Given an L^2-accurate score estimate, SGMs can sample from (essentially) any data distribution

Haitong Lan

Academy of Mathematics and Systems Science

December 2,2024

Outline

Mathematical formulation of diffusion models

Denoising diffusion probabilistic model (DDPM) [Song, Ermon '19; *etc.*]

- Forward SDE: $\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X_0 \sim p_{\mathrm{data}}$
- Backward SDE:

$$\mathrm{d}X^{\leftarrow}_t = \{X^{\leftarrow}_t + 2\, s_{T-kh}(X^{\leftarrow}_{kh})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad X^{\leftarrow}_0 \sim \mathcal{N}(0, I)$$

- **Sources of error:**
  - ⇝ Initialization
  - ⇝ Score matching
  - ⇝ Discretization

Theory for DDPM

Theorem: Assume that

- $\nabla \log p_t$ is L-Lipschitz for all $t \geq 0$, and
- $\|s_t - \nabla \log p_t\|_{L^2(p_t)} \leq \varepsilon$ for all $t \geq 0$.

Then, DDPM can output a distribution which is $\widetilde{O}(\varepsilon)$-close in TV distance to p_{data} using $\widetilde{O}(L^2 d/\varepsilon^2)$ iterations.

Sitan Chen, **S.C.**, Jerry Li, Yuanzhi Li, Adil Salim, Anru R. Zhang '23,
Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions.

Polynomial-time guarantees without assuming log-concavity!

Proof

Let $\gamma = \mathcal{N}(0, I)$.

$$\begin{aligned}
&\mathrm{TV}(\text{ROU started at } p_T,\ \text{ALG started at } \gamma) \\
&\quad \le \mathrm{TV}(\text{ALG started at } p_T,\ \text{ALG started at } \gamma) \\
&\qquad + \mathrm{TV}(\text{ROU started at } p_T,\ \text{ALG started at } p_T)\,.
\end{aligned}$$

By *data processing* and convergence of OU, the first term is at most

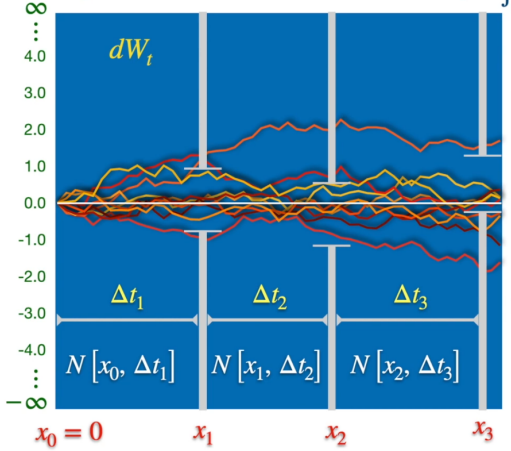
$$\mathrm{TV}(p_T, \gamma) \lesssim (d + \mathbb{E}_{p_{\text{data}}}[\|\cdot\|^2])\exp(-T)\,.$$

This handles the *initialization error*.

ROU:Reverse OU-process. ALG:Algorithm

Outline

$$P\left[\alpha_1 \le W_{t_1} \le \beta_1,\ \alpha_2 \le W_{t_2} \le \beta_2, \alpha_3 \le X_{t_3} \le \beta_3\right] = \int_{\alpha_1}^{\beta_1}\int_{\alpha_2}^{\beta_2}\int_{\alpha_3}^{\beta_3}\prod_{i=1}^{3}\frac{1}{\sqrt{2\pi\Delta t_i}}\, e^{-\frac{1}{2\Delta t_i}\left(x_i - x_{i-1}\right)^2} dx_i$$



$$W_t \sim N\left[0,\ t\right]$$

$$dP\left(W_t\right) = \frac{1}{\sqrt{2\pi t}} e^{-W_t^2/2t}\, dW_t$$

$$dP\left(x\right) = \frac{dP}{dx}dx \ = f(x)\, dx$$

Theorem (Kolmogorov's extension theorem).

For all $t_1, \cdots, t_k \in T, k \in \mathrm{N}$ let $\nu_{t,\cdots,t_k}$ be probability measures on $\mathbf{R}^{nk}$ s.t.

$$\nu_{t_{\sigma(1)},\cdots,t_{\sigma(k)}}(F_1 \times \cdots \times F_k) = \nu_{t_1,\cdots,t_k}\left(F_{\sigma^{-1}(1)} \times \cdots \times F_{\sigma^{-1}(k)}\right)$$

for all permutations σ on $\{1, 2, \cdots, k\}$ and

$$\nu_{t_1,\ldots,t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1,\cdots,t_k,t_{k+1},\cdots,t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbf{R}^n \times \cdots \times \mathbf{R}^n)$$

for all $m \in \mathbf{N}$, where (of course) the set on the right hand side has a total of $k+m$ factors. Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $\{X_t\}$ on $\Omega, X_t : \Omega \to \mathbf{R}^n$, s.t.

$$\nu_{t_1},\cdots,t_k(F_1 \times \cdots \times F_k) = P[X_{t_1} \in F_1, \cdots, X_{t_k} \in F_k]$$

for all $t_i \in T, k \in \mathbf{N}$ and all Borel sets F_i.

Theorem(Kolmogorov's extension theorem for markov process)

Let $\nu \in Prob(S, \mathcal{B})$ and let $\{Q_{s,t}\}_{s \leq t}$ be Markov trsansition operaters on $(S, \mathcal{B})^2$.Then there exist a unique probability measure

$$\mathbb{P}^{\nu} \in Prob(S^T, \mathcal{B}^{\otimes T})$$

s.t.$X_t(\omega) = \omega(t)$ is a Markov process on$(S^T, \mathcal{B}^{\otimes T}, \mathbb{P}^{\nu})$ with trsanition operaters$\{Q_{s,t}\}_{s \leq t}$and $X_0 \sim \nu$.

