Diffusion Generative Flow Samplers: Improving Learning Signals Through Partial Trajectory Optimization

Dinghuai Zhang*, Ricky T. Q. Chen, Cheng-Hao Liu, Aaron Courville & Yoshua Bengio

Thomas Mousseau

October 20, 2025

Overview

1. Introduction
1.1 Problem Statement
1.2 Limitations of Prior Work

2. Methodology
2.1 DGFS framework

3. Results and Limitations
3.1 Results

4. Conclusion
4.1 Key Insights
4.2 Future Directions and Usage since its release

Sampling from Unnormalized Densities

Goal. Sample from a D-dimensional target with unnormalized density $\mu(x)$:

$$\pi(x) = \frac{\mu(x)}{Z}, \qquad Z = \int_{\mathbb{R}^D} \mu(x)\, dx \text{ (unknown)}.$$

We assume we can evaluate $\mu(x)$, but we have no samples from π and do not know Z.

Context. We seek a *sampler* (similar to MCMC/VI) that produces calibrated samples and, ideally, estimates of $\log Z$, *without* any dataset from π.

Chemistry (small-molecule conformers). Different 3D conformations have a formation energy from force-field terms (bonds, angles, dihedrals, nonbonded); lower energy $\Rightarrow$ higher Boltzmann probability. A well-calibrated sampler is needed to draw conformers in proportion to these probabilities, which is important in binding-pose ranking/free-energy estimation, Boltzmann-weighted property prediction (e.g., NMR shifts), and generating diverse realistic 3D conformers for screening.

From sampling to control: reference dynamics

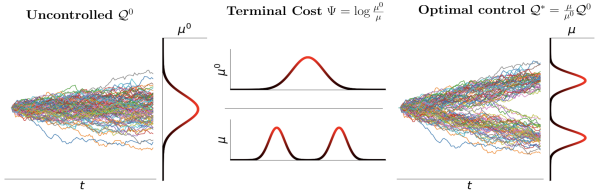
Reference (uncontrolled) diffusion.

$$dx_t = \sigma\, dW_t, \qquad x_0 \sim \nu \text{ (simple, e.g. Gaussian).}$$

More generally, introduce a *feedback control* $u_t = u(x_t, t)$:

$$dx_t = u(x_t, t)\, dt + \sigma\, dW_t, \qquad x_0 \sim \nu.$$

Let Q^{ref} be the path law of the uncontrolled process ($u \equiv 0$), with known terminal marginal p_T^{ref}. Our goal is to *choose a control u so that the terminal law of the controlled process matches the target* $\pi(x) \propto \mu(x)$.



SOC formulation

SOC objective. Choose a control u to minimize

$$\mathbb{E}_{Q^u}\left[\int_0^T \frac{1}{2\sigma^2}\|u(x_t,t)\|^2\,dt \;+\; \Psi(x_T)\right].$$

Pick the terminal cost to encode the target.

$$\Psi(x_T) \;=\; \log\frac{p_T^{\mathrm{ref}}(x_T)}{\mu(x_T)} \quad \big(\text{equivalently } -\log\mu(x_T) \text{ up to } \log p_T^{\mathrm{ref}}\big).$$

Key theorem (path–KL $\Leftrightarrow$ SOC). Define the *target path measure* by terminal reweighting:

$$P(x_{0:T}) \;\propto\; Q^{\mathrm{ref}}(x_{0:T})\;\frac{\mu(x_T)}{p_T^{\mathrm{ref}}(x_T)}.$$

Then, under standard regularity,

$$\boxed{D_{\mathrm{KL}}(Q^u\|P) = \mathbb{E}_{Q^u}\Big[\int_0^T \tfrac{1}{2\sigma^2}\|u\|^2\,dt + \log\tfrac{p_T^{\mathrm{ref}}(x_T)}{\mu(x_T)}\Big] \;+\; \text{const}}$$

Steps 1–2: Build forward & reference diffusions

Step 1 (controlled forward process). Parameterize a sampler as a controlled diffusion:

$$dx_t = f(x_t, t)\,dt + \sigma\,dW_t, \qquad x_0 \sim p_0^{\text{ref}},$$

with drift $f(\cdot,\cdot)$ (the control) to be learned. Let Q denote the path law of this process.

