DIFFUSION PROBABILISTIC MODELS

WENPIN TANG

1. Setup

The key takeaway of *diffusion probabilistic models* (DPMs) is to *reverse* some meticulously chosen stochastic dynamics to create meaningful distributions from noise, thus achieving generative modeling. Given a data in $\mathbb{R}^n$ with distribution $p_{\text{data}}(\cdot)$, the task is to generate multiple (many) data sets whose distributions are $p_{\text{data}}(\cdot)$, or close to $p_{\text{data}}(\cdot)$.

Reverse SDE. Let us explain DPMs. Consider a (forward) stochastic differential equation (SDE):

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, \quad \text{with } X_0 \sim p_{\text{data}}(\cdot), \tag{1.1}$$

where $(B_t,\, t \ge 0)$ is d-dimensional Brownian motion, and $b : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n \times d}$ are model parameters to be designed. Some conditions on $b(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are required so that the SDE (1.1) is at least well-defined (see [15, Chapter 5]), and we will elaborate this later. Let $p(t, \cdot)$ be the probability density of X_t.

Set $T > 0$ to be fixed, and run the SDE (1.1) until time T to get $X_T \sim p(T, \cdot)$. The idea is that if we start with $p(T, \cdot)$ and run the process X backward, then we can generate copies of $p(0,\, \cdot) = p_{\text{data}}(\cdot)$. To be more precise, consider the time reversal $\overline{X}_t := X_{T-t}$ for $0 \le t \le T$. Assuming that $\overline{X}$ also satisfies an SDE, we can implement the backward procedure by

$$d\overline{X}_t = \bar{b}(t, \overline{X}_t)dt + \bar{\sigma}(t, \overline{X}_t)dB_t, \quad \text{with } \overline{X}_0 \sim p(T, \cdot). \tag{1.2}$$

So we generate the desired $\overline{X}_T \sim p_{\text{data}}(\cdot)$ at time T. There are two questions:

(1) How can we sample the initial distribution $p(T, \cdot)$ in (1.2)?

(2) What are the parameters $\bar{b}(\cdot, \cdot)$ and $\bar{\sigma}(\cdot, \cdot)$ (in terms of $b(\cdot, \cdot)$, $\sigma(\cdot, \cdot)$ and the distribution of X)?

For (1), the distribution $p(T, \cdot)$ of X_T depends on $p(0, \cdot) = p_{\text{data}}(\cdot)$, and it is generally hard to compute $p(T, \cdot)$ in closed-form. One way is to choose suitable model parameters $b(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ so that X_t converges to a target or prior distribution $p_\infty(\cdot)$. If $b(t, x) = b(x)$ and $\sigma(t, x) = \sigma(x)$ are time-independent, $p_\infty(\cdot)$ can be the stationary distribution of the SDE (1.1). Now instead of taking $p(T, \cdot)$ as the initial distribution for the backward process (1.2), we set

$$\overline{X}_0 \sim p_\infty(\cdot), \tag{1.3}$$

which is independent of $p(0, \dot{)} = p_{\text{data}}(\cdot)$. This explains why DPMs generate distributions from "noise" $p_\infty(\cdot)$. It should also be kept in mind that the model parameters $b(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are

Date: August 26, 2023.

chosen so that "noise" $p_\infty(\cdot)$ is easy to sample, e.g. uniform or Gaussian distributed, or as a product measure. The backward process then becomes

$$d\overline{X}_t = \bar{b}(t, \overline{X}_t)dt + \bar{\sigma}(t, \overline{X}_t)dB_t, \quad \text{with } \overline{X}_0 \sim p_\infty(\cdot). \tag{1.4}$$

Comparing (1.2) with (1.4), we see that closer the distributions $p(T, \cdot)$ and $p_\infty(\cdot)$ are, closer the distribution of $\overline{X}_T$ governed by (1.4) is to $p_{\text{data}}(\cdot)$. It requires either taking T to be sufficiently large, or performing a time-change or conditioning which we will illustrate with examples later.

For (2), it relies on the following general result on the time reversal of SDEs.

Theorem 1.1 (Time reversal formula)**.** [12] *Let $a(t,x) := \sigma(t,x)\sigma(t,x)^\top$. Under suitable conditions on $b(\cdot,\cdot)$, $\sigma(\cdot,\cdot)$ and $\{p(t,\cdot)\}_{0 \le t \le T}$, we have*

$$\overline{\sigma}(t,x) = \sigma(T-t,x), \quad \overline{b}(t,x) = -b(T-t,x) + \frac{\nabla \cdot (p(T-t,x)a(T-t,x))}{p(T-t,x)}, \tag{1.5}$$

where $\nabla \cdot f = divf$ (the divergence of f).

Proof. Here we give a heuristic derivation of the time reversal formula (1.5). First, the infinitesimal generator of X is $\mathcal{L} := \frac{1}{2}\nabla \cdot a(t,x)\nabla + b_a \cdot \nabla$, where $b_a := b - \frac{1}{2}\nabla \cdot a$. It is known that the density $p(t,x)$ satisfies the the FokkerPlanck equation:

$$\frac{\partial}{\partial t} p(t,x) = \mathcal{L}^* p(t,x), \tag{1.6}$$

where $\mathcal{L}^* := \frac{1}{2}\nabla \cdot a(t,x)\nabla - \nabla \cdot b_a$ is the adjoint of $\mathcal{L}$. Let $\overline{p}(t,x) := p(T-t,x)$ be the probability density of the time reversal $\overline{X}$. By (1.6), we get

$$\frac{\partial}{\partial t}\overline{p}(t,x) = -\frac{1}{2}\nabla \cdot \left(a(T-t,x)\,\nabla \overline{p}(t,x)\right) + \nabla \cdot \left(b_a(T-t,x)\,\overline{p}(t,x)\right). \tag{1.7}$$

On the other hand, we expect the generator of $\overline{X}$ to be $\overline{\mathcal{L}} := \frac{1}{2}\nabla \cdot \overline{a}(t,x)\nabla + \overline{b}_{\overline{a}} \cdot \nabla$. The Fokker-Planck equation for $\overline{p}(t,x)$ is

$$\frac{\partial}{\partial t}\overline{p}(t,x) = \frac{1}{2}\nabla \cdot \left(\overline{a}(t,x)\,\nabla \overline{p}(t,x)\right) - \nabla \cdot \left(\overline{b}_{\overline{a}}(t,x)\,\overline{p}(t,x)\right). \tag{1.8}$$

Comparing (1.7) and (1.8), we set $\overline{a}(t,x) = a(T-t,x)$ and then get

$$\left(b_a(T-t,x) + \overline{b}_{\overline{a}}(t,x)\right)\overline{p}(t,x) = a(T-t,x)\,\nabla \overline{p}(t,x),$$

which is rewritten as

$$\left(b(T-t,x) + \overline{b}(t,x)\right)\overline{p}(t,x) - \nabla \cdot a(T-t,x)\,\overline{p}(t,x) = a(T-t,x)\,\nabla \overline{p}(t,x). \tag{1.9}$$

This yields the desired result. □

Let us comment on Theorem 1.1. [12] proved the result by assuming that $b(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ are globally Lipschitz, and the density $p(t,x)$ satisfies an a priori H^1 bound. The implicit condition on $p(t,x)$ is guaranteed if $\partial_t + \mathcal{L}$ is hypoelliptic, or $\nabla^2 a(t,x)$ is uniformly bounded. These conditions were relaxed in [23], where only the boundedness of $\nabla a(t,x)$ in some L^2 norm is required. In another direction, [10, 11] used an entropy argument to prove the time reversal formula in the case $\sigma(t,x) = \sigma I$. This approach was further developed in [4] which made connections to optimal transport.

By Theorem 1.1, the backward procedure is written as

$$d\overline{X}_t = \left(-b(T-t,\overline{X}_t) + a(T-t,\overline{X}_t)\nabla \log p(T-t,\overline{X}_t) + \nabla \cdot a(T-t,\overline{X}_t)\right) dt \\ + \sigma(T-t,\overline{X}_t)dB_t. \quad (1.10)$$

Since $b(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ are chosen in advance, all but the term $\nabla \log p(T-t,\overline{X}_t)$ in (1.10) are available. So in order to implement the backward procedure (1.10), we need to compute $\nabla p(t,x)$ – known as *Stein's score* of X_t.

Examples. Now let us illustrate reverse SDE (1.10) with several concrete examples of DPMs.

Example 1.2 (Constant diffusion)**.** *Let $\sigma(t,x) = \sigma I$. Theorem 1.1 specifies to Föllmer's time reversal formula:*

$$d\overline{X}_t = \left(-b(T-t,\overline{X}_t) + \sigma^2\nabla \log p(T-t,\overline{X}_t)\right) dt + \sigma dB_t. \quad (1.11)$$

(a) Orstein-Ulenback (OU) process $b(t,x) = \theta(\mu - x)$ with $\theta > 0$, $\mu \in \mathbb{R}^n$. It is known that given $X_0 = x$, the distribution of X_t is $p(t,\cdot;x) = \mathcal{N}(\mu + (x-\mu)e^{-\theta t}, \frac{\sigma^2}{2\theta}(1-e^{-2\theta t})I)$. So the stationary distribution of the OU process is $p_\infty(\cdot) \sim \mathcal{N}(\mu, \frac{\sigma^2}{2\theta}I)$. The backward procedure of the OU process is

$$d\overline{X}_t = \left(\theta(\overline{X}_t - \mu) + \sigma^2\nabla \log p(T-t,\overline{X}_t)\right)dt + \sigma dB_t, \\ \text{with } \overline{X}_0 \sim \mathcal{N}\left(\mu + (x-\mu)e^{-\theta T}, \frac{\sigma^2}{2\theta}(1-e^{-2\theta T})I\right), \; x \sim p_{data}(\cdot), \quad (1.12)$$

or

$$d\overline{X}_t = \left(\theta(\overline{X}_t - \mu) + \sigma^2\nabla \log p(T-t,\overline{X}_t)\right) dt + \sigma dB_t, \\ \text{with } \overline{X}_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\theta}I\right). \quad (1.13)$$