Proof of consistency(informal). Assume $B_0, B_2 \in S$,then the marginal measure on t_0 and t_2 is

$$\begin{aligned} &\mathbb{P}^{\nu}_{t_0,t_1,t_2}(B_0 \times S \times B_2) \\ &= \int_{B_0} \nu(dx_0) \int_S q_{t_0,t_1(x_0,dx_1)} \int_{B_2} q_{t_1,t_2}(x_1, dx_2) \end{aligned}$$

$$
\begin{aligned}
&= \int_{B_0 \times B_2} \nu(dx_0) \int_S q_{t_0,t_1}(x_0, dx_1) q_{t_1,t_2}(x_0, dx_2) (\text{Tonelli's theorem}) \\
&= \int_{B_0 \times B_2} \nu(dx_0) q_{t_0,t_2}(x_0, dx_2) (\text{Chapman–Kolmogorov equation}) \\
&= \mathbb{P}^{\nu}_{t_0,t_2}(B_0 \times B_2)
\end{aligned}
$$

□

The Chanpman-kolmogorov equation imply the consistency. **It is not the C-K equation satisfies the Markov process, but the Markov process satisfies the C-K equation!**[1] In fact, the more general C-K equation describes exactly this consistency:

Suppose that $\{ f_i \}$ is an indexed collection of random variables, that is, a stochastic process. Let

$$p_{i_1,\ldots,i_n}(f_1, \ldots, f_n)$$

be the joint probability density function of the values of the random variables f_1 to f_n. Then, the Chapman–Kolmogorov equation is

$$p_{i_1,\ldots,i_{n-1}}(f_1, \ldots, f_{n-1}) = \int_{-\infty}^{\infty} p_{i_1,\ldots,i_n}(f_1, \ldots, f_n)\, df_n$$

i.e. a straightforward marginalization over the nuisance variable.

(Note that nothing yet has been assumed about the temporal (or any other) ordering of the random variables—the above equation applies equally to the marginalization of any of them.)

[1] for details: Introduction to Stochastic Integration,Hui-Hsiung Kuo,Section 10.5

- From this theorem,we see that a Markov process is determined by its initial distribution and transition probabilities satisfying the Chapman-Kolmogorov equation.
- Conversely,if ν is aprobability measure on $\mathbb{R}$ and $\{P_{s,t}(x,\cdot); s<t, x\in\mathbb{R}\}$ is a collection of probability measures satisfying the Chapman-Kolmogorov equation,then there exists a Markov process X_t with initical distribution ν .

Theorem(Solution of SDE is Markov process)

Let $\sigma(t,x)$ and $f(t,x)$ be measurable functions on $[a,b]\times\mathbb{R}$ satisfying the Lipschitz and linear growth conditions in x. Suppose ξ is an $\mathcal{F}_a$-measurable random variable with $E\left(\xi^2\right)<\infty$. Then the unique continuous solution of the stochastic integral equation

$$X_t=\xi+\int_a^t \sigma\left(s,X_s\right)dB(s)+\int_a^t f(s,X_s)\,ds,\quad a\le t\le b$$

is a Markov process.

BTW,if ξ is not a r.v.,but is a real number that $\xi\in\mathbb{R}$,Then X_t is a stationary Markov process.

Ex. Transition probability of OU process

Examples

What is the transition probability (transition kernel) of the Langevin equation:

$$dX_t = \alpha dB_t - \beta X_t dt$$

First, we need the ito's lemma[2][3] to solve this SDE.

Theorem(Ito's lemma).

Let X_t be an Ito process given by $dX_t = udt + vdB_t$. Let $g(t,x) \in C^2([0,\infty) \times \mathbf{R})$ (i.e. g is twice continuously differentiable on $[0,\infty) \times \mathbf{R}$). Then $Y_t = g(t, X_t)$ is again an Ito process, and

$$dY_t = \frac{\partial g}{\partial t}(t, X_t)\, dt + \frac{\partial g}{\partial x}(t, X_t)\, dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t) \cdot (dX_t)^2$$

[2]Strict proof: Introduction to Stochastic Integration,Hui-Hsiung Kuo,Section 7.1

[3]Normal proof:Stochastic Differential Equations,Bernt Oksendal,Section 4.1

Solution.

$$\begin{aligned} d\left(e^{\beta t}X_t\right) &= \beta e^{\beta t}X_t dt + e^{\beta t}dX(t) \\ &= \beta e^{\beta t}X_t dt + e^{\beta t}\left(\alpha dB(t) - \beta X_t dt\right) \\ &= \alpha e^{\beta t}dB(t) \end{aligned}$$

Then convert this stochastic differential into a stochastic integral,

$$e^{\beta t}X_t = e^{\beta s}X_s + \int_s^t \alpha e^{\beta u}dB(u), \quad s \leq t$$

Therefore, X_t is given by

$$X_t = e^{-\beta(t-s)}X_s + \alpha \int_s^t e^{-\beta(t-u)}dB(u), \quad s \leq t$$

In particular, when $s = 0$, we get the solution of the Langevin equation :

$$X_t = e^{-\beta t} x_0 + \alpha \int_0^t e^{-\beta(t-u)} dB(u)$$

Then,we can use the ito isometric:

$$X_t = e^{-\beta t} x_0 + N\left(0, \frac{\alpha^2}{2\beta}\left(1 - e^{-2\beta(t-0)}\right)\right)$$

We've already figured out the transfer kernel:

$$\begin{aligned} P(X_t \in dx_t \mid X_s = x_s) &= p_{s,t}(x_s, dx_t) \\ &= \int_{dx_t} N\left(x; e^{-\beta t}, \frac{\alpha^2}{2\beta}\left(1 - e^{-2\beta(t-s)}\right)\right) dx \end{aligned}$$

□

Remark. Although the initial value of the OU process may not be constant, its transition probability is only related to $(t - s)$, so the OU process is a stationary Markov process.