Step 2 (uncontrolled reference). Use the same noise but zero drift:

$$dx_t = \sigma\,dW_t, \qquad x_0 \sim p_0^{\text{ref}},$$

whose marginals p_t^{ref} are known in closed form. Let Q^{ref} be its path law.

Goal. Choose f so that the terminal marginal of Q matches the target $\pi(x) = \mu(x)/Z$ (no data, Z unknown).

Step 3: Target path measure via terminal reweighting

Construct a target measure on paths by reweighting only the terminal state:

$$P(x_{0:T}) \;\propto\; Q^{\text{ref}}(x_{0:T})\;\frac{\mu(x_T)}{p_T^{\text{ref}}(x_T)}.$$

Key property (endpoint correctness). By construction, the terminal marginal of P satisfies

$$P(x_T) \;\propto\; \mu(x_T) \quad\Longrightarrow\quad P(x_T) = \pi(x_T) \text{ up to the (unknown) normalizer } Z.$$

Takeaway. We have a principled path-space target P whose endpoint is exactly the desired distribution, without ever using Z along the way.

Step 4: Path–KL equals an SOC objective

Match the target on path space by minimizing a KL divergence:

$$\min_f \; D_{\mathrm{KL}}\big(Q \parallel P\big).$$

KL–control identity (Girsanov + chain rule). Under standard regularity,

$$\boxed{D_{\mathrm{KL}}(Q\|P) = \mathbb{E}_Q\Big[\int_0^T \tfrac{1}{2\sigma^2}\|f(x_t,t)\|^2\,dt \;+\; \log \tfrac{p_T^{\mathrm{ref}}(x_T)}{\mu(x_T)}\Big] \;+\; \text{const}}$$

Interpretation (SOC).

- *Running cost:* $\frac{1}{2\sigma^2}\|f\|^2$ (control effort).
- *Terminal cost:* $\log p_T^{\mathrm{ref}}(x_T) - \log \mu(x_T)$ (rewards high-μ endpoints).
- The additive constant absorbs $+\log Z \Rightarrow Z$ **cancels**.

Therefore: Minimizing $D_{\mathrm{KL}}(Q\|P)$ is *equivalent* to solving a stochastic optimal control problem.

Step 5 & intuition: why this works, what remains hard

Continuous-time view (optional). With T fixed and finer discretization, the discrete sum becomes

$$\min_f \ \mathbb{E}_Q\Big[\int_0^T \tfrac{1}{2\sigma^2}\|f(x_t,t)\|^2\,dt \ + \ \log \tfrac{p_T^{\text{ref}}(x_T)}{\mu(x_T)}\Big],$$

the classic entropy-regularized SOC form.

Intuition (why it works).

- We *steer* a simple diffusion toward regions where μ is large, trading control effort vs. terminal reward.
- The endpoint of the ideal optimizer matches $\pi = \mu/Z$ (calibrated sampling), while never needing Z explicitly.
- Training only requires evaluating $\mu(x_T)$ (and optionally $\nabla \log \mu$), not intermediate-time scores.

What remains hard (motivation for DGFS).

- The loss provides signal *only at terminal time* $\Rightarrow$ poor credit assignment, high-variance gradients.

Diffusion Process

Idea. Let the target "diffuse to Gaussian" via a reference process (VP/VE SDE). The *reverse-time* dynamics can, in principle, generate target samples if we know the *score* $\nabla_x \log p_t(x)$:

$$\underbrace{dx_t = \sigma\, dW_t}_{\text{forward/noising}} \quad \Longleftrightarrow \quad \underbrace{dx_t = \big[f_{\text{ref}}(x,t) - \sigma^2 \nabla_x \log p_t(x)\big]\, dt + \sigma\, d\bar{W}_t}_{\text{reverse/generative}}.$$

What is score matching? Learn a network $s_\theta(x,t) \approx \nabla_x \log p_t(x)$ by regressing on *noised data*:

$$\min_\theta \; \mathbb{E}_t\, \mathbb{E}_{x_0 \sim p_{\text{data}}}\, \mathbb{E}_\varepsilon \big\| s_\theta(x_t,t) - \nabla_x \log p_t(x_t) \big\|^2,$$

which is equivalent to denoising a corrupted sample x_t back toward x_0.

