SCORE-BASED DETERMINISTIC DENSITY SAMPLING

Vasily Ilin, Bamdad Hosseini, Jingwei Hu

ABSTRACT

We propose and analyze a deterministic sampling framework using Score-Based Transport Modeling (SBTM) for sampling an unnormalized target density π. While diffusion generative modeling relies on pre-training the score function $\nabla \log f_t$ using samples from π, SBTM addresses the more general and challenging setting where only the $\nabla \log \pi$ is known. SBTM approximates the Wasserstein gradient flow on $\mathrm{KL}(f_t \| \pi)$ by learning the time-varying score $\nabla \log f_t$ on the fly using score matching. The learned score gives immediate access to relative Fisher information, providing a built-in convergence diagnostic. The deterministic trajectories are smooth, interpretable, and free of Brownian-motion noise, while having the same distribution as ULA. We prove that SBTM dissipates relative entropy at the same rate as the exact gradient flow, provided sufficient training. We further extend our framework to annealed dynamics, to handle non log-concave targets. Numerical experiments validate our theoretical findings: SBTM converges at the optimal rate, has smooth trajectories, and is easily integrated with annealed dynamics. We compare to the baselines of ULA and annealed ULA.

1 INTRODUCTION

Diffusion generative modeling (DGM) (Song & Ermon, 2019; Song et al., 2020b) has emerged as a powerful set of techniques to generate "more of the same thing," i.e., given many samples from some unknown distribution π, train a model to generate more samples from π. While the SDE-based generation is most commonly used in practice, its deterministic counterpart, termed "probability flow ODE" by Song et al. (2020b), offers several practical advantages – higher order solvers (Huang et al., 2024), better dimension dependence (Chen et al., 2024), and interpolation in the latent space (Song et al., 2020a). The two processes are given by the reverse of the Ornstein-Uhlenbeck (OU) process and the corresponding ODE, respectively:

$$\mathrm{d}X_t = (X_t + 2\nabla \log f_t(X_t))\ \mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{DGM SDE}$$

$$\mathrm{d}X_t = (X_t + \nabla \log f_t(X_t))\ \mathrm{d}t, \tag{DGM ODE}$$

where f_t is the law of the OU process. The score $\nabla \log f_t$ is learned by running the OU process from π to $\mathcal{N}(0, I_d)$, which crucially depends on having many samples from π. In this work we pose and attempt to answer the following question: **How to deterministically unnormalized density π in the absence of samples, using techniques of diffusion generative modeling?** This is a much harder problem than DGM. Indeed, DGM can be reduced to unnormalized density sampling by estimating the score $\nabla \log \pi$ on the samples, but the converse is not true. Hence, we tackle a strictly harder and more general problem than DGM.

Our main contribution is a deterministic sampling algorithm and a proof of exponentially fast convergence to log-concave distributions. Unlike the classic Langevin dynamics, our algorithm produces smooth deterministic trajectories, and gives access to the otherwise intractable score $\nabla \log f_t$. The convergence analysis is based on a system of coupled gradient flows – a novel framework that can be of independent interest as a general technique for analyzing convergence of NN-based approximations of gradient flows. In particular, we are not aware of any other work that uses the neural tangent kernel for analyzing dynamically changing loss. Our main result is theorem 3.7.

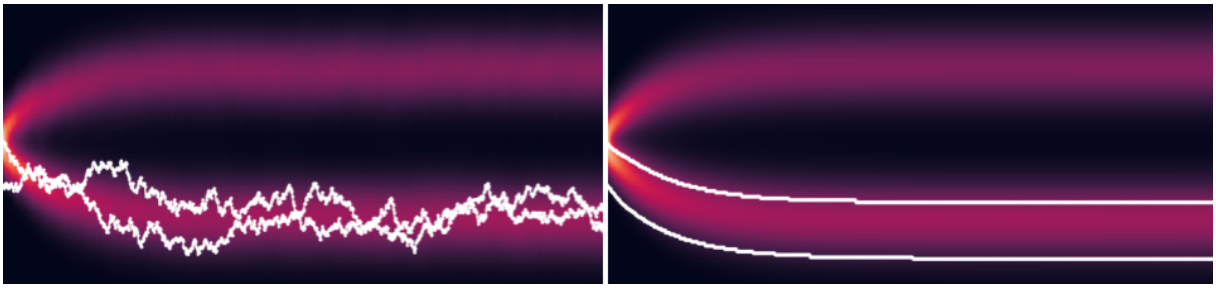

Figure 1: Left: Langevin dynamics (stochastic). Right: ours (deterministic). The deterministic algorithm has the same density as the stochastic one but with smooth trajectories.

In Section 2 we introduce the system of coupled gradient flows, building on Score-Based Transport Modeling (Boffi & Vanden-Eijnden, 2023). In Section 3 we prove that the resulting dynamics achieve the optimal convergence rate. Section 4 adapts some of the results to annealed dynamics. Finally, in Section 5 we verify convergence in several numerical experiments. We defer proofs to the Appendix.

