AN INTRODUCTION TO DENOISING DIFFUSION PROBABILISTIC MODELS

DANNY CHEN (LAST UPDATED DECEMBER 6, 2024)

Contents

Suppose you receive a sample X_T from a trajectory of the Ornstein-Uhlenbeck (OU) process

$$dX_t = -X_t dt + \sqrt{2} dB_t, \ X_0 \sim \mu_0 \tag{1}$$

for some $T > 0$. Is it possible for you to reconstruct the initial conditions X_0? Unless T happens to be very small or you have a priori knowledge on μ_0—which we will rule out for the remainder of the note—your guess for X_0 is likely not very good. In particular, for large T, $\mu_T := \mathcal{L}(X_T) \approx \gamma$ where γ is a standard Gaussian.

We can ask a weaker question. Let $X^{(i)}$ be independent samples from the OU process and let $\tau^{(i)}$ be independent uniform draws from the interval $[0, T]$. Now, you're given instead the collection $\{X^{(i)}_{\tau^{(i)}}\}_{i=1}^n$, can we reconstruct the initial distribution μ_0 suitably well? This problem is more feasible; here's an idea: suppose there is a way to reverse time, i.e., construct a function $u : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$ such that the reversed SDE

$$d\bar{X}_t = (\bar{X}_t + u(t, \bar{X}_t))dt + \sqrt{2} d\bar{B}_t, \ \bar{X}_0 \sim \mu_T \tag{2}$$

has the same law (on path space) as the forward SDE (1). Then, if we can 1) estimate u with the samples at hand, and 2) approximate μ_T with γ, we can approximately draw samples from μ_0 by simulating (2).

It turns out that not only does this method work, it works miraculously well! This general approach—with engineering tricks sprinkled on top—is now the backbone of the most impressive image generation tools like `DALLE-2`. We would begin by investigating the existence of this time-reversal function u. Then, we will sketch out the real problem we are interested in, i.e., generative modeling, and some implementation details. We will conclude with a quick convergence proof and why proving convergence alone does not say much about statistical behavior.

1. Going in back in time

Reversing deterministic dynamics is easy. Take a general ODE

$$\frac{d}{dt} x_t = b(t, x_t), \ x_0 = x.$$

Time reversal follows from a change-in-variable $t \mapsto T - t$, from which we get

$$\frac{d}{dt} \bar{x}_t = -b(T - t, \bar{x}_t), \ \bar{x}_0 = x_T.$$

However, life is not so simple for stochastic dynamics. Take a generic SDE with constant diffusion coefficient

$$dX_t = b(t, X_t)dt + \sqrt{2}dB_t,\ X_0 = x. \tag{3}$$

An immediate problem comes from specifying the terminal condition $\bar{X}_T = x$ almost surely. Something about randomness requires adding some non-trivial forcing to the dynamics to get the dynamics to behave statistically the same. This suggests that it is a good idea to start by working on the distributional level rather than trying to get a pathwise description.

1.1. **A deterministic argument.** We begin with a heuristic derivation of the time reversal by reversing the Fokker-Planck equation, which we've determined is easy. Start by writing the associated forward equation of the density $(\mu_t)_{t\geq 0}$ for (3):

$$\partial_t \mu_t(x) = -\nabla \cdot (b(t,x)\mu_t(x)) + \Delta \mu_t(x).$$

Time reversing μ_t as a deterministic function then gives

$$\partial_t \bar{\mu}_t(x) = \nabla \cdot (b(T-t,x)\bar{\mu}_t(x)) - \Delta \bar{\mu}_t(x). \tag{4}$$

However, this does not look like a Fokker-Planck equation because of the $-\Delta\bar{\mu}_t$ term. So, we will be clever and notice that

$$\Delta \bar{\mu}_t = \nabla \cdot \nabla \bar{\mu}_t = \nabla \cdot (\bar{\mu}_t \nabla \log \bar{\mu}_t)$$

so we can move this inside the divergence term. Now, we can rewrite (4) to

$$\partial_t \bar{\mu}_t(x) = \nabla \cdot ((b(T-t,x) - 2\nabla \log \bar{\mu}_t(x))\bar{\mu}_t(x)) + \Delta \bar{\mu}_t(x),$$

which corresponds to the time marginals of the SDE

$$d\bar{X}_t = (-b(T-t,\bar{X}_t) - 2\nabla \log \mu_{T-t}(\bar{X}_t))dt + \sqrt{2}d\bar{B}_t. \tag{5}$$

