CS 2243 Fall 2024 Scribes: <names>
11/20 Based on notes by: Tianze Jiang

Lecture 22: Diffusion models

This is the last lecture of the semester. We will conclude past discussions on Bayesian inference and connect them to the setting of learning a distribution via diffusion models, a successful approach both empirically and theoretically. The material is largely adapted from [CCL+23] and references therein.

1 Primer on Diffusion Models

In generative models, one typically seeks to approximate a transport map from $\mathcal{S} \to \mathcal{T}$ where $\mathcal{S}$ is some natural source distribution (Gaussian, uniform, etc.) and $\mathcal{T}$ is the target distribution. In the focus of diffusion models, we will refer to the process $\mathcal{T} \to \mathcal{S}$ as the "Forward Process", adding noise to data, and the reversed $\mathcal{S} \to \mathcal{T}$ map as the "Backward Process", where sampling actually happens.

1.1 Forward Process

Conceptually, denoising diffusion probabilistic modeling (DDPM) starts with a forward "noising" process characterized by a stochastic differential equation (SDE). For clarity, we consider the simplest possible choice, which is the Ornstein-Uhlenbeck (OU) process:

$$\mathrm{d}\bar{X}_t = -\bar{X}_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t, \quad \bar{X}_0 \sim q, \tag{1}$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The OU process is the unique time-homogeneous Markov process which is also a Gaussian process, with a stationary distribution equal to the standard Gaussian distribution on $\mathbb{R}^d$. In practice, it is also common to introduce a positive smooth function $g : \mathbb{R}_+ \to \mathbb{R}^2$ and consider the time-rescaled OU process

$$\mathrm{d}\bar{X}_t = -g(t)^2 \bar{X}_t \, \mathrm{d}t + \sqrt{2} g(t) \mathrm{d}B_t, \quad X_0 \sim q,$$

but we stick with the choice $g \equiv 1$. In fact, it is easy to solve in this case that

$$\mathrm{Law}(\bar{X}_t) = \mathrm{Law}(e^{-t}\bar{X}_0 + \sqrt{1 - e^{-2t}} G), \quad G \sim \mathcal{N}(0, \mathbb{I}_d) \perp\!\!\!\perp \bar{X}.$$

The forward process has the interpretation of transforming samples from the data distribution q into pure noise. From the well-developed theory of Markov diffusions, one sees that the law converges to a standard Gaussian exponentially fast in various divergences and metrics such as the 2-Wasserstein metric W_2.

1.2 Backward Process

Given the "noising" forward process, one can thus try to *invert* the forward map and construct the corresponding "de-noising" process. Conceptually, this reversed process is similar to reversing a Markov Chain with time steps going to zero. We write down the reverse formula and defer the justification to the next subsection.

In general, suppose that we have an SDE of the form

$$\mathrm{d}\bar{X}_t = b_t \left(\bar{X}_t\right) \mathrm{d}t + \sigma_t \, \mathrm{d}B_t,$$

where $(\sigma_t)_{t \geq 0}$ is a deterministic matrix-valued process. Then the reverse process also admits an SDE description. Namely, if we fix the terminal time $T > 0$ and set

$$\bar{X}_t^{\leftarrow} := \bar{X}_{T-t}, \quad \text{for } t \in [0, T]$$

then the process $\left(\bar{X}_t^{\leftarrow}\right)_{t \in [0,T]}$ satisfies the SDE

$$\mathrm{d}\bar{X}_t^{\leftarrow} = b_t^{\leftarrow} \left(\bar{X}_t^{\leftarrow}\right) \mathrm{d}t + \sigma_{T-t} \, \mathrm{d}B_t$$

where the backwards drift satisfies the relation

$$b_t + b_{T-t}^{\leftarrow} = \sigma_t \sigma_t^{\top} \nabla \ln q_t, \quad q_t := \mathrm{Law}\left(\bar{X}_t\right).$$

Applying this to the forward process (1), we obtain the reverse process

$$\mathrm{d}\bar{X}_t^{\leftarrow} = \left\{\bar{X}_t^{\leftarrow} + 2\nabla \ln q_{T-t}\left(\bar{X}_t^{\leftarrow}\right)\right\} \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t. \tag{2}$$

where $\nabla \ln q_{T-t}\left(\bar{X}_t^{\leftarrow}\right)$ is "score": the gradient of the density at time $T-t$ of the forward (and time t of the reverse) process.