2 Wasserstein Gradient Flow

Diffusion generative modeling (DGM SDE) and (DGM ODE) mixes fast because it is the reverse of a fast process, namely the OU process. In the absence of samples from π, there is no known numerically tractable process that can be run from π to f_0 and reversed. Thus, the standard approach to classical sampling is to greedily minimize some divergence, such as relative entropy $\mathrm{KL}(f_t||\pi) := \mathbb{E}_{f_t} \log \frac{f_t}{\pi}$, between f_t and π at every time step. In continuous time, this is called a Wasserstein gradient flow (GF). The Wasserstein GF on $\mathrm{KL}(\cdot||\pi)$ can be implemented stochastically or deterministically, mimicking equations (DGM SDE) and (DGM ODE):

$$\mathrm{d}X_t = \nabla \log \pi(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad B_t := \text{Brownian motion}, \tag{GF SDE}$$

$$\mathrm{d}X_t = \nabla \log \pi(X_t)\,\mathrm{d}t - \nabla \log f_t(X_t)\,\mathrm{d}t, \quad f_t := \mathrm{law}(X_t),. \tag{GF ODE}$$

See Jordan et al. (1998) for derivation and connection to Optimal Transport. For a deterministic process we focus on (GF ODE). Since we assume access to the unnormalized density π, its score $\nabla \log \pi$ is known. The score $\nabla \log f_t$ must be computed from samples $X_t^1, ..., X_t^n \sim f_t$ alone. This is the idea behind score-based transport modeling.

Remark 2.1. *While equations* (DGM ODE) *and* (GF ODE) *look similar, there is a big difference in how* $\nabla \log f_t$ *is obtained. In* (DGM ODE)*,* $\nabla \log f_t$ *is given by the OU process, and is approximated by a NN* ***pre-trained*** *on the OU process. In* (GF ODE)*, there is only one process – the gradient flow. Hence,* $\nabla \log f_t$ *is given by the gradient flow, and is approximated by a NN trained on the current sample* ***without the costly pre-training****.*

2.1 Score-Based Transport Modeling

Score-based transport modeling (SBTM) was introduced by Boffi & Vanden-Eijnden (2023) as a method of solving Fokker-Planck equations, of which (GF ODE) is a special case. Additionally, SBTM was successfully employed to other Fokker-Planck-type equations (Ilin et al., 2024; Lu et al., 2024; Huang & Wang, 2024).

The core idea of SBTM is to approximate $\nabla \log f_t$ with a neural network s^{Θ_t} trained on the sample $X_t^1, ..., X_t^n$. The training is done the same way as in DGM – by minimizing the score matching loss (Hyvärinen, 2005):

$$\begin{aligned} L(s, f) :&= \mathbb{E}_f \|s - \nabla \log f\|^2 \\ &= \mathbb{E}_f \left(\|s\|^2 + 2\nabla \cdot s \right) + \text{const}(s) && \text{(expand } \|\cdot\|^2\text{, integrate by parts)} \\ &= \frac{1}{n} \sum_{i=1}^n \left(\|s(X_t^i)\|^2 + 2\nabla \cdot s(X_t^i) \right) + \text{const}(s) \quad \text{if} \quad f = \frac{1}{n} \sum_{i=1}^n \delta_{X_t^i}. \end{aligned}$$

Thus the SBTM dynamics are given by the coupled set of equations

$$\begin{aligned} \frac{\mathrm{d}X_t}{\mathrm{d}t} &= \nabla \log \pi(X_t) - s^{\Theta_t}(X_t), \quad X_0 \sim f_0 \text{ iid} \\ \frac{\mathrm{d}\Theta_t}{\mathrm{d}t} &= -\eta \nabla_{\Theta_t} L(s^{\Theta_t}, f_t), \quad X_t \sim f_t, \end{aligned} \tag{SBTM}$$

where $\eta(t)$ is a time rescaling factor that controls the amount of NN training relative to sampling dynamics. Following Karras et al. (2022), we use the second-order Heun time integrator, which allows for larger time steps without losing stability.

Remark 2.2. *The neural network trains on the particles that were produced following the same neural network, as opposed to the true solution of the Wasserstein GF* (GF ODE)*. The commonsense intuition makes one suspect that local errors will accumulate uncontrollably. Miraculously, this does not happen, see remark 3.3.*

3 Convergence Analysis

How quickly does f_t converge to π in (SBTM)**?** To answer this question we must study decay rate of the Lyapunov functional $\mathrm{KL}(f_t||\pi)$. See Chewi (2023) for an excellent exposition.

3.1 Entropy Dissipation

Since (SBTM) approximates the Wasserstein GF on $\mathrm{KL}(f_t||\pi)$, it is natural to ask for a simple sufficient condition to guarantee the optimal rate of entropy dissipation. First, recall the following classical result

Theorem 3.1. *If f_t follows the Wasserstein GF on $KL(\cdot||\pi)$ then relative entropy dissipates at the rate of the relative Fisher information:*

$$-\frac{d}{dt} KL(f_t||\pi) = F(f_t||\pi) := \mathbb{E}_{f_t} \left\| \nabla \log \frac{f_t}{\pi} \right\|^2.$$

Additionally, if π satisfies the log-Sobolev inequality with constant α (e.g. if π is α-log-concave) then

$$KL(f_t||\pi) \leq KL(f_0||\pi) e^{-\frac{1}{\alpha}t}.$$

Thus, the best we can hope for when approximating the true Wasserstein GF with (SBTM) is to recover the original rate of entropy dissipation $-\frac{d}{dt}\mathrm{KL}(f_t||\pi) = \mathrm{F}(f_t||\pi)$. Indeed, this holds if the score matching loss is small:

