Theoretical Foundations for Diffusion Models

Holden Lee

Johns Hopkins University

February 23, 2024

Outline

1. Introduction
   - Diffusion models in generative modeling
   - The score function
2. Convergence theory given an accurate score function
   - Convergence for general distributions without smoothness
   - Faster convergence with the probability flow ODE
3. Learning the score function
   - Gaussian mixture

What are diffusion models?

Problem (Generative Modeling)

Learn a probability distribution from samples, and generate additional samples.

- Diffusion models (Hyvärinen 2005; Sohl-Dickstein, Weiss, Maheswaranathan, et al. 2015; Y. Song and Ermon 2019) are a modern paradigm for generative modeling with state-of-the-art performance on image, audio, video generation.

- Core component of DALL·E, Imagen, Stable Diffusion...

Pictures from Ramesh, Dhariwal, Nichol, et al. 2022.

an espresso machine that makes coffee from human souls, artstation — panda mad scientist mixing sparkling chemicals, artstation — a corgi's head depicted as an explosion of a nebula

What **theoretical guarantees** can we obtain for diffusion models?

What are diffusion models?

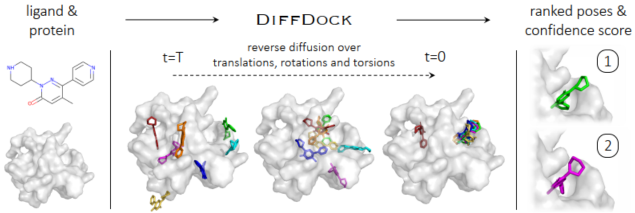
Problem (Generative Modeling)

Learn a probability distribution from samples, and generate additional samples.

- Diffusion models (Hyvärinen 2005; Sohl-Dickstein, Weiss, Maheswaranathan, et al. 2015; Y. Song and Ermon 2019) are a modern paradigm for generative modeling with scientific applications including:

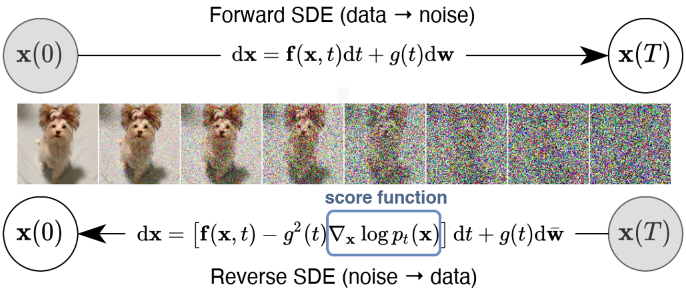
- Inverse problems
- Physics simulations
- Molecular modeling
- Protein design

Corso, Stärk, Jing, Barzilay, and Jaakkola 2022



What **theoretical guarantees** can we obtain for diffusion models?

How diffusion models work

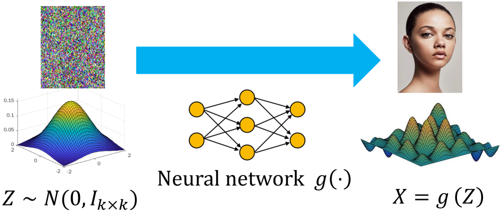


Picture from Y. Song, Sohl-Dickstein, Kingma, et al. 2020

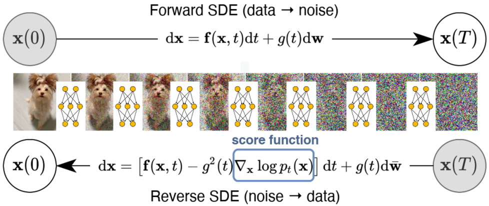
- Define a forward process, e.g., stochastic differential equation (SDE) that converts data into pure noise.
  - Can also be Markov process on discrete space or (deterministic) ODE.
  - General framework: Montanari 2023. ($\leftrightarrow$ Stochastic localization)
- Transform pure noise into samples from learned data distribution via reverse process.
- The SDE involves the (Stein) **score function**, often estimated with a neural network.

Diffusion models vs. other generative models

Generative adversarial networks (GAN's), variational auto-encoders, normalizing flows...



Diffusion models:



Two steps to diffusion models

1. Estimate the score function from data.

Definition

The **(Stein) score function** of a probability distribution with density $p(x) \propto e^{-V(x)}$ is

$$s(x) = \nabla \ln p(x) = -\nabla V(x).$$

2. Draw samples given a score estimate.
   - **Langevin Monte Carlo**: Algorithm for drawing samples from $p \propto e^{-V}$ given the score.

Langevin diffusion	$\to$	Langevin Monte Carlo
$dx_t = -\nabla V(x_t) + \sqrt{2}\, dw_t$	$\to$	$x_{t+h} = h \cdot \underbrace{\cancel{-\nabla V(x_t)}}_{s(x_t)} + \sqrt{2h} \cdot \xi_t,$

   - Other diffusion processes based on **reverse SDE's**.

Estimating the score function: Bayesian inference problem

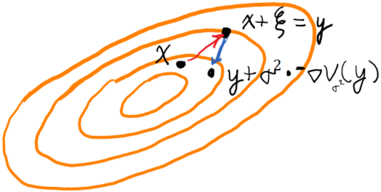
Reduction to supervised learning problem: The score function can be estimated in $L^2(p)$ by minimizing the denoising auto-encoder (DAE) objective. Let φ_{σ^2} be the density of $N(0, \sigma^2 I_d)$.

Proposition (Vincent 2011)

Suppose $p * \varphi_{\sigma^2} \propto e^{-V_{\sigma^2}}$. The minimum of the DAE objective

$$L_{\mathrm{DAE}}(r) = \mathbb{E}_{X\sim p}\mathbb{E}_{\xi\sim N(0,\sigma^2)}\left[\|r(X+\xi) - X\|^2\right]$$

is $r(y) = y - \sigma^2 \nabla V_{\sigma^2}(y)$, i.e., $\nabla V_{\sigma^2}(y) = \frac{y - r(y)}{\sigma^2}$.



Estimating the score function: Bayesian inference problem

Proposition (Vincent 2011, equivalently)

Suppose $p * \varphi_{\sigma^2} \propto e^{-V_{\sigma^2}}$. The minimum of the objective

$$L(g) = \mathbb{E}_{X\sim p}\mathbb{E}_{\xi\sim N(0,\sigma^2)}\left[\|g(X+\xi)-\xi\|^2\right] \quad \text{is} \quad g(y) = \sigma^2\nabla V_{\sigma^2}(y)$$

Consider $X \sim p$, $\xi \sim N(0, \sigma^2 I_d)$, $Y = X + \xi$. By Bayes's Rule,

$$\begin{aligned}
\nabla V_{\sigma^2}(y) &= -\nabla \ln(p * \varphi_{\sigma^2}(y)) = -\nabla_y \ln \int_{\mathbb{R}^d} e^{-V_0(x)} e^{-\frac{\|y-x\|^2}{2\sigma^2}}\,dx \\
&= \frac{\int_{\mathbb{R}^d} \frac{y-x}{\sigma^2} e^{-V_0(x)} e^{-\frac{\|y-x\|^2}{2\sigma^2}}\,dx}{\int_{\mathbb{R}^d} e^{-V_0(x)} e^{-\frac{\|y-x\|^2}{2\sigma^2}}\,dx} = \frac{\int_{\mathbb{R}^d} \frac{y-x}{\sigma^2} p(x) P(y|x)\,dx}{\int_{\mathbb{R}^d} p(x)P(y|x)\,dx} \\
&= \frac{1}{\sigma^2}\mathbb{E}[y - X | Y = y] = \frac{1}{\sigma^2}\mathbb{E}[\xi | Y = y].
\end{aligned}$$

Estimating the score function: Bayesian inference problem

Proposition (Vincent 2011, equivalently)

Suppose $p * \varphi_{\sigma^2} \propto e^{-V_{\sigma^2}}$. The minimum of the objective

$$L(g) = \mathbb{E}_{X\sim p}\mathbb{E}_{\xi\sim N(0,\sigma^2)}\left[\|g(X+\xi)-\xi\|^2\right] \quad \text{is} \quad g(y) = \sigma^2\nabla V_{\sigma^2}(y)$$

Consider $X \sim p$, $\xi \sim N(0, \sigma^2 I_d)$, $Y = X + \xi$.