Definition (f-divergence)

Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a convex function such that:
(i) $f(1) = 0$ (ii) f is strictly convex around 1, i.e.,

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y)$$

for all $x, y \in \mathbb{R}_{\geq 0}$ and $\alpha \in [0,1]$ such that $\alpha x + (1-\alpha)y = 1$. Let $P, Q \in \mathcal{P}(\mathcal{X})$ be two probability measures on $\mathcal{X}$, and let $\lambda \in \mathcal{M}_+(\mathcal{X})$ be a measure that dominates them both, i.e., $P, Q \ll \lambda$. The f-divergence between Q and P is defined as

$$D_f(P\|Q) := \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P/\mathrm{d}\lambda}{\mathrm{d}Q/\mathrm{d}\lambda}\right)\right] = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P/\mathrm{d}\lambda}{\mathrm{d}Q/\mathrm{d}\lambda}\right)\mathrm{d}Q(x)$$

where $\mathrm{d}P/\mathrm{d}\lambda$ and $\mathrm{d}Q/\mathrm{d}\lambda$ are the Radon-Nikodym derivatives of P and Q, respectively, w.r.t λ.

If $P \ll Q$, then $D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of P w.r.t. Q.

The **Kullback-Leibler (KL) divergence** (also sometimes referred to as relative entropy or information divergence) is the f-divergence induced by $f(x) = x\log x$. Namely, the KL divergence of Q from P is

$$D_{\mathrm{KL}}(P\|Q) = D_{x\log x}(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP/d\lambda}{dQ/d\lambda}\right)\right] = \mathbb{E}_P\left[\log\left(\frac{dP/d\lambda}{dQ/d\lambda}\right)\right].$$

The **Total Variation (TV) distance**, is the f-divergence induced by $f(x) = \frac{1}{2}|x-1|$. Namely, the TV distance between Q and P is

$$\delta_{\mathrm{TV}}(P,Q) = D_{\frac{1}{2}|x-1|}(P\|Q) = \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{\mathrm{d}P/\mathrm{d}\lambda}{\mathrm{d}Q/\mathrm{d}\lambda} - 1\right|\right] = \frac{1}{2}\int_{\mathcal{X}}\left|\frac{\mathrm{d}P}{\mathrm{d}\lambda} - \frac{\mathrm{d}Q}{\mathrm{d}\lambda}\right|.$$

Equivalent Definition (TV distance).

Consider a measurable space $(\Omega, \mathcal{F})$ and probability measures P and Q defined or $(\Omega, \mathcal{F})$. The total variation distance between P and Q is defined as

$$\mathrm{TV}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

This is the largest absolute difference between the probabilities that the two probability distributions assign to the same event.

Theorem(Pinsker's inequality).

If P and Q are two probability distributions on a measurable space $(\Omega, \mathcal{F})$, then

$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P \| Q)}$$

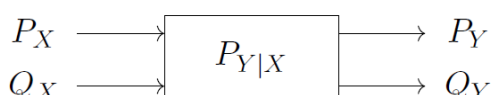
Data Processing Inequality

Theorem (Data Processing Inequality)

Let $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ and $P_{Y|X}$ be a **transition kernel** from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$. Let $P_Y, Q_Y \in \mathcal{P}(\mathcal{Y})$ be the transformation of P_X and Q_X, respectively, when pushed through $P_{Y|X}$, i.e., $P_X(B) = \int_{\mathcal{X}} P_{Y|X}(B \mid x) \mathrm{d}P_X(x)$. Then, for any f-divergence, we have that

$$D_f(P_X \| Q_X) \geq D_f(P_Y \| Q_Y) .$$

This can be thought of as follows: pushing two observations X and Y through a channel will only make it **harder to distinguish** between them.



Outline

Definition(Diffusion process).

An $\mathbb{R}^n$-valued Markov process X_t, $a \leq t \leq b$, is called a diffusion process if satisfy the following three conditions for any $t \in [a, b], x \in \mathbb{R}^n$, and $c > 0$:

(1) $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} P\left(|X_{t+\varepsilon} - x| > c \mid X_t = x\right) = 0.$

(2) $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}\left[X_{t+\varepsilon} - x \mid X_t = x\right] = \rho(t, x)$ exists.

(3) $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}\left[\left(X_{t+\varepsilon} - x\right)\left(X_{t+\varepsilon} - x\right)^T \mid X_t = x\right] = Q(t, x)$ exists.

Theorem (Solution of SDE is a diffusion process).

Let $\sigma(t, x)$ and $f(t, x)$ be functions satisfie Lipschitz's conditions and linear growth condition. Assume that $\sigma(t, x)$ and $f(t, x)$ are continuous on $[a, b] \times \mathbb{R}^n$. Then the solution X_t of $X_t = x_a + \int_a^t \sigma(s, X_s) dB_s + \int_a^t f(s, X_s) ds$ is a diffusion process with diffusion coefficient $Q(t, x)$ and drift $\rho(t, x)$ given by

$$Q(t, x) = \sigma(t, x)\sigma(t, x)^T, \quad \rho(t, x) = f(t, x)$$

Diffusion model

Forward process. In DDPM, we start with a SDE. For clarity, we consider the simplest possible choice, which is the Ornstein-Uhlenbeck (OU) process

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

where $(B_t)_{t>0}$ is a standard Brownian motion in $\mathbb{R}^d$.
Reverse process.If we reverse the forward process in time, then we obtain a process that transforms noise into samples from q, which is the aim of generative modeling.

$$\mathrm{d}\bar{X}_t^{\leftarrow} = \left\{ \bar{X}_t^{\leftarrow} + 2\nabla \ln q_{T-t}\left(\bar{X}_t^{\leftarrow}\right) \right\} \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t, \quad \bar{X}_0^{\leftarrow} \sim q_T$$

where now $(B_t)_{t\in[0,T]}$ is the reversed Brownian motion. Here, $\nabla \ln q_t$ is called the score function for q_t.