Theorem 3.2 (Small loss guarantees optimal entropy dissipation)**.** *If f_t is the density of X_t, which follows*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \nabla \log \pi(X_t) - s_t(X_t), \quad X_0 \sim f_0,$$

for any time-dependent vector field s_t, then

$$\begin{aligned} -\frac{d}{dt} KL(f_t||\pi) &= F(f_t||\pi) - \mathbb{E}_{f_t} \left\langle s_t - \nabla \log f_t, \nabla \log \frac{f_t}{\pi} \right\rangle \\ &\geq \frac{1}{2} F(f_t||\pi) - \frac{1}{2} L(s_t, f_t). \end{aligned} \tag{3.1}$$

In particular, if $L(s_t, f_t) \leq \frac{1}{2} F(f_t||\pi)$, then

$$-\frac{d}{dt} KL(f_t||\pi) \geq \frac{1}{4} F(f_t||\pi).$$

While theorem 3.2 looks deceptively simple, it has remarkable properties that form the intuition about SBTM.

Remark 3.3. *Integrating* (3.1) *in time, one obtains*

$$KL(f_T||\pi) \leq -\frac{1}{2}\int_0^T F(f_t||\pi)\,\mathrm{d}t + \frac{1}{2}\int_0^T L(s_t, f_t)\,\mathrm{d}t.$$

Since there no exponential term e^T in the bound, local errors do not accumulate.

Remark 3.4. *Theorem 3.2 applies to any time dependent vector field $s_t(x)$, such as the RKHS parametrization in Maoutsa et al. (2020). But, unlike kernels, the neural network parametrization can achieve any desired accuracy by scaling compute.*

3.2 Bounding Score Matching Loss

John von Neumann famously said: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." Similarly, with sufficient number of weights and training a neural network can fit arbitrary (finite) data.

Theorem 3.5. *Assume that f_t is any time-dependent density, $X_t^1, \ldots, X_t^n$ are distributed according to f_t, and the neural network $s = s_t^\Theta$ is trained with gradient descent*

$$\frac{\mathrm{d}\Theta_t}{\mathrm{d}t} = -\eta\nabla_{\Theta_t} L^n(s_t, f_t), \quad L^n(s_t, f_t) := \frac{1}{n}\sum_{i=1}^n \|s(X_t^i) - \nabla \log f(X_t^i)\|^2$$

and is such that the Neural Tangent Kernel (NTK)

$$H_{\alpha,\beta}^{i,j}(t) = \sum_{k=1}^N \nabla_{\theta_k} s_t^\alpha(X_t^i)\nabla_{\theta_k} s_t^\beta(X_t^j), \quad s(x) = (s^1(x), \ldots, s^d(x))$$

is lower bounded by $\lambda = \lambda(t) > 0$, i.e. $\|Hv\|^2 \geq \lambda\|v\|^2$. Then as long as

$$\eta(t) \geq \frac{n}{\lambda}\frac{\partial}{\partial\tau}\Big|_{\tau=t} \log L^n(s_t, f_\tau),$$

the loss $L^n(s_t, f_t)$ is non-increasing, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} L^n(s_t, f_t) \leq 0.$$

Remark 3.6. *The assumption that the NTK is lower bounded in theorem 3.5 is non-trivial, but holds under relatively mild assumptions (Karhadkar et al., 2024); the most restrictive one is that the NN size is superlinear in the number of datapoints. We expect that even this can be weakened under additional regularity assumptions of the target function.*

Combining theorems 3.1 and 3.5 shows that the convergence rate of SBTM matches the convergence rate of the true GF (GF ODE).

Theorem 3.7. *Suppose that X_t and Θ_t follow* (SBTM) *and η and H are as in theorem 3.5. If the initial loss is small and the true loss L is well-approximated by the training loss L^n*

$$L^n(s_0, f_0) \leq \frac{1}{4}\varepsilon, \quad \varepsilon := \inf_{t\leq T} F(f_t||\pi), \tag{3.2}$$

$$|L^n(s_t, f_t) - L(s_t, f_t)| \leq \frac{1}{4}\varepsilon, \tag{3.3}$$

then SBTM dissipates relative entropy at the same rate, up to factor $\frac{1}{4}$, as the true Wasserstein GF on $KL(\cdot||\pi)$:

$$-\frac{d}{dt}KL(f_t||\pi) \geq \frac{1}{4}F(f_t||\pi).$$

If π satisfies the log-Sobolev inequality with constant α then convergence is exponential:

$$KL(f_t||\pi) \leq KL(f_0||\pi)e^{-\frac{1}{4\alpha}t}.$$

Remark 3.8. *Since $s_t \approx \nabla \log f_t$, relative Fisher information may be approximated by*

$$F(f_t||\pi) \approx F^n(f_t||\pi) := \frac{1}{n}\sum_{i=1}^{n} \|s(X^i) - \nabla \log \pi(X^i)\|^2.$$

One may stop the sampling when $F^n(f_t||\pi) \leq \varepsilon$ for some predefined ε. This makes condition (3.2) *practical.*

If particles $X^1(t), \ldots, X^n(t)$ were independent, the law of large numbers would imply

$$\lim_{n\to\infty} |L^n(s_t, f_t) - L(s_t, f_t)| = 0,$$

satisfying condition (3.3) *for large enough n. Further, one may hope for propagation of chaos, namely that as $n \to \infty$, any finite subset of particles becomes independent. We have not attempted to find the set of assumptions to guarantee this. However, in numerical experiments, we treat L^n as L, and observe the optimal rate of relative entropy dissipation (e.g. in the left panel of figure 2).*