Key identity (Tweedie's formula)

$$\sigma^2\nabla V_{\sigma^2}(y) = \mathbb{E}[y - X|Y = y] = \mathbb{E}[\xi|Y = y].$$

Identity $\Rightarrow$ Proposition: The minimal mean square estimator (MMSE) g is exactly the mean of the posterior $\mathbb{E}[\xi|Y = y]$.

Two steps to SGM

1. Estimate the score function from data.

$$\|\nabla \ln p - s\|_{L^2(p)}^2 = \mathbb{E}_p \|\nabla \ln p(x) - s(x)\|^2 \le \varepsilon^2.$$

2. Draw samples given a score estimate, using **reverse SDE**.

Two steps to SGM

1. Estimate the score function from data.

$$\|\nabla \ln p - s\|_{L^2(p)}^2 = \mathbb{E}_p \|\nabla \ln p(x) - s(x)\|^2 \le \varepsilon^2.$$

2. Draw samples given a score estimate, using **reverse SDE**.

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

Key differences from usual setting of sampling algorithms (Langevin Monte Carlo):

- We only have $L^2(p)$-accurate score function.
- Non-time-homogeneous process ("non-equilibrium thermodynamics").

Two steps to SGM

1. Estimate the score function from data.

$$\|\nabla \ln p - s\|^2_{L^2(p)} = \mathbb{E}_p \|\nabla \ln p(x) - s(x)\|^2 \le \varepsilon^2.$$

2. Draw samples given a score estimate, using **reverse SDE**.

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

- Why not just Langevin Monte Carlo? Efficient sampling relies on mixing.
- Reverse process acts like annealing to allow sampling from multimodal distributions.

Two steps to SGM

1. Estimate the score function from data.

Question 2

When can we obtain a $L^2(p)$-accurate score estimate?

$$\|\nabla \ln p - s\|_{L^2(p)}^2 = \mathbb{E}_p \|\nabla \ln p(x) - s(x)\|^2 \le \varepsilon^2.$$

2. Draw samples given a score estimate, using **reverse SDE**.

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

- Why not just Langevin Monte Carlo? Efficient sampling relies on mixing.
- Reverse process acts like annealing to allow sampling from multimodal distributions.

Questions

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

1. Can we get guarantees for general data distributions **without smoothness**?

2. Can we obtain **better dimension dependence**?

Questions

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

1. Can we get guarantees for general data distributions **without smoothness**?
   [H. Chen, **L**, and Lu 2023], Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions.
   http://www.arxiv.org/abs/2211.01916
   - Smoothing properties of the forward process lead to good convergence rates for arbitrary data distributions.
2. Can we obtain **better dimension dependence**?
   [S. Chen, Chewi, **L**, Li, Lu, and Salim 2023], The probability flow ODE is provably fast.
   http://www.arxiv.org/abs/2305.11798
   - Using an ODE instead of SDE, in conjunction with a corrector step, can reduce dimension dependence from $O(d)$ to $O(\sqrt{d})$.

DDPM with exponential integrator

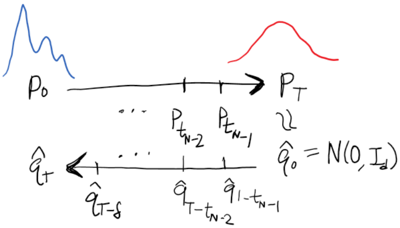
Suppose unit speed: $g \equiv 1$.

$$\text{Forward SDE:} \qquad dx_t = -\frac{1}{2}x_t\,dt + dw_t$$

$$\text{Backward SDE:} \qquad dx_t^{\leftarrow} = \frac{1}{2}\left[x_t + 2\nabla \ln p_{T-t}(x_t^{\leftarrow})\right]dt + dw_t$$

Start with

$$z_0 \sim N(0, I_d) \approx p_T.$$



DDPM with exponential integrator

Suppose unit speed: $g \equiv 1$.

$$\text{Forward SDE:} \qquad dx_t = -\frac{1}{2} x_t \, dt + dw_t$$

$$\text{Backward SDE:} \qquad dx_t^{\leftarrow} = \frac{1}{2} \left[x_t + 2\nabla \ln p_{T-t}(x_t^{\leftarrow}) \right] dt + dw_t$$

Start with

$$z_0 \sim N(0, I_d) \approx p_T.$$

Exponential integrator: letting $h_k = t_k - t_{k-1}$,

$$z_{T-t_{k-1}} = z_{T-t_k} + (e^{h_k/2} - 1)(z_{T-t_k} + 2s(z_k, t_k)) + \sqrt{e^{h_k} - 1} \underbrace{\eta_k}_{\sim N(0, I_d)}.$$

Summary of results

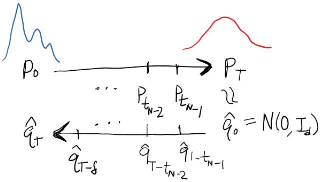
Theorem (H. Chen, L, and Lu 2023, informal)

Suppose p_0 has bounded 2nd moment and average L^2 score error is $\le \varepsilon_{sc}$. Guarantees for DDPM hold under the following smoothness assumptions:

Smoothness assumption	*Error guarantee*	*Steps to get $\widetilde{O}(\varepsilon_{sc}^2)$ error*
$\forall t, \nabla \ln p_t$ *L-Lipschitz*	$\mathrm{KL}(p_0\|\|\hat{q}_T)$	$O\left(\frac{dL^2}{\varepsilon_{sc}^2}\right)$
$\nabla \ln p_0$ *L-Lipschitz*	$\mathrm{KL}(p_0\|\|\hat{q}_T)$	$O\left(\frac{d^2(\ln L)^2}{\varepsilon_{sc}^2}\right)$
None	$\mathrm{KL}(p_\delta\|\|\hat{q}_{T-\delta})$	$O\left(\frac{d^2\ln(1/\delta)^2}{\varepsilon_{sc}^2}\right)$

Intuition:

Lipschitz constant of $\nabla \ln p_t$ for $t = \Omega(1)$ is "effectively" bounded by $\sqrt{d}$.



Summary of results

Theorem (H. Chen, **L**, and Lu 2023, informal)

Suppose p_0 *has bounded 2nd moment and average* L^2 *score error is* $\le \varepsilon_{sc}$. *Guarantees for DDPM hold under the following smoothness assumptions:*

Smoothness assumption	*Error guarantee*	*Steps to get* $\widetilde{O}(\varepsilon_{sc}^2)$ *error*
$\forall t, \nabla \ln p_t$ *L-Lipschitz*	$\mathrm{KL}(p_0 \| \hat{q}_T)$	$O\left(\frac{dL^2}{\varepsilon_{sc}^2}\right)$
$\nabla \ln p_0$ *L-Lipschitz*	$\mathrm{KL}(p_0 \| \hat{q}_T)$	$O\left(\frac{d^2(\ln L)^2}{\varepsilon_{sc}^2}\right)$
None	$\mathrm{KL}(p_\delta \| \hat{q}_{T-\delta})$	$O\left(\frac{d^2\ln(1/\delta)^2}{\varepsilon_{sc}^2}\right)$

Sampling is as easy as learning the score function. (S. Chen, Chewi, J. Li, et al. 2023)

DDPM with estimated score: smooth distributions

Assumption

1. p_0 has second moment $\mathbb{E}_{p_0} \|x\|^2 = M_2$.
2. The score estimate s has *average* error

$$\frac{1}{T}\sum_{k=1}^{T} \|\nabla \ln p_{t_k} - s(\cdot, t_k)\|^2_{L^2(p_{t_k})} \le \varepsilon_{sc}^2.$$

3. $\nabla \ln p_t$ is L-Lipschitz for every t.