(b) Sign-drifted process $b_i(t,x) = -\operatorname{sgn}(x_i)b_i$ for $|x_i| \ge \varepsilon$ and smooth, with $b_i, \varepsilon > 0$. The stationary distribution $p_\infty(\cdot) \propto \otimes_{i=1}^n \nu_i(\cdot)$, where $\nu_i(\cdot) \propto \exp(-\frac{2b_i|x_i|}{\sigma^2})$ for $|x_i| \ge \varepsilon$. If we take ε small, then $\nu_i(\cdot)$ is close to Laplace$(0, \frac{\sigma^2}{2b_i}) \overset{d}{=} \frac{\sigma^2}{2b_i}$Laplace$(0,1)$ (Laplace$(0,1)$ is known as the double exponential). So the the backward procedure of the sign-drifted process is considered as:

$$d\overline{X}_t = (\text{“}\operatorname{sgn}(\overline{X}_t)\cdot b\text{”} + \sigma^2\nabla \log p(T-t,\overline{X}_t))dt + \sigma dB_t, \\ \text{with } \overline{X}_0 \sim \otimes_{i=1}^n \mathit{Laplace}\left(0, \frac{\sigma^2}{2b_i}\right). \quad (1.14)$$

in which “$\operatorname{sgn}(\overline{X}_t)\cdot b$”$_i = -b_i(T-t,\overline{X}_t) = \operatorname{sgn}((\overline{X}_t)_i)b_i$

(c) (Overdamped) Langevin process $b(t,x) = -\nabla U(x)$ for a suitable function $U: \mathbb{R}^n \to \mathbb{R}$. We know that the Langevin process has the stationary distribution $p_\infty(\cdot) \propto \exp\left(-\frac{\sigma^2 U(\cdot)}{2}\right)$ which is a Gibbs measure. The backward procedure of the Langevin

process is

$$d\overline{X}_t = \left(\nabla U(\overline{X}_t) + \sigma^2 \nabla \log p(T-t, \overline{X}_t)\right) dt + \sigma dB_t, \qquad \text{with } \overline{X}_0 \propto \exp\left(-\frac{2\,U(\cdot)}{\sigma^2}\right). \tag{1.15}$$

By taking $U(x) = \frac{\theta}{2}|x-\mu|^2$, *we recover the OU process in (a); and by taking* $U(x) = (-b_1|x_1|, \dots, -b_n|x_n|)$, *we recover the sign-drifted process in (b). Generally, we choose* $U(\cdot)$ *so that the Gibbs measure* $\overline{X}_0 \propto \exp\left(-\frac{2\,U(\cdot)}{\sigma^2}\right)$ *is easy to sample.*

Example 1.3 (State-dependent diffusion)**.** *Let* $\sigma(t,x) = \sigma(x)$*, and* $a(x) := \sigma(x)\sigma(x)^\top$*. Theorem 1.1 reduces to*

$$d\overline{X}_t = \left(-b(T-t, \overline{X}_t) + a(\overline{X}_t)\nabla \log p(T-t, \overline{X}_t) + \nabla \cdot a(\overline{X}_t)\right) dt + \sigma(\overline{X}_t) dB_t. \tag{1.16}$$

(a) *1-dimensional Geometric Brownian Motion* $b(t,x) = \mu x$ *and* $\sigma(x) = \sigma x$ *with* $\sigma > 0$*. It is known that given* $X_0 = x$, $X_t = x \exp\left((\mu - \frac{\sigma^2}{2})t + \sigma B_t\right)$*, i.e.* X_t/x *follows a log-normal distribution. Thus the backward procedure of the Geometric Brownian Motion process becomes*

$$d\overline{X}_t = \left(-(\mu - 2\sigma^2)\overline{X}_t + \sigma^2 \nabla \log p(T-t, \overline{X}_t)\right) dt + \sigma \overline{X}_t dB_t, \qquad \text{with } \overline{X}_0 \sim x \cdot \exp(\mathcal{N}((\mu - \frac{\sigma^2}{2})T, \sigma^2)) \tag{1.17}$$

Example 1.4 (Time-dependent diffusion)**.** *Let* $\sigma(t,x) = \sigma(t)$ *and* $a(t) := \sigma(t)\sigma(t)^\top$*. Theorem 1.1 simplifies to*

$$d\overline{X}_t = \left(-b(T-t, \overline{X}_t) + a(T-t)\nabla \log p(T-t, \overline{X}_t)\right) dt + \sigma(T-t) dB_t. \tag{1.18}$$

If we take $\sigma(t) = \sigma(t)I$ *(where* $\sigma(t)$ *on the l.h.s. is a matrix, while that on the r.h.s. is a scalar), the backward procedure reads as*

$$d\overline{X}_t = \left(-b(T-t, \overline{X}_t) + \sigma^2(T-t)\nabla \log p(T-t, \overline{X}_t)\right) dt + \sigma(T-t) dB_t. \tag{1.19}$$

(a) *Simulated annealing (SA)* $\sigma(t) = \frac{E}{\log(2+t)}$ *and* $b(t,x) = -\nabla U(x)$ *for* $E > 0$ *sufficiently large, and a suitable function* $U : \mathbb{R}^n \to \mathbb{R}$ *with the (global) minimum at* x_**. It is known that the SA process converges to the point mass* $p_\infty(\cdot) = \delta_{x_*}$*, see e.g.* [28]*. So the backward procedure of the SA process is*

$$d\overline{X}_t = \left(\nabla U(\overline{X}_t) + \frac{E^2}{\log^2(2+T-t)} \nabla \log p(T-t, \overline{X}_t)\right) dt + \frac{E}{\log(2+T-t)} dB_t, \text{ with } \overline{X}_0 \sim \delta_{x_*}. \tag{1.20}$$

The time reversal (1.20) *is a noise boosting process. The benefit of SA is that the initial distribution for the backward procedure is a point mass, and so is easy to "sample".*

(b) *Variance exploding (VE) SDE* [27] $b(t,x)=0$ *and*

$$\sigma(t)=\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{t}{T}}\sqrt{\frac{2}{T}\log\frac{\sigma_{\max}}{\sigma_{\min}}},\quad \textit{with } \sigma_{\min}\ll\sigma_{\max}, \tag{1.21}$$

which is the continuum limit of score matching with Langevin dynamics (SMLD) [26]. *Note that* $X_t = X_0+\int_0^t\sigma(s)dB_s$ *is the Paley-Wiener integral. Given* $X_0=x$*, the distribution of* X_t *is given by*

$$\begin{aligned} p(t,\cdot;x) &= \mathcal{N}\left(x_0,\left(\int_0^t\sigma^2(s)ds\right)I\right)\\ &= \mathcal{N}\left(x,\sigma_{\min}^2\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{2t}{T}}-1\right)I\right). \end{aligned} \tag{1.22}$$

The name "variance exploding" comes from the fact that $\mathrm{Var}(X_0)\ll\mathrm{Var}(X_T)$ *since* $\sigma_{\min}\ll\sigma_{\max}$*. Moreover,* $p(T,\cdot;x_0)$ *is close to the uniform distribution (in the support of data). So the backward procedure of the VE SDE is*

$$\begin{aligned} d\overline{X}_t = \sigma^2(T-t))\nabla\log p(T-t,\overline{X}_t)+\sigma(T-t)dB_t,&\\ \textit{with } \overline{X}_0\sim\mathcal{N}\left(x,(\sigma_{\max}^2-\sigma_{\min}^2)I\right),\ x\sim p_{data}(\cdot),& \end{aligned} \tag{1.23}$$

or

$$d\overline{X}_t = \sigma^2(T-t))\nabla\log p(T-t,\overline{X}_t)+\sigma(T-t)dB_t,\quad \textit{with } \overline{X}_0\sim \textit{Unif.} \tag{1.24}$$

(c) *Variance preserving (VP) SDE* [27]. *Let*

$$\beta(t):=\beta_{\min}+\frac{t}{T}(\beta_{\max}-\beta_{\min}),\quad \textit{with } \beta_{\min}\ll\beta_{\max}. \tag{1.25}$$

Set

$$\sigma(t)=\sqrt{\beta(t)}\quad \textit{and}\quad b(t,x)=-\frac{1}{2}\beta(t)x, \tag{1.26}$$

which is the continuum limit of the denoising diffusion probabilistic models (DDPM) [13]. *By applying Itô's formula to* $e^{\frac{1}{2}\int_0^t\beta(s)ds}X_t$*, we get the distribution of* X_t *given* $X_0=x$*:*

$$\begin{aligned} p(t,\cdot;x) &= e^{-\frac{1}{2}\int_0^t\beta(s)ds}\mathcal{N}\left(x,(e^{\int_0^t\beta(s)ds}-1)I\right)\\ &= \mathcal{N}\left(e^{-\frac{1}{2}\int_0^t\beta(s)ds}x,(1-e^{-\int_0^t\beta(s)ds})I\right)\\ &= \mathcal{N}\left(e^{-\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})-\frac{t}{2}\beta_{\min}}x,(1-e^{-\frac{t^2}{2T}(\beta_{\max}-\beta_{\min})-t\beta_{\min}})I\right). \end{aligned} \tag{1.27}$$

Thus, $p(T,\cdot;x_0)=\mathcal{N}(e^{-\frac{T}{4}(\beta_{\max}+\beta_{\min})}x_0,(1-e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})})I)$*, which is close to* $\mathcal{N}(0,I)$ *if* $\beta_{\max}$ *is set to be large. This is why the SDE is called variance preserving. So the backward procedure of the VP SDE is*

$$\begin{aligned} d\overline{X}_t = \left(\frac{1}{2}\beta(T-t)\overline{X}_t+\beta(T-t)\nabla\log p(T-t,\overline{X}_t)\right)&dt\\ +\sqrt{\beta(T-t))}&dB_t,\\ \textit{with } \overline{X}_0\sim\mathcal{N}(e^{-\frac{T}{4}(\beta_{\max}+\beta_{\min})}x,(1-e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})})I),\ x\sim p_{data}(\cdot),& \end{aligned} \tag{1.28}$$

or

$$d\overline{X}_t = \left(\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)\nabla \log p(T-t, \overline{X}_t)\right)dt \\ + \sqrt{\beta(T-t)}dB_t, \quad \textit{with } \overline{X}_0 \sim \mathcal{N}(0, I). \tag{1.29}$$