This gives our candidate u as the gradient of the log-density, i.e.,

$$u(t,x) = 2\nabla \log \mu_{T-t}(x).$$

A downside of working with PDEs is the fact that regularity becomes important. It turns out the standard Lipschitz, linear-growth conditions in addition to Hormander-type regularity on the density is enough for the reversal formula to hold [7]. For those of us who are not well-versed in regularity, there is a more probabilistic approach.

Remark 1.1 (Nonlinear Markov processes)**.** The reversed SDE (5) is an example of a *nonlinear Markov process* in the sense that the evolution of the system depends on its own marginal. Alterantively, the associated Fokker-Planck equation is a nonlinear PDE. The study of mean-field theory—in some sense, the theme of the research group—dedicates itself to understanding properties of similar nonlinear Markov processes that arise from interacting particle systems.

Remark 1.2 (Time reversal gives weak solutions)**.** It is important to note that, while the forward process might admit a unique strong solution, the reversed process is only given as weak solutions! For example, Haussmann and Pardoux [7] showed the time-reversal formula in terms of the solution to the appropriate martingale problem, which gives weak solutions. Föllmer's approach [5, 6] in the next section will also only rely on distributional properties on path space rather than pathwise properties. We will come back to this point later.

1.2. **A stochastic argument.** The key to the stochastic argument is to interpret the drift coefficient as certain stochastic derivatives. We begin with a proposition that makes this precise.

Proposition 1.3 ([6, Proposition 2.5])**.** *Suppose that the drift coefficient in (3) is such that*

$$\mathbb{E}\left[\int_0^T |b(t,X_t)|^p dt\right] < \infty$$

for some $p \geq 1$. Then, for almost all $t \in [0,T]$, we have

$$b(t,X_t) = \lim_{h\to 0} \frac{1}{h}\mathbb{E}[X_{t+h} - X_t | \mathcal{F}_t]$$

where $(\mathcal{F}_t)_{t\geq 0}$ is the natural filtration of X and the limit is in L^p.

Let's first carry out some computation, assuming all limits exist, to obtain the same time-reversed drift. Thinking on the canonical space $\mathcal{C}([0,T];\mathbb{R}^d)$, let $\mathbb{P}$ be the law of X. Moreover, denote $R:\mathcal{C}([0,T];\mathbb{R}^d)\to\mathcal{C}([0,T];\mathbb{R}^d)$ to be the reversal map $R(x)_t = x_{T-t}$. Then, we are interested in the drift coefficient $\bar{b}$ for the diffusion process characterized by the $\bar{\mathbb{P}} = R_{\#}\mathbb{P}$. By Proposition 1.3 and assuming the desired integrability, we know that $\bar{b}$ can be written as the limit

$$\bar{b}(t,\bar{X}_t) = \lim_{h\to 0}\frac{1}{h}\bar{\mathbb{E}}[\bar{X}_{t+h}-\bar{X}_t|\bar{\mathcal{F}}_t]$$

where $\bar{\mathcal{F}}_t$ is the natural filtration of $\bar{X}$. By Itô's formula, we also know that for all $f\in C_0^2(\mathbb{R}^d;\mathbb{R})$,

$$\begin{aligned}
\bar{\mathbb{E}}\left[\bar{b}(t,\bar{X}_t)f(\bar{X}_t)\right] &= \lim_{h\to 0}\frac{1}{h}\bar{\mathbb{E}}\left[(\bar{X}_{t+h}-\bar{X}_t)f(\bar{X}_t)\right]\\
&= -\lim_{h\to 0}\frac{1}{h}\,\mathbb{E}\left[(X_t-X_{t-h})\left[f(X_{t-h})+\int_{t-h}^{t}\nabla f(X_s)\cdot dX_s+\int_{t-h}^{t}\Delta f(X_s)ds\right]\right]\\
&= -\,\mathbb{E}[b(t,X_t)f(X_t)]-2\,\mathbb{E}[\nabla f(X_t)]\\
&= -\bar{\mathbb{E}}[b(T-t,\bar{X}_t)f(\bar{X}_t)]-2\,\mathbb{E}[\nabla f(\bar{X}_t)].
\end{aligned}$$