DDPM with estimated score: smooth distributions

Theorem (S. Chen, Chewi, J. Li, et al. 2023; H. Chen, L, and Lu 2023)

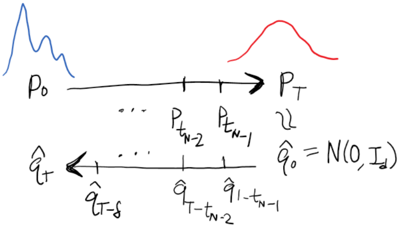
Under these assumptions, the error of DDPM with exponential integrator and N discretization steps satisfies (for $T = \Omega(1)$)

$$\mathrm{KL}(p_0 \| \widehat{q}_T) \lesssim (M_2 + d)e^{-T} + T\varepsilon_{\mathrm{sc}}^2 + \frac{T^2 L^2 d}{N}.$$

Choosing $T = \ln\left(\frac{M_2+d}{\varepsilon_{\mathrm{sc}}^2}\right)$ and $N = \Theta\left(\frac{dT^2L^2}{\varepsilon_{\mathrm{sc}}^2}\right)$ makes this $\widetilde{O}(\varepsilon_{\mathrm{sc}}^2)$.

Terms quantify

1. convergence of forward process,
2. score estimation error, and
3. discretization error.



Term #1: Convergence of forward process

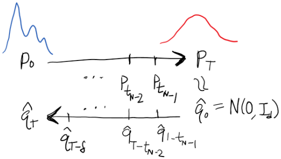
By chain rule for KL divergence,

$$\mathrm{KL}(p_0\|\widehat{q}_T) \le \mathrm{KL}(p_T\| \underbrace{\widehat{q}_0}_{N(0,I_d)}) + \mathbb{E}_{p_T(a)}\,\mathrm{KL}(p_{0|T}(\cdot|a)\|\widehat{q}_{T|0}(\cdot|a)).$$

$\mathrm{KL}(p_T\|\widehat{q}_0)$ is bounded by convergence of the forward process:

$$\mathrm{KL}(p_T\|\widehat{q}_0) \lesssim \underbrace{(d+M_2)}_{(1)}\underbrace{e^{-T}}_{(2)}$$

1. The KL divergence after $\Theta(1)$ time is $O(d+M_2)$.
2. Exponential mixing of the forward (Ornstein-Uhlenbeck) process.

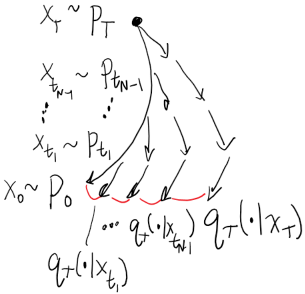


Term #2: score estimation error

$$\mathrm{KL}(p_0 \| \widehat{q}_T) \le \mathrm{KL}(p_T \| \underbrace{\widehat{q}_0}_{N(0,I_d)}) + \mathbb{E}_{p_T(a)} \mathrm{KL}(p_{0|T}(\cdot|a) \| \widehat{q}_{T|0}(\cdot|a)).$$

By chain rule for KL divergence and Girsanov*,

$$\begin{aligned}
&\mathbb{E}_{p_T(a)} \mathrm{KL}(p_{0|T}(\cdot|a) \| \widehat{q}_{T|0}(\cdot|a)) \\
&= \sum_{k=1}^{N} \mathbb{E}_{p_{t_k}(a)} \mathrm{KL}(p_{t_{k-1}|t_k}(\cdot|a) \| \widehat{q}_{T-t_{k-1}|T-t_k}(\cdot|a)) \\
&\le \sum_{k=1}^{N} \frac{1}{2} \int_{t_{k-1}}^{t_k} \mathbb{E}_{x_t \sim p_t} \left\| s(x_{t_k}, t_k) - \nabla \ln p_t(x_t) \right\|^2 \, dt \\
&\le \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}_{x_t \sim p_t} \left\| s(x_{t_k}, t_k) - \nabla \ln p_{t_k}(x_{t_k}) \right\|^2 \\
&\quad + \mathbb{E} \left\| \nabla \ln p_{t_k}(x_{t_k}) - \nabla \ln p_t(x_t) \right\|^2 \, dt.
\end{aligned}$$



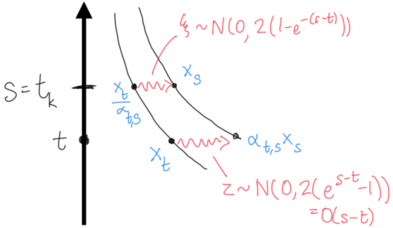
$$\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}_{x_t \sim p_t} \left\| s(x_{t_k}, t_k) - \nabla \ln p_{t_k}(x_{t_k}) \right\|^2 dt \le T \varepsilon_{\mathrm{sc}}^2$$

(Key) Term #3: Discretization error

$$\text{Need to bound} \quad \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E} \left\| \nabla \ln p_{t_k}(x_{t_k}) - \nabla \ln p_t(x_t) \right\|^2 \, dt.$$

Let $\alpha = \alpha_{t,s} = e^{s-t}$, $s = t_k$. Split up

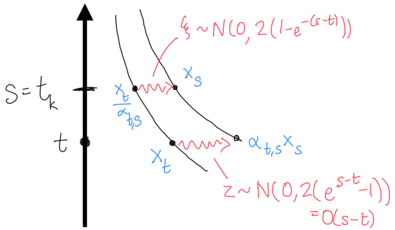
$$\begin{aligned} &\mathbb{E} \left\| \nabla \ln p_s(x_s) - \nabla \ln p_t(x_t) \right\|^2 \\ &\lesssim \mathbb{E} \left\| \nabla \ln p_s(x_s) - \nabla \ln p_t(\alpha_{s,t} x_s) \right\|^2 && \text{(time)} \\ &\quad + \mathbb{E} \left\| \nabla \ln p_t(\alpha_{s,t} x_s) - \nabla \ln p_t(x_t) \right\|^2 && \text{(space)} \end{aligned}$$



- It turns out to be sufficient to bound space discretization error $\mathbb{E} \left\| \nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha_{t,s} x_s) \right\|^2$.
- $\alpha_{t,s} x_s = x_t + z$, where z is Gaussian of variance $O(s - t)$.

Bounding the space discretization error with smoothness

Goal: Bound $\mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2$, where $\alpha = e^{s-t}$.



- Note $\alpha x_s = x_t + z$, z Gaussian of variance $O(s-t)$. By **Lipschitzness**,

$$\mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2 \lesssim L^2 \mathbb{E}\left\|z\right\|^2 \lesssim dL^2(s-t).$$

- Hence

$$\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \ln p_{t_k}(x_{t_k}) - \nabla \ln p_t(x_t)\right\|^2 \, dt \lesssim TdL^2 h = \frac{T^2 L^2 d}{N}.$$

DDPM w/ estimated score: no smoothness, early stopping

Assumption

1. p_0 has second moment $\mathbb{E}_{p_0} \|x\|^2 = \mathfrak{m}_2^2$.
2. The score estimate s has *weighted average* error (h_k are step sizes)

$$\frac{1}{T-\delta} \sum_{k=1}^{T} h_k \left\| \nabla \ln p_{t_k} - s_{t_k} \right\|_{L^2(p_{t_k})}^2 \le \varepsilon_{\mathrm{sc}}^2.$$

Note: If

$$\left\| \nabla \ln p_{t_k} - s_{t_k} \right\|_{L^2(p_{t_k})}^2 \le \frac{\varepsilon^2}{\min\{t_k, 1\}},$$

then (2) is satisfied with a log factor:

$$\frac{1}{T-\delta} \sum_{k=1}^{T} h_k \left\| \nabla \ln p_{t_k} - s_{t_k} \right\|_{L^2(p_{t_k})}^2 \lesssim \frac{1}{T-\delta} \int_\delta^T \frac{\varepsilon^2}{t \wedge 1}\, dt \lesssim \varepsilon^2 \ln\left(\frac{1}{\delta}\right).$$

DDPM w/ estimated score: no smoothness, early stopping

Theorem (H. Chen, L, and Lu 2023)

Given these assumptions, the error of DDPM with exponential integrator and N (exponentially decaying) discretization steps satisfies

$$\mathrm{KL}(p_\delta \| \widehat{q}_{T-\delta}) \lesssim (\mathfrak{m}_2^2 + d)e^{-2T} + T\varepsilon_{\mathrm{sc}}^2 + \frac{\left(\ln\left(\frac{1}{\delta}\right) + T\right)^2 d^2}{N}.$$

Choosing $T = \ln\left(\frac{\mathfrak{m}_2^2 + d}{\varepsilon_{\mathrm{sc}}^2}\right)$, $N = \Theta\left(\frac{\left(\ln\left(\frac{1}{\delta}\right)+T\right)^2 d^2}{\varepsilon_{\mathrm{sc}}^2}\right)$ *makes this* $\widetilde{O}(\varepsilon_{\mathrm{sc}}^2)$.