4 Annealed Langevin Dynamics

Classical sampling can be broken down into three distinct parts: mode discovery, mode weighting, and mode approximation. While Langevin dynamics performs mode approximation very fast, both mode discovery and mode weighting are challenging. For example, even without any dynamics, in the absence of samples from π, the mere discovery of a mode with support of small width h in d dimensions requires $\Omega\left(h^{-d}\right)$ function evaluations. This is in contrast to DGM, where the reverse of the OU process converges to π fast, without any log-concavity assumptions, see Chen et al. (2022). While the problem of mode discovery is fundamentally hard, empirically it helps to pick an interpolation or annealing between f_0 and π to "guide" f_t, similar to how the forward OU process guides the reverse process in DGM. The standard annealing path is the geometric interpolation $\pi_t \propto f_0^{1-t}\pi^t$ (Neal, 2001). While classical, this interpolation may require teleportation of mass, because the velocity $\dot{X}_t$ can become singular, as noted by Máté & Fleuret (2023); Chemseddine et al. (2024). Recently, the dilation interpolation was proposed by Chehab & Korba (2024), which also suffers from an exploding velocity at $t = 0$ but the singularity can be controlled numerically.

One obtains a similar entropy dissipation estimate in annealed dynamics. Taking $\pi_t = \pi$ recovers theorem 3.2.

Theorem 4.1 (Entropy Dissipation in Annealed Dynamics)**.** *If π_t is any time-dependent density, s_t any time-dependent vector field, and*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \nabla \log \pi_t(X_t) - s(X_t)$$

then

$$-\frac{d}{dt}KL(f_t||\pi) = \mathbb{E}_{f_t}\left\langle s_t - \nabla \log \pi, \nabla \log \frac{f_t}{\pi_t}\right\rangle - \mathbb{E}_{f_t}\left\langle s_t - \nabla \log f_t, \nabla \log \frac{f_t}{\pi}\right\rangle.$$

In particular, if $L(s_t, f_t) = 0$, then

$$-\frac{d}{dt}KL(f_t||\pi) = \mathbb{E}_{f_t}\left\langle \nabla \log \frac{f_t}{\pi}, \nabla \log \frac{f_t}{\pi_t}\right\rangle. \tag{4.1}$$

Theorem 4.1 gives an indirect way to test for $L(s_t, f_t) = 0$ by testing equality (4.1). This is important, because the loss $L(s_t, f_t)$ cannot be computed from a finite sample $X_t^1, \ldots, X_t^n$ of f_t. Numerically, we observe that (4.1) nearly holds (see figures 4 and 6).

5 Experiments

We demonstrate the optimal rate of relative entropy dissipation in several experiments in low dimensions, including challenging non log-concave targets. We compare SBTM to its stochastic counterpart given by (GF SDE). Additionally, we demonstrate the flexibility of SBTM by simulating annealed Langevin dynamics with the geometric

and the dilation annealing paths (Chehab & Korba, 2024) on the two most challenging examples, and compare to the corresponding SDEs such as the ULA used in (Chehab & Korba, 2024). We do not compare SBTM to SVGD (Liu & Wang, 2016) because without additional tricks such as momentum and cherry-picking the kernel bandwidth we were unable to obtain even remotely good performance from SVGD. We emphasize while the numerical experiments presented below would be trivial in the context of DGM, they are quite challenging in our context.

5.1 Linear Fokker-Planck Equation

First, we consider an example that admits an analytic solution, f_t in (5.1), which allows us to compute the true entropy dissipation and the L^2 distance to the true solution.

$$\pi = \mathcal{N}(0,1), \quad f_t = \mathcal{N}\left(0, 1 - e^{-2(t+0.1)}\right) \tag{5.1}$$

The SDE solution is more noisy than SBTM. Moreover, SBTM exhibits close-to-optimal relative entropy dissipation rate, as evidenced by the close alignment of the orange, green, and red lines in the left panel of figure 2.

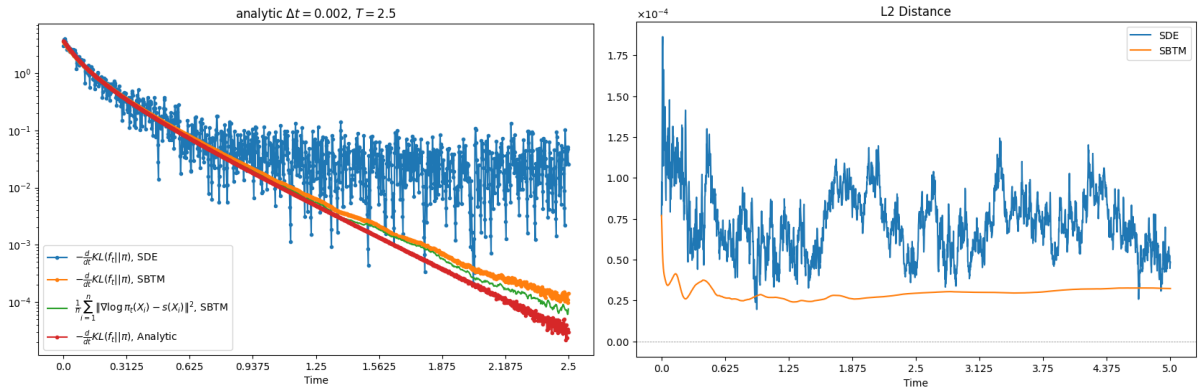

Figure 2: Left: entropy dissipation of SBTM (ours) and SDE (stochastic). SBTM approximates entropy decay rate perfectly, while SDE is extremely noisy. Right: L2 error to the true solution. SBTM produces lower error with much smoother trajectory.