Rearranging the above while assuming the density $\mu_t(x)$ exists, we get

$$2\int_{\mathbb{R}^d}\nabla f(x)\mu_{T-t}(x)dx = -2\int_{\mathbb{R}^d}f(x)\nabla\mu_{T-t}(x)dx = -\int_{\mathbb{R}^d}(b(T-t,x)+\bar{b}(t,x))f(x)\mu_{T-t}(x)dx,$$

alternatively, it means that

$$2\nabla\mu_{T-t}(x) = \mu_{T-t}(x)(b(T-t,x)+\bar{b}(t,x)).$$

weakly. Rearranging the above yields the same time-reversed drift.

In order for Proposition 1.3 to apply, we will take the particular case of $p=2$ due to its connection with entropy via Cameron-Martin-Girsanov. Recall the definition of relative entropy

$$\mathcal{H}(\mu\|\nu) = \begin{cases}\displaystyle\int\log\frac{d\mu}{d\nu}d\mu & \text{if } \mu\ll\nu\\ +\infty & \text{if } \mu\not\ll\nu.\end{cases}$$

In particular, if we assume 1) the initial distribution admits a density with respect to Lebesgue measure $\mu_0\ll\lambda$, 2) the entropy $\mathcal{H}(\mathbb{P}\|\mathbb{W})$ is finite where $\mathbb{W}$ is the σ-finite measure

$$\mathbb{W} = \int_{\mathbb{R}^d}\mathbb{W}_x dx \tag{6}$$

and $\mathbb{W}_x$ is the law of $\sqrt{2}B$ with initial conditions x. The first condition will guarantee the density $\mu_t(x)$ exists and the weak derivative of the density makes sense. The second condition will give the finite L^2-energy that we need to make sense of the stochastic derivative, that is,

$$\begin{aligned}
\mathcal{H}(\mathbb{P}\|\mathbb{W}) &= \mathcal{H}(\mu_0\|\lambda)+\int_{\mathcal{C}([0,T];\mathbb{R}^d)}\left[\int_0^T b(t,x_t)dx_t-\frac{1}{2}\int_0^T|b(t,x_t)|^2dt\right]\mathbb{P}(dx)\\
&= \mathcal{H}(\mu_0\|\lambda)+\int_{\mathcal{C}([0,T];\mathbb{R}^d)}\left[\int_0^T b(t,x_t)d\beta_t+\frac{1}{2}\int_0^T|b(t,x_t)|^2dt\right]\mathbb{P}(dx)\\
&= \mathcal{H}(\mu_0\|\lambda)+\mathbb{E}\left[\frac{1}{2}\int_0^T|b(t,X_t)|^2dt\right]
\end{aligned}$$

where $\beta_t = x_t-\int_0^t b(s,x_s)ds$ is a (scaled) Brownian motion under $\mathbb{P}$. In particular, since $\mathbb{W} = R_{\#}\mathbb{W}$,

$$\mathcal{H}(\bar{\mathbb{P}}\|R_{\#}\mathbb{W}) = \mathcal{H}(\mathbb{P}\|\mathbb{W}) < \infty,$$

which implies the reversed process has finite L^2-energy and the use of Proposition 1.3 is justified.

Remark 1.4 (Local finite entropy)**.** The reason for using (6) instead of a probability measure is for simplicity of the argument. Take for example $\mu_0=\delta_0$. Then, reversing $\mathbb{W}_0$ leads to singular behavior at terminal time, leading one to believe that the finite energy only holds only locally when we drop the condition $\mu_0\ll\lambda$.

However, finite entropy of the forward process indeed is enough to guarantee the time-reversal formula in $(0,T)$ and the argument lengthens; see [6, Lemma 3.1].

2. Mapping noise to data

Now, we return to the main task of this note: *generative modeling*. In some sense, this is a density estimation problem, but we will see soon that we care about something different. Concisely put, we have data $\{X^{(i)}\}_{i=1}^n$ drawn i.i.d. from some distribution μ^* and we want to generate more statistically similar samples. These problems enjoy a wide range of applications. Most notably so, it generates fun images and videos. But it is also used for drug design where the data is existing protein configurations and we use generative models to guess new protein structures.