(d) *Sub-variance preserving (Sub VP) SDE* [27]. *Let $\beta(t)$ be defined by* (1.25). *Set*

$$\sigma(t) = \sqrt{\beta(t)(1 - e^{-2\int_0^t \beta(s)ds})} \quad \textit{and} \quad b(t,x) = -\frac{1}{2}\beta(t)x. \tag{1.30}$$

Let

$$\gamma(t) := e^{-2\int_0^t \beta(s)ds} = e^{-\frac{t^2}{T}(\beta_{\max}-\beta_{\min})-2t\beta_{\min}}, \tag{1.31}$$

so $\sigma(t) = \sqrt{\beta(t)(1-\gamma(t))}$. The same reasoning as in (c) shows that given $X_0 = x$, the distribution of X_t is

$$\begin{aligned} p(t,\cdot;x) &= \mathcal{N}\left(e^{-\frac{1}{2}\int_0^t \beta(s)ds}x, (1 - e^{-\int_0^t \beta(s)ds})^2 I\right) \\ &= \mathcal{N}\left(e^{-\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})-\frac{t}{2}\beta_{\min}}x, (1 - e^{-\frac{t^2}{2T}(\beta_{\max}-\beta_{\min})-t\beta_{\min}})^2 I\right) \end{aligned} \tag{1.32}$$

Note that $\mathrm{Var}_{\textit{sub VP}}(X_t) \le \mathrm{Var}_{VP}(X_t)$*, thus the name "sub VP" SDE. The backward procedure of the VP SDE is*

$$d\overline{X}_t = \left(\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)(1-\gamma(T-t))\nabla \log p(T-t, \overline{X}_t)\right)dt \\ + \sqrt{\beta(T-t)(1-\gamma(T-t))}dB_t, \\ \textit{with } \overline{X}_0 \sim \mathcal{N}(e^{-\frac{T}{4}(\beta_{\max}+\beta_{\min})}x, (1 - e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})})^2 I),\ x \sim p_{data}(\cdot), \tag{1.33}$$

or

$$d\overline{X}_t = \left(\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)(1-\gamma(T-t))\nabla \log p(T-t, \overline{X}_t)\right)dt \\ + \sqrt{\beta(T-t)(1-\gamma(T-t))}dB_t, \quad \textit{with } \overline{X}_0 \sim \mathcal{N}(0, I). \tag{1.34}$$

Score matching. The idea from recently developed score-based generative modeling [13, 26, 27] consists of estimating $\nabla \log p(t,x)$ by function approximations, i.e. *score matching.* More precisely, denote by $\{s_\theta(t,x)\}_\theta$ a family of functions on $\mathbb{R}_+ \times \mathbb{R}^n$ parametrized by θ, e.g. kernel or neural networks. The goal is to solve the score matching problem:

$$\min_\theta \mathcal{J}(\theta) := \mathbb{E}_{p(t,\cdot)}|s_\theta(t,X) - \nabla \log p(t,X)|^2. \tag{1.35}$$

Again the stochastic optimization (1.35) *seems to* be far-fetched since $\nabla \log p(t,x)$'s are not available. Interestingly, this problem has been studied in the context of estimating statistical models with unknown normalizing constant. (It is easily seen that if $p(\cdot)$ is a Gibbs measure, then its Stein's score $\nabla \log p(\cdot)$ does not depend on the normalizing constant). The following result due to Hyvärinen shows that the (implicit) score matching problem (1.35) can be recast into a feasible stochastic optimization with no $\nabla \log p(t,X)$-term.

Theorem 1.5 (Equivalent score matching)**.** [14] *Let*

$$\widetilde{\mathcal{J}}(\theta) := \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2 + 2\,\nabla \cdot s_\theta(t,X)\right]. \tag{1.36}$$

Under suitable conditions on s_θ, we have $\widetilde{\mathcal{J}}(\theta) = \mathcal{J}(\theta) + C$ for some C independent of θ. Consequently, the minimum point of $\widetilde{\mathcal{J}}$ and that of $\mathcal{J}$ coincide.

Proof. We have

$$\begin{aligned}
\nabla_\theta \mathcal{J}(\theta) &= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] - 2\mathbb{E}_{p(t,\cdot)}\left[\nabla_\theta s_\theta(t,X)\cdot\nabla\log p(t,X)\right] \\
&= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] - 2\int \nabla_\theta s_\theta(t,x)\cdot\nabla p(t,x)dx \\
&= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] - 2\,\nabla_\theta \int s_\theta(t,x)\cdot\nabla p(t,x)dx \\
&= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] + 2\nabla_\theta \int \nabla\cdot s_\theta(t,x)\,p(t,x)dx \\
&= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2 + 2\,\nabla\cdot s_\theta(t,X)\right] = \nabla_\theta\widetilde{\mathcal{J}}(\theta),
\end{aligned}$$

where we use the divergence theorem in the fourth equation. □

Now assume that θ_* solves the *equivalent score matching problem*:

$$\min_\theta \widetilde{\mathcal{J}}(\theta) = \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2 + 2\,\nabla\cdot s_\theta(t,X)\right]. \tag{1.37}$$

The backward procedure in the generative SDE modeling is:

$$\begin{aligned}
d\overline{X}_t = \big(-b(T-t,\overline{X}_t) + a(T-t,\overline{X}_t)\,s_\theta(T-t,\overline{X}_t) + \nabla\cdot a(T-t,\overline{X}_t)\big)\,dt& \\
+\,\sigma(T-t,\overline{X}_t)dB_t,&
\end{aligned} \tag{1.38}$$

with $\overline{X}_0 \sim p(T,\cdot)$ or $\overline{X}_0 \sim p_\infty(\cdot)$. By running the SDE (1.38) until time T multiple (many) times, we generate copies of $\overline{X}_T$ whose distribution is expected to be close to $p_{\mathrm{data}}(\cdot)$.

2. Wasserstein Score Matching Error bound

Now we consider how $\overline{X}_T$ in the generative SDE model (1.38) approaches $p_{\text{data}}(\cdot)$, with $\overline{X}_0 \sim p(T, \cdot)$ or $\overline{X}_0 \sim p_\infty(\cdot)$. We focus on the time-dependent case $\sigma(t, x) = \sigma(t)$, which covers Examples 1.2 and 1.4. Recall that $a(t) := \sigma(t)\sigma(t)^\top$. We make the following assumptions.

Assumption 2.1. *The following conditions hold:*

(1) There exists $r_\sigma : [0, T] \to \mathbb{R}_+$ such that $||a(t)||_2 \le r_\sigma(t)$ for $0 \le t \le T$.

(2) There exists $r_b : [0, T] \to \mathbb{R}$ such that $(x - x') \cdot (b(t, x) - b(t, x')) \ge r_b(t)|x - x'|^2$ for all $0 \le t \le T$ and $x, x' \in \mathbb{R}^n$.

(3) There exists $L > 0$ such that $|s_\theta(t, x) - s_\theta(t, x')| \le L|x - x'|$ for all $0 \le t \le T$ and $x, x' \in \mathbb{R}^n$.

(4) There exists $\varepsilon > 0$ such that $\mathbb{E}_{p(t,\cdot)}|s_\theta(t, x) - \nabla \log p(t, x)|^2 \le \varepsilon^2$ for all $0 \le t \le T$.

The condition (1) assumes the boundedness of the diffusion parameter from above; (2) assumes the monotonicity of the drift parameter; and (3) assumes the (uniform) Lipschitz property of the score matching functions. (As a comparison, in all other existing works, the Lipschitz continuity assumption is for the score function itself instead of the matching function, e.g. [7, 5]). These conditions are used to quantify how a perturbation of the model parameters in an SDE will affect its distribution. The condition (4) specifies how accurate Stein's score is estimated by function approximations. Note that the constants L, ε may depend on T. In most aforementioned examples, the drift parameter $b(t, x)$ is affine in x. So the density $p(t, \cdot)$ is Gaussian-like, and its Stein's score $\nabla \log p(t, \cdot)$ is almost affine. Thus, it is reasonable (and consistent) to assume (3), i.e. the score matching $s_\theta(t, x)$ is (uniform) Lipschitz in x.

The following theorem quantifies how close $p_{\text{data}}(\cdot)$ and the distribution of $\overline{X}_T$ are.

Theorem 2.2. *Let $(\overline{X}_t,\, 0 \le t \le T)$ be defined by (1.38), and let Assumption 2.1 hold. Define $\eta := 0$ if $\overline{X}_0 \sim p(T, \cdot)$, or $W_2(p(T, \cdot), p_\infty(\cdot))$ if $\overline{X}_0 \sim p_\infty(\cdot)$, and*

$$u(t) := \int_{T-t}^{T} \left(1 - 2r_b(s) + 2L^2 r_\sigma^2(s)\right) ds. \tag{2.1}$$

Then we have

$$W_2(p_{data}(\cdot), \overline{X}_T) \le \sqrt{\eta^2 e^{u(T)} + 2\varepsilon^2 \int_0^T r_\sigma^2(t) e^{u(T) - u(T-t)} dt}. \tag{2.2}$$

Moreover, if $r_b(t) \le \frac{1}{2}$ holds for all $t \in [0, T]$ (e.g. $r_b(t) \le 0$), then the bound (2.2) can be simplified as:

$$W_2(p_{data}(\cdot), \overline{X}_T) \le \sqrt{\eta^2 e^{u(T)} + \frac{\varepsilon^2}{L^2}(e^{u(T)} - 1)}. \tag{2.3}$$