5.2 Gaussian Mixture

This and all other examples do not admit an analytic solution but we can still compare the quality of the final sample as well as compare the trajectories of the SDE and SBTM. Here we sample from the Gaussian mixture $\pi = \frac{1}{4}\mathcal{N}(-2,1) + \frac{3}{4}\mathcal{N}(2,1)$. As evidences by the change of slope in the right panel of figure 3, around $t = 2.5$ the Markov chain enters metastability, which plagues the convergence to non log-concave targets. Here the non-log-concavity is mild, so the Markov chain still converges in a reasonable time frame.

5.3 Gaussian Mixture with Geometric Annealing

When the target is non log-concave, its annealed score can be used. Here we use the classical geometric annealing.

$$\pi = \frac{1}{4}\mathcal{N}(-4,1) + \frac{3}{4}\mathcal{N}(4,1)$$
$$\nabla \log \pi_t = (1-t)\nabla \log f_0 + t\nabla \log \pi^t.$$

The match between the orange and blue lines in the right panel of figure 4 indirectly indicates very good score approximation, as per (4.1).

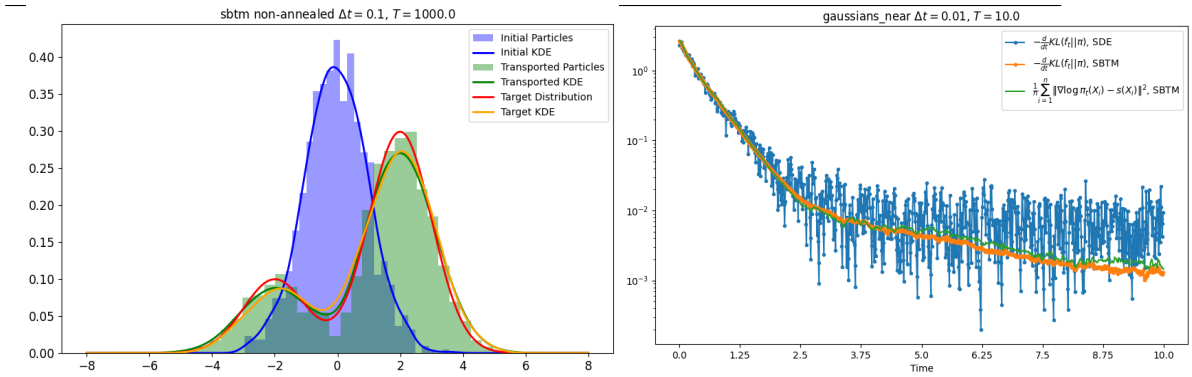

Figure 3: Left: reconstructed density of SBTM. It approximates the solution pefectly. Right: entropy dissipation of SBTM (ours) and SDE (stochastic). SBTM approximates entropy decay rate perfectly, while SDE is extremely noisy.

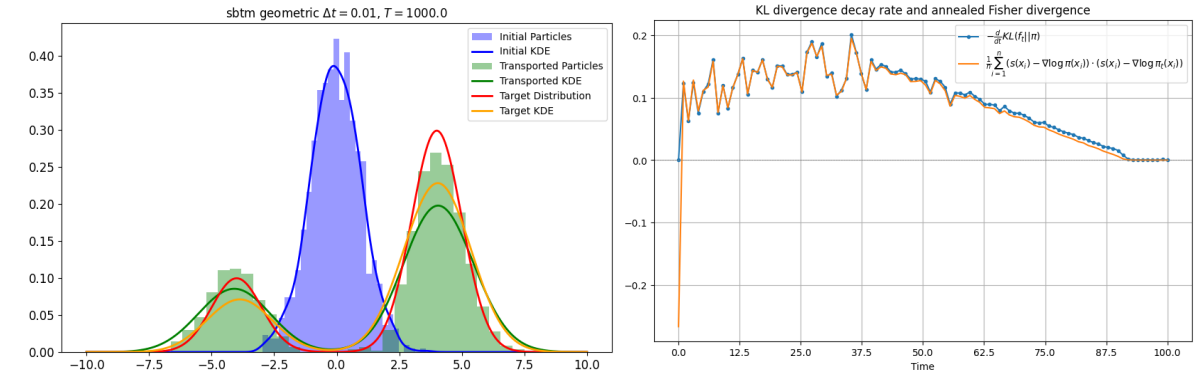

Figure 4: Left: reconstructed density of SBTM. It approximates the solution well despite the non-log-concavity. Right: entropy dissipation of SBTM (ours) and SDE (stochastic). SBTM approximates entropy decay rate perfectly even in annealed dynamics.

5.4 Gaussian Mixture with Dilation Annealing

Sampling from a Gaussian mixture with 16 modes requires a better annealing path. We employ the annealing schedule $\pi_t(x) = \pi\left(\frac{T}{t}x\right)$ from Chehab & Korba (2024). Figure 5 shows the densities at different time points, and figure 6 shows the entropy dissipation as in (4.1). This is a very challenging example for classical sampling due to the extreme non-log-concavity of the target. While the dilation annealing performs much better on this example than the geometric annealing, both perform equally sub-par on example 5.3.