Remark 2.1 (Terminology issues)**.** The term generative modeling has evolved significantly. Traditionally, we think of generative models as some sort of (hierarchical) Bayesian models where the generation comes from the posterior predictive distributions. More recently, a theoretically unjustified method—a.k.a. machine learning—took over the term "generative modeling" to refer to deep-learning-based Boltzmann machines, variational autoencoders, and generative adversarial networks.

2.1. Diffusion models and score-matching.

Denoising diffusion probabilistic models (DDPMs) are fairly recent methods. It first appeared in discrete-time and derived within the context of variational inference [12,13]. However, the method did not take off until Song et al. [14] put it into the continuous-time framework. Subsequent works following the footsteps are enumerable and I will not attempt to provide a "complete" (in whatever sense) list of related work, instead, I will only refer to ones that are relevant; interested readers can find further references therein.

$$\begin{array}{ccc} X^{(i)} \sim \mu^* & \xrightarrow{\text{Simulate } dX_t = -X_t dt + \sqrt{2} dB_t \text{ with } X_0 = X^{(i)}} & \mu_T \\ \wr & & \wr \\ \hat{X}_T \sim \hat{\mu}_T \approx & \xrightarrow[\text{Simulate } d\hat{X}_t = (\hat{X}_t + \hat{u}(t,\hat{X}_t))dt + \sqrt{2} d\bar{B}_t]{} & \gamma \end{array}$$

The idea of DDPMs was sketched out in the introduction, but we reiterate it here. For each data point $X^{(i)}$, we associate a forward process—usually the OU process (1)—with initial condition $X_0 = X^{(i)}$ that converges to some easy-to-sample invariant measure γ. Using these trajectories, we can learn the *score function*

$$u(t,x) = 2\nabla \log \mu_{T-t}(x)$$

for all $(t,x) \in [0,T] \times \mathbb{R}^d$. Then, using the estimated score $\hat{u}$, we can simulate $\bar{X}$, the reverse SDE (5), with initial distribution γ and obtain an approximate sample from $\mu^* \approx \bar{X}_T$.

The only problem here is how should u be estimated and this is more broadly—beyond DDPM, though certainly most well-known in the context of DDPM—known as *score matching*. And the solution is really not so clever: take your favorite function class $\mathcal{F}$, e.g., artificial neural network, and perform the optimization

$$\min_{\hat{u} \in \mathcal{F}} \mathbb{E}[\|u(t,x) - \hat{u}(t,x)\|^2]$$

where the expectation is over random times $t \sim \text{Uniform}([0,T])$ and random positions from the forward process $x \sim X_t$. However, it is impossible to evaluate this loss function because u is unknown. So, we need to do a little more work: for any fixed $t \in [0,T]$,

$$\begin{aligned} \arg\min_{\hat{u} \in \mathcal{F}} \mathbb{E}[\|u(t,x) - \hat{u}(t,x)\|^2] &= \arg\min_{\hat{u} \in \mathcal{F}} \int_{\mathbb{R}^d} (2\hat{u}(t,x) \cdot \nabla \log \mu_t(x) + \|\hat{u}(t,x)\|^2) \mu_t(x) dx \\ &= \arg\min_{\hat{u} \in \mathcal{F}} \int_{\mathbb{R}^d} (-2\nabla \cdot \hat{u}(t,x) + \|\hat{u}(t,x)\|^2) \mu_t(x) dx \\ &= \arg\min_{\hat{u} \in \mathcal{F}} \mathbb{E}[\|\hat{u}(t,x)\|^2 - 2\nabla \cdot \hat{u}(t,x)]. \end{aligned}$$

Of course, instead of evaluating the expectation, we simply minimize the empirical risk. Moreover, the sample will be obtained by simulating the SDE via (almost) Euler-Maruyama. To distinguish the process run with the true score u and the estimated, discretized score $\hat{u}$, we will use $\hat{X}$ and $\hat{\mu}_t = \mathcal{L}(\hat{X}_t)$ to denote the

approximate backward process. More explicitly, for discretization step $h = T/N$, the process $\hat{X}$ follows the solution of an SDE with piecewise-constant drift

$$d\hat{X}_t = (\hat{X}_t + \hat{u}(kh, \hat{X}_{kh}))dt + \sqrt{2}d\bar{B}_t \tag{7}$$

for $t \in [kh, (k+1)h)$.

