Assignment 1: Tempered MALA with Global Jumps

Langbiao Mou — CDS DS 595
Boston University
bowenmou@bu.edu

Abstract

I modify Metropolis-adjusted Langevin dynamics with two simple tricks: brief high-temperature bursts and rare global Gaussian jumps. The sampler is easy to code in JAX yet handles Neal's Funnel far better than a plain random walk. Four tempered chains reach ESS ≈ 640 at $\sim 80\%$ acceptance. The same design still lags behind HMC on Rosenbrock (ESS ≈ 50 per chain), so I treat it as a case study in when lightweight tempering helps.

1 Introduction

RWMH struggles on anisotropic targets; HMC is powerful but harder to tune when geometry changes rapidly. I therefore keep the lightweight MALA kernel but add two ingredients—short high-temperature phases and occasional global jumps—to see whether modest tempering can fix Neal's Funnel without the full complexity of HMC. The focus is understanding trade-offs rather than beating every baseline. The goal was to keep the code compact (<100 lines of JAX), expose interpretable knobs, and learn how diagnostics such as ESS and $\hat{R}$ react when we gradually add parallel chains.

The two benchmark targets stress any sampler that relies on a single global step size: the Rosenbrock "banana" is narrow and curved, so isotropic proposals either fall off the ridge or move too timidly; Neal's Funnel ties the latent scale v to x, so step sizes that work in the wide mouth instantly fail in the neck. Even crude temperature cycling might help by temporarily flattening narrow regions, enabling proposals that traverse otherwise forbidden valleys.

2 Method

Tempered Langevin step. Each cycle of length L contains a hot segment of length $\lfloor \rho L \rfloor$ using temperature $\tau > 1$ and a cold segment with $\tau = 1$. During hot steps the Langevin drift and noise are scaled by τ, producing proposals $y = x_t + \tau\eta\nabla \log p(x_t) + \sqrt{2\tau\eta}\,\epsilon$ with the standard MALA accept ratio. Cold steps revert to $\tau = 1$.

Global jump. With probability p_g the algorithm bypasses the Langevin move and proposes $y = x_t + \sigma_g \epsilon_g$ using a symmetric random-walk test. This move occasionally drags the chain through the Funnel neck.

Implementation. Everything is written in JAX. A helper `run_tempered_mala_chains` launches C chains ($C = 4$ in my runs) and reshapes results for ArviZ. Both targets reuse one hand-tuned schedule $(\eta, \tau, \rho, L, p_g, \sigma_g) = (0.02, 3.5, 0.4, 80, 0.18, 1.8)$ to keep the comparison simple; tuning separate schedules is left for future work. Each cycle checks whether it sits in the hot window ($\tau_t = \tau$) or cold window ($\tau_t = 1$), flips a Bernoulli(p_g) to decide between a global random walk jump and a Langevin step, and then applies the standard MALA acceptance rule. Optional warm-up chains can estimate a diagonal mass matrix, but I keep $M = I$ to isolate temperature effects.

Sampler	Acceptance (%)	ESS_{bulk}	$\hat{R}_{\max}$
RWMH — Rosenbrock	50.3	17 924	n/a
HMC — Rosenbrock	81.3	455	n/a
Tempered MALA — Rosenbrock	60.7	51.5	1.07
RWMH — Funnel	33.9	11 049	n/a
HMC — Funnel	97.5	1 370	n/a
Tempered MALA — Funnel	79.9	638	1.01

Table 1: Diagnostics aggregated over 50 000 draws per chain. Tempered MALA uses four independent chains; the RWMH and HMC baselines are still single-chain, so $\hat{R}$ is undefined for them and their ESS numbers likely overstate true efficiency.

3 Experiments

3.1 Setup

I compare against the provided RWMH and HMC baselines on the Rosenbrock and Funnel targets. Each sampler draws 50 000 samples (per chain) after warm-up; Tempered MALA runs four chains while the baselines remain single-chain. All runs share identical random seeds to avoid cherry-picking trajectories, and diagnostics are computed with ArviZ. Following the assignment instructions, I keep the baseline implementations untouched to highlight how coding time constraints can leave experiments partially mismatched (a recurring theme in Section 4).

3.2 Baselines and tuning effort

RWMH uses isotropic proposals with $\sigma = 1.0$ on Rosenbrock and $\sigma = (1.2, 3.0)$ on Funnel, values copied from the starter notebook. HMC uses an identity mass matrix, step size 0.2 with 10 leapfrog steps on Rosenbrock, and 0.08 with 25 steps on Funnel. I spent roughly 45 minutes trying to tune Tempered MALA schedules; the best shared schedule came from a coarse grid-search over $\tau \in \{2.0, 3.5, 5.0\}$, $\rho \in \{0.2, 0.4, 0.7\}$, $p_g \in \{0.0, 0.1, 0.2\}$ combined with quick visual inspections. This limited search emphasizes reproducibility over optimality and explains why the Rosenbrock results still lag.