Proof. The idea relies on coupling, which is similar to [29, Lemma 4]. Consider the coupled SDEs:

$$\begin{cases} dY_t = (-b(T - t, Y_t) + a(T - t)\nabla \log p(T - t, Y_t))\, dt + \sigma(T - t) dB_t, \\ dZ_t = (-b(T - t, Z_t) + a(T - t)s_\theta(T - t, Z_t))\, dt + \sigma(T - t) dB_t, \end{cases}$$

where 1) $Y_0 = Z_0 \sim p(T,\cdot)$ if $\overline{X}_0 \sim p(T,\cdot)$, and 2) (Y_0, Z_0) are coupled to achieve $W_2(p(T,\cdot), p_\infty(\cdot))$ if $\overline{X}_0 \sim p_\infty(\cdot)$, i.e. $Y_0 \sim p(T,\cdot)$, $Z_0 \sim p(\infty,\cdot)$, and $\mathbb{E}|Y_0 - Z_0|^2 = W_2(p(T,\cdot), p_\infty(\cdot))$. Then it is easy to see that

$$W_2^2(p_{\text{data}}(\cdot), \overline{X}_T) \leq \mathbb{E}|Y_T - Z_T|^2. \tag{2.4}$$

So the goal is to bound $\mathbb{E}|Y_T - Z_T|^2$. By Itô's formula, we get

$$\begin{aligned} d|Y_t - Z_t|^2 = 2(Y_t - Z_t) \cdot (&-b(T-t, Y_t) + a(T-t)\nabla \log p(T-t, Y_t) \\ &+ b(T-t, Z_t) - a(T-t)s_\theta(T-t, Z_t))dt \end{aligned}$$

which implies that

$$\begin{aligned} \frac{d\,\mathbb{E}|Y_t - Z_t|^2}{dt} &= -2\underbrace{\mathbb{E}((Y_t - Z_t)\cdot(b(T-t,Y_t) - b(T-t,Z_t)))}_{(a)} \\ &\quad + 2\underbrace{\mathbb{E}((Y_t - Z_t)\cdot a(T-t)(\nabla \log p(T-t,Y_t) - s_\theta(T-t,Z_t)))}_{(b)}. \end{aligned} \tag{2.5}$$

By Assumption 2.1 (2), we get

$$(a) \geq r_b(T-t)\,\mathbb{E}|Y_t - Z_t|^2. \tag{2.6}$$

Moreover,

$$\begin{aligned} (b) &\leq \frac{1}{2}\mathbb{E}|Y_t - Z_t|^2 + \frac{1}{2}\mathbb{E}|a(T-t)(\nabla \log p(T-t,Y_t) - s_\theta(T-t,Z_t))|^2 \\ &\leq \frac{1}{2}\mathbb{E}|Y_t - Z_t|^2 + \frac{1}{2}r_\sigma^2(T-t)\,\mathbb{E}|\nabla \log p(T-t,Y_t) - s_\theta(T-t,Z_t)|^2 \\ &\leq \frac{1}{2}\mathbb{E}|Y_t - Z_t|^2 + r_\sigma^2(T-t)\Big(\mathbb{E}|\nabla \log p(T-t,Y_t) - s_\theta(T-t,Y_t)|^2 \\ &\qquad\qquad + \mathbb{E}|s_\theta(T-t,Y_t) - s_\theta(T-t,Z_t)|^2\Big) \\ &\leq \left(\frac{1}{2} + L^2 r_\sigma^2(T-t)\right)\mathbb{E}|Y_t - Z_t|^2 + \varepsilon^2 r_\sigma^2(T-t), \end{aligned} \tag{2.7}$$

where we use Assumption 2.1 (1) in the second inequality, and (3)(4) in the last inequality. Combining (2.5), (2.6) and (2.7), we have

$$\frac{d\,\mathbb{E}|Y_t - Z_t|^2}{dt} \leq \left(-2r_b(T-t) + 1 + 2L^2 r_\sigma^2(T-t)\right)\mathbb{E}|Y_t - Z_t|^2 + 2\varepsilon^2 r_\sigma^2(T-t). \tag{2.8}$$

Applying Grönwall's inequality, we have:

$$\mathbb{E}|Y_T - Z_T|^2 \leq e^{u(T)}\mathbb{E}|Y_0 - Z_0|^2 + \varepsilon^2\int_0^T r_\sigma^2(T-t)e^{u(T)-u(t)}dt.$$

Combining (2.4) yield (2.2). Notice that denoting $f(t) = u(T) - u(T-t)$, then

$$\frac{df}{dt}(s) = -2r_b(s) + 2L^2 r_\sigma^2(s) + 1.$$

Thus if $r_b(t) \leq \frac{1}{2}$, then combining the following inequality leads to (2.3), as

$$\int_0^T r_\sigma^2(t)e^{u(T)-u(T-t)}dt \leq \int_0^T \frac{1}{2L^2}f'(t)e^{f(t)}dt = \frac{1}{2L^2}(e^{f(T)} - e^{f(0)}) = \frac{1}{2L^2}(e^{u(T)} - 1).$$

□

Looking at the bound (2.2), the *data generation error* $W_2(p_{\text{data}}(\cdot), \overline{X}_T)$ is linear in the *score matching error* ε and the *initialization error* η, and these errors propagate exponentially in time T (in most examples, $r_b(t) \le 0$ so $u(t)$ is positive and at least linear). This implies that the data generation error may explode if T is set to be large even at the continuous level, not to mention discretization errors. If $\overline{X}_0 \sim p(T, \cdot)$ (exact time reversal), then $\eta = 0$ and we can simply take $T = 1$, say. If $\overline{X}_0 \in p_\infty(\cdot)$, then $\eta > 0$ and is generally decreasing in T. In this case, we can choose T to tradeoff the term "$\eta e^{u(T)}$".

Now let's specify the error bound (2.2) to the examples. (Example 1.2 constant diffusion and Example 1.4 time-dependent diffusion)

- Example 1.2 (a) OU process. We have $r_\sigma(t) = \sigma > 0$ and $r_b(t) = -\theta < 0$. So $u(t) = (2\theta + \frac{L^2\sigma^2}{2} + 1)t := C_{\text{OU}} t$, and
$$W_2(p_{\text{data}}(\cdot), \overline{X}_T) \le \sqrt{\left(\eta^2 + \frac{\varepsilon^2\sigma^2}{2C_{\text{OU}}}\right) e^{C_{\text{OU}}T} - \frac{\varepsilon^2\sigma^2}{2C_{\text{OU}}}}. \tag{2.9}$$
Recall that
$$W_2(\mathcal{N}(m_1, \Sigma_1), \mathcal{N}(m_2, \Sigma_2)) = \sqrt{|m_1 - m_2|^2 + \operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\right)}.$$
If $\overline{X}_0 \sim p_\infty(\cdot) = \mathcal{N}(\mu, \frac{\sigma^2}{2\theta} I)$, because of convexity of Wasserstein Distance we have
$$\eta \le \mathbb{E}_{X \sim p_{\text{data}}(\cdot)} \sqrt{|X - \mu|^2 e^{-2\theta T} + \frac{\sigma^2}{2\theta}\left(2 - e^{-2\theta T} - 2\sqrt{1 - e^{-2\theta T}}\right) n}, \tag{2.10}$$
which decays as $e^{-\theta T}$. Thus, $\eta^2 e^{C_{\text{OU}}T} \asymp e^{(\frac{L^2\sigma^2}{2} + 1)T}$ which suggests to pick a small T.
- Example 1.2 (c) Langevin process. Assume that there exists $L' > 0$ such that
$$(x - x') \cdot (\nabla U(x) - \nabla U(x')) \le L'|x - x'|^2 \quad \text{for all } x, x' \in \mathbb{R}^n, \tag{2.11}$$
(which is obviously satisfied by the OU process). We have $r_\sigma(t) = \sigma > 0$ and $r_b(t) = -L' < 0$. So we get the same bound as in (2.9) by replacing C_{OU} with $C_{\text{Lang}} := 2L' + \frac{L^2\sigma^2}{2} + 1$.

  Assume further that the Gibbs measure $p_\infty(\cdot) \propto \exp(\frac{-2U(\cdot)}{\sigma^2})$ satisfies the log-Sobolev inequality (LSI) with constant α. For instance, if $\nabla^2 U \ge \kappa I$ (known as Bakry-Émery condition [1]), then the LSI constant $\alpha = \frac{2\kappa}{\sigma^2}$. By [25, Theorem 1.7],
$$H(p(T, \cdot)\,|\,p_\infty(\cdot)) \le e^{-\sigma^2\alpha T} H(p(0, \cdot)\,|\,p_\infty(\cdot)),$$
where $H(\mu\,|\,\nu)$ denotes the relative entropy (or KL divergence) between μ and ν. Recall that a probability measure ν is said to satisfy Talagrands inequality with constant $\gamma > 0$, if for all probability measure μ with $H(\mu\,|\,\nu) < \infty$,
$$W_2^2(\mu, \nu) \le \frac{2}{\gamma} H(\mu\,|\,\nu).$$

It follows from [22, Theorem 1] that LSI implies Talagrands inequality with the same constant. Thus, if $\overline{X}_0 \sim p_\infty(\cdot)$, we get

$$\eta = W_2(p(T,\cdot), p_\infty(\cdot)) \le \sqrt{\frac{2\,H(p(0,\cdot)\,|\,p_\infty(\cdot))}{\alpha}} e^{-\frac{\sigma^2\alpha}{2}T}. \tag{2.12}$$

Therefore, the term $\eta^2 e^{C_{\text{Lang}}T} \asymp e^{(-\sigma^2\alpha + 2L' + \frac{L^2\sigma^2}{2} + 1)T}$.

- Example 1.4 (a) SA process. We also assume (2.11). We have $r_\sigma(t) = \frac{E}{\log(2+T-t)}$ and $r_b(t) = -L' < 0$, so

$$\begin{aligned} u(t) &= (2L'+1)t + \frac{L^2E^2}{2}\int_0^t \frac{1}{\log^2(2+s)}ds \\ &= (2L'+1)t + \frac{L^2E^2}{2}\left[li(s) - \frac{s}{\log s}\right]_2^{t+2}, \end{aligned}$$

where $li(t) := \int_0^t \frac{1}{\log s}ds$ is the logarithmic integral. It is easy to see that for T sufficiently large, $u(T) \le (2L'+1.1)T$, and $u(T) - u(T-t) \le (2L'+1.1)t$ for all $0 \le t \le T$. Thus, we get the bound

$$W_2(p_{\text{data}}(\cdot), \overline{X}_T) \le \sqrt{\left(\eta^2 + \frac{\varepsilon^2E^2}{2(2L'+1.1)}\right)e^{(2L'+1.1))T} - \frac{\varepsilon^2E^2}{2(2L'+1.1)}}. \tag{2.13}$$

But the dependence of $\eta = W_2(p(T,\cdot), \delta_{x_*})$ in T seems to be hard. In practice, we may try a simple $U(\cdot)$ (e.g. quadratic) and a suitable time-decaying $\sigma(t)$.