Score error assumption:

$$\frac{1}{T-\delta}\sum_{k=1}^{T} h_k \left\|\nabla \ln p_{t_k} - s_{t_k}\right\|_{L^2(p_{t_k})}^2 \le \varepsilon_{\mathrm{sc}}^2.$$

DDPM w/ estimated score: no smoothness, early stopping

Theorem (H. Chen, L, and Lu 2023)

Given these assumptions, the error of DDPM with exponential integrator and N (exponentially decaying) discretization steps satisfies

$$\mathrm{KL}(p_\delta \| \widehat{q}_{T-\delta}) \lesssim (\mathfrak{m}_2^2 + d)e^{-2T} + T\varepsilon_{\mathrm{sc}}^2 + \frac{\left(\ln\left(\frac{1}{\delta}\right) + T\right)^2 d^{2}}{N}.$$

Choosing $T = \ln\left(\frac{\mathfrak{m}_2^2 + d}{\varepsilon_{\mathrm{sc}}^2}\right)$, $N = \Theta\left(\frac{\left(\ln\left(\frac{1}{\delta}\right)+T\right)^2 d^{2}}{\varepsilon_{\mathrm{sc}}^2}\right)$ *makes this* $\widetilde{O}(\varepsilon_{\mathrm{sc}}^2)$.

Corollary (Pure Wasserstein guarantee)

For $\delta = \Theta\left(\frac{\varepsilon^2}{d}\right)$, $N = \widetilde{\Theta}\left(\frac{d^2 R^4}{\varepsilon^4}\right)$ *(R the "high-probability" radius of* p_0*), the rescaled & truncated output satisfies*

$$W_2(p_0, \widehat{q}_{T-\delta}^{\mathrm{trunc}}) = \widetilde{O}(\varepsilon).$$

Non-smooth setting: Bounding the space discretization error

Goal

Bound $\varepsilon_{\text{space}} = \mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2$.

$$\mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2 \lesssim L^2 \mathbb{E}\left\|z\right\|^2 \lesssim dL^2(s-t).$$

How to bound without smoothness assumption on p_t?

Non-smooth setting: Bounding the space discretization error

Goal

Bound $\varepsilon_{\text{space}} = \mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2$.

$$\mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2 \lesssim L^2 \mathbb{E}\left\|z\right\|^2 \lesssim dL^2(s-t).$$

How to bound without smoothness assumption on p_t?

- **Previous approach (S. Chen, Chewi, J. Li, et al. 2023):** Use global Lipschitzness of $\nabla \ln p_t$. If p_0 is supported on ball of radius $R \geq 1$ and $t \leq 1$, then $\left\|\nabla^2 \ln p_t\right\| = O\left(\frac{R^2}{t^2}\right)$.

$$L \approx \frac{R^2}{t^2} \qquad \text{disc. error} : \frac{T^2 L^2 d}{N}$$

Non-smooth setting: Bounding the space discretization error

Goal

Bound $\varepsilon_{\mathrm{space}} = \mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2$.

$$\mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2 \lesssim L^2 \mathbb{E}\left\|z\right\|^2 \lesssim dL^2(s-t).$$

How to bound without smoothness assumption on p_t?

- **New approach:** Note that $x_t = \alpha x_s + z$, where $x_s \sim p_s$, z z Gaussian of variance $O(s-t)$. Suffices to bound Hessian $\left\|\left\|\nabla^2 \ln p_t(x)\right\|_F\right\|_{\psi_1} \lesssim \frac{d}{\min\{t,1\}}$
  - at a random point
  - in a random direction, i.e., in Frobenius norm.
  - To deal with diverging bound, take **geometrically decreasing** step size. Stopping at $t_0 = \delta$, integrating gives $\int_\delta^T \frac{d}{t \wedge 1}\,dt = d \cdot \left(\ln\left(\frac{1}{\delta}\right) + T\right)$.

$$L \approx \frac{R^2}{t^2} \rightsquigarrow \frac{\sqrt{d}}{t} \qquad \text{disc. error} : \frac{T^2L^2d}{N} \rightsquigarrow \frac{\left(\ln\left(\frac{1}{\delta}\right) + T\right)^2 d^2}{N}$$

Non-smooth setting: Bounding the space discretization error

Goal

Bound $\varepsilon_{\text{space}} = \mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2$.

$$\mathbb{E}\left\|\nabla \ln p_t(x_t) - \nabla \ln p_t(\alpha x_s)\right\|^2 \lesssim L^2 \mathbb{E}\|z\|^2 \lesssim dL^2(s-t).$$

How to bound without smoothness assumption on p_t?

- **Subsequent work** (Benton, De Bortoli, Doucet, et al. 2023): Bound $\mathbb{E}\left\|\nabla^2 \ln p_t\right\|_F^2$ by deriving ODE for the expected value using **stochastic localization**.

$$L \approx \frac{R^2}{t^2} \rightsquigarrow \frac{\sqrt{d}}{t} \rightsquigarrow \frac{1}{t} \qquad \text{disc. error}: \frac{T^2 L^2 d}{N} \rightsquigarrow \frac{\left(\ln\left(\frac{1}{\delta}\right) + T\right)^2 d^2}{N} \rightsquigarrow \frac{\left(\ln\left(\frac{1}{\delta}\right) + T\right)^2 d}{N}$$

SDE vs. ODE formulation

Denoising Diffusion Probabilistic Modeling (SDE)	Probability Flow (ODE)
$dx_t^{\to} = -x_t^{\to}\,dt + \sqrt{2}\,dW_t$ $dx_t^{\leftarrow} = x_t^{\leftarrow}\,dt + 2\underbrace{\nabla \ln p_{T-t}(x_t^{\leftarrow})}_{\approx s_{T-t}(x_t^{\leftarrow})}\,dt + \sqrt{2}\,dW_t.$	$dx_t^{\to} = -x_t^{\to}\,dt - \nabla \ln p_t(x_t^{\to})\,dt$ $dx_t^{\leftarrow} = x_t^{\leftarrow}\,dt + \underbrace{\nabla \ln p_{T-t}(x_t^{\leftarrow})}_{\approx s_{T-t}(x_t^{\leftarrow})}\,dt.$
• Convergence guarantees with $O(d)$ steps. (S. Chen, Chewi, J. Li, et al. 2023; H. Chen, **L**, and Lu 2023; Benton, De Bortoli, Doucet, et al. 2023) • Lower bound $\Omega(d)$ for trajectory-wise analysis, even for critically damped Langevin diffusion (S. Chen, Chewi, J. Li, et al. 2023).	• Much faster (10x–50x) in practice (J. Song, Meng, and Ermon 2020)... • ...but can sometimes be less stable. • **This work:** $O(\sqrt{d})$ steps using corrector steps, assuming smoothness.