To see why (2) is indeed the reversed SDE of (1), we need to apply the Fokker-Planck equation, which states that:

Proposition 1 (Fokker-Planck). *For any smoothly varying family of smooth vector fields $v_t : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, the iterates x_t of the SDE:*

$$dx_t = v_t\left(x_t\right) dt + \sqrt{2} dB_t$$

are distributed according to q_t satisfying the PDE

$$\frac{\partial q_t}{\partial t} = -\operatorname{div}\left(q_{\mathrm{t}} \cdot v_t\right) + \Delta q_t.$$

We omit the proof in this note, as it can be found in many textbooks of stochastic analysis (e.g., pp. 47-49 of [PB13]). Instead, under Proposition 1, we observe that (1) and (2) are indeed reversals of each other by pattern-matching the terms.

1.3 Score Matching

Consider the following sampling procedure: pick a large T such that q_T is close to a standard Gaussian, and run (2) with $\bar{X}_0^{\leftarrow}$ distributed according to $\mathcal{N}(0, \mathbb{I}_d)$. While this is the most natural way to go, the biggest caveat is that one does not typically know $s_t := \nabla \ln q_t$ without knowing the distribution q_0 a priori. Fortunately, such a score function can be computed via the below lemma:

Lemma 1 (Tweedie's Formula). *Given $\tilde{x} = x + e$ for $x \sim p$ and $e \sim \mathcal{N}\left(0, \sigma^2 \cdot \mathbb{I}_d\right)$,*

$$\mathbb{E}[x \mid \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x})$$

where $\tilde{p}$ is the density for $\tilde{x}$.

The above lemma connects the score function to the Bayesian-optimal estimation of the noise. In other words, estimating the gradient of log density is equivalent to estimating the noise. In practice, one runs the following optimization over a class of neural nets $\mathcal{F}$ given samples $\{x_i\} \sim \mu$:

$$\hat{s}_t = \arg\min_{\mathrm{NN} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left\|\sigma_t^{-1} g_i + \mathrm{NN}\left(\lambda_t x_i + \sigma_t g_i; t\right)\right\|_2^2, \quad g_i \sim_{iid} \mathcal{N}\left(0, \mathbb{I}_d\right) \tag{3}$$

where NN is a function with input (x, t) for some explicit scaling of $t \longrightarrow (\lambda_t, \sigma_t)$. Under Lemma 1, it is not hard to check that the minimizing $\hat{s}_t$ of (3) is indeed equivalent to $\nabla \ln q_t$ for $\left(\lambda_t, \sigma_t^2\right) = \left(e^{-t}, 1 - e^{-2t}\right)$. This allows us an optimization form that depends on samples x_i only and not the distribution μ.

Proof of Lemma 1. By Bayes' rule,

$$\mathbb{P}[x \mid \tilde{x}] = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\tilde{x}-x)^2}{2\sigma^2}\right) \cdot p(x)}{\tilde{p}(\tilde{x})}$$

so

$$\mathbb{E}\left[\frac{x - \tilde{x}}{\sigma^2} \mid \tilde{x}\right] = \frac{\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(\tilde{x}-x)^2}{2\sigma^2}\right) \cdot \frac{x-\tilde{x}}{\sigma^2} \cdot p(x) \mathrm{d}x}{\tilde{p}(\tilde{x})}.$$

Observe that on the other hand

$$\begin{aligned}\tilde{p}(\tilde{x}) &= \int_{-\infty}^{\infty} \exp\left(-\frac{(\tilde{x}-x)^2}{2\sigma^2}\right)\cdot p(x)\mathrm{d}x \\ \nabla\tilde{p}(\tilde{x}) &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{(\tilde{x}-x)^2}{2\sigma^2}\right)\cdot\frac{x-\tilde{x}}{\sigma^2}\cdot p(x)\mathrm{d}x,\end{aligned}$$

and therefore

$$\mathbb{E}\left[\frac{x-\tilde{x}}{\sigma^2} \mid \tilde{x}\right] = \frac{\nabla\tilde{p}(\tilde{x})}{\tilde{p}(\tilde{x})} = \nabla\ln\tilde{p}(\widetilde{x})$$

concluding the proof. □

2 Discretization Analysis

Recall that (3) is a non-convex, over-parameterized optimization task on empirical samples. Whether such optimization generalizes from $\{x_i\}$ to μ usually depends highly on the structure of $\mathcal{F}$ (the class of neural nets) with even less known about whether algorithms like SGD guarantees ERM in the first place. We divert from such discussions and assume instead access to an oracle $t \longrightarrow s_t$ such that for all t:

$$\mathbb{E}_{q_t}\left[\|s_t\left(X_t\right) - \nabla\ln q_t\left(X_t\right)\|^2\right] \leq \varepsilon_{\mathrm{sc}}^2 \tag{4}$$