Remark 2.2 (Stochastic localization and message-passing)**.** It is worth noting that there is another way to think about DDPMs through the lens of *stochastic localization*. This is a technique originally used to study high-dimensional convex bodies and Markov chain mixing, which Lee and Vampala [8] and Chen [2] used to tackle the famous KLS conjecture. Using the mimicking theorem, Montanari [11] gave a stochastic localization approach to solving the same problem. Instead of estimating the score, the main hurdle is to estimate the conditional expectation term from the mimicking theorem, which he attacked with the rich theory of message-passing algorithms. Similar ideas were used by Huang, Montanari, and Pham [11] for sampling from mean-field spin glasses.

2.2. **Convergence in total variation.** Seeing the success of diffusion models, mathematicians were excited: this is perhaps the biggest victory for stochastic analysis since mathematical finance! However, the idea of DDPMs is not well-studied, and with great interest (citations) comes great number of papers trying to establish theoretical underpinnings of these diffusion-based, score-based generative models. Below, we give one that is particularly liked by the community.

Theorem 2.3 ([1, Theorem 2])**.** *Assume the following:*

(1) for all $t \geq 0$, $u(t, \cdot)$ is L-Lipschitz;
(2) the data distribution μ^* has finite second moment $\mathfrak{m} := \int_{\mathbb{R}^d} \|x\|^2 \mu^*(dx)$;
(3) the estimated score is close to the true score, i.e., $\mathbb{E}[\|u(t, \cdot) - \hat{u}(t, \cdot)\|^2] \leq \epsilon^2_{\text{SCORE}}$ for all $t \geq 0$.

Then, running DDPM *with time steps $h = T/N$ gives*

$$d_{\text{TV}}(\hat{\mu}_T, \mu^*) \leq e^{-T}\mathcal{H}(\mu^* \| \gamma)^{1/2} + (L\sqrt{dh} + L\mathfrak{m}h)\sqrt{T} + \epsilon_{\text{SCORE}}\sqrt{T}.$$

Remark 2.4 (On sore estimation error)**.** Getting ϵ_{SCORE} not as an assumption is difficult because the typical function estimator is a deep neural network and we barely have a grasp on statistical properties of them. However, under the assumption that μ^* is a Markov random field, Mei and Wu [10] were able to establish properties of deep-neural-network estimation for the score function using techniques from message-passing.

Remark 2.5 (Interpreting the convergence)**.** Suppose that $\mathcal{H}(\mu^* \| \gamma) = \mathcal{O}(\text{poly}(d))$ and $\mathfrak{m} = \mathcal{O}(d)$. Then, we can choose $T = \Theta(\mathcal{H}(\mu^* \| \gamma)/\epsilon)$ and $h = \Theta(\epsilon^2/L^2 d)$, we can get

$$d_{\text{TV}}(\hat{\mu}_T, \mu^*) = \tilde{\mathcal{O}}(\epsilon + \epsilon_{\text{SCORE}}) \text{ for } N = \tilde{\Theta}\left(\frac{L^2 d}{\epsilon}\right).$$

This complexity bound in fact matches the state-of-the-art bound for Langevin Monte Carlo [3].

Proof. We only give a sketch and the full proof can be found in [1, Section 5]. Let $\bar{\mathbb{P}}$ and $\hat{\mathbb{P}}$ denote the law of $\bar{X}$ from (2) and $\hat{X}$ from (7) respectively, both with initial distribution $\mathcal{L}(X_T)$.

(1) If we can apply Girsanov's theorem, then the relative entropy would take the form

$$\mathcal{H}(\hat{\mathbb{P}} \| \bar{\mathbb{P}}) = \frac{1}{2}\sum_{k=0}^{N-1} \mathbb{E}\left[\int_{kh}^{(k+1)h} \|\hat{u}(hk, X_{kh}) - u(t, X_t)\|^2 dt\right].$$

We will proceed by bounding the error accumulated from discretization, then carefully applying Girsanov.