Score matching. In order to estimate the score function $\nabla \ln q_t$, consider minimizing the $L^2(q_t)$ loss over a function class $\mathcal{F}$,

$$\underset{s_t \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}_{q_t}\left[\|s_t - \nabla \ln q_t\|^2\right],$$

where $\mathcal{F}$ could be, e.g., a class of neural networks.

Score-based Generative Model(SGM). In order to approximately implement the reverse SDE , 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

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

In particular, conditionally on $X_{kh}^{\leftarrow}$, the next iterate $X_{(k+1)h}^{\leftarrow}$ has an explicit Gaussian distribution.

Notation

Stochastic processes and their laws.

- The data distribution is $q = q_0$.
- The forward process is denoted $\left(\bar{X}_t\right)_{t\in[0,T]}$, and $\bar{X}_t \sim q_t$.
- The reverse process is denoted $\left(\bar{X}_t^{\leftarrow}\right)_{t\in[0,T]}$, where $\bar{X}_t^{\leftarrow} := \bar{X}_{T-t} \sim q_{T-t}$.
- The SGM algorithm is denoted $\left(X_t^{\leftarrow}\right)_{t\in[0,T]}$, and $X_t^{\leftarrow} \sim p_t$. Recall that we initialize at $p_0 = \gamma^d$, the standard Gaussian measure.
- The process $\left(X_t^{\leftarrow,q_T}\right)_{t\in[0,T]}$ is the same as $\left(X_t^{\leftarrow}\right)_{t\in[0,T]}$, except that we initialize this process at q_T rather than at γ^d. We write $X_t^{\leftarrow,q_T} \sim p_t^{q_T}$.

Assumption

- Assumption 1 (Lipschitz score). For all $t \geq 0$, the score $\nabla \ln q_t$ is L-Lipschitz.
- Assumption 2 (second moment bound). We assume that $\mathfrak{m}_2^2 := \mathbb{E}_q \left[\| \cdot \|^2 \right] < \infty$.
- Assumption 3 (score estimation error). For all $k = 1, \ldots, N$,

$$\mathbb{E}_{q_{kh}} \left[\| s_{kh} - \nabla \ln q_{kh} \|^2 \right] \leq \varepsilon_{\text{score}}^2$$

Outline

Sampling assurance for DDPM

Theorem 2 (DDPM).

Suppose that Assumptions 1, 2, and 3 hold. Let p_T be the output of the DDPM algorithm (Section 2.1) at time T, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$, where $L \geq 1$. Then, it holds that

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

To interpret this result, suppose that $\mathrm{KL}\left(q \| \gamma^d\right) \leq \mathrm{poly}(d)$ and $\mathfrak{m}_2 \leq d$. 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,

$$\mathrm{TV}\left(p_T, q\right) \leq \widetilde{O}\left(\varepsilon + \varepsilon_{\text{score}}\right), \qquad \text{for } N = \widetilde{\Theta}\left(\frac{L^2 d}{\varepsilon^2}\right)$$

In particular, in order to have $\mathrm{TV}\left(p_T, q\right) \leq \varepsilon$, it suffices to have score error $\varepsilon_{\text{score}} \leq \widetilde{O}(\varepsilon)$.

Proof (informal). By data processing inequality

$$\mathrm{TV}\left(p_T, q\right) \leq \mathrm{TV}\left(P_T, Q_T^{\leftarrow}\right) \tag{1}$$

Here, the distributions p_T, q can be viewed as marginal distributions of the distributions $P_T, Q_T^{\leftarrow}$ on the path space. Given the set $A \subseteq \mathbb{R}^d$, there is

$$p_t(A) = P_T\left(\left\{\omega \in C\left([0, T], \mathbb{R}^d\right) : X_t(\omega) \in A\right\}\right) \tag{2}$$

The state distribution can then be obtained from the path measure by an integral, i.e

$$p_t(A) = \int_{\left\{\omega \in C\left([0,T],\mathbb{R}^d\right) : X_t(\omega) \in A\right\}} P_T(d\omega) \tag{3}$$

Using triangle inequality and data processing inequality

$$\begin{aligned}
\mathrm{TV}\left(p_T, q\right) &\leq \mathrm{TV}\left(P_T, Q_T^{\leftarrow}\right) \\
&\leq \mathrm{TV}\left(P_T, P_T^{q_T}\right) + \mathrm{TV}\left(P_T^{q_T}, Q_T^{\leftarrow}\right) \\
&\leq \mathrm{TV}\left(q_T, \gamma^d\right) + \mathrm{TV}\left(P_T^{q_T}, Q_T^{\leftarrow}\right) \\
&\leq \sqrt{\mathrm{KL}\left(q \| \gamma^d\right)} \exp(-T) + \mathrm{TV}\left(P_T^{q_T}, Q_T^{\leftarrow}\right)
\end{aligned} \tag{4}$$

Where $\mathrm{TV}\left(P_T, P_T^{q_T}\right) \leq \mathrm{TV}\left(q_T, \gamma^d\right)$is because of the data processing inequality, The path distribution $P_T, P_T^{q_T}$is induced by **the same stochastic process evolution**(i.e. through the same transfer kernel), and the path measure can be determined by Kolmogorov's expansion theorem.

