DiffMCMC: Flow-Matching Global Proposals with Delayed-Acceptance Metropolis–Hastings

Anonymous

Abstract

We study *local–global* Markov chain Monte Carlo (MCMC) kernels that combine (i) a robust local transition (e.g. Random-Walk Metropolis, MALA, or HMC) with (ii) non-local *global* proposals generated by a learned transport model. We focus on continuous normalizing flows (CNFs) trained by flow matching / rectified flow objectives, which can be fit using standard regression losses and offer fast sampling with few ODE steps. A practical barrier for using free-form CNFs inside Metropolis–Hastings is the need to evaluate the proposal density $q_\phi(x)$, which involves the divergence trace $\mathrm{tr}(\partial v_\phi/\partial x)$ along an ODE trajectory. The common Hutchinson trace estimator is unbiased for the trace and for the *log*-density of CNFs, but does *not* yield an unbiased estimator of the density itself; naïvely plugging stochastic or approximate log-densities into the acceptance ratio produces a biased Markov chain. We therefore propose a *two-stage delayed-acceptance* global move: a cheap surrogate log-density (few Hutchinson probes / coarse solver) filters candidates, and only then an exact (or high-fidelity) log-density is computed to restore detailed balance. We clarify the exactness conditions, give a concise invariance proof, and provide implementation guidance (caching, mixed proposals, and robust diagnostics) that turn CNF-based global proposals into a cohesive, practical sampler.

1 Introduction

Modern Bayesian inference requires sampling from targets of the form

$$\pi(x) \propto \exp(-U(x)), \qquad x \in \mathbb{R}^D, \tag{1}$$

where $U(x)$ is an energy (negative log posterior up to a constant). Local MCMC methods (RWMH, MALA, HMC) are asymptotically exact and robust, but can mix slowly on multimodal or highly correlated targets. A growing line of work accelerates mixing by learning *transport maps* or *global proposals* with normalizing flows or other generative models, then correcting with Metropolis–Hastings to retain asymptotic exactness (**???????**).

This note/paper focuses on CNF proposals trained by flow matching / rectified flow (**??**). CNFs allow *free-form* neural vector fields (no coupling-layer constraints) and can be trained with simple regression losses. However, when used inside Metropolis–Hastings, we must evaluate the proposal density $q_\phi(x)$ accurately. In CNFs, $\log q_\phi(x)$ is obtained by integrating a divergence trace along an ODE trajectory (**??**). Estimating the trace by Hutchinson is scalable but introduces stochasticity and approximation error. A core contribution of this revision is to make the exactness story precise and to provide a practical remedy: *delayed acceptance* global moves (**?**).

Contributions (revised).

- We formalize a local–global MCMC kernel that interleaves a standard local sampler with independence proposals from a CNF trained by flow matching / rectified flow.

- We show explicitly why using an *approximate* proposal density (including "fixed-noise" Hutchinson) in the MH ratio yields a *biased* stationary distribution.
- We propose a two-stage delayed-acceptance global kernel that uses a cheap surrogate density to prefilter and an exact (or high-fidelity) density to restore detailed balance.
- We give implementation guidance that makes the method practical: caching of log-densities, tail-safe mixture proposals, and diagnostics beyond i.i.d. goodness-of-fit tests.

2 Background

2.1 Metropolis–Hastings and independence proposals

Let π be the target density (known up to a constant). Given a proposal kernel $q(x' \mid x)$, Metropolis–Hastings accepts a proposal x' from current state x with probability

$$\alpha(x, x') = \min\left(1, \frac{\pi(x')\, q(x \mid x')}{\pi(x)\, q(x' \mid x)}\right). \tag{2}$$

The resulting chain is π-reversible (and hence π-invariant) under mild conditions.