The trouble with SDE's

DDPM:

$$dx_t^{\leftarrow} = [x_t^{\leftarrow} + 2\nabla \ln p_{T-t}(x_t^{\leftarrow})]\, dt + \sqrt{2}\, dw_t$$

$$x_{t+h}^{\leftarrow} \approx x_t^{\leftarrow} + h\,[x_t^{\leftarrow} + 2\nabla \ln p_{T-t}(x_t^{\leftarrow})] + \sqrt{2h}\,\xi, \ \xi \sim N(0, I_d).$$

Discretization error from...

- Drift term (order 1): $O(Lh\sqrt{d}) \to$ can take $h = O\left(\frac{1}{L\sqrt{d}}\right)$.
- Diffusion term (order 1/2): $O(L\sqrt{hd}) \to$ need to take $h = O\left(\frac{1}{L^2 d}\right)$.
  Trajectories of Brownian motion are not smooth!

Probability flow ODE:

$$dx_t^{\leftarrow} = [x_t^{\leftarrow} + \nabla \log p_{T-t}(x_t^{\leftarrow})]\, dt.$$

Assumptions

Assumption

1. p_0 has second moment $\mathbb{E}_{p_0}\|x\|^2 = \mathfrak{m}_2^2$.
2. For each t_k, the score estimate s has error

$$\|\nabla \ln p_{t_k} - s_{t_k}\|^2_{L^2(p_{t_k})} \le \varepsilon_{sc}^2.$$

3. $\nabla \ln p_t$ is L-Lipschitz for every t.
4. The score estimate s_{t_k} is L-Lipschitz for every t_k.

DPUM (Diffusion Predictor + Underdamped Modeling)

Theorem (DPUM, S. Chen, Chewi, **L**, Li, Lu, and Salim 2023)

Suppose that Assumptions hold. If $\widehat{q}$ denotes output of DPUM with $\delta \asymp \frac{\varepsilon^2}{L^2(d+m_2^2)}$, then

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim \underbrace{(\sqrt{d} + m_2^2)e^{-T}}_{(1)} + \underbrace{L^{1/2} T \varepsilon_{\mathrm{sc}}}_{(2)} + \underbrace{L^2 T d^{1/2} h_{\mathrm{pred}}}_{(3a)} + \underbrace{L^{3/2} T d^{1/2} h_{\mathrm{corr}}}_{(3b)} + \underbrace{\varepsilon}_{(4)} .$$

Setting $T = \Theta(\ln(\frac{d+m_2^2}{\varepsilon^2}))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\varepsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\varepsilon}{L^{3/2} d^{1/2}})$, if $\varepsilon_{\mathrm{sc}} \le \widetilde{O}(\frac{\varepsilon}{\sqrt{L}})$, then we obtain TV error $O(\varepsilon)$ with number of steps

$$N = \widetilde{\Theta}\left(\frac{L^2 d^{1/2}}{\varepsilon}\right).$$

DPUM (Diffusion Predictor + Underdamped Modeling)

Theorem (DPUM, S. Chen, Chewi, L, Li, Lu, and Salim 2023)

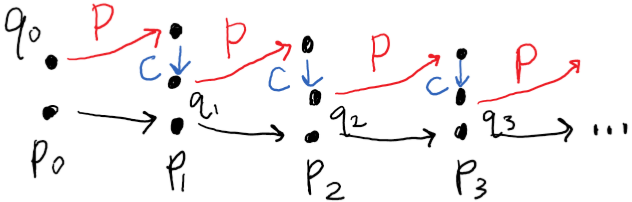
Suppose that Assumptions hold. If $\widehat{q}$ denotes output of DPUM with $\delta \asymp \frac{\varepsilon^2}{L^2(d+m_2^2)}$, then

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim \underbrace{(\sqrt{d} + m_2^2)e^{-T}}_{(1)} + \underbrace{L^{1/2} T \varepsilon_{sc}}_{(2)} + \underbrace{L^2 T d^{1/2} h_{pred}}_{(3a)} + \underbrace{L^{3/2} T d^{1/2} h_{corr}}_{(3b)} + \underbrace{\varepsilon}_{(4)} .$$

1. Convergence of forward process
2. Score estimation error
3. Discretization error (predictor/corrector)
4. Early stopping

Predictors and correctors (Y. Song, Sohl-Dickstein, Kingma, et al. 2020)

- **Predictor (P):** Simulate the reverse SDE/ODE to track a *time-varying* distribution.
- **Corrector (C):** Run MCMC (e.g., Langevin Monte Carlo) to converge towards a *stationary* distribution.
- **Predictor-corrector (PC):** Intersperse P & C steps.



- **Predictor (P):** Simulate the reverse SDE/ODE to track a *time-varying* distribution.
- **Corrector (C):** Run MCMC (e.g., Langevin Monte Carlo) to converge towards a *stationary* distribution.
- **Predictor-corrector (PC):** Intersperse P & C steps.

	Variance Exploding SDE (SMLD)				Variance Preserving SDE (DDPM)			
FID↓ Sampler / Predictor	P1000	P2000	C2000	PC1000	P1000	P2000	C2000	PC1000
ancestral sampling	4.98 ± .06	4.88 ± .06		**3.62 ± .03**	3.24 ± .02	3.24 ± .02		**3.21 ± .02**
reverse diffusion	4.79 ± .07	4.74 ± .08	20.43 ± .07	**3.60 ± .02**	3.21 ± .02	3.19 ± .02	19.06 ± .06	**3.18 ± .01**
probability flow	15.41 ± .15	10.54 ± .08		**3.51 ± .04**	3.59 ± .04	3.23 ± .03		**3.06 ± .03**

DPUM (Diffusion Predictor + Underdamped Modeling)

Algorithm

- Draw $\widehat{x}_0 \sim N(0, I_d)$.
- For $n = 0, \ldots, LT - 1$:
  - **Predictor:** Starting from $\widehat{x}_{n/L}$, run the discretized probability flow ODE from time $\frac{n}{L}$ to $\frac{n+1}{L}$ with step size h_{pred} to obtain $\widehat{x}'_{\frac{n+1}{L}}$.

$$x^{\leftarrow}_{t+h} = e^h x^{\leftarrow}_t + (e^h - 1) s_{T-t}(x^{\leftarrow}_t).$$

  - **Corrector:** Starting from $\widehat{x}'_{\frac{n+1}{L}}$, run underdamped LMC for time $\frac{1}{\sqrt{L}}$ with step size h_{corr} to obtain $\widehat{x}_{\frac{n+1}{L}}$.
- Return $\widehat{x}_T$.

Note: For technical reasons, we need to modify the above algorithm to use geometrically decreasing step sizes in the last stage and employ early stopping.

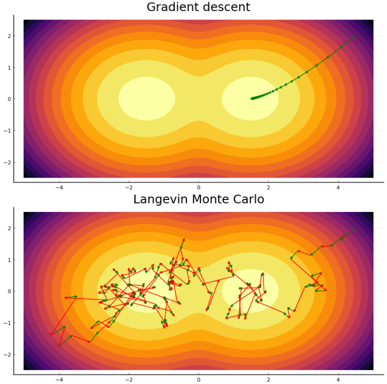
Challenges

Problem: Cannot use Girsanov's Theorem with ODE's.
Solution: Use **Wasserstein analysis** with coupling.

Problem: Distance grows exponentially with rate L; can only run for time $O(1/L)$.
Solution: Convert Wasserstein to TV error with a **corrector** step (short-time regularization).
Using data processing inequality for TV distance, we can restart coupling.

Corrector: Langevin dynamics

SDE-based method to sample from $p(x) \propto e^{-f(x)}$:



- Overdamped:

$$dx_t = -\nabla f(x_t)\,dt + \sqrt{2}\,dB_t$$

- Underdamped:

$$dx_t = v_t\,dt$$
$$dv_t = -\nabla f(x_t)\,dt - \gamma v_t\,dt + \sqrt{2\gamma}\,dB_t$$

Problem: Overdamped Langevin needs $O(d)$ steps.
Solution: Use **underdamped Langevin** (Langevin "with acceleration"), which needs $O(\sqrt{d})$ steps.