(2) For $t \in [kh, (k+1)h)$, we can write

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{u}(hk, X_{kh}) - u(t, X_t)\|^2\right] \\
&\leq \mathbb{E}\left[\|\hat{u}(hk, X_{kh}) - u(hk, X_{hk})\|^2\right] + \mathbb{E}\left[\|u(hk, X_{kh}) - u(t, X_{hk})\|^2\right] + \mathbb{E}\left[\|u(t, X_{kh}) - u(t, X_t)\|^2\right] \\
&\leq \epsilon^2_{\text{SCORE}} + \mathbb{E}\left[\|u(hk, X_{kh}) - u(t, X_{hk})\|^2\right] + L^2\,\mathbb{E}\left[\|X_{kh} - X_t\|^2\right].
\end{aligned}$$

The second term can be bounded by exploiting the choice of an OU process, which gives

$$\mathbb{E}\left[\|u(hk, X_{kh}) - u(t, X_{hk})\|^2\right] \lesssim L^2 dh + L^2 h^2\,\mathbb{E}[\|X_{kh}\|^2] + L^2 h^2\,\mathbb{E}[\|u(t, X_{kh}\|^2].$$

Lastly, the expected norm of the score and X can be bounded by L, d, $\mathfrak{m}$, and h. In particular, we will arrive at the estimate

$$\mathbb{E}\left[\|\hat{u}(hk, X_{kh}) - u(t, X_t)\|^2\right] \lesssim \epsilon_{\text{SCORE}}^2 + L^2 dh + L^2\mathfrak{m}^2 h^2.$$

(3) A priori, we do not have enough integrability for Girsanov to hold exactly. However, the estimate in the previous step implies that there is a sequence of stopping times $\tau_n \to \infty$ a.s. such that if we define X^n to be

$$dX_t^n = (X_t^n + 2\hat{u}(kh, X_{kh}^n)1_{[0,\tau_n]}dt + (X_t^n + 2\hat{u}(t, X_t^n))1_{(\tau_n,T]}dt + \sqrt{2}d\bar{B}_t$$

and denote $P^n \coloneqq \mathcal{L}(X^n)$, we can apply Girsanov for each n. Lastly, by lower semicontinuity and an additional approximation argument (which crucially uses a coupling argument), we have

$$\mathcal{H}(\bar{\mathbb{P}}\|\hat{\mathbb{P}}) \le \liminf_{n\to\infty} \mathcal{H}(\bar{\mathbb{P}}\|P^n) \lesssim \epsilon_{\text{SCORE}}^2 + L^2 dh + L^2\mathfrak{m}^2 h^2.$$

(4) Lastly, we convert entropy estimates to total variation. Let $\mathbb{Q}$ be the law of the backward process (2) with initial distribution γ. By Pinsker's inequality,

$$d_{\text{TV}}(\hat{\mu}_T, \mu^*) \le d_{\text{TV}}(\hat{\mathbb{P}}, \bar{\mathbb{P}}) + d_{\text{TV}}(\bar{\mathbb{P}}, \mathbb{Q}) \lesssim e^{-T}\mathcal{H}(\mu^*\|\gamma)^{1/2} + (\epsilon_{\text{SCORE}} + L\sqrt{dh} + L\mathfrak{m}h)\sqrt{T}. \tag{8}$$

□

Remark 2.6 (Lipschitz score functions?)**.** A crucial part of the approximation argument was established using a coupling argument, which requires the existence of strong solutions of the reversed process. Hence, this motivated the authors to add the Lipschitz condition to u, which is not uncommon in the machine learning literature. However, as pointed out before, time-reversal in inherently a statement on the law. A more refined argument that respects this subtlety was offered by Conforti, Durmus, and Silveri [4] using ideas from stochastic control.

2.3. **Why we shouldn't care about convergence.** Sure, convergence is nice, but are good convergence rates the end of the story? In particular, what is the fundamental difference between DDPM and some more traditional density estimation methods like the kernel density estimate (KDE)? For a fixed with σ, the KDE of μ^* is simply saying

$$\mu^* \approx \rho_\sigma \coloneqq \frac{1}{n}\sum_{i=1}^n \gamma_{X^{(i)},\sigma}$$

where $\gamma_{x,\sigma}$ is a mean x, variance σ^2 Gaussian.