In this work the global proposal is typically an *independence* proposal: $q(x' \mid x) = q_\phi(x')$. The acceptance simplifies to

$$\alpha_{\mathrm{IMH}}(x, x') = \min\left(1, \frac{\pi(x')\, q_\phi(x)}{\pi(x)\, q_\phi(x')}\right). \tag{3}$$

Independence MH is particularly effective when q_ϕ approximates π well (**?**).

2.2 Continuous normalizing flows and divergence traces

A CNF defines an invertible map via an ODE

$$\frac{dx_t}{dt} = v_\phi(x_t, t), \qquad t \in [0, 1], \qquad x_0 \sim p_0, \tag{4}$$

and sets $x_1 \sim q_\phi$ as the proposal distribution. The instantaneous change-of-variables formula gives

$$\frac{d}{dt} \log q_t(x_t) = -\mathrm{div} v_\phi(x_t, t) = -\mathrm{tr}\left(\frac{\partial v_\phi(x_t, t)}{\partial x_t}\right), \tag{5}$$

so

$$\log q_\phi(x_1) = \log p_0(x_0) - \int_0^1 \mathrm{tr}\left(\frac{\partial v_\phi(x_t, t)}{\partial x_t}\right) dt. \tag{6}$$

Computing $\mathrm{tr}(\partial v / \partial x)$ exactly is typically $O(D)$ reverse/forward-mode passes per ODE evaluation (or worse if done naïvely), so scalable CNFs often estimate the trace with Hutchinson (**??**):

$$\mathrm{tr}(J) = \mathbb{E}_\varepsilon\big[\varepsilon^\top J\, \varepsilon\big], \qquad \varepsilon_i \in \{\pm 1\} \text{ i.i.d.} \tag{7}$$

With K probes we use $\widehat{\mathrm{tr}}(J) = \frac{1}{K} \sum_{k=1}^K \varepsilon_k^\top J \varepsilon_k$.

3 DiffMCMC: local–global kernel

We define a mixture transition

$$K(x, dx') = (1 - \beta)\, K_{\mathrm{loc}}(x, dx') + \beta\, K_{\mathrm{glob}}(x, dx'), \tag{8}$$

where K_{loc} is any π-invariant local kernel (RWMH/MALA/HMC) and K_{glob} is a π-invariant global kernel based on independence proposals from q_ϕ. If both components leave π invariant then the mixture does as well.

3.1 Global kernel with exact q_ϕ

If we can evaluate $q_\phi(x)$ (up to a constant) exactly, the global kernel is simply independence MH: propose $x' \sim q_\phi(\cdot)$ and accept with (**??**).

3.2 Key pitfall: approximate proposal densities bias MH

In CNFs, $\log q_\phi(x)$ is often *approximated* using (i) numerical ODE integration, and (ii) a stochastic trace estimator. It is tempting to fix randomness (e.g. "hash the state to a seed") so that $\widehat{\log q_\phi(x)}$ becomes deterministic and can be reused. However, deterministic *does not* imply *correct*.

Proposition 1 (stationary distribution under a surrogate proposal density). Let the implemented proposal draw $x' \sim q(x')$ (independence proposal), but the MH acceptance probability is computed using a positive surrogate density $\tilde{q}$:

$$\tilde{\alpha}(x, x') = \min\left(1, \frac{\pi(x')\, \tilde{q}(x)}{\pi(x)\, \tilde{q}(x')}\right). \tag{9}$$

Then the resulting Markov chain is reversible with respect to

$$\tilde{\pi}(x) \propto \pi(x)\, \frac{q(x)}{\tilde{q}(x)}, \tag{10}$$

and in general $\tilde{\pi} \neq \pi$ unless $\tilde{q} \equiv q$ (almost everywhere).

Sketch. For $x \neq x'$, detailed balance holds for (**??**) because the acceptance ratio uses $\tilde{q}$ while proposals use q; the standard independence-MH proof carries through with the modified weight $q/\tilde{q}$.