Next,we use **Girsanov's theorem** to bound $\mathrm{TV}\left(P_T^{q_T}, Q_T^{\leftarrow}\right)$.

Girsanov's theorem

Theorem 8(Girsanov's theorem).

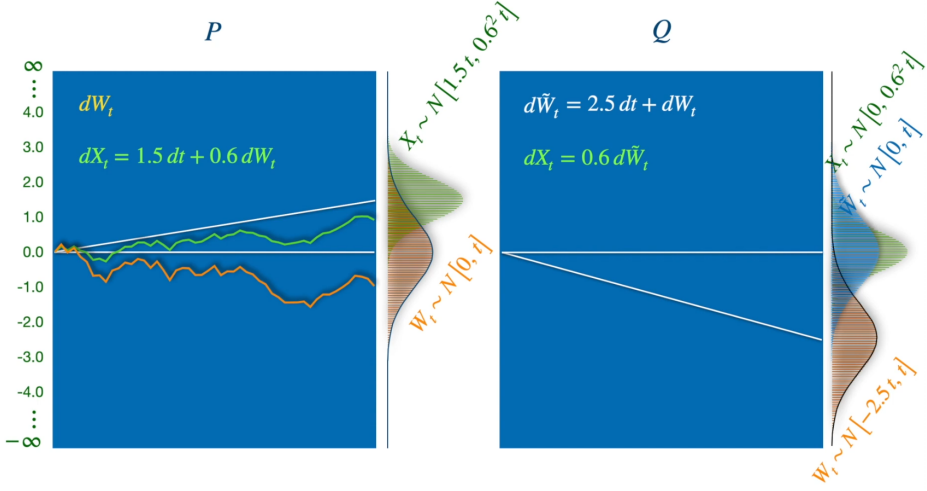
For $t \in [0, T]$, let $\mathcal{L}_t = \int_0^t b_s \, \mathrm{d}B_s$ where B_t is a Q-Brownian motion. Assume $\mathbb{E}_Q \int_0^T \|b_s\|^2 \, \mathrm{d}s < \infty$. Then, $\mathcal{L}_t$ is a Q-martingale in $L^2(Q)$. Moreover, if

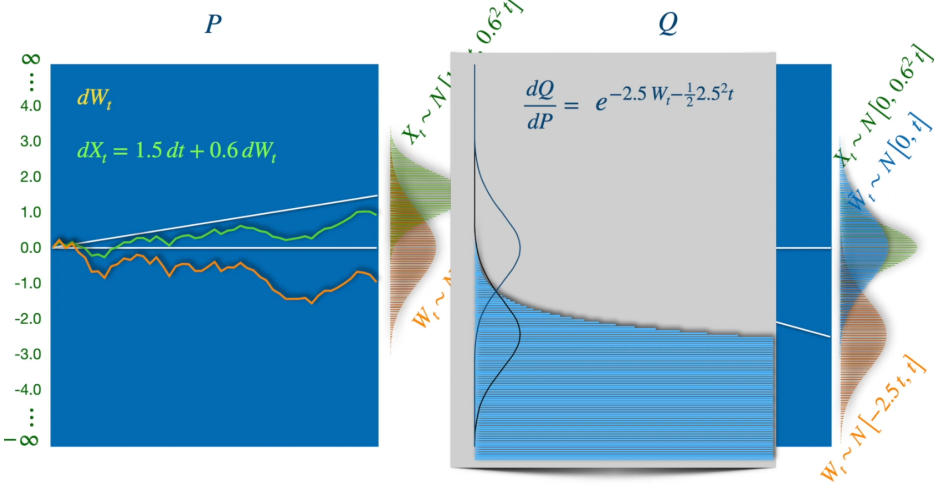
$$\mathbb{E}_Q \mathcal{E}(\mathcal{L})_T = 1, \quad \text{where} \quad \mathcal{E}(\mathcal{L})_t := \exp\left(\int_0^t b_s \, \mathrm{d}B_s - \frac{1}{2} \int_0^t \|b_s\|^2 \, \mathrm{d}s \right)$$

then $\mathcal{E}(\mathcal{L})$ is also a Q-martingale and the process

$$t \mapsto B_t - \int_0^t b_s \, \mathrm{d}s$$

is a Brownian motion under $P := \mathcal{E}(\mathcal{L})_T Q$, the probability distribution with density $\mathcal{E}(\mathcal{L})_T$ w.r.t. Q.


P
Q
dW_t
$dX_t = 1.5\,dt + 0.6\,dW_t$
$d\tilde{W}_t = 2.5\,dt + dW_t$
$dX_t = 0.6\,d\tilde{W}_t$
$X_t \sim N[1.5t, 0.6^2 t]$
$W_t \sim N[0, t]$
$X_t \sim N[0, 0.6^2 t]$
$\tilde{W}_t \sim N[0, t]$
$W_t \sim N[-2.5t, t]$
∞ ⋯
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
⋯ −∞


P
Q
dW_t
$dX_t = 1.5\,dt + 0.6\,dW_t$
$\frac{dQ}{dP} = e^{-2.5\,W_t - \frac{1}{2}2.5^2 t}$
$X_t \sim N[\ldots, 0.6^2 t]$
$W_t \sim N[0, t]$
$X_t \sim N[0, 0.6^2 t]$
$W_t \sim N[-2.5\,t, t]$
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0

Conventions for Girsanov's theorem. When we apply Girsanov's theorem, it is convenient to instead think about a single stochastic process, which for ease of notation we denote simply via $(X_t)_{t\in[0,T]}$, and we consider different measures over the path space $\mathcal{C}\left([0,T];\mathbb{R}^d\right)$.