6 Conclusion

In this work we use tools and intuition from diffusion generative modeling to tackle the harder problem of sampling π given only $\nabla \log \pi$ but not samples from π. Our method allows for deterministic sampling with smooth trajectories and optimal rate of entropy dissipation. Additionally, SBTM gives access to the score, which allows for the computation of relative Fisher info to estimate convergence, making the dynamics interpretable. Finally, SBTM integrates well with annealed dynamics to sample from challenging non log-concave densities.

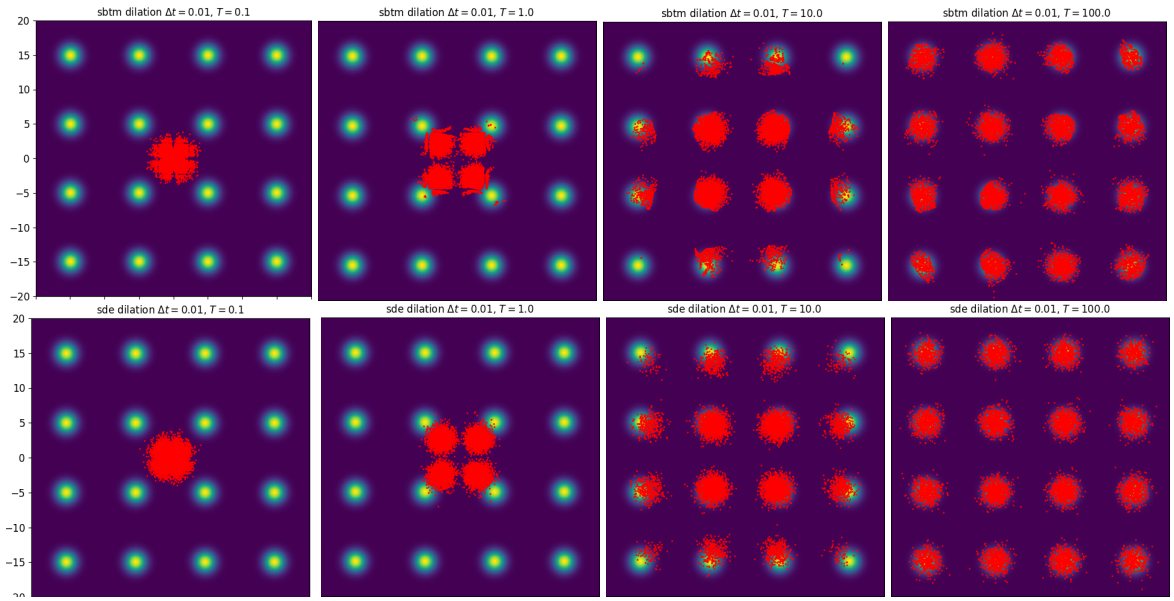

Figure 5: Top: SBTM (ours). Bottom: SDE (Chehab & Korba, 2024).

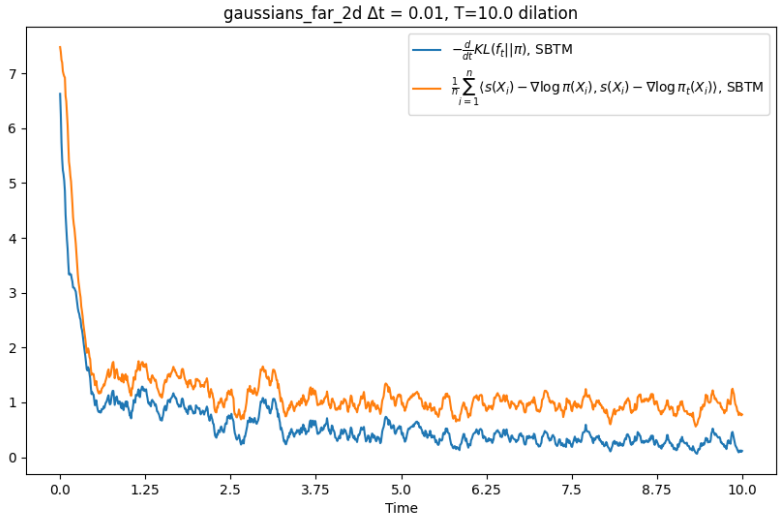

Figure 6: Relative entropy decay in SBTM. The estimate in (4.1) nearly holds, indicating good score approximation and allowing for convergence monitoring.

REFERENCES

Nicholas M Boffi and Eric Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, 2023.

Omar Chehab and Anna Korba. A practical diffusion path for sampling. *arXiv preprint arXiv:2406.14040*, 2024.

Jannis Chemseddine, Christian Wald, Richard Duong, and Gabriele Steidl. Neural sampling from boltzmann densities: Fisher-rao curves in the wasserstein geometry. *arXiv preprint arXiv:2410.03282*, 2024.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024.

Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github. io*, 9:17–18, 2023.

Miguel Corrales, Sean Berti, Bertrand Denel, Paul Williamson, Mattia Aleardi, and Matteo Ravasi. Annealed stein variational gradient descent for improved uncertainty estimation in full-waveform inversion. *Geophysical Journal International*, 241(2):1088–1113, 2025.

Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

A Duncan, N Nuesken, and L Szpruch. On the geometry of stein variational gradient descent, 775. *arXiv preprint arXiv:1912.00894*, 374, 2019.