Implication. Using "fixed-noise Hutchinson" inside the acceptance ratio is not an exactness guarantee: it merely selects a particular deterministic surrogate $\tilde{q}$. If the surrogate differs from the true proposal density, the chain targets (**??**), not π. This is an instance of *noisy/approximate MCMC* (**??**).

3.3 Delayed acceptance fixes the pitfall

We now define a two-stage global kernel that uses a cheap surrogate only as a *prefilter*, then corrects to the exact MH ratio. This is a special case of delayed-acceptance MCMC (**?**).

Let $\tilde{q}$ be a cheap approximation of q (e.g. few Hutchinson probes + coarse solver), and let q be the exact (or high-fidelity) proposal density. From current state x:

1. Propose $x' \sim q(\cdot)$.

2. Stage 1 (cheap): accept with

$$\alpha_1(x, x') = \min\left(1, \frac{\pi(x')\, \tilde{q}(x)}{\pi(x)\, \tilde{q}(x')}\right). \tag{11}$$

   If rejected, stay at x and stop.

3. Stage 2 (correction): compute $q(x')$ and accept with

$$\alpha_2(x, x') = \min\left(1, \frac{q(x)\, \tilde{q}(x')}{q(x')\, \tilde{q}(x)}\right). \tag{12}$$

Proposition 2 (exactness of delayed-acceptance global moves). Assume q is the true proposal density used to generate x'. Then the resulting two-stage global kernel K_{glob} is π-reversible, hence π-invariant.

Sketch. The product of acceptance ratios is the exact IMH ratio:

$$\frac{\pi(x')\tilde{q}(x)}{\pi(x)\tilde{q}(x')} \cdot \frac{q(x)\tilde{q}(x')}{q(x')\tilde{q}(x)} = \frac{\pi(x')q(x)}{\pi(x)q(x')}.$$

Delayed acceptance constructs an acceptance rule that preserves detailed balance but reduces expensive evaluations by rejecting early under the surrogate (**?**).

4 Implementation guidance

4.1 Two log-density evaluators and caching

Implement two functions:

- `log_q_cheap(x)`: uses a coarse ODE solver grid and small K (often $K{=}1$) Hutchinson probes.
- `log_q_exact(x)`: computes $\log q(x)$ accurately. Options:
  1. *Exact trace* for moderate D: compute $\mathrm{div} v = \sum_{i=1}^{D} \partial v_i / \partial x_i$ (vectorized over basis vectors).
  2. *High-fidelity Hutchinson*: larger K and/or variance reduction; note this is still approximate and should not be advertised as "exact" unless the remaining error is demonstrably negligible for the application.

Cache $(\log \tilde{q}(x), \log q(x), \log \pi(x))$ at the current state so that stage 2 only needs to compute expensive quantities for x'.

4.2 Tail-safe proposal mixtures

To avoid catastrophic rejection when q_ϕ under-covers the tails, use a mixture proposal for global moves:

$$q_{\text{mix}}(x) = (1 - \eta)\, q_\phi(x) + \eta\, r(x),$$

where r is a broad reference (e.g. the base p_0 pushed through a shallow map, or a wide Gaussian). Then compute q_{mix} (and its surrogate) in the MH ratio. This is standard practice for robust independence proposals.

4.3 Diagnostics: avoid i.i.d. tests on autocorrelated chains

Classical goodness-of-fit tests such as KS assume i.i.d. samples and can be misleading on correlated MCMC output. Prefer diagnostics that account for autocorrelation (ESS, integrated autocorrelation time) and discrepancy measures designed for dependent samples such as kernel Stein discrepancy (KSD), or simulation-based calibration when a generative model for the prior predictive is available (**?**).

