears to have very high ESS, but it only explores a small neighborhood around the origin (e.g., $\hat{\sigma}_x = 0.64$) and does not truly traverse the banana-shaped ridge; hence this "high ESS" is more reflective of fast local mixing in a restricted region. In contrast, HMC and the Hybrid cover a much broader region of parameter space (e.g., $\hat{\sigma}_x \approx 2$), indicating better exploration along the curved valley.

As shown by the ablation study under the same $(\epsilon, \sigma_{\text{global}})$, the Hybrid can achieve substantially higher ESS near an optimal mixture rate (e.g., $q = 0.15$). This suggests that on Rosenbrock, carefully tuning $q$ is important to fully leverage the benefit of global jumps.

### 3.2 Benchmark 2: Neal's Funnel

Neal's Funnel is defined by $v \sim \mathcal{N}(0, 9)$ and $x \mid v \sim \mathcal{N}(0, e^v)$. It is a classic stress test for fixed-step-size samplers: because the marginal scale of $x$ changes exponentially with $v$, a single step size cannot work well everywhere. Hyperparameters: RWMH uses $\sigma = 1.5$; HMC uses $\epsilon = 0.3$, $L = 5$; Hybrid uses $\epsilon = 0.3$, $\sigma_{\text{global}} = 3.0$, $q = 0.15$.

Table 2: Neal's Funnel benchmark ($N = 50{,}000$). True $\sigma_v = 3$.

| Method | Acc. rate | $\hat{\sigma}_v$ | $\hat{\sigma}_x$ | $\text{ESS}_v$ | $\text{ESS}_x$ |
|---|---|---|---|---|---|
| RWMH | 49.1% | 0.907 | 0.675 | 18,384 | 31,415 |
| HMC | 78.5% | 3.263 | 11.137 | 32 | 223 |
| Hybrid (ours) | 79.0% | 3.283 | 10.323 | 58 | 57 |

Although RWMH reports extremely large ESS, it substantially underestimates the marginal scale of $v$ ($\hat{\sigma}_v \approx 0.91 \ll 3$), indicating that the chain is trapped near the funnel "neck" and thus produces inflated ESS due to fast local mixing. HMC explores the funnel more broadly ($\hat{\sigma}_v \approx 3.26$) but achieves only $\text{ESS}_v = 32$ in 50,000 steps. The Hybrid attains a similar marginal scale for $v$ ($\hat{\sigma}_v \approx 3.28$) while improving $\text{ESS}_v$ to 58 (about $1.8\times$ HMC).

### 3.3 Ablation: Effect of Jump Probability q

We ablate $q \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.50, 0.70, 1.0\}$ on the Rosenbrock distribution with $\varepsilon = 0.3$ and $\sigma_{\text{global}} = 3.0$, using $N = 50{,}000$ samples per chain and $4$ independent chains. As $q$ increases, the acceptance rate generally decreases (from $63.1\%$ at $q = 0$ to $8.9\%$ at $q = 1$), since large isotropic jumps rarely land on the narrow curved ridge. In contrast, sampling efficiency is highly non-monotonic: pure MALA ($q = 0$) achieves relatively low ESS (minESS $\approx 35$), while a small but nonzero jump rate yields a substantial improvement. The best performance occurs at $q = 0.15$, where ESS is high and balanced across coordinates (ESS$_x \approx 196$, ESS$_y \approx 197$; minESS $\approx 196$). According to the result, we set $q = 0.15$ in subsequent experiments.

## 4 Discussion

**Where the Hybrid works well.** On Rosenbrock, the Hybrid (MALA + global jumps) performs strongly: mixing local MALA updates with occasional global proposals substantially improves sampling efficiency. The ablation study shows that performance is highly sensitive to the jump probability $q$. With $\epsilon = 0.3$ and $\sigma_{\text{global}} = 3.0$ fixed, using a small but nonzero $q$ yields a large increase in ESS. In particular, $q = 0.15$ achieves high and well-balanced ESS across both dimensions (minESS $\approx 196$), while maintaining a reasonable acceptance rate ($57.1\%$). By contrast, pure MALA ($q = 0$) has a higher acceptance rate ($63.1\%$) but a much lower ESS (minESS $\approx 35$), indicating strong autocorrelation and slow exploration along the narrow curved ridge. This suggests that occasional global proposals help the chain traverse the banana-shaped valley more quickly than local moves alone. Compared with HMC, the Hybrid remains competitive and avoids a leapfrog loop, so each iteration is cheaper.

**Where the Hybrid struggles.** Performance is clearly non-monotonic in $q$, and the method can degrade sharply when the mixture is poorly balanced. For example, increasing $q$ too much does not necessarily help: at $q = 1$ (pure global jumps), the acceptance rate drops to $8.9\%$ and mixing becomes uneven (e.g., ESS$_y \approx 48$), because large isotropic proposals rarely land on the thin Rosenbrock ridge. Even within the intermediate range, we observe failures such as $q = 0.20$, where ESS becomes very low (minESS $\approx 22$) despite a moderate acceptance rate, suggesting an interaction between the fixed global scale $\sigma_{\text{global}}$ and the mixture rate $q$. More broadly, the Hybrid still requires manual tuning of $\epsilon$, $\sigma_{\text{global}}$, and $q$ across targets, which limits robustness.

**Future directions.** A natural next step is to incorporate adaptive tuning so the sampler is less sensitive to target geometry. For instance, dual averaging could adjust $\epsilon$ to target a desired MALA acceptance rate, and a simple adaptation scheme could tune $\sigma_{\text{global}}$ (or $q$) to avoid regimes where global proposals are almost always rejected.

for example using dual averaging as in NUTS.

## 5 AI Collaboration

- **Tools and interfaces.** I mainly used **ChatGPT (chat)** for brainstorming, concept explanations, and writing help, and **Claude (code agent)** for code-level debugging and implementation support.
- **What I used AI for.** AI helped me propose sampler ideas (e.g., adding a **MALA** component and using a mixture with global jumps), interpret diagnostics (especially what **ESS** means and why it can be misleading), and improve the clarity of my method/experiment text.
- **Prompting strategies that worked.** The best approach was: first ask AI to **generate multiple options and improvement suggestions**, then provide concrete context (target distribution, current acceptance rates/ESS tables, and small code snippets) so the advice became specific and actionable.
- **Where AI helped vs. where I corrected it.** AI was strongest at method improvement ideas, tuning directions, and catching small issues (missing terms, inconsistent notation, minor logic bugs). I still needed to guide it on details that must match my code and results, and to verify that final claims were supported by the actual experiments. I still use chatGPT to help me correct the format of latex, it helps a lot in formating.

- **What I learned.** Treating AI as a **tool and a suggestion-oriented teammate** works best: let it propose plans and spot problems, then I verify, implement, and decide what to keep.