Karthik Elamvazhuthi, Xuechen Zhang, Matthew Jacobs, Samet Oymak, and Fabio Pasqualetti. A score-based deterministic diffusion algorithm with smooth scores for general distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11866–11873, 2024.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*, 2024.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

Yan Huang and Li Wang. A score-based particle method for homogeneous landau equation. *arXiv preprint arXiv:2405.05187*, 2024.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL `http://jmlr.org/papers/v6/hyvarinen05a.html`.

Vasily Ilin, Jingwei Hu, and Zhenfu Wang. Transport based particle methods for the fokker-planck-landau equation. *arXiv preprint arXiv:2405.10392*, 2024.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Kedar Karhadkar, Michael Murray, and Guido Montúfar. Bounds for the smallest eigenvalue of the ntk for arbitrary spherical data of arbitrary dimension. *arXiv preprint arXiv:2405.14630*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Jianfeng Lu, Yue Wu, and Yang Xiang. Score-based transport modeling for mean-field fokker-planck equations. *Journal of Computational Physics*, 503:112859, 2024.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker–planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, 2020.

Bálint Máté and François Fleuret. Learning interpolations between boltzmann densities. *arXiv preprint arXiv:2301.07388*, 2023.

Aimee Maurais and Youssef Marzouk. Sampling in unit time with kernel fisher-rao flow. *arXiv preprint arXiv:2401.03892*, 2024.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

Chris J Oates. Minimum kernel discrepancy estimators. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pp. 133–161. Springer, 2022.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. 1996.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Amirhossein Taghvaei, Jana De Wiljes, Prashant G Mehta, and Sebastian Reich. Kalman filter and its modern extensions for the continuous-time nonlinear filtering problem. *Journal of Dynamic Systems, Measurement, and Control*, 140(3):030904, 2018.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7): 1661–1674, 2011.

Linfeng Wang and Nikolas Nüsken. Measure transport with kernel mean embeddings. *arXiv preprint arXiv:2401.12967*, 2024.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on learning theory*, pp. 2093–3027. PMLR, 2018.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by jko scheme. *arXiv preprint arXiv:2212.14424*, 2022.

7 APPENDIX

7.1 RELATED WORK

Stochastic-gradient MCMC. Classical samplers such as the overdamped and unadjusted Langevin algorithms can be interpreted as Wasserstein gradient flows (WGF) of the KL functional. Their convergence is now well understood under log-Sobolev or isoperimetric conditions (Roberts & Tweedie, 1996; Wibisono, 2018; Dalalyan & Karagulyan, 2017; Vempala & Wibisono, 2019; Chewi, 2023). Although provably fast for log-concave targets, these stochastic methods mix slowly in multimodal landscapes, motivating variance reduction, tempering, and control-variate extensions.

Deterministic interacting particles. Stein Variational Gradient Descent (SVGD), introduced by Liu & Wang (2016), performs steepest descent of $\mathrm{KL}(f_t \parallel \pi)$ in a Stein-RKHS and admits a mean-field limit that solves a nonlinear Fokker–Planck PDE (Duncan et al., 2019). Follow-ups improved uncertainty quantification via annealing (Corrales et al., 2025), proposed kernel mean-embedding flows (Wang & Nüsken, 2024), introduced unit-time sampling with the Kernel Fisher–Rao flow (Maurais & Marzouk, 2024), and connected these ideas to ensemble Kalman transport maps (Taghvaei et al., 2018). While fully deterministic, such methods still rely on handcrafted kernels and can suffer in high dimensions.

Score matching for unnormalised models. Score matching, proposed by Hyvärinen (2005), estimates the score without the partition function. Its link to denoising autoencoders was clarified by Vincent (2011). In the sampling context, minimum-discrepancy and energy-based variants have been developed (Oates, 2022; Chen et al., 2022), but none analyse the feedback loop created when the score network is trained *online* from the evolving particle cloud—an issue central to our setting.

Score-driven deterministic flows. Probability-flow ODEs are exact deterministic counterparts of score-based diffusions (Song & Ermon, 2019; Song et al., 2020b). Recent analyses establish fast convergence when the score is pre-trained from data (Huang et al., 2024; Chen et al., 2024). Learning the score *on the fly* for an unnormalised target was first proposed in Score-Based Transport Modeling (SBTM) by Boffi & Vanden-Eijnden (2023) and later extended via neural JKO schemes (Xu et al., 2022) and smooth deterministic diffusions (Elamvazhuthi et al., 2024). Our work tightens these results by proving entropy decay for the coupled particle–network dynamics.

Annealing and tempering. Bridging a simple reference distribution and a rugged target through an inverse-temperature path is a staple of importance sampling and sequential Monte Carlo (Neal, 2001). Deterministic analogues include annealed SVGD (Corrales et al., 2025) and, on the diffusion side, temperature-scheduled probability-flow ODEs (Huang et al., 2025).

Summary. We unify and advance these strands by (i) coupling a transport ODE with *online* score matching for an unnormalised target, (ii) proving dimension-free exponential decay of $\mathrm{KL}(f_t \parallel \pi)$ under standard log-Sobolev assumptions, and (iii) extending the analysis to annealed schedules for non-convex densities—thereby pushing deterministic sampling closer to practical applicability.