The three measures we consider over path space are:

- $Q_T^{\leftarrow}$, under which $(X_t)_{t\in[0,T]}$ has the law of the reverse process

$$\mathrm{d}X_t = \{X_t + 2\nabla \ln q_{T-t}(X_t)\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad X_0 \sim q_T$$

- $P_T^{q_T}$, under which $(X_t)_{t\in[0,T]}$ has the law of the SGM algorithm initialized at q_T (corresponding to the process $\left(X_t^{\leftarrow,q_T}\right)_{t\in[0,T]}$ defined above).

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

The SDE described by ROU and ALG is valid under different path measures, but the path evolution of Brownian motion is not consistent, now we consider the measure transformation from $Q_T^{\leftarrow}$ to P_T.
Consider the SDE described by ALG:

$$\begin{aligned} dX_t &= \{X_t + 2s_{T-kh}(X_{kh})\}\, dt + \sqrt{2} d\beta_t \\ &= \{X_t + 2\nabla \ln q_{T-t}(X_t)\}\, dt + \{2s_{T-kh}(X_{kh}) - 2\nabla \ln q_{T-t} + (X_t)\}\, dt + \sqrt{2} d\beta_t \end{aligned} \tag{5}$$

The SDE described by ROU:

$$\mathrm{d}X_t = \{X_t + 2\nabla \ln q_{T-t}(X_t)\}\, \mathrm{d}t + \sqrt{2}\ \mathrm{d}B_t, \quad X_0 \sim q_T \tag{6}$$

Where $d\beta_t$is the differential of the standard **Brownian motion on the measure** P. If we can make the blue parts equal. Then it is possible to find an adaptive process connecting two Brownian motions according to Girsanov's theorem.

Let

$$\left\{2 s_{T-kh}\left(X_{kh}\right)-2 \nabla \ln q_{T-t}+\left(X_t\right)\right\} d t+\sqrt{2} d \beta_t=\sqrt{2} d B_t \tag{7}$$

By Girsanov's theorem, it can be obtained

$$b_t=\sqrt{2}\left(s_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(X_t\right)\right) \tag{8}$$

We can then calculate the KL divergence between path measures.

$$\begin{aligned} KL\left(Q_T^{\leftarrow} \| P_T^{q_T}\right) & =\mathbb{E}_{Q_T^{\leftarrow}} \ln \frac{d Q_T^{\leftarrow}}{d P_T^{q_T}} \\ & =\mathbb{E}_{Q_T^{\leftarrow}} \ln \mathcal{E}(\mathcal{L})_T^{-1} \\ & =\sum_{k=0}^{N-1} \mathbb{E}_{Q_T^{\leftarrow}} \int_{kh}^{(k+1) h}\left\|s_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(x_t\right)\right\|^2 d t . \end{aligned} \tag{9}$$

For $t \in [kh, (k+1)h]$, decompose this formula

$$
\begin{aligned}
&\mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|s_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(X_t\right)\right\|^2\right] \\
&\lesssim \mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|s_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-kh}\left(X_{kh}\right)\right\|^2\right] \\
&\quad+\mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|\nabla \ln q_{T-kh}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(X_{kh}\right)\right\|^2\right] \\
&\quad+\mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|\nabla \ln q_{T-t}\left(X_{kh}\right)-\nabla \ln q_{T-t}\left(X_t\right)\right\|^2\right] \\
&\lesssim \varepsilon_{\text{score}}^2+\mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|\nabla \ln \frac{q_{T-kh}}{q_{T-t}}\left(X_{kh}\right)\right\|^2\right]+L^2 \mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|X_{kh}-X_t\right\|^2\right]
\end{aligned}
\tag{10}
$$

The third term can be easily bound, and then we deal with the blue formula, which we call the **perturbation error**.

The OU process:

$$\mathrm{d}\bar{X}_t = -\bar{X}_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t$$

can be solved to obtain

$$d(e^t X_t) = \sqrt{2} e^t dB_t \tag{11}$$

Integrate it from s to t

$$X_t = e^{s-t} X_s + \int_s^t \sqrt{2} e^t dB_t \tag{12}$$

The ito integral of the second term is the integral of a deterministic function, which follows the **Gaussian distribution**, and the variance can be obtained using **ito isometric**, and X_t can be rewritten as

$$X_t = e^{s-t} X_s + \sqrt{1 - e^{2(s-t)}} \xi, \xi \sim N(0, 1) \tag{13}$$

If we integrate the formula (11) from $(T-t)$ to $(T-kh)$, we can relate the numerator and denominator in the perturbation error, i.e

$$\begin{aligned} X_{T-kh} &= e^{T-t-(T-kh)} X_{T-t} + \sqrt{1-e^{2(T-t-(T-kh))}}\xi \\ &= e^{kh-t} X_{T-t} + \sqrt{1-e^{2(kh-t)}}\xi \\ &= S_{\#} q_{T-t} * N(0, 1-e^{2(kh-t)}) \end{aligned} \tag{14}$$

Where $S : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the mapping, $S(x) := exp(-(t-kh))x$