Summary of convergence bounds

Algorithm	Assumptions		to get error ε_0	
	Lipschitzness	$\varepsilon_{sc} \le ?$	Error guarantee	Steps
DDPM	$\forall t, \nabla \ln p_t$	$\widetilde{O}(\varepsilon_0)$	$\sqrt{\mathrm{KL}(p_0 \| \hat{q}_T)}$	$O\left(\frac{dL^2}{\varepsilon_0^2}\right)$
DDPM*	$\nabla \ln p_0$	$\widetilde{O}(\varepsilon_0)$	$\sqrt{\mathrm{KL}(p_0 \| \hat{q}_T)}$	$O\left(\frac{d^2(\ln L)^2}{\varepsilon_0^2}\right)$
DDPM*	None	$\widetilde{O}(\varepsilon_0)$	$\sqrt{\mathrm{KL}(p_\delta \| \hat{q}_{T-\delta})}$	$O\left(\frac{d^2 \ln(1/\delta)^2}{\varepsilon_0^2}\right)$
DDPM♡	None	$\widetilde{O}(\varepsilon_0)$	$\sqrt{\mathrm{KL}(p_\delta \| \hat{q}_{T-\delta})}$	$O\left(\frac{d \ln(1/\delta)^2}{\varepsilon_0^2}\right)$
PF♢	Jacobian error	$\widetilde{O}(\varepsilon_0/\sqrt{d})$	$\mathrm{TV}(p_0, \hat{q}_T)$	$O\left(\frac{d^2}{\varepsilon_0}\right)$
DPUM†	$\forall t, \nabla \ln p_t$ & s_t	$\widetilde{O}(\varepsilon_0/\sqrt{L})$	$\mathrm{TV}(p_0, \hat{q}_T)$	$O\left(\frac{d^{1/2}L^2}{\varepsilon_0}\right)$

∗: H. Chen, **L**, and Lu 2023.

†: **Probability flow + underdamped corrector**, S. Chen, Chewi, **L**, Li, Lu, and Salim 2023.

♡: Benton, De Bortoli, Doucet, and Deligiannidis 2023

♢: G. Li, Wei, Y. Chen, and Chi 2023

We've reduced the problem to learning the score, now what?

Sampling is as easy as learning the score function

We can efficiently sample from the data distribution if we have $L^2(p)$-accurate score estimates at the different noise levels.

Question 2

When can we obtain a $L^2(p)$-accurate score estimate?

Can we come up with *any* nontrivial problem where diffusion models provably learn a distribution better than (or as well as) other known methods?

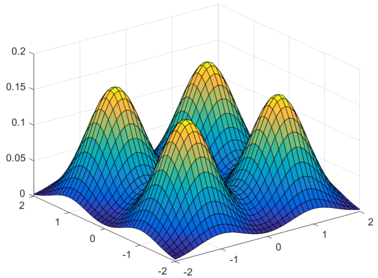
Gaussian mixture

Problem

Learn a mixture of gaussians from samples:

$$X \sim \sum_{i=1}^{k} p_i \mathcal{N}(\mu_i, I_n), \quad \text{i.e.,} \quad p(x) \propto \sum_{i=1}^{k} p_i \exp\left(-\frac{\|x-\mu_i\|^2}{2}\right).$$

- Efficient algorithms rely on parameter learning, which fail without $\Omega(\sqrt{\ln k})$ separation (Regev and Vijayaraghavan 2017).
- Diakonikolas and Kane 2020: Algorithm based on algebraic methods with time/sample complexity $\operatorname{poly}\left(n, k, \frac{1}{\varepsilon}\right) + \left(\frac{k}{\varepsilon}\right)^{O(\ln^2 k)}$.
- We show that diffusion models can also learn with quasi-polynomial time and samples, giving a *completely different* approach to this problem!



Gaussian mixture

Problem

Learn a mixture of gaussians from samples:

$$X \sim \sum_{i=1}^{k} p_i \mathcal{N}(\mu_i, I_n), \quad \text{i.e.,} \quad p(x) \propto \sum_{i=1}^{k} p_i \exp\left(-\frac{\|x-\mu_i\|^2}{2}\right).$$

- **Observation**: Score function is exactly a softmax neural network with 1 hidden layer (and skip-connection).

$$\nabla \ln p(x) = \frac{\sum_{i=1}^{k} \mu_i \exp\left(\langle x, \mu_i\rangle\right)}{\sum_{i=1}^{k} \exp\left(\langle x, \mu_i\rangle\right)} - x = \mathbb{E}[\mu | x] - x. \quad (1)$$

- Shah, S. Chen, and Klivans 2023: Gradient descent with diffusion models learns a mixture of 2 gaussians, or K gaussians with separation (with warm start)—does *as well as* EM.
- Gatmiry, Kelner, and **L** 2024: (1) can be learned with quasi-polynomial complexity without separation conditions!

Learning gaussian mixture with diffusion model

Problem

Learn $P_0 = Q_0 * \mathcal{N}(0, I_n)$ from samples where Q_0 is made up of k clusters:

- The support of Q_0 can be covered with k balls of radius $O(1)$,
- each with probability $\geq \alpha_{\min}$ under Q_0.

Theorem (Gatmiry, Kelner, and **L** 2024)

For $\varepsilon < \alpha_{\min}$, diffusion models can learn a distribution that is ε-close in TV distance to P_0 with time and sample complexity $n^{\text{poly}\log(n,k,\frac{1}{\varepsilon})}$.

Learning gaussian mixture with diffusion model

Problem

Learn $P_0 = Q_0 * \mathcal{N}(0, I_n)$ from samples where Q_0 is made up of k clusters:

- The support of Q_0 can be covered with k balls of radius $O(1)$,
- each with probability $\geq \alpha_{\min}$ under Q_0.

Theorem (Gatmiry, Kelner, and **L** 2024)

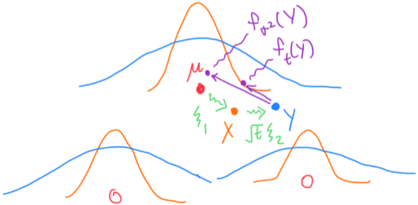
For $\varepsilon < \alpha_{\min}$, diffusion models can learn a distribution that is ε-close in TV distance to P_0 with time and sample complexity $n^{\text{poly}\log(n,k,\frac{1}{\varepsilon})}$.

Corollary (Manifold learning)

Suppose that Q_0 is supported on a set M that can be covered with C^d balls of constant radius, each with Q_0-probability $\geq \frac{1}{C^d}$. Then diffusion models can learn a distribution ε-close in TV distance to P_0 with time and sample complexity $n^{\text{poly}(d,\ln n,\ln C,\ln(\frac{1}{\varepsilon}))}$

Score function of gaussian mixture

$$\mu \sim Q_0, \quad \xi_1, \xi_2 \sim \mathcal{N}(0, I_n),$$
$$X = \mu + \xi_1,$$
$$Y = X + \sqrt{t}\xi_2 = \mu + \xi_1 + \sqrt{t}\xi_2.$$



Consider the case of 1 cluster ($\text{Supp}(Q_0) \subseteq B_R(0)$).
It suffices to learn (as a **supervised problem**)

$$f_{\sigma^2}(y) := y + \sigma^2 \nabla \ln p_t(y) = \mathbb{E}[\mu | Y = y] = \frac{1}{\sigma^2}\left(-y + \frac{\int_{\mathbb{R}^n} \mu \exp\left(\frac{\langle y, \mu\rangle}{\sigma^2} - \frac{\|\mu\|^2}{2\sigma^2}\right) dQ_0(\mu)}{\int_{\mathbb{R}^n} \exp\left(\frac{\langle y, \mu\rangle}{\sigma^2} - \frac{\|\mu\|^2}{2\sigma^2}\right) dQ_0(\mu)}\right)$$

where $t = 1 + \sigma^2$.