Why Denoising Score Matching is *not* applicable here

- We have *no dataset* from π, only the unnormalized $\mu(x)$ (and maybe $\nabla \log \mu$).

Alternative to DSM: learn the vector field (control)

Reverse SDE drift (generative side).

$$dx_t = \left[f_{\text{ref}}(x,t) - \sigma^2 \nabla_x \log p_t(x)\right] dt + \sigma \, d\bar{W}_t.$$

DSM route (data world). Learn the *score* $s_\theta(x,t) \approx \nabla_x \log p_t(x)$ from noised *data*, then plug it into the reverse drift.
Vector-field route (our setting). Directly learn the *control/drift* $u_\theta(x,t)$ instead of the score. The two are *equivalent* via:

$$u_\theta(x,t) \;=\; f_{\text{ref}}(x,t) \;-\; \sigma^2 \, s_\theta(x,t) \quad\Longleftrightarrow\quad s_\theta(x,t) \;=\; \tfrac{f_{\text{ref}}(x,t) - u_\theta(x,t)}{\sigma^2}.$$

Why do this here?

- We have no dataset from π, so DSM can't form expectations over p_t; scores $\nabla \log p_t$ are unavailable.
- Instead, treat u_θ as a *control* and *learn it* by minimizing a Z-free *path-space* KL that uses only the given $\mu(\cdot)$.
- This sets up diffusion samplers à la PIS/DDS and enables DGFS's improvements (intermediate, subtrajectory credit).

Notation tip. Use u_θ (or f) for the vector field to avoid clashing with $\mu(\cdot)$, which denotes the unnormalized density.

Sampling as a Stochastic Optimal Control problem

Forward (controlled) process Q. A Markov chain with Gaussian transitions:

$$Q(x_{0:N}): \quad x_0 \sim p_0^{\text{ref}}, \quad x_{n+1} \sim P_F(\cdot \mid x_n) = \mathcal{N}\big(x_n + h\, f(x_n, n),\ h\sigma^2 I\big).$$

Reference process Q^{ref}. Same covariance, zero drift:

$$x_{n+1} \sim P_F^{\text{ref}}(\cdot \mid x_n) = \mathcal{N}\big(x_n,\ h\sigma^2 I\big), \qquad x_0 \sim p_0^{\text{ref}}, \quad p_n^{\text{ref}} \text{ known}.$$

Target process P. Tie the terminal marginal to π via the reference:

$$P(x_{0:N}) \ := \ Q^{\text{ref}}(x_{0:N})\ \frac{\pi(x_N)}{p_N^{\text{ref}}(x_N)}.$$

Then $P(x_N) \propto \mu(x_N)$, making P a valid path-space target.

Sampling as a Stochastic Optimal Control problem

Learning objective (discrete-time SOC). Learn f by minimizing the path KL:

$$\min_f \; D_{\mathrm{KL}}(Q \parallel P) \iff \min_f \; \mathbb{E}_Q\Big[\sum_{n=0}^{N-1} \frac{h}{2\sigma^2} \|f(x_n, n)\|^2 + \log \Psi(x_N)\Big],$$

with $\Psi(x_N) = \dfrac{p_N^{\mathrm{ref}}(x_N)}{\mu(x_N)}$. (Continuous-time limit recovers the classic VE-SDE SOC formulation.)

Why SOC, and what still hurts (motivation for DGFS)

Why SOC/control-as-inference?

- Principled *path-space* objective; Z cancels, so only μ (and optionally $\nabla \log \mu$) is needed.
- Calibrated sampling by *steering* a simple reference process toward the target.

Pain point in prior SOC samplers (PIS/DDS).

- Training signal sits *only at terminal time* N and losses use *full trajectories* $\Rightarrow$ poor credit assignment, high variance, weaker mode coverage.

DGFS in one line.

- Keep the same SOC/path-KL setup, but introduce a learned *flow function* $F_n(x_n)$ and enforce *subtrajectory balance* to inject *intermediate* learning signals and enable *partial-trajectory* training.

Conclusion