To approximately implement the reverse (2), we first replace the score function $\nabla\ln q_{T-t}$ with the estimate s_{T-t}. Then, for $t \in [kh,(k+1)h]$ we freeze the value of this coefficient in the SDE at time kh. It yields the new SDE:

$$\mathrm{d}X_t^{\leftarrow} = \left\{X_t^{\leftarrow} + 2s_{T-kh}\left(X_{kh}^{\leftarrow}\right)\right\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad t \in [kh,(k+1)h]. \tag{5}$$

In this sense, for every k, conditioned on $X_{kh}^{\leftarrow}$, the next iterate $X_{(k+1)h}^{\leftarrow}$ has an explicit Gaussian distribution where we integrate $\mathrm{d}B_t$ directly.

As mentioned before: although the reverse SDE (2) should be started at q_T, we do not have access to q_T directly. Instead, we instead initialize the algorithm at $X_0^{\leftarrow} \sim \mathcal{N}(0,\mathbb{I}_d)$, i.e., from pure noise.

Let $p_t := \operatorname{law}\left(X_t^{\leftarrow}\right)$ denote the law of the algorithm at time t. The goal of this work is to bound $\mathrm{TV}\left(p_T, q\right)$, taking into account three sources of error: (1) the estimation of the score function; (2) the discretization of the SDE with step size $h > 0$; and (3) the initialization of the algorithm at pure noise rather than at q_T.

2.1 Assumptions

In this lecture, we will assume that our distribution and score estimation satisfy the following fundamental assumptions.

A1: (Lipschitz score). For all $t \geq 0$, the score $\nabla\ln q_t$ is L-Lipschitz.

A2: (Second moment bound). We assume that $\mathfrak{m}_2^2 := \mathbb{E}_q\left[\|\cdot\|^2\right] < \infty$.

A3: (Estimation error). We assume access to s_t satisfying (4) with some $\varepsilon_{\mathrm{sc}} > 0$.

A1 and **A2** are satisfied by most natural distributions and appear in existing works as well (e.g., [LLT22]). It is worth noting that no assumptions on the Lipschitz-ness of *score estimates* is required. Intuitively, it may appear as surprising since one would expect to bound terms like $|\hat{s}(\hat{X}) - s(X)|$ along the way. Here our results can bypass it without making this sometimes demanding assumption.

Furthermore, we do not assume that q satisfies a log-Sobolev inequality. Hence, our assumptions cover a wide range of highly non-log-concave distributions. Even **A1** could be relaxed by considering a different time-change in (1), although we focus on the simplest setting in order to better illustrate the conceptual significance.

Finally, it is worth noting that we assume the estimation error (4) to be small ($o(1)$) in L_2, as opposed to L_∞, which matches with the optimization objective (3) and more closely reflects the error setting.

2.2 DDPM Convergence Theorem

Theorem 1 (Theorem 2 in [CCL+23])**.** *Assuming assumptions **A1**, **A2**, and **A3** hold, Let p_T be the output of the DDPM discretization* (5) *at time T, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$, where $L > 1$. Then, it holds that (let $\gamma^d = \mathrm{Law}(\mathcal{N}(0, \mathbb{I}_d))$):*

$$\mathrm{TV}\left(p_T, q\right) \lesssim \underbrace{\sqrt{\mathrm{KL}\left(q \| \gamma^d\right)} \exp(-T)}_{\textit{convergence of forward process}} + \underbrace{\left(L\sqrt{dh} + L\mathfrak{m}_2 h\right)\sqrt{T}}_{\textit{discretization error}} + \underbrace{\varepsilon_{\mathrm{sc}}\sqrt{T}}_{\textit{score estimation error}} \quad . \tag{6}$$

To interpret this result, suppose that $\mathrm{KL}\left(q \| \gamma^d\right) \le \mathrm{poly}(d)$ and $\mathfrak{m}_2 \le d$. Choosing $T \asymp \log\left(\mathrm{KL}\left(q \| \gamma^d\right)/\varepsilon\right)$ and $h \asymp \frac{\varepsilon^2}{L^2 d}$, and hiding logarithmic factors, we get

$$\mathrm{TV}\left(p_T, q\right) \le \widetilde{O}\left(\varepsilon + \varepsilon_{\mathrm{sc}}\right)$$

for $N = \widetilde{\Theta}\left(\frac{L^2 d}{\varepsilon^2}\right)$. In particular, in order to have $\mathrm{TV}\left(p_T, q\right) \le \varepsilon$, it suffices to have score error $\varepsilon_{\mathrm{sc}} \le \widetilde{O}(\varepsilon)$ in the said parameter setup.