Decompose the perturbation error, term

$$
\begin{aligned}
&\left\| \nabla \ln \frac{q_{T-kh}}{q_{T-t}}\left(x_{kh}\right) \right\| \\
=&\left\| \nabla \ln \frac{S_{\#} q_{T-t} * N}{q_{T-t}}\left(x_{kh}\right) \right\| \\
=&\left\| \nabla \ln \frac{S_{\#} q_{T-t} * N \cdot S_{\#} q_{T-t}}{q_{T-t} \cdot S_{\#} q_{T-t}}\left(x_{kh}\right) \right\| \\
\leq&\left\| \nabla \ln \frac{S_{\#} q_{T-t} * N}{S_{\#} q_{T-t}}\left(x_{kh}\right) \right\| + \left\| \nabla \ln \frac{S_{\#} q_{T-t}}{q_{T-t}}\left(x_{kh}\right) \right\|
\end{aligned}
\tag{15}
$$

The second term can be solved directly, and now let's analyze the first term.

$$\nabla \ln S_{\#} q_{T-t} * N_{\sigma^2}(x) = \frac{\int_{\mathbb{R}^d} -\nabla V(y) e^{-V(y)} e^{-\frac{\|x-y\|^2}{2\sigma^2}} dy}{\int_{\mathbb{R}^d} e^{-V(y)} e^{-\frac{\|x-y\|^2}{2\sigma^2}} dy} = -\mathbb{E}_{p_{x,\sigma^2}} \nabla V(y),$$

Which takes advantage of the $S_{\#} q_{T-T} \propto e^{-V(x)}$and convolution formula, p_{x,σ^2} is probability density:

$$p_{x,\sigma^2}(y) \propto p(y) e^{-\frac{\|y-x\|^2}{2\sigma^2}}$$

According to the hypothesis, the gradient of the potential function $\nabla V(x)$ is L-Lipschitz (i.e. $\nabla \ln q_t L$-Lipschitz),

$$\begin{aligned} \|\nabla \ln S_{\#} q_{T-t} * N_{\sigma^2}(x) - \nabla \ln S_{\#} q_{T-t}\| &= \left\| \mathbb{E}_{p_{x,\sigma^2}} [\nabla V(y) - \nabla V(x)] \right\| \\ &\leq \mathbb{E}_{p_{x,\sigma^2}} [L \|y - x\|] \end{aligned}$$

Next, by the lemma12 in *Convergence for score-based generative modeling with polynomial complexity*[4] ,We can finally bound it. □

[4] https://arxiv.org/abs/2206.06227

Advantages of using Girsanov's theorem

- Handle score errors under **real distribution**.
- The proof requires **the weakest hypothesis**.

Disadvantages of using Girsanov's theorem

- the final result is unsatisfying because we have **moved to a weaker metric** (TV rather than KL) for a seemingly silly reason (the failure of the triangle inequality for the KL divergence).
- we only able to establish a bound on KL which **grows with the iteration number** N,this is unsatisfying because "running the Markov chain too long" should not be a problem.[5]

[5] Log-Concave Sampling, https://chewisinho.github.io/main.pdf

Manifold hypothesis

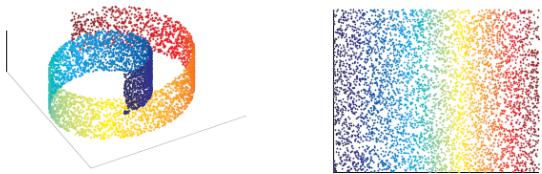


Figure 15.1
The "Swiss roll" dataset. (a) high-dimensional representation. (b) lower-dimensional representation.

The **manifold hypothesis**[6] posits that many high-dimensional data sets that occur in the real world actually lie along **low-dimensional latent manifolds** inside that high-dimensional space.

[6] Foundations of Machine Learning,Mehryar Mohri,Chapter 15

Consequences for arbitrary data distributions with bounded support

In general we cannot obtain non-trivial guarantees[7] for $TV(p_T, q)$, because p_T has full support and therefore $TV(p_T, q) = 1$ under the manifold hypothesis.

Corollary 3 (compactly supported data)

Suppose that q is supported on the ball of radius $R \geq 1$. Let $t \asymp \varepsilon_{W_2}^2/(\sqrt{d}(R \vee \sqrt{d}))$. Then, the output p_{T-t} of DDPM is $\varepsilon_{\mathrm{TV}}$-close in TV to the distribution $\bar{q}_t$, which is ε_{W_2}-close in W_2 to q, provided that the step size h is chosen appropriately according to Theorem 2 and

$$N = \widetilde{\Theta}\left(\frac{d^3 R^4 (R \vee \sqrt{d})^4}{\varepsilon_{\mathrm{TV}}^2 \varepsilon_{W_2}^8}\right) \quad \text{and} \quad \varepsilon_{\mathrm{score}} \leq \widetilde{O}(\varepsilon_{\mathrm{TV}})$$

[7] Convergence of denoising diffusion models under the manifold hypothesis, https://arxiv.org/abs/2208.05314

If the output p_{T-t} of DDPM at time $T-t$ is projected onto $\mathrm{B}(0, R_0)$ for an appropriate choice of R_0, then we can also translate our guarantees to the standard W_2 metric, which we state as the following corollary.

Corollary 5 (compactly supported data, W_2 metric)

Suppose that q is supported on the ball of radius $R \geq 1$. Let $t \asymp \varepsilon^2/(\sqrt{d}(R \vee \sqrt{d}))$, and let p_{T-t,R_0} denote the output of DDPM at time $T-t$ projected onto $\mathrm{B}(0, R_0)$ for $R_0 = \widetilde{\Theta}(R)$. Then, it holds that

$$W_2\left(p_{T-t,R_0}, q\right) \leq \varepsilon$$

, provided that the step size h is chosen appropriately according to Theorem 2, $N = \widetilde{\Theta}\left(d^3 R^8 (R \vee \sqrt{d})^4/\varepsilon^{12}\right)$, and $\varepsilon_{\text{score}} \leq \widetilde{O}(\varepsilon)$.