- Example 1.4 (b) VE SDE. We have $r_\sigma(t) = \sigma^2(t)$ with $\sigma(t)$ defined by (1.21), and $r_b(t) = 0$, so

$$u(t) = \frac{L^2}{2}\int_{T-t}^T \sigma^4(s)ds + t = \frac{L^2\sigma_{\min}^4}{2T}\log\frac{\sigma_{\max}}{\sigma_{\min}}\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{4t}{T}} - 1\right) + t.$$

Hanyang: There looks to be some calculation error here as I redo the computation, the result should be

$$u(t) = \frac{L^2}{2}\int_{T-t}^T \sigma^4(s)ds + t = \frac{L^2\sigma_{\max}^4}{2T}\log\frac{\sigma_{\max}}{\sigma_{\min}}\left(1 - \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{-\frac{4t}{T}}\right) + t.$$

but the inequality of $u(T) - u(T-t)$ below is still right.

We have $u(T) - u(T-t) \le t + \frac{L^2(\sigma_{\max}^4 - \sigma_{\min}^4)}{2T}\log\frac{\sigma_{\max}}{\sigma_{\min}}$. Therefore,

$$\begin{aligned} W_2(p_{\text{data}}(\cdot), \overline{X}_T) \le &\Bigg(\eta^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{L^2}{2T}(\sigma_{\max}^4 - \sigma_{\min}^4)} e^T + \frac{2\varepsilon^2}{T(T + 4\log\frac{\sigma_{\max}}{\sigma_{\min}})} \\ &\left(\log\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{L^2}{2T}(\sigma_{\max}^4 - \sigma_{\min}^4)}(e^T\sigma_{\max}^4 - \sigma_{\min}^4)\Bigg)^{\frac{1}{2}}. \end{aligned} \tag{2.14}$$

- Example 1.4 (c) VP SDE. We have $r_\sigma(t) = \beta(t)$ and $r_b(t) = -\frac{1}{2}\beta(t)$, with $\beta(t)$ defined by (1.25) as:

$$\beta(t) := \beta_{\min} + \frac{t}{T}(\beta_{\max} - \beta_{\min}) = \beta_{\max} - \frac{T-t}{T}(\beta_{\max} - \beta_{\min}) =: \gamma(T-t),$$

thus $\gamma(0) = \beta_{\max}$, $\gamma(t) - \gamma(0) = -\frac{\beta_{\max}-\beta_{\min}}{T} t$ and

$$\begin{aligned}
u(t) &= \int_{T-t}^{T} \left(\beta(s) + \frac{L^2 \beta^2(s)}{2} \right) ds + t = \int_0^t \left(\gamma(s) + \frac{L^2 \gamma^2(s)}{2} \right) ds + t \\
&= \int_0^t \left((L^2 \beta_{\max} + 1)(\gamma(s) - \beta_{\max}) + \frac{L^2 (\gamma(s) - \beta_{\max})^2}{2} + \frac{L^2 \beta_{\max}^2}{2} + \beta_{\max} \right) ds + t \\
&= \frac{L^2 \beta_{\max} + 1}{2} (\gamma(t) - \beta_{\max}) t + \frac{L^2 (\gamma(s) - \beta_{\max})^2 t}{6} + \left(\frac{L^2 \beta_{\max}^2}{2} + \beta_{\max} + 1 \right) t \\
&= \frac{L^2 (\beta_{\max} - \beta_{\min})^2 t^3}{6T^2} - \frac{(L^2 \beta_{\max} + 1)(\beta_{\max} - \beta_{\min})}{2T} t^2 + \left(\frac{L^2 \beta_{\max}^2}{2} + \beta_{\max} + 1 \right) t
\end{aligned}$$

thus $u(T) = C \cdot T$, where $C = \frac{L^2(\beta_{\max}-\beta_{\min})^2}{6} - \frac{(L^2\beta_{\max}+1)(\beta_{\max}-\beta_{\min})}{2} + \left(\frac{L^2 \beta_{\max}^2}{2} + \beta_{\max} + 1 \right)$. Notice that $r_b(t) < 0$, thus by the bound (2.3), we have:

$$W_2(p_{\text{data}}(\cdot), \overline{X}_T) \le \sqrt{(\eta^2 + \frac{\varepsilon^2}{L^2}) e^{C \cdot T} - \frac{\varepsilon^2}{L^2}}. \tag{2.15}$$

We also compute the initialization error η as:

$$\begin{aligned}
\eta &\le \mathbb{E}_{x \sim p_{\text{data}}(\cdot)} W_2 \left(\mathcal{N}(e^{-\frac{T}{4}(\beta_{\max}+\beta_{\min})} x, (1 - e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})}) I), \mathcal{N}(0, I) \right) \\
&= \mathbb{E}_{x \sim p_{\text{data}}(\cdot)} \left(e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})} \|x\|^2 + (1 - \sqrt{1 - e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})}})^2 \times n \right)^{\frac{1}{2}}
\end{aligned} \tag{2.16}$$

- Example 1.4 (d) sub-VP SDE. We have $r_\sigma(t) = \beta(t)(1 - e^{-2\int_0^t \beta(s) ds})$ and $r_b(t) = -\frac{1}{2}\beta(t)$, with $\beta(t)$ defined by (1.25). It is easy to see that the error bound for the VP SDE is also valid.

3. Discretization Issues

We further investigate the discretization error imposed by the time discretization.

3.1. L^2 error for general SDEs.

We first focus on a general non-autonomous SDE result for the L^2 error presented by [17] for Euler-Maruyama approximation, defined as below:

For a given discretization $0 = t_0 < t_1 < \cdots < t_n < \cdots < t_N = T$ of the time interval $[0, T]$ (not necessarily a uniform one), an Euler approximation is a continuous time stochastic process $Y = \{Y(t), 0 \leq t \leq T\}$ satisfying the iterative scheme

$$Y_{n+1} = Y_n + b\left(t_n, Y_n\right)\Delta_n + \sigma\left(t_n, Y_n\right)\Delta W_n, \tag{3.1}$$

for $n = 0, 1, 2, \ldots, N-1$ with initial value $Y_0 = X_0$, and in (3.1) we denote $Y_n = Y(t_n)$, $\Delta_n = t_{n+1} - t_n$, and $\Delta W_n \sim \mathcal{N}(0, \Delta_n I)$

Denote $\delta = \max_n \Delta_n$, we have the following L^2 error bound of the discretization error, mainly revised from Theorem 10.2.2 of [17]:

Theorem 3.1. *Suppose that $E(|X_0|^2) < \infty$ and assume the smoothness and the growth condition as:*

$$\begin{aligned}&(i)\ |b(t,x) - b(t,y)| \leq K|x-y|, |\sigma(t,x) - \sigma(t,y)| \leq K|x-y|,\\ &(ii)\ |b(t,x)|^2 \leq K^2\left(1+|x|^2\right), |\sigma(t,x)|^2 \leq K^2\left(1+|x|^2\right),\\ &(iii)\ |b(s,x) - b(t,x)| + |\sigma(s,x) - \sigma(t,x)| \leq K(1+|x|)|s-t|^{\frac{1}{2}},\end{aligned} \tag{3.2}$$

for all $s, t \in [0, T]$ and $x, y \in \mathbb{R}^n$, where the constants K do not depend on δ. Then, for the Euler approximation Y^δ, we have

$$E\left(\left|X_T - Y_T^\delta\right|^2\right) \leq K(T) \cdot \delta. \tag{3.3}$$

Proof. By Lemma A.3, we have:

$$E\left(\sup_{0\leq s\leq t} |X_s|^2\right) \leq (4t+16)\left\{E\left|X_0\right|^2 + (1 + E\left|X_0\right|^2)te^{(2K^2+1)t}\right\} \tag{3.4}$$

Consider the Euler Approximation interpolated continuously by

$$Y_t^\delta = Y_n^\delta + \int_{t_n}^t b\left(t_n, Y_n^\delta\right) \mathrm{d}s + \int_{t_n}^t \sigma\left(t_n, Y_n^\delta\right) \mathrm{d}W_s, \tag{3.5}$$

for $t \in [t_n, t_{n+1}]$ and $n = 0, \cdots, N-1$. By Lemma A.5, we have:

$$E\left(\sup_{0\leq s\leq t} |Y_s|^2\right) \leq (4t+16)\left\{E\left|X_0\right|^2 + (1 + E\left|X_0\right|^2)te^{(2K^2+1)t}\right\}. \tag{3.6}$$

(Proof to be finished) □

3.2. L^2 error for time-dependent-only diffusion.

For the time-dependent case $\sigma(t, x) = \sigma(t)$ which is of our interest to DPM, we can derive a sharper bound by the results of Theorem 10.3.5 of [17]:

Theorem 3.2. *Suppose that $E(|X_0|^2) < \infty$ and under appropriate smoothness and the growth condition of $b(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ (stronger than (3.2)), then for the Euler approximation Y^δ the estimate*

$$E\left(\left|X_T - Y_T^\delta\right|^2\right) \leq K(T) \cdot \delta^2 \tag{3.7}$$

holds, where the constant $K(T)$ does not depend on δ.

As a remark, actually for Theorem 3.1 and 3.2, both the error bound is uniform over T, i.e. we actually have:

$$E\left(\sup_{0\leq t\leq T}\left|X_t - Y_t^\delta\right|^2\right) \leq K_5(T) \cdot \delta \tag{3.8}$$

for general case of the diffusion term and for an only time-dependent diffusion term,

$$E\left(\sup_{0\leq t\leq T}\left|X_t - Y_t^\delta\right|^2\right) \leq K(T) \cdot \delta^2 \tag{3.9}$$

which can be directly see from the proof.