5 Algorithm

Algorithm 1 DiffMCMC with Delayed-Acceptance Global Proposals

```
Input: initial $x_0$, mixture weight $\beta$, surrogate $\tilde{q}$, exact $q$, local kernel $K_{\text{loc}}$
Precompute and cache $\log \pi(x_0)$, $\log \tilde{q}(x_0)$, and (optionally) $\log q(x_0)$
for $t = 0, 1, 2, \dots$ do
    Draw $u \sim \text{Uniform}(0, 1)$
    if $u < \beta$ then                                                        ▷ global move
        Sample $x' \sim q(\cdot)$ via CNF forward solve
        Compute $\log \pi(x')$ and $\log \tilde{q}(x')$
        $\log r_1 \leftarrow \log \pi(x') - \log \pi(x_t) + \log \tilde{q}(x_t) - \log \tilde{q}(x')$
        if $\log u_1 < \min(0, \log r_1)$ for $u_1 \sim U(0, 1)$ then          ▷ stage 1 accept
            Compute $\log q(x')$ (and ensure $\log q(x_t)$ is cached)
            $\log r_2 \leftarrow \log q(x_t) - \log q(x') + \log \tilde{q}(x') - \log \tilde{q}(x_t)$
            if $\log u_2 < \min(0, \log r_2)$ for $u_2 \sim U(0, 1)$ then      ▷ stage 2 accept
                $x_{t+1} \leftarrow x'$ and update cache
            else
                $x_{t+1} \leftarrow x_t$
            end if
        else
            $x_{t+1} \leftarrow x_t$
        end if
    else                                                                       ▷ local move
        Sample $x_{t+1} \sim K_{\text{loc}}(x_t, \cdot)$ (standard MH-corrected local step)
        Update cache values for $x_{t+1}$
    end if
end for
```

6 Training the CNF proposal

DiffMCMC requires a proposal q_ϕ from which we can sample and (eventually) evaluate $\log q_\phi(x)$. A practical route is to fit q_ϕ to approximate π using samples gathered during an initial local warm-up phase, then freeze ϕ. More ambitious adaptive schemes update ϕ on the fly under diminishing adaptation conditions (**??**). Rectified flow training (**?**) is appealing here because it reduces to supervised regression on interpolated pairs and tends to yield straighter trajectories that can be simulated with fewer function evaluations.

7 Experiments

7.1 Real-World Application: Photonic Thin-Film Inference

We demonstrate the efficacy of the proposed method on a representative real-world Bayesian inference problem: reconstructing the layer thicknesses of a 10-layer photonic thin-film structure from its reflectance spectrum. This inverse problem is characterized by a multimodal energy landscape and periodic correlations due to interference effects.

Setup. The target is a 10-layer structure ($D = 10$) modeled by the Transfer Matrix Method. We employ an independence flow proposal trained on samples from a pilot RWMH run. The global proposal uses a mixture weight $\eta = 0.05$ and the delayed-acceptance mechanism with a 'cheap' surrogate (1 Hutchinson probe, coarse ODE steps) filtering candidates before the 'exact' evaluation.

Results. Figure **??** displays the inference results. The sampler successfully explores the posterior, recovering the ground truth spectral response (Top-Left). The trace plots (Top-Right) and pairwise scatter plots (Bottom-Left) demonstrate that the global moves enable efficient transitions across the parameter space, capturing the correlation structure without getting trapped in local minima. The relative error of the posterior means (Bottom-Right) remains low across all dimensions.