Comparison to Langevin MCMC Another popular approach towards sampling via SDE is the Langevin dynamics ([SE19]), where instead one considers the following differential equation with stationary distribution $Y \sim q$:

$$\mathrm{d}Y_t = -\nabla \ln q\left(Y_t\right) + \sqrt{2}\,\mathrm{d}B_t. \tag{7}$$

The pro of this approach is obvious, as it only asks for access to $\nabla \ln q$ as opposed to $\nabla \ln q_t$ for many different t's. However, the drawback is that such a process may take a long time to mix for "multimodal" distributions (e.g. Gaussian Mixture with two imbalanced clusters, see Figure 3 in [SE19]). In light of this, our diffusion process can also be pictured as Langevin dynamics with varying noise levels to accelerate mixing by passing freely between different modes.

Finally, we remark that the iteration complexity of $N = \tilde{\Theta}\left(\frac{L^2 d}{\varepsilon^2}\right)$ matches SOTA complexity bounds for the Langevin when sampling under a log-Sobolev inequality ([VW19]). This provides some evidence of the correct order one should expect.

2.3 Discretization Error of Theorem 1

Let us briefly discuss the proof of Theorem 1, focusing on the most interesting (non-trivial) part which is the discretization error. The key intuition is that, instead of comparing the distance between the *end* of the true reverse process versus the discretized reverse process, we compare the distance between the *entire trajectory* of the true versus discretized reverse process which upper bounds the terminal distance due to the data processing inequality. To do that, we need the following:

Lemma 2 (Girsanov's Theorem)**.** *For $t \in [0, T]$, let $\mathcal{L}_t = \int_0^t b_s\, \mathrm{d}B_s$. Assume that $\mathbb{E}\int_0^T \|b_s\|^2\, \mathrm{d}s < \infty$. Then, $\mathcal{L}$ is a martingale in L^2. Moreover, if*

$$\mathbb{E}\mathcal{E}(\mathcal{L})_T = 1, \quad \textit{where} \quad \mathcal{E}(\mathcal{L})_t := \exp\left(\int_0^t b_s\, \mathrm{d}B_s - \frac{1}{2}\int_0^t \|b_s\|^2\, \mathrm{d}s\right),$$

then $\mathcal{E}(\mathcal{L})$ is also a martingale and the process $t \mapsto B_t - \int_0^t b_s\, \mathrm{d}s$ is a Brownian motion under some explicit change of measures.

The key idea is that for every possible discretized path with N steps, the probability assigned to this path is easy to compute as it is just N Gaussian densities multiplied together. The remaining steps can be roughly summarized as (let $Q_T^{\leftarrow}$ be the true law of process (2) and $P_T^{\leftarrow}$ be our estimated law of (5))[1]:

[1] The following parts are crude over-simplifications and not technically correct. A rigorous proof can be found at Section 5 in [CCL+23]

Bound on the discretization error To apply Lemma 2, we need to bound the error between the estimated score at discretized endpoints to the actual score. Specifically,

$$\mathbb{E}_{Q_T^{\leftarrow}}\left[\|s_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(X_t\right)\|^2\right] \lesssim \varepsilon_{\mathrm{sc}}^2+L^2 dh+L^2 \mathfrak{m}_2^2 h^2.$$

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

Approximation argument For $t \in [0, T]$, let $\mathcal{L}_t = \int_0^t b_s \, \mathrm{d}B_s$ where

$$b_t = \sqrt{2}\left\{s_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(X_t\right)\right\},$$

if $t \in [kh, (k+1)h]$. The previous part translates to:

$$\mathbb{E}_{Q_T^{\leftarrow}} \frac{1}{2} \int_0^T \|b_s\|^2 \, \mathrm{d}s \lesssim \left(\varepsilon_{\mathrm{sc}}^2+L^2 dh+L^2 \mathfrak{m}_2^2 h^2\right) T$$

We conclude by showing that the left-hand quantity upper bounds $\mathrm{KL}(Q_T^{\leftarrow} \| P_T^{\leftarrow})$ via Lemma 2 and finish with Pinsker's inequality.