Key technique: Noise stability

Smooth functions on $\mathcal{N}(0, \sigma^2 I_n)$ can be efficiently learned via low-degree polynomials.
Measure smoothness using the generator of the Ornstein-Uhlenbeck process:

$$\mathscr{L}f(x) = -\frac{1}{\sigma^2}\langle x, \nabla f(x)\rangle + \Delta f(x).$$

Theorem (Noise stability implies low-degree approximability)

Suppose that $\|\mathscr{L}f\|_{L^2(\mathcal{N}(0,\sigma^2 I_n))} \le L$. Then there exists a polynomial g of degree $< d$ such that

$$\|f - g\|_{L^2(\mathcal{N}(0,\sigma^2 I_n))} \le \frac{L\sigma^2}{d}.$$

Problem: Requires degree $\frac{1}{\varepsilon}$ degree to get ε accuracy.

Key technique: Noise stability

Smooth functions on $\mathcal{N}(0, \sigma^2 I_n)$ can be efficiently learned via low-degree polynomials.
Measure smoothness using the generator of the Ornstein-Uhlenbeck process:

$$\mathscr{L} f(x) = -\frac{1}{\sigma^2} \langle x, \nabla f(x) \rangle + \Delta f(x).$$

Theorem (Noise stability implies low-degree approximability)

Suppose that $\|\mathscr{L}^m f\|_{L^2(\mathcal{N}(0,\sigma^2 I_n))} \le L^m$. *Then there exists a polynomial g of degree $< d$ such that*

$$\|f - g\|_{L^2(\mathcal{N}(0,\sigma^2 I_n))} \le \left(\frac{L\sigma^2}{d}\right)^m.$$

Problem: Requires degree $\frac{1}{\varepsilon}$ degree to get ε accuracy.
Solution: Iterate $\mathscr{L}$! (Take $m = \ln\left(\frac{1}{\varepsilon}\right)$.)

Calculation

Taking derivatives of $\nabla \ln p_t$ gives moments under the posterior distribution $\langle\cdot\rangle := \mathbb{E}_{X|Y}[\cdot]$.

Lemma

Let $f(y) = y + \nabla \ln p(y)$. We have

$$\mathscr{L}^d f(y) = \left\langle x^{(1)} \sum_{s+t\le d} \sum_{i,i'\in[2d+1]^s, j\in[2d+1]^t} a_{i,i',j}\sigma^{-2(s+t+d)} \prod_{\ell=1}^{s} \left\langle x^{(i_\ell)}, x^{(i'_\ell)}\right\rangle \prod_{m=1}^{t} \left\langle x^{(j_m)}, y\right\rangle \right\rangle$$

where $\sum_{i,i',j} |a_{i,i',j}| \le 30^d d!^2$, and $x^{(i)}$ are independent draws from the posterior $X|Y = y$.

Lemma

Suppose Q_0 is supported on $B_R(0)$, $R \ge 1$. Then

$$\|\mathscr{L}^m f\|_{L^2(P_0)} = R \cdot O\left(m^2 R^2 \left(1 + \frac{m}{R}\right)\right)^m .$$

The rest of the proof

Goal: Learn $P_0 = Q_0 * \mathcal{N}(0, I_n)$ from samples where Q_0 is made up of k clusters.
Problem: Multiple clusters.

- Do piecewise polynomial regression on Voronoi cells around warm starts.
- Inductively maintain warm starts (from high to low noise level) by using score estimates:

$$f_{\sigma^2}(y) = y + \sigma^2 \nabla \ln p_t(y) = \mathbb{E}[\mu | Y = y] = \mu + O(\sigma \sqrt{\ln(1/\alpha_{\min})}).$$

Problem: Change of measure between $\mathcal{N}(\hat{\mu}_i, \sigma^2 I_n)$ and $P_0|_{V_i}$, where $\hat{\mu}_i$ is center of Voronoi cell V_i.

- Surprisingly delicate: Need to take $d = \Omega(\ln^2\left(\frac{1}{\varepsilon}\right))$-degree polynomial to get ε error. Polynomials grow quickly: Naive change of measure gives $\varepsilon \cdot \Omega(1)^{\sqrt{d}} \gg 1$.
- First compare f to smoothed version of f, so Hermite coefficients now decay exponentially.

Conclusion

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

- [H. Chen, **L**, and Lu 2023]. Smoothing properties of the forward process lead to good convergence rates for arbitrary data distributions.
- [S. Chen, Chewi, **L**, Li, Lu, and Salim 2023] Using an ODE instead of SDE, in conjunction with a corrector step, can reduce dimension dependence from $O(d)$ to $O(\sqrt{d})$.

Question 2

When can we obtain a $L^2(p)$-accurate score estimate?

- [Gatmiry, Kelner, and **L** 2024] The score function can be efficiently learned for mixtures of (identity-covariance) gaussians—including when the mixing measure is close to a low-dimensional manifold—giving an end-to-end learning result.

Thanks to collaborators: Hongrui Chen, Sitan Chen, Sinho Chewi, Khashayar Gatmiry, Jonathan Kelner, Yuanzhi Li, Jianfeng Lu, Adil Salim, Yixin Tan.

Open directions

Question 1

What guarantees can we obtain for sampling from p given a $L^2(p)$-accurate score estimate?

- Use insights from numerical analysis, geometry of probability distributions, structure of distributions (low-dimensionality, Fourier/multiscale, coming from function space...), etc.
- One/few-step generation: progressive distillation, consistency models, ...

Question 2

When can we obtain a $L^2(p)$-accurate score estimate?

- Other families of distributions that allow efficient learning?
- Tradeoffs compared to other generative models? (Sometimes makes problems intractable! Ghio, Dandi, Krzakala, et al. 2023)
- Efficient learning with neural networks?

"Learning" beyond $L^2(p)$-accurate score and ε-closeness in distributional distance?

Bibliography I

- Benton, Joe et al. (2023). "Linear convergence bounds for diffusion models via stochastic localization". In: *arXiv preprint arXiv:2308.03686*.
- Chen, Hongrui, Holden **L**, and Jianfeng Lu (2023). "Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions". In: *arXiv preprint arXiv:2211.01916*.
- Chen, Sitan et al. (2023). "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *arXiv preprint arXiv:2209.11215*.
- Chen, Sitan et al. (2023). *The probability flow ODE is provably fast*. arXiv: 2305.11798 [cs.LG].
- Corso, Gabriele et al. (2022). "DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking". In: *arXiv preprint arXiv:2210.01776*.
- Diakonikolas, Ilias and Daniel M Kane (2020). "Small covers for near-zero sets of polynomials and learning latent variable models". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 184–195.

Bibliography II

Gatmiry, Khashayar, Jonathan Kelner, and Holden **L** (2024). "Learning Mixtures of Gaussians with Diffusion Models". In: *In progress.*

Ghio, Davide et al. (2023). "Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective". In: *arXiv preprint arXiv:2308.14085.*

Guillin, Arnaud and Feng-Yu Wang (2012). "Degenerate Fokker–Planck equations: Bismut formula, gradient estimate and Harnack inequality". In: *Journal of Differential Equations* 253.1, pp. 20–40.

Hyvärinen, Aapo (2005). "Estimation of non-normalized statistical models by score matching.". In: *Journal of Machine Learning Research* 6.4.

L, Holden, Jianfeng Lu, and Yixin Tan (2022). "Convergence for score-based generative modeling with polynomial complexity". In: *Advances in neural information processing systems* 35.

Li, Gen et al. (2023). "Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models". In: *arXiv preprint arXiv:2306.09251.*

Bibliography III

Montanari, Andrea (2023). "Sampling, Diffusions, and Stochastic Localization". In: *arXiv preprint arXiv:2305.10690*.

Ramesh, Aditya et al. (2022). "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125*.

Regev, Oded and Aravindan Vijayaraghavan (2017). "On learning mixtures of well-separated gaussians". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 85–96.

Shah, Kulin, Sitan Chen, and Adam Klivans (2023). "Learning Mixtures of Gaussians Using the DDPM Objective". In: *arXiv preprint arXiv:2307.01178*.