3.3. **Summarized Results for DPM.**

4. The power of contraction

Next we consider the discretization of the backward SDE $(\overline{X}_t,\ 0 \le t \le T)$ defined by (1.38). We show that under certain contraction assumptions, the constant $K(T)$ in the discretization error derived in the previous section will 'vanish' and become a constant independent of T.

To simplify the presentation, we still focus on the time-dependent case $\sigma(t,x) = \sigma(t)$. Fix $\delta > 0$ as the step size, and set $t_k := k\delta$ for $k = 0, \dots, N := T/\delta$. Let $\widehat{X}_0 = \overline{X}_0$, and

$$\begin{aligned}\widehat{X}_k := \widehat{X}_{k-1} + (-b(T-t_k, \widehat{X}_{k-1}) + &a(T-t_{k-1}) s_\theta(T-t_{k-1}, \widehat{X}_{k-1}))\delta \\ &+ \sigma(T-t_{k-1})(B_{t_k} - B_{t_{k-1}}), \quad \text{for } k = 1, \dots, N,\end{aligned} \tag{4.1}$$

be the Euler-Maruyama discretization of $\overline{X}$ over $[0,T]$. (Here $B_{t_k} - B_{t_{k-1}}$ can be realized as $\mathcal{N}(0,\delta)$.) By triangle inequality,

$$W_2(p_{\text{data}}(\cdot), \widehat{X}_N) \le W_2(p_{\text{data}}(\cdot), \overline{X}_T) + \left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}}, \tag{4.2}$$

where in the second term on the right side $\overline{X}$ and $\widehat{X}$ are coupled (i.e. driven by the same Brownian motion B). The term $W_2(p_{\text{data}}(\cdot), \overline{X}_T)$ has been studied in Theorem 2.2. In what follows, we deal with the *global discretization error* $\left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}}$.

The key idea is to make the backward SDE (1.38) be "contractive", so the discretization error will not expand (or ampilify) over the time. To this end, we need some extra assumptions.

Assumption 4.1. *The following conditions hold:*

1. *There exists $L_\sigma > 0$ such that $|\sigma(t) - \sigma(t')| \le L_\sigma |t - t'|$ for all t, t'.*
2. *There exists $R_\sigma > 0$ such that $r_\sigma(t) \le R_\sigma$.*
3. *There exists $L_b > 0$ such that $|b(t,x) - b(t',x')| \le L_b(|t-t'| + |x-x'|)$ for all t, t' and x, x'.*
4. *There exists $L_s > 0$ such that $|s_\theta(t,x) - s_\theta(t',x')| \le L_s(|t-t'| + |x-x'|)$ for all t, t' and x, x'.*
5. *There exists $R_s > 0$ such that $|s_\theta(T,x)| \le R_s(1 + |x|)$ for all x.*

Wenpin: Comment on these assumptions...

The analysis of the error $\left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}}$ relies on the following lemmas. The lemma below proves the contraction of the backward SDE $\overline{X}$.

Assumption 4.2 (Contraction)**.** *There exists $\beta > 0$ such that*

$$\int_{T-t}^{T} (r_b(s) - L r_\sigma^2(s))\, ds \ge \beta t, \quad \text{for all } t, \tag{4.3}$$

or simply

$$\beta := \inf_{0 \le t \le T} \left(r_b(t) - L r_\sigma^2(t)\right) > 0.$$

Lemma 4.3. *Let $(\overline{X}^x_t,\, 0 \le t \le T)$ be defined by* (1.38) *with $\overline{X}^x_0 = x$. Let Assumption 2.1 (1)–(3) and Assumption 4.2 (4) hold, then*

$$\left(\mathbb{E}|\overline{X}^x_t - \overline{X}^y_t|^2\right)^{\frac{1}{2}} \le \left(\mathbb{E}|x-y|^2\right)^{\frac{1}{2}} \exp(-\beta t), \quad \textit{for all } t, \tag{4.4}$$

where $\overline{X}^x$ and $\overline{X}^y$ be coupled, i.e. they are driven by the same Brownian motion with (different) initial values x and y respectively (x and y represent two random variables).

Proof. Recall that $\sigma(t,x) = \sigma(t)$ does not depend on the state x. We have

$$\begin{aligned} d|\overline{X}^x_s - \overline{X}^y_s|^2 = 2\left(\overline{X}^x_s - \overline{X}^y_s\right) \cdot \Bigg(& - b(T-s, \overline{X}^x_s) + a(T-s) s_\theta(T-s, \overline{X}^x_s) \\ & + b(T-s, \overline{X}^y_s) - a(T-s) s_\theta(\overline{X}^y_s)\Bigg) ds. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{d}{ds}\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2 = & -2\underbrace{\mathbb{E}\left[(\overline{X}^x_s - \overline{X}^y_s)\cdot(b(T-s,\overline{X}^x_s) - b(T-s,\overline{X}^y_s))\right]}_{(a)} \\ & + 2\underbrace{\mathbb{E}\left[(\overline{X}^x_s - \overline{X}^y_s) a(T-s)(s_\theta(T-s,\overline{X}^x_s) - s_\theta(T-s,\overline{X}^y_s))\right]}_{(b)}. \end{aligned} \tag{4.5}$$

By Assumption 2.1 (2), we get

$$(a) \ge r_b(T-s)\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2. \tag{4.6}$$

By Assumption 2.1 (1)(3), we obtain

$$(b) \le L r^2_\sigma(T-s)\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2. \tag{4.7}$$

Combining (4.5), (4.6) and (4.7) yields

$$\frac{d}{ds}\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2 \le -2(r_b(T-s) - L r^2_\sigma(T-s))\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2.$$

By Grönwall's inequality, we have

$$\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2 \le \mathbb{E}|x-y|^2 \exp\left(-2\int_0^t (r_b(T-s) - L r^2_\sigma(T-s))ds\right),$$

which, by the condition (4.3), yields (4.4) □

Next we deal with the *local (one-step) discretization error* of the process $\overline{X}$. Fixing $t_\star \le T - \delta$, the (one-step) discretization of $\overline{X}$ starting at $\overline{X}_{t_\star} = x$ is:

$$\widehat{X}^{t_\star,x}_1 = x + (-b(T-t_\star,x) + a(T-t_\star)s_\theta(T-t_\star,x))\delta + \sigma(T-t_\star)(B_{t_\star+\delta} - B_{t_\star}). \tag{4.8}$$

The following lemma provides an estimate of the local discretization error.

Lemma 4.4. *Let $(\overline{X}^{t_\star,x}_t,\, t_\star \le t \le T)$ be defined by* (1.38) *with $\overline{X}^{t_\star,x}_{t_\star} = x$, and $\widehat{X}^{t_\star,x}_1$ be given by* (4.8)*. Let Assumption 4.1 hold. Then for δ sufficiently small (i.e. $\delta \le \overline{\delta}$ for some $\overline{\delta} < 1$), there exists $C_1, C_2 > 0$ independent of δ and x such that*

$$\left(\mathbb{E}|\overline{X}^{t_\star,x}_{t_\star+\delta} - \widehat{X}^{t_\star,x}_1|^2\right)^{\frac{1}{2}} \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^{\frac{3}{2}}, \tag{4.9}$$

$$|\mathbb{E}(\overline{X}^{t_\star,x}_{t_\star+\delta} - \widehat{X}^{t_\star,x}_1)| \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^{\frac{3}{2}}. \tag{4.10}$$

Proof. For ease of presentation, we write $\overline{X}_t$ (resp. $\widehat{X}_1$) for $\overline{X}^{t_\star,x}_t$ (resp. $\widehat{X}^{t_\star,x}_1$). Without loss of generality, set $t_\star = 0$. We have

$$\overline{X}_\delta = x + \int_0^\delta -b(T-t,\overline{X}_t) + a(T-t)s_\theta(T-t,\overline{X}_t)dt + \int_0^\delta \sigma(T-t)dB_t,$$
$$\widehat{X}_1 = x + \int_0^\delta -b(T,x) + a(T)s_\theta(T,x)dt + \int_0^\delta \sigma(T)dB_t.$$

So

$$\begin{aligned}
&\mathbb{E}|\overline{X}_\delta - \widehat{X}_1|^2 \\
&= \mathbb{E}\bigg|\int_0^\delta b(T,x) - b(T-t,\overline{X}_t)dt + \int_0^\delta a(T-t)s_\theta(T-t,\overline{X}_t) - a(T)s_\theta(T,x)dt \\
&\qquad\qquad + \int_0^\delta \sigma(T-t) - \sigma(T)dB_t\bigg|^2 \\
&\le 3\mathbb{E}\bigg(\bigg|\int_0^\delta b(T,x) - b(T-t,\overline{X}_t)dt\bigg|^2 + \bigg|\int_0^\delta a(T-t)s_\theta(T-t,\overline{X}_t) - a(T)s_\theta(T,x)dt\bigg|^2 \\
&\qquad\qquad + \bigg|\int_0^\delta \sigma(T-t) - \sigma(T)dB_t\bigg|^2\bigg) \\
&\le 3\bigg(\delta\underbrace{\int_0^\delta \mathbb{E}|b(T,x) - b(T-t,\overline{X}_t)|^2dt}_{(a)} + \delta\underbrace{\int_0^\delta \mathbb{E}|a(T-t)s_\theta(T-t,\overline{X}_t) - a(T)s_\theta(T,x)|^2dt}_{(b)} \\
&\qquad\qquad + \underbrace{\int_0^\delta |\sigma(T-t) - \sigma(T)|^2dt}_{(c)}\bigg),
\end{aligned} \tag{4.11}$$

where we use the CauchySchwarz inequality and Itô's isometry in the last inequality. By Assumption 4.1 (1), we get

$$(c) \le \int_0^\delta L_\sigma^2 t^2 dt = \frac{L_\sigma^2}{3}\delta^3. \tag{4.12}$$

By Assumption 4.1 (3), we have

$$(a) \le \int_0^\delta 2L_b^2(t^2 + \mathbb{E}|\overline{X}_t - x|^2)dt = 2L_b^2\left(\frac{\delta^3}{3} + \int_0^\delta \mathbb{E}|\overline{X}_t - x|^2dt\right).$$

According to [16, Theorem 4.5.4], we have $\mathbb{E}|\overline{X}_t - x|^2 \le C(1 + \mathbb{E}|x|^2)te^{Ct}$ for some $C > 0$ (independent of x). Consequently, for $t \le \delta$ sufficiently small (bounded by $\overline{\delta} < 1$),

$$\mathbb{E}|\overline{X}_t - x|^2 \le C'(1 + \mathbb{E}|x|^2)t, \quad \text{for some } C' > 0 \text{ (independent of } \delta, x).$$

We then get

$$(a) \le 2L_b^2\left(\frac{\delta^3}{3} + \frac{C'(1+\mathbb{E}|x|^2)}{2}\delta^2\right) \le 2L_b^2\left(\frac{1}{3} + \frac{C'}{2} + \frac{C'}{2}\mathbb{E}|x|^2\right)\delta^2. \tag{4.13}$$

Similarly, we obtain by Assumption 4.1 (1)(2)(4)(5):

$$(b) \le C''(1+\mathbb{E}|x|^2)\delta^2, \quad \text{for some } C''>0 \text{ (independent of } \delta, x). \tag{4.14}$$

Combining (4.11), (4.12), (4.13) and (4.14) yields the estimate (4.9).