3.3 Computational cost

Running four tempered chains for 50 000 iterations takes ≈ 40 seconds on my M2 MacBook Air (pure JAX CPU backend), similar to HMC. The global jump branch introduces negligible overhead because it reuses the same gradient evaluation that MALA would have computed. Memory footprint stays under 1 GB even when storing full traces for plotting.

3.4 Diagnostics

Table 1 reports acceptance, ESS, and $\hat{R}$. Tempered MALA dramatically lifts Funnel ESS relative to my earlier single-chain attempt (now ≈ 640 with $\hat{R} = 1.01$). Rosenbrock remains challenging: ESS hovers near 50 per chain despite a reasonable 61% acceptance rate, and the baselines still lack reliable $\hat{R}$ estimates because they use single chains.

3.5 Qualitative Behavior

Scatter plots (see notebook) show full Funnel coverage, whereas single-chain RWMH hugs the neck. Trace plots also confirm that the high-temperature bursts let the chains visit both wide and narrow regions. On Rosenbrock the chain still oscillates slowly along the banana arms, so the qualitative behavior matches the modest (≈ 50) ESS. Autocorrelation plots reveal that tempering dramatically shortens the lag-1 correlation on Funnel but barely touches Rosenbrock, reinforcing the need for geometry-aware proposals there.

Setting	η	τ	ρ	L	p_g	ESS_{bulk}
Base (short hot bursts)	0.02	3.5	0.4	80	0.18	3.5
No temperature cycling	0.02	1.0	0.0	80	0.18	11.0
Long hot phase (70%)	0.02	3.5	0.7	80	0.18	10.5
No global jumps	0.02	3.5	0.4	80	0.00	15.0
Short cycle (40)	0.02	3.5	0.4	40	0.18	14.5

Table 2: Rosenbrock ablation (two chains, $15\,000$ draws each). All variants still mix poorly, but the "base" schedule under-performs simpler alternatives, emphasizing the need for further tuning.

3.6 Multi-chain vs. single-chain diagnostics

Even though the baselines are still single-chain, I experimented with running Tempered MALA using $C \in \{1, 2, 4\}$ chains. Increasing C primarily shrinks $\hat{R}$ towards 1.0 and stabilizes ESS estimates; the per-chain ESS is almost unchanged, which is encouraging because it means parallel chains can be combined without loss of efficiency. The remaining warning emitted by ArviZ for RWMH/HMC ("minimum shape chains=2") is a reminder that my comparison is incomplete until I rerun the baselines with the new multi-chain helper.

3.7 Ablation Study

Table 2 varies a few knobs on Rosenbrock (two chains, $15\,000$ draws). The current tuning is clearly suboptimal: the "base" schedule yields the lowest ESS, and shortening the cycle or disabling global jumps actually helps. This highlights how sensitive the method is on Rosenbrock and why matching baselines remains future work.

4 Discussion

Strengths. Simple tweaks make MALA competitive on the Funnel: high acceptance, ESS ≈ 640, and no elaborate integrators.

Limitations. Rosenbrock still defeats this sampler; ESS stays near 50 because the chain crawls along the curved ridge. Baseline chains are also single-chain, so comparisons remain imperfect and likely overstate RWMH efficiency.

Failure analysis. Examining rejected proposals shows that hot steps often leave the Rosenbrock ridge altogether, causing long sequences of rejections that reset the global-jump momentum. Simply raising τ or σ_g makes this worse, so the next lever is a position-dependent mass matrix or local quadratic fit that bends proposals along the ridge. Another avenue is to let the global jumps target previously visited states (tempered transitions style) instead of drawing isotropic noise.

Next steps. Try per-coordinate preconditioning or short HMC substeps during the cold phase, and re-run all methods with equal numbers of chains/burn-in to clean up diagnostics once the updated code paths are fully working. I also plan to profile the sampler under JAX's `pmap` to see whether we can pay for additional chains without extending wall-clock time, which would make the comparison to multi-chain baselines much tighter.

5 AI Collaboration

- Refactoring `run_rwmh` and `run_hmc` to support multiple chains. The assistant highlighted how to broadcast initial states and vmap BlackJAX kernels.
- Investigating ArviZ warnings about single-chain ESS. I asked why the warnings appeared and received the reminder that ArviZ expects at least two chains; I then adjusted my data reshaping logic accordingly.