We know basically all there is to know about KDE—bias, variance, finite-sample concentration. Most importantly, we know it is a terrible method for generative modeling because I simply place a mass around the data points I've previously seen. In essence, I would be seeing replicas of the data. So, DDPM must be better... right?

Wrong.

Theorem 2.7 ([9, Theorem 4.3])**.** *Consider the* DDPM *output $\hat{\mu}_T$ with an empirically optimal score function. Moreover, suppose $\|X^{(i)}\| \le d$ for all i. For any $\epsilon > 0$, setting $T = \log d/\epsilon$ and $\delta = \epsilon^2/d$, we have*

$$d_{\text{TV}}(\hat{\mu}_{T-\delta}, \rho_\sigma) \le \epsilon$$

where $\sigma = \sqrt{1 - e^{-2\delta}}$. Moreover, taking $T \to \infty$ and $\delta = 0$, we have $\hat{\mu}_\infty = \rho_0$.

The proof follows from the observation that the empirically optimal score function is exactly that of the theoretical one when initialized with empirical measures. Moreover, the theoretical forward process is nothing but a KDE. From which, convergence results guarantee that we're not far optimal given that the empirically optimal one is not far from a KDE. The computations are short and, one could argue, not insightful [9, Appendix D]; however, I think it is exactly the conciseness that emphasizes the dearth of true theoretical understanding on DDPMs.

This problem of *memorization effects* has surfaced recently as text-to-image generators started generating copyrighted content [15]. Perhaps, it is important to start making the mathematical distinction that

density estimate/sampling ≠ generative modeling.

In generative modeling, we not only want closeness in distribution, but we also want "new" samples that are not simply corrupted replicas of previous data. And for this reason, convergence proofs are not the end of the story; in fact, the core reason why DDPMs enjoyed such success remains fundamentally open.

References

[1] 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).

[2] Tianqi Chen, Emily Fox, and Carlos Guestrin, *Stochastic gradient hamiltonian monte carlo*, International conference on machine learning, 2014, pp. 1683–1691.

[3] Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang, *Analysis of langevin monte carlo from poincare to log-sobolev*, Foundations of Computational Mathematics (2024), 1–51.

[4] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri, *Score diffusion models without early stopping: finite fisher information is all you need*, arXiv preprint arXiv:2308.12240 (2023).

[5] Hans Föllmer, *An entropy approach to the time reversal of diffusion processes*, Stochastic differential systems filtering and control: Proceedings of the ifip-wg 7/1 working conference marseille-luminy, france, march 12–17, 1984, 2005, pp. 156–163.

[6] ———, *Time reversal on wiener space*, Stochastic processes—mathematics and physics: Proceedings of the 1st bibos-symposium held in bielefeld, west germany, september 10–15, 1984, 2006, pp. 119–129.

[7] Ulrich G Haussmann and Etienne Pardoux, *Time reversal of diffusions*, The Annals of Probability (1986), 1188–1205.

[8] Yin Tat Lee and Santosh S Vempala, *Eldan's stochastic localization and the kls conjecture: Isoperimetry, concentration and mixing*, Annals of Mathematics **199** (2024), no. 3, 1043–1092.

[9] Sixu Li, Shi Chen, and Qin Li, *A good score does not lead to a good generative model*, arXiv preprint arXiv:2401.04856 (2024).

[10] Song Mei and Yuchen Wu, *Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models*, arXiv preprint arXiv:2309.11420 (2023).

[11] Andrea Montanari, *Sampling, diffusions, and stochastic localization*, arXiv preprint arXiv:2305.10690 (2023).

[12] Alexander Quinn Nichol and Prafulla Dhariwal, *Improved denoising diffusion probabilistic models*, International conference on machine learning, 2021, pp. 8162–8171.

[13] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, International conference on machine learning, 2015, pp. 2256–2265.

[14] 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 (2020).

[15] Stuart A Thompson, *We asked ai to create the joker. it generated a copyrighted image*, The New York Times (Jan 2024). https://www. nytimes. com/interactive/2024/01/25/business/ai-imagegenerators-openai-microsoft-midjourney-copyright. html (2024).