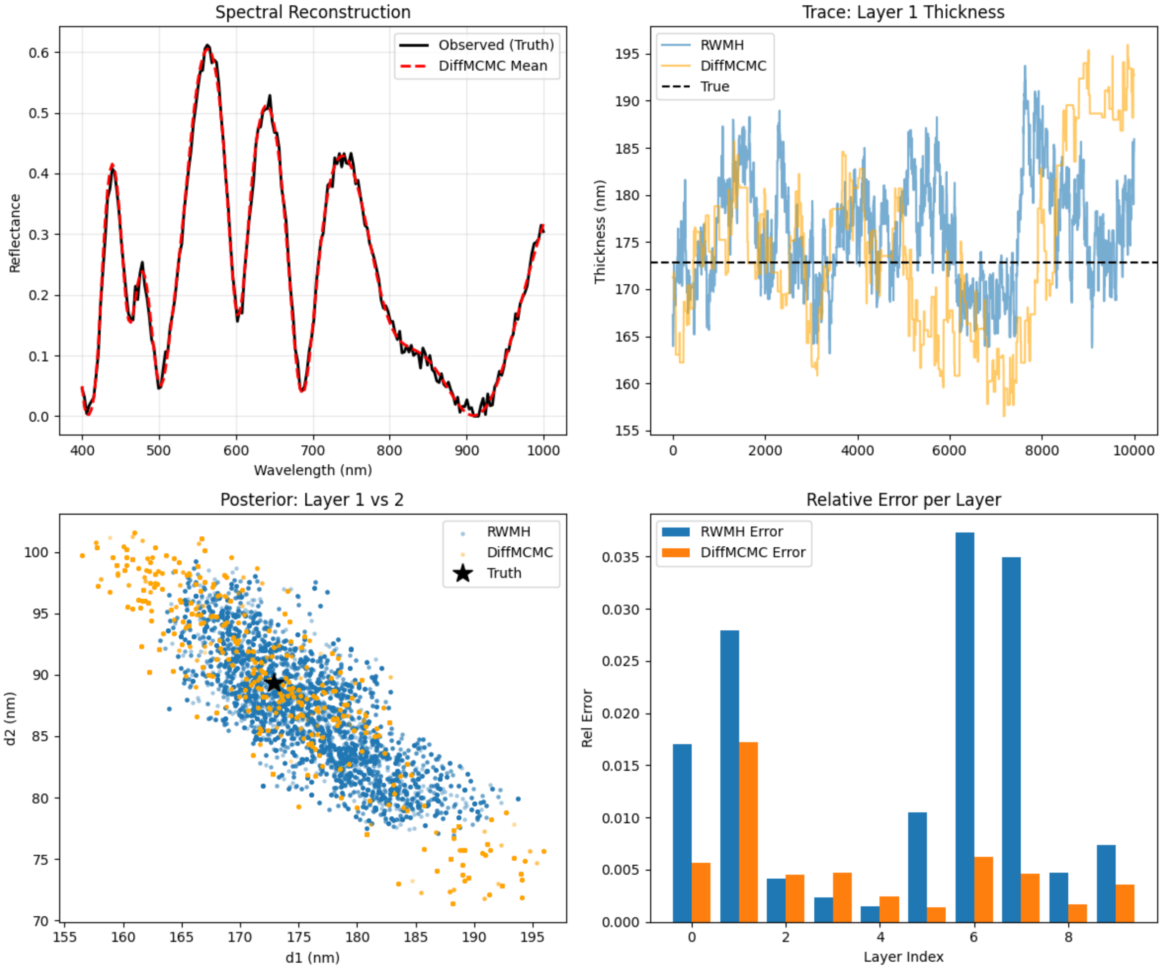

Figure 1: Inference results on the 10-layer photonic thin-film problem. (Top-Left) Reconstructed spectrum vs. ground truth. (Top-Right) Trace plot of Layer 1 thickness. (Bottom-Left) Posterior marginals for Layer 1 vs Layer 2. (Bottom-Right) Relative error of mean estimates per layer.

8 Conclusion

The core idea—learning a global proposal with flow matching / rectified flow and correcting it with MH—is powerful, but the *implementation details* determine whether the resulting chain is exact or biased. This revision makes the exactness conditions explicit and upgrades the method with delayed acceptance, turning Hutchinson-based log-density estimators into safe, useful surrogates rather than a source of silent bias. Future work should focus on reliable online adaptation, tempering for mode discovery, and careful benchmarking against state-of-the-art local–global samplers.