3 Provable Score Estimation

In the last part of this note, we list several examples where diffusion models succeed in achieving state-of-the-art sampling/learning algorithms in theoretical models. This is one of the very few examples where empirically motivated methodology inspires progress in highly theoretical problems.

At a high level, those examples succeed in the context that (approximately) optimal algorithms for score estimation are known, and thus by connecting with existing diffusion bounds we can sample approximately in polynomial time.

Sampling from the Sherrington-Kirkpatrick Model ([EAMS22]) For a given matrix $W \in \mathbb{R}^{n\times n}$ with entries i.i.d. sampled from $\mathcal{N}(0, 1/n)$, the goal is to efficiently sample from:

$$\mathbb{P}_W(x) \propto \exp(-\frac{\beta}{2}\langle x, Wx\rangle)$$

with high probability. As is well-known that sampling from a worst-case Ising model is #P-hard, the problem asks for the complexity of sampling from the *average-case* Ising model. Classical sampling techniques such as the Glauber dynamics guarantee sampling when the inverse temperature $\beta < 1/4$. In [EAMS22] and subsequently [Cel22], it was shown that sampling all the way to $\beta < 1$ is possible in normalized W_2 via DDPM, and that sampling when $\beta > 1$ is geometrically hard for algorithmically stable samplers. The result relies on showing that the denoising function $m(y, \sigma) := \mathbb{E}[x|W; x + \sigma g = y]$ can be approximated by AMP for all σ. Furthermore, an addition procedure of Natural Gradient Descent is applied on the AMP estimate of denoising m such that certain Lipschitz conditions are met.

Posterior Sampling from the Spiked Model ([MW23]) Similar to the SK model, suppose now one has a $\theta \sim \{-1, 1\}^n$ sampled from some $\mathbb{P}_\theta$ and observes:

$$A = \frac{\beta}{n}\theta\theta^T + W$$

and tries to sample θ given A. Such posterior distribution can be written as:

$$\mathbb{P}(\theta|A) \propto \exp(-\frac{\beta}{2}\langle x, Ax\rangle).$$

The more interesting setting is when $\mathbb{P}_\theta$ is uniform, in which case the problem is known as the $\mathbb{Z}_2$-synchronization. Similar to the SK model, Glauber Dynamics was not known to mix beyond $\beta > 1/4$, which in this case is especially un-interesting since A is statistically indistinguishable from pure noise W when $\beta < 1$.

Similar to the above approach, the key lies in approximating the denoising function $m(y, \sigma) = \mathbb{E}[\theta | A, \theta + \sigma g = y]$. For $\mathbb{Z}_2$-synchronization, AMP is known to achieve Bayes optimal for all $\beta > 0$. However, it is unclear whether the algorithmic output $\hat{m}$ is Lipschitz and thus the guarantee cannot extend to all $\beta > 0$, but only for large enough β.

Learning Mixtures of Gaussians ([SCK23]) Let us now turn to a problem with different flavors: estimation of Gaussian Mixture Models. Consider:

$$q = \frac{1}{K} \sum_{i=1}^{K} \mathcal{N}(\mu_i, \mathbb{I}_d),$$

our goal is to given samples estimate $\{\mu_i\}$. While not a sampling task by itself, we show that gradient descent (with a warm start) on the DDPM objective (3) recovers $\{\mu_i\}$ up to additive error ε in $\mathrm{poly}(d, \varepsilon^{-1})$ when $K \in O(1)$. The key lies in observing that minimizing (3) in different regimes resembles classical learning algorithms such as the Expectation-Maximization and Spectral algorithms. Specifically, gradient descent closely approximates the power method on a large noise level, and E&M on a small noise level. This result gives guarantees of effectiveness in learning scores via commonly used methods such as Gradient Descent.

References

[CCL+23] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.

[Cel22] Michael Celentano. Sudakov-fernique post-amp, and a new proof of the local convexity of the tap free energy. *arXiv preprint arXiv:2208.09550*, 2022.

[EAMS22] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

[MW23] Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

[PB13] Wolfgang Paul and Jörg Baschnagel. *A Brief Survey of the Mathematics of Probability Theory*, pages 17–61. 01 2013.

[SCK23] Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *arXiv preprint arXiv:2307.01178*, 2023.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.