Next we have

$$\begin{aligned}
&|\mathbb{E}(\overline{X}_\delta - \widehat{X}_1)| \\
&= \left|\mathbb{E}\int_0^\delta b(T,x) - b(T-t,\overline{X}_t)dt + \mathbb{E}\int_0^\delta a(T-t)s_\theta(T-t,\overline{X}_t) - a(T)s_\theta(T,x)dt\right| \\
&\le \int_0^\delta \mathbb{E}|b(T,x) - b(T-t,\overline{X}_t)|dt + \int_0^\delta \mathbb{E}|a(T-t)s_\theta(T-t,\overline{X}_t) - a(T)s_\theta(T,x)|dt \\
&\le C''' \int_0^\delta \big(t(1+\mathbb{E}|x|) + \mathbb{E}|\overline{X}_t - x|\big)\, dt \\
&\le C''''(1+\sqrt{\mathbb{E}|x|^2})\delta^{\frac{3}{2}}, \quad \text{for some } C''''>0 \text{ (independent of } \delta, x).
\end{aligned}$$

where the third inequality follows from Assumption 4.1, and the last inequality is due to the fact that $\mathbb{E}|\overline{X}_t - x| \le \big(\mathbb{E}|\overline{X}_t - x|^2\big)^{\frac{1}{2}} \le \sqrt{C'(1+\mathbb{E}|x|^2)t}$. This yields the estimate (4.10). □

Now we state the result for the global discretization error $\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2$.

Theorem 4.5. *Let* $(\overline{X}_t,\, 0 \le t \le T)$ *be defined by* (1.38)*, and* $(\widehat{X}_k,\, 0 \le k \le N)$ *be defined by* (4.1)*. Let Assumption 2.1 (1)–(3) and Assumption 4.1 hold. Then there exists* $C > 0$ *(independent of* δ, T*) such that*

$$\Big(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\Big)^{\frac{1}{2}} \le C\sqrt{\delta}. \tag{4.15}$$

Proof. The proof is split into four steps.

Step 1. Recall that $t_k = k\delta$ for $k = 0, \dots, N$. Denote $\overline{X}_k := \overline{X}_{t_k}$, and let

$$e_k := \Big(\mathbb{E}|\overline{X}_k - \widehat{X}_k|^2\Big)^{\frac{1}{2}}.$$

The idea is to build a recursion for the sequence $(e_k)_{k=0,\dots,N}$. Also write $(\overline{X}_t^{t_\star,x},\, t_\star \le t \le T)$ to emphasize that the reversed SDE (1.38) starts at $\overline{X}_{t_\star}^{t_\star,x} = x$, so $\overline{X}_{k+1} = \overline{X}_{t_{k+1}}^{t_k,\overline{X}_k}$. We have

$$\begin{aligned}
e_{k+1}^2 &= \mathbb{E}\left|\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} + \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1}\right|^2 \\
&= \underbrace{\mathbb{E}|\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k}|^2}_{(a)} + \underbrace{\mathbb{E}|\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1}|^2}_{(b)} + 2\underbrace{\mathbb{E}\left[(\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k})(\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1})\right]}_{(c)}.
\end{aligned} \tag{4.16}$$

Step 2. We analyze the term (a) and (b). By Lemma 4.3 (the contraction property), we get

$$(a) = \mathbb{E}|\overline{X}_{t_{k+1}}^{t_k,\overline{X}_k} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k}|^2 \le e_k^2 \exp(-2\beta\delta). \tag{4.17}$$

By (4.9) (in Lemma 4.4), we have

$$(b) \le \left(C_1 + C_2 \mathbb{E}|\widehat{X}_k|^2\right) \delta^3. \tag{4.18}$$

Step 3. We analyze the cross-product (c). By splitting

$$\overline{X}_{k+1} - \overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} = (\overline{X}_k - \widehat{X}_k) + \underbrace{\left[(\overline{X}_{k+1} - \overline{X}_k) - (\overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} - \widehat{X}_k)\right]}_{:= d_\delta(\overline{X}_k, \widehat{X}_k)},$$

we obtain

$$(c) = \underbrace{\mathbb{E}\left[(\overline{X}_k - \widehat{X}_k)(\overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1})\right]}_{(d)} + \underbrace{\mathbb{E}\left[d_\delta(\overline{X}_k, \widehat{X}_k)(\overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1})\right]}_{(e)}. \tag{4.19}$$

For the term (d), we have

$$\begin{aligned}(d) &= \mathbb{E}\left[(\overline{X}_k - \widehat{X}_k)\, \mathbb{E}(\overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1} | \mathcal{F}_k)\right] \\ &\le e_k \left(\mathbb{E}|\mathbb{E}(\overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1} | \mathcal{F}_k)|^2\right)^{\frac{1}{2}} \\ &\le e_k \left(C_1 + C_2 \sqrt{\mathbb{E}|\widehat{X}_k|^2}\right) \delta^{\frac{3}{2}},\end{aligned} \tag{4.20}$$

where we use the tower property (of the conditional expectation) in the first equation, the Cauchy-Schwarz inequality in the second inequality, and (4.10) in the final inequality. According to [21, Lemma 1.3], there exists $C_0 > 0$ (independent of $\delta, \widehat{X}_k$) such that

$$\left(\mathbb{E} d^2_\delta(\overline{X}_k, \widehat{X}_k)\right)^{\frac{1}{2}} \le C_0 e_k \sqrt{\delta}. \tag{4.21}$$

Thus,

$$\begin{aligned}(e) &\le \left(\mathbb{E} d^2_\delta(\overline{X}_k, \widehat{X}_k)\right)^{\frac{1}{2}} \left(\mathbb{E}|\overline{X}^{t_k, \widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1}|^2\right)^{\frac{1}{2}} \\ &\le C_0 e_k \left(C_1 + C_2 \sqrt{\mathbb{E}|\widehat{X}_k|^2}\right) \delta^2.\end{aligned} \tag{4.22}$$

where we use (4.9) and (4.21) in the last inequality. Combining (4.19), (4.20) and (4.22) yields for δ sufficiently small,

$$(c) \le e_k \left(C'_1 + C'_2 \sqrt{\mathbb{E}|\widehat{X}_k|^2}\right) \delta^{\frac{3}{2}}, \quad \text{for some } C'_1, C'_2 > 0 \text{ (independent of } \delta, \widehat{X}_k). \tag{4.23}$$

Step 4. Combining (4.16) with (4.17), (4.18) and (4.23) yields

$$e^2_{k+1} \le e^2_k \exp(-2\beta\delta) + \left(C_1 + C_2 \mathbb{E}|\widehat{X}_k|^2\right) \delta^3 + e_k \left(C'_1 + C'_2 \sqrt{\mathbb{E}|\widehat{X}_k|^2}\right) \delta^{\frac{3}{2}}.$$

A standard argument shows that Lemma 4.3 (the contraction property) implies $\mathbb{E}|\overline{X}_t|^2 \le C$ for some $C > 0$. Thus, $\mathbb{E}|\widehat{X}_k|^2 \le 2(C + e_k^2)$. As a result, for δ sufficiently small,

$$e_{k+1}^2 \le e_k^2 \left(1 - \frac{3}{4}\beta\delta\right) + D_1 \delta^3 + D_2 e_k^2 \left(\delta^3 + \delta^{\frac{3}{2}}\right) + D_3 e_k \delta^{\frac{3}{2}}, \tag{4.24}$$

for some $D_1, D_2, D_3 > 0$ (independent of δ). Note that

$$D_2 e_k^2 \left(\delta^3 + \delta^{\frac{3}{2}}\right) \le \frac{1}{4} e_k^2 \beta\delta, \quad \text{for } \delta \text{ sufficiently small,}$$

and

$$D_3 e_k \delta^{\frac{3}{2}} \le \frac{1}{4} e_k^2 \beta\delta + \frac{2D_3^2}{\beta}\delta^2.$$

Thus, the estimate (4.24) leads to

$$e_{k+1}^2 \le e_k^2 \left(1 - \frac{1}{4}\beta\delta\right) + D\delta^2, \quad \text{for some } D > 0 \text{ (independent of } \delta). \tag{4.25}$$

Unfolding the inequality (4.25) yields the estimate (4.15). □

As a remark, if we can improve the estimate in (4.10):

$$|\mathbb{E}(\overline{X}_{t_\star+\delta}^{t_\star,x} - \widehat{X}_1^{t_\star,x})| \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^2, \tag{4.26}$$

(i.e. δ^2 local error instead of $\delta^{\frac{3}{2}}$), then the discretization error is $C\delta$.