Outline

Critically damped Langevin diffusion (CLD)

The critically damped Langevin diffusion (CLD) is based on the forward process

$$\begin{aligned} \mathrm{d}\bar{X}_t &= -\bar{V}_t \, \mathrm{d}t \\ \mathrm{d}\bar{V}_t &= -\left(\bar{X}_t + 2\bar{V}_t\right) \mathrm{d}t + 2 \, \mathrm{d}B_t \end{aligned}$$

The corresponding reverse process is

$$\begin{aligned} \mathrm{d}\bar{X}_t^{\leftarrow} &= -\bar{V}_t^{\leftarrow} \mathrm{d}t \\ \mathrm{d}\bar{V}_t^{\leftarrow} &= \left(\bar{X}_t^{\leftarrow} + 2\bar{V}_t^{\leftarrow} + 4\nabla_v \ln \boldsymbol{q}_{T-t}\left(\bar{X}_t^{\leftarrow}, \bar{V}_t^{\leftarrow}\right)\right) \mathrm{d}t + 2 \, \mathrm{d}B_t \end{aligned}$$

where $\boldsymbol{q}_t := \mathrm{law}\left(\bar{X}_t, \bar{V}_t\right)$ is the law of the forward process at time t. Note that the gradient in the score function is only taken w.r.t. the velocity coordinate.

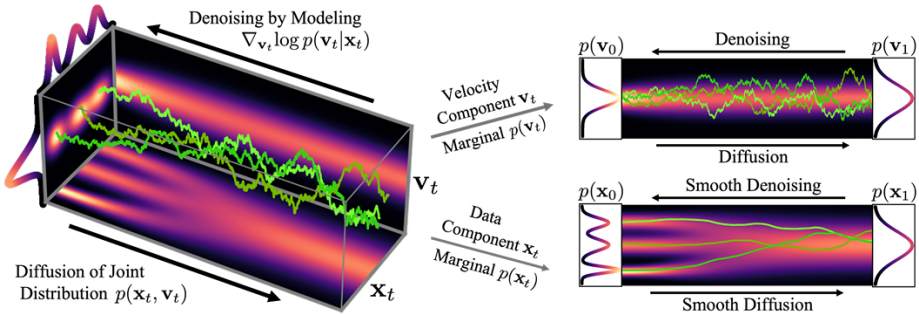


Figure 1: In critically-damped Langevin diffusion, the data $\mathbf{x}_t$ is augmented with a velocity $\mathbf{v}_t$. A diffusion coupling $\mathbf{x}_t$ and $\mathbf{v}_t$ is run in the joint data-velocity space (probabilities in **red**). Noise is injected only into $\mathbf{v}_t$. This leads to smooth diffusion trajectories (**green**) for the data $\mathbf{x}_t$. Denoising only requires $\nabla_{\mathbf{v}_t} \log p(\mathbf{v}_t|\mathbf{x}_t)$.

Assumption 4. For all $t \geq 0$, the score $\nabla_v \ln \boldsymbol{q}_t$ is L-Lipschitz.

Assumption 5. For all $k = 1, \ldots, N$,

$$\mathbb{E}_{\boldsymbol{q}_{kh}} \left[\left\| s_{kh} - \nabla_v \ln \boldsymbol{q}_{kh} \right\|^2 \right] \leq \varepsilon_{\text{score}}^2 \,.$$

Theorem 6 (CLD).

Suppose that Assumptions 2, 4, and 5 hold. Let $\boldsymbol{p}_T$ be the output of the SGM algorithm based on the CLD (Section 2.2) at time T, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$, where $L \geq 1$. Then, there is a universal constant $c > 0$ such that

$$\mathrm{TV}\left(\boldsymbol{p}_T, q \otimes \gamma^d\right) \lesssim \underbrace{\sqrt{\mathrm{KL}\left(q \| \gamma^d\right) + \mathrm{FI}\left(q \| \gamma^d\right)} \exp(-cT)}_{\text{convergence of forward process}}$$
$$+ \underbrace{\left(L\sqrt{dh} + L\mathfrak{m}_2 h\right)\sqrt{T}}_{\text{discretization error}} + \underbrace{\varepsilon_{\text{score}}\sqrt{T}}_{\text{score estimation error}}$$

where $\mathrm{FI}\left(q \| \gamma^d\right)$ is the relative Fisher information $\mathrm{FI}\left(q \| \gamma^d\right) := \mathbb{E}_q \left[\left\| \nabla \ln \left(q/\gamma^d\right) \right\|^2 \right]$.

Remark.

- Under our assumptions, the CLD does not improve the complexity of SGMs over DDPM.
- "When comparing models with similar network capacity and under NFE(number of function—neural network—evaluations) budgets < 500,our CLD-SGM **outperforms all published results** in terms of FID.We attribute these positive results to our **easier score matching task**."[8]
- In fact,if the distribution is assumed to satisfy the **log-concave condition**, it can be shown that CLD can effectively reduce the computational complexity and thus obtain better bound.

[8] https://arxiv.org/pdf/2112.07068

References

- High-Dimensional Statistics and Inference,Speaker: Sinho CHHEWI (Institute for Advanced Study, USA),https://www.youtube.com/watch?v=dtre50qhaZQ
- Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions,https://arxiv.org/abs/2209.11215
- Convergence for score-based generative modeling with polynomial complexity,https://arxiv.org/abs/2206.06227
- Simplified: Girsanov Theorem for Brownian Motion (Change of Probability Measure),https://www.youtube.com/watch? v=57iWm1ZfdyU&list=LL&index=1
- Log-Concave Sampling, https://chewisinho.github.io/main.pdf
- F-divergence: https://people.ece.cornell.edu/zivg/ECE_5630_Lectures6.pdf

The End