Sohl-Dickstein, Jascha et al. (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265.

Song, Jiaming, Chenlin Meng, and Stefano Ermon (2020). "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502*.

Bibliography IV

Song, Yang and Stefano Ermon (2019). "Generative Modeling by Estimating Gradients of the Data Distribution". In: *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*.

Song, Yang et al. (2020). "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.

Vincent, Pascal (2011). "A connection between score matching and denoising autoencoders". In: *Neural computation* 23.7, pp. 1661–1674.

Proof outline for Probability flow ODE

1. Predictor analysis
   (a) Score perturbation lemma
2. Corrector analysis
   (a) Short-time regularization
3. Combining bounds for predictor and corrector

1a. Score perturbation lemma

Lemma (Score perturbation)

Suppose $q_t^{\rightarrow}$ is the density of the OU process at time t, started at $q_0^{\rightarrow}$, and y_t follows the probability flow ODE. Suppose for all t and all x that $\left\|\nabla^2 \ln q_t^{\rightarrow}(x)\right\|_{\mathrm{op}} \le L$, where $L \ge 1$. Then,

$$\mathbb{E}[\|\partial_t \nabla \ln q_t^{\rightarrow}(y_t)\|^2] \lesssim L^2 d \left(L + \frac{1}{t}\right).$$

Previous work (**L**, Lu, and Tan 2022) only gave a $\frac{1}{2}$-Hölder continuity bound.

1a. Score perturbation lemma

Lemma (Score perturbation)

Suppose $q_t^{\rightarrow}$ is the density of the OU process at time t, started at $q_0^{\rightarrow}$, and y_t follows the probability flow ODE. Suppose for all t and all x that $\left\|\nabla^2 \ln q_t^{\rightarrow}(x)\right\|_{\mathrm{op}} \le L$, where $L \ge 1$. Then,

$$\mathbb{E}[\|\partial_t \nabla \ln q_t^{\rightarrow}(y_t)\|^2] \lesssim L^2 d \left(L + \frac{1}{t}\right).$$

Proof sketch: Consider simpler setting where $p_t = p_0 * N(0,t)$, $p_0 \propto e^{-V}$

$$\text{Key identity:} \quad \nabla \ln p_t(y) = -\mathbb{E}_{P_{0|t}(\cdot|y)}(\nabla V)$$

where $P_{0,t}$ is the joint distribution of $X_0 \sim p_0$ and $X_t = X_0 + \sqrt{t}Z$, $Z \sim N(0, I)$.

1a. Score perturbation lemma

Lemma (Score perturbation)

Suppose $q_t^{\rightarrow}$ is the density of the OU process at time t, started at $q_0^{\rightarrow}$, and y_t follows the probability flow ODE. Suppose for all t and all x that $\left\|\nabla^2 \ln q_t^{\rightarrow}(x)\right\|_{\mathrm{op}} \le L$, where $L \ge 1$. Then,

$$\mathbb{E}[\|\partial_t \nabla \ln q_t^{\rightarrow}(y_t)\|^2] \lesssim L^2 d \left(L + \frac{1}{t}\right).$$

Key identity: $\nabla \ln p_t(y) = -\mathbb{E}_{P_{0|t}(\cdot|y)}(\nabla V) \quad p_t = p_0 * N(0,t),\ p_0 \propto e^{-V}.$

- Bound in terms of $L^2 \cdot W_1^2(P_{0|t+\Delta t}(\cdot|x_t), P_{0|t}(\cdot|x_t))$.
- Bound W_1^2 by $\mathrm{KL}(P_{0|t+\Delta t}(\cdot|x_t)\|P_{0|t}(\cdot|x_t))$ by Talagrand's transport cost inequality and strong log-concavity of posterior (for $t \le \frac{1}{2L}$).
- Bound by $\mathrm{KL}(P_{0,t+\Delta t}\|P_{0,t})$, which can be explicitly calculated.

1. Predictor analysis

For simplicity, consider $t > \frac{1}{L}$.

$$\begin{aligned}
&& \mathbb{E}[\|\partial_t \nabla \ln q_t^{\rightarrow}(y_t)\|^2] &\lesssim L^3 d \\
\overset{\int_s^t,\ \text{C-S}}{\Longrightarrow} && \mathbb{E}\left[\|\nabla \ln q_t^{\rightarrow}(x_t) - \nabla \ln q_s^{\rightarrow}(x_s)\|^2\right] &\lesssim L^3 d h^2 \\
\overset{\text{Grönwall}}{\Longrightarrow} && W_2(qP_{\text{ODE}}^{t_0,h}, q\widehat{P}_{\text{ODE}}^{t_0,h}) &\lesssim L^{3/2} d^{1/2} h^2 + h\varepsilon_{\text{sc}} \\
\overset{\frac{1}{Lh}\ \text{steps}}{\Longrightarrow} && W_2(qP_{\text{ODE}}^{t_0,h\times\frac{1}{Lh}}, q\widehat{P}_{\text{ODE}}^{t_0,h\times\frac{1}{Lh}}) &\lesssim L^{1/2} d^{1/2} h + \frac{\varepsilon_{\text{sc}}}{L}.
\end{aligned}$$

Last step uses Lipschitzness of score estimate. Because distance is multiplied by e^{LT}, we need to take $T = O(1/L)$.

Lemma (Short-time regularization, Guillin and Wang 2012)

For $T_{\text{corr}} = \frac{1}{\sqrt{L}}$, $q =$ stationary distribution, the continuous dynamics satisfies

$$\mathrm{TV}(pP^N_{\text{ULD}}, q) \lesssim \sqrt{\text{KL}(pP^N_{\text{ULD}} \| q)} \lesssim \sqrt{L} W_2(p, q).$$

This converts Wasserstein distance to TV distance (with an extra $\sqrt{L}$ factor). Combined with a discretization analysis,

$$\mathrm{TV}(p\widehat{P}^N_{\text{ULMC}}, q) \lesssim \underbrace{\sqrt{L} W_2(p, q)}_{(1)} + \underbrace{\frac{\varepsilon_{\text{sc}}}{\sqrt{L}}}_{(2)} + \underbrace{\sqrt{Ld}\, h}_{(3)}.$$

1. Short-time Regularization Lemma.
2. Score estimation error (for time $\frac{1}{\sqrt{L}}$).
3. Discretization error of underdamped Langevin.

3. End-to-end analysis

Predictor:

$$W_2(q\widehat{P}_{\text{ODE}}^{N_{\text{pred}}}, qP_{\text{ODE}}^{N_{\text{pred}}}) \lesssim \underbrace{\sqrt{Ld}\,h_{\text{pred}} + \frac{\varepsilon_{\text{sc}}}{L}}$$

Corrector:

$$\mathrm{TV}(p\widehat{P}_{\text{ULMC}}^{N}, q) \lesssim \sqrt{L}W_2(p,q) + \sqrt{Ld}\,h_{\text{corr}} + \frac{\varepsilon_{\text{sc}}}{\sqrt{L}}.$$

1-stage of predictor (time $1/L$) and corrector (time $1/\sqrt{L}$):

$$\begin{aligned}&\mathrm{TV}(p\widehat{P}_{\text{ODE}}^{N_{\text{pred}}}\widehat{P}_{\text{ULMC}}^{N_{\text{corr}}}, q_{t_0+T_{\text{pred}}})\\ &\le \mathrm{TV}(p, q_{t_0}) + O\left(\sqrt{L}\sqrt{Ld}\,h_{\text{pred}} + \sqrt{Ld}\,h_{\text{corr}} + \frac{\varepsilon_{\text{sc}}}{\sqrt{L}}\right).\end{aligned}$$

Predictor + corrector for time T: $(\times LT)$

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim (\sqrt{d} + m_2^2)e^{-T} + TL^2 d^{1/2} h_{\text{pred}} + TL^{3/2} d^{1/2} h_{\text{corr}} + TL^{1/2}\varepsilon_{\text{sc}}.$$