Now by (4.2), Theorem 2.2 and Theorem 4.5, we have

$$W_2(p_{\text{data}}(\cdot), \widehat{X}_N) \le W_2(p_{\text{data}}(\cdot), \overline{X}_T) + C\sqrt{\delta}, \tag{4.27}$$

where a bound for $W_2(p_{\text{data}}(\cdot), \overline{X}_T)$ is given by (2.2) or (2.3). The term $W_2(p_{\text{data}}(\cdot), \overline{X}_T)$ depends on T, and may decay (exponentially) in T as discussed in Section 2. The discretization error $C\sqrt{\delta}$ is now independent of T – it is an "infill" discretization error (by fixing the horizon T). If we choose

$$N = T^\gamma \text{ steps with } \gamma > 1,$$

then $\delta = T^{1-\gamma}$. The discretization error scales as $T^{-\frac{\gamma-1}{2}} = N^{-\frac{\gamma-1}{2\gamma}}$, which is roughly $N^{-\frac{1}{2}}$ for large γ.

Note that as mentioned before, if we can improve the estimate in (4.10):

$$|\mathbb{E}(\overline{X}_{t_\star+\delta}^{t_\star,x} - \widehat{X}_1^{t_\star,x})| \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^2, \tag{4.28}$$

(i.e. δ^2 local error instead of $\delta^{\frac{3}{2}}$), then the discretization error is $C\delta$. In this case, the error scales as $N^{-\frac{\gamma-1}{\gamma}}$, which is roughly $1/N$ for large γ.

5. Examples and Experiments

We consider some concrete examples which may lead to the contraction property of the reversed SDE. There's an issue justifying the key assumption Assumption 4.2, as typically we do not know L thus we could not verify (4.4). But a necessary condition implied by (4.4) is that there exists $\inf_{t\in[0,T]} r_b(t) > 0$. Recall that $r_b(t)$ such that:

$$(x-x')\cdot(b(t,x)-b(t,x')) \geq r_b(t)|x-x'|^2 \text{ for all } 0 \leq t \leq T$$

(a) 'Reverse' Orstein-Ulenback (OU) process $b(t,x) = \theta(x-\mu)$ with $\theta > 0$, $\mu \in \mathbb{R}^n$ and $\sigma(t) \equiv \sigma$. Given $X_0 = x$, the distribution of X_T is

$$p(T,\cdot;x) = \mathcal{N}(\mu + (x-\mu)e^{\theta T}, \frac{\sigma^2}{2\theta}(e^{2\theta T}-1)I), \tag{5.1}$$

which enables us to conduct exact sampling.

(b) 'Reverse' Variance preserving (VP) SDE [27]. Let

$$\beta(t) := \beta_{\min} + \frac{t}{T}(\beta_{\max} - \beta_{\min}), \quad \text{with } \beta_{\min} \ll \beta_{\max}. \tag{5.2}$$

Set

$$\sigma(t) = \sqrt{\beta(t)} \quad \text{and} \quad b(t,x) = \frac{1}{2}\beta(t)x, \tag{5.3}$$

By applying Itô's formula to $e^{-\frac{1}{2}\int_0^t \beta(s)ds} X_t$, we get the distribution of X_t given $X_0 = x$:

$$\begin{aligned} p(t,\cdot;x) &= e^{\frac{1}{2}\int_0^t \beta(s)ds}\mathcal{N}\left(x, (1-e^{-\int_0^t \beta(s)ds})I\right) \\ &= \mathcal{N}\left(e^{\frac{1}{2}\int_0^t \beta(s)ds}x, (e^{\int_0^t \beta(s)ds}-1)I\right) \\ &= \mathcal{N}\left(e^{\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})+\frac{t}{2}\beta_{\min}}x, (e^{\frac{t^2}{2T}(\beta_{\max}-\beta_{\min})+t\beta_{\min}}-1)I\right). \end{aligned} \tag{5.4}$$

Thus, $p(T,\cdot;x_0) = \mathcal{N}(e^{\frac{T}{4}(\beta_{\max}+\beta_{\min})}x_0, (e^{\frac{T}{2}(\beta_{\max}+\beta_{\min})}-1)I)$.

(c) 'Reverse' sub-variance preserving (Sub VP) SDE [27]. Let $\beta(t)$ be defined by (1.25). Set

$$\sigma(t) = \sqrt{\beta(t)(e^{2\int_0^t \beta(s)ds}-1)} \quad \text{and} \quad b(t,x) = \frac{1}{2}\beta(t)x. \tag{5.5}$$

Let

$$\gamma(t) := e^{2\int_0^t \beta(s)ds} = e^{\frac{t^2}{T}(\beta_{\max}-\beta_{\min})+2t\beta_{\min}}, \tag{5.6}$$

so $\sigma(t) = \sqrt{\beta(t)(\gamma(t)-1)}$. The same reasoning as in (b) shows that given $X_0 = x$, the distribution of X_t is

$$\begin{aligned} p(t,\cdot;x) &= \mathcal{N}\left(e^{\frac{1}{2}\int_0^t \beta(s)ds}x, (e^{\int_0^t \beta(s)ds}-1)^2 I\right) \\ &= \mathcal{N}\left(e^{\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})+\frac{t}{2}\beta_{\min}}x, (e^{\frac{t^2}{2T}(\beta_{\max}-\beta_{\min})+t\beta_{\min}}-1)^2 I\right) \end{aligned} \tag{5.7}$$

The backward procedure of the VP SDE is

$$\begin{aligned} d\overline{X}_t = &\left(\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)(1-\gamma(T-t))\nabla \log p(T-t, \overline{X}_t)\right)dt \\ &\qquad\qquad\qquad\qquad + \sqrt{\beta(T-t)(1-\gamma(T-t))}dB_t, \\ &\text{with } \overline{X}_0 \sim \mathcal{N}\left(e^{\frac{T}{4}(\beta_{\max}+\beta_{\min})}x, (e^{\frac{T}{2}(\beta_{\max}+\beta_{\min})}-1)^2 I\right),\ x \sim p_{\text{data}}(\cdot), \end{aligned} \tag{5.8}$$

6. Literature Review

We briefly review some key assumptions and bounds developed in recent works for the theoretical analysis of DPMs. We limit the analysis to diffusion instead of latent diffusion first. We also focus on the direct analysis for [27], with an earlier analysis on SGM like [3].

6.1. TV distance with L^{∞} error.

The first quantitative convergence results for the methodology of [27] is [7] (extra assumptions on time-homogeneous diffusion term contrary to the exact form of [27]), which provides the TV distance bound with respect to the horizon length T under a L^{∞} score matching error. The precise result is listed as below.

Theorem 6.1. *Assume that $b(t,x) = -\alpha x$ for $\alpha \geq 0$ with a constant diffusion term $\sigma(t,x) = I$ wlog. Further assume that there exists $\mathrm{M} \geq 0$ such that for any $t \in [0,T]$ and $x \in \mathbb{R}^d$*

$$\|s_{\theta^\star}(t,x) - \nabla \log p_t(x)\| \leq \varepsilon,$$

with $s_{\theta^\star} \in \mathrm{C}\left([0,T] \times \mathbb{R}^d, \mathbb{R}^d\right)$. Assume that $p_{\mathrm{data}} \in \mathrm{C}^3\left(\mathbb{R}^d, (0,+\infty)\right)$ is bounded and that there exist $d_1, A_1, A_2, A_3 \geq 0, \beta_1, \beta_2, \beta_3 \in \mathbb{N}$ and $\mathrm{m}_1 > 0$ such that for any $x \in \mathbb{R}^d$ and $i \in \{1,2,3\}$

$$\left\|\nabla^i \log p_{data}\,(x)\right\| \leq A_i \left(1 + \|x\|^{\beta_i}\right), \langle \nabla \log p_{data}\,(x), x\rangle \leq -\mathrm{m}_1 \|x\|^2 + d_1\|x\|, \tag{6.1}$$

with $\beta_1 = 1$. Then for any $\alpha \geq 0$, there exist $B_\alpha, C_\alpha, D_\alpha \geq 0$ such that for any $N \in \mathbb{N}$ and $\{\gamma_k\}_{k=1}^N$ with $\gamma_k > 0$ for any $k \in \{1,\ldots,N\}$ denoting the EulerMaruyama discretization parameters such that $T = \sum_{k=1}^N \gamma_k, \bar{\gamma} = \sup_{k\in\{1,\ldots,N\}} \gamma_k$, the following bounds on the total variation distance hold:
(a) if $\alpha > 0$, we have $\left\|\mathcal{L}\left(\bar{X}_T\right) - p_{data}\right\|_{\mathrm{TV}} \leq C_\alpha\left(\varepsilon + \bar{\gamma}^{1/2}\right)\exp\left[D_\alpha T\right] + B_\alpha \exp\left[-\alpha^{1/2} T\right]$;
(b) if $\alpha = 0$, we have $\left\|\mathcal{L}\left(\bar{X}_T\right) - p_{data}\right\|_{\mathrm{TV}} \leq C_0\left(\varepsilon + \bar{\gamma}^{1/2}\right)\exp\left[D_0 T\right] + B_0\left(T^{-1} + T^{-1/2}\right)$;
in which $\mathcal{L}\left(\bar{X}_T\right)$ denotes the true distribution of $\bar{X}_T$ after descritization and score-matching error.

Proof. A sketch of proof is that we bound the distance by two terms through the following inequality:

$$\begin{aligned} \|\mathcal{L}\left(\bar{X}_T\right) - p_{data}\|_{\mathrm{TV}} &\leq \|\mathcal{L}\left(\bar{X}_T\right) - \bar{X}_T\|_{\mathrm{TV}} + \|\bar{X}_T - p_{data}\|_{\mathrm{TV}} && (6.2)\\ &\leq \underbrace{\|\mathcal{L}\left(\bar{X}_T\right) - \bar{X}_T\|_{\mathrm{TV}}}_{(a)} + \underbrace{\|\bar{X}_0 - X_T\|_{\mathrm{TV}}}_{(b)}, && (6.3)\end{aligned}$$

in which the second inequality follows from the data processing inequality (for the TV distance) that:

$$\|\mu_0 \mathrm{P} - \mu_1 \mathrm{P}\|_{\mathrm{TV}} \leq \|\mu_0 - \mu_1\|_{\mathrm{TV}}$$