7.2 PROOFS

Theorem 7.1 (Entropy Dissipation in Annealed Dynamics)**.** *If π_t is any time-dependent density, s_t any time-dependent vector field, and*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \nabla \log \pi_t(X_t) - s(X_t)$$

then

$$-\frac{d}{dt}KL(f_t||\pi) = \mathbb{E}_{f_t}\left\langle s_t - \nabla \log \pi, \nabla \log \frac{f_t}{\pi_t}\right\rangle - \mathbb{E}_{f_t}\left\langle s_t - \nabla \log f_t, \nabla \log \frac{f_t}{\pi}\right\rangle.$$

In particular, if $L(s_t, f_t) = 0$*, then*

$$-\frac{d}{dt}KL(f_t||\pi) = \mathbb{E}_{f_t}\left\langle \nabla \log \frac{f_t}{\pi}, \nabla \log \frac{f_t}{\pi_t}\right\rangle.$$

Proof. If f_t is the density of X_t, where X_t satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t = \nabla \log \pi_t(X_t) - s_t(X_t)$$

then f_t satisfies the Fokker-Planck equation

$$\partial_t f_t + \nabla \cdot (f_t(\nabla \log \pi_t - s_t)) = 0.$$

Thus, we may explicitly compute the relative entropy dissipation rate as

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} f_t \log \frac{f_t}{\pi}\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \partial_t f_t(\log f_t - \log \pi)\,\mathrm{d}x + \int_{\mathbb{R}^d} f_t \partial_t \log f_t\,\mathrm{d}x \\
&= -\int_{\mathbb{R}^d} \langle s_t - \nabla \log \pi_t, \nabla \log f_t - \nabla \log \pi\rangle\, f_t\,\mathrm{d}x,
\end{aligned}$$

where we used integration by parts and that $\int_{\mathbb{R}^d} f_t \partial_t \log f_t\,\mathrm{d}x$ is zero for the last equality. □

Proofs of 3.1 and 3.2. By choosing $\pi_t = \pi$ in the proof of theorem 4.1, we obtain

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} f_t \log \frac{f_t}{\pi}\,\mathrm{d}x \\
&= -\int_{\mathbb{R}^d} \langle s_t - \nabla \log \pi, \nabla \log f_t - \nabla \log \pi\rangle\, f_t\,\mathrm{d}x.
\end{aligned}$$

By taking $s_t = \nabla \log f_t$ exactly, we recover the proof of the classical result 3.1. Otherwise, adding and subtracting $\nabla \log f_t$ in the integrand, we get

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} f_t \log \frac{f_t}{\pi}\,\mathrm{d}x \\
&= -\int_{\mathbb{R}^d} \|\nabla \log f_t - \nabla \log \pi\|^2 f_t\,\mathrm{d}x + \int_{\mathbb{R}^d} \langle \nabla \log f_t - s_t, \nabla \log f_t - \nabla \log \pi\rangle\, f_t\,\mathrm{d}x \\
&\le -\frac{1}{2}\int_{\mathbb{R}^d} \|\nabla \log f_t - \nabla \log \pi\|^2 f_t\,\mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}^d} \|\nabla \log f_t - s_t\|^2 f_t\,\mathrm{d}x
\end{aligned}$$

The last line is by Young's inequality. □

Proof of theorem 3.5. We start with an elementary computation based on chain rule.

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} L(s_t, f_t) &= \frac{\partial}{\partial \tau}\Big|_{\tau=t} L(s_\tau, f_t) + \frac{\partial}{\partial \tau}\Big|_{\tau=t} L(s_t, f_\tau) \\
\frac{\partial}{\partial \tau}\Big|_{\tau=t} L(s_\tau, f_t) &= \nabla_\Theta L \cdot \frac{\mathrm{d}}{\mathrm{d}t}\Theta \\
&= -\eta \sum_{k=1}^{N} (\nabla_{\theta_k} L)^2 \\
&= -\frac{\eta}{n^2} \sum_{k=1}^{N} \left([s(X^i) - \nabla \log f(X^i)] \cdot \nabla_{\theta_k} s(X^i) \right)^2 \\
&= -\frac{\eta}{n^2} \sum_{i,j=1}^{n} \sum_{\alpha,\beta=1}^{d} [s_\alpha(X^i) - \nabla_\alpha \log f(X^i)] H^{i,j}_{\alpha,\beta} [s_\beta(X^j) - \nabla_\beta \log f(X^j)] \\
&= -\frac{\eta}{n^2} \|s - \nabla \log f\|_H^2.
\end{aligned}
$$

where

$$
H^{i,j}_{\alpha,\beta} = \sum_{k=1}^{N} \nabla_{\theta_k} s_\alpha(X^i) \nabla_{\theta_k} s_\beta(X^j)
$$

is called the Neural Tangent Kernel. Its lowest eigenvalue determines the convergence speed of gradient descent. If $\|Hv\|^2 \geq \lambda \|v\|^2$, then

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} L(s_t, f_t) &= \frac{\partial}{\partial \tau}\Big|_{\tau=t} L(s_\tau, f_t) + \frac{\partial}{\partial \tau}\Big|_{\tau=t} L(s_t, f_\tau) \\
&\leq -\frac{\eta\lambda}{n} L(s_t, f_t) + \frac{\partial}{\partial \tau}\Big|_{\tau=t} L(s_t, f_\tau).
\end{aligned}
$$

Thus, the loss is non-increasing if

$$
\eta(t) \geq \frac{n}{\lambda} \frac{\partial}{\partial \tau}\Big|_{\tau=t} \log L(s_t, f_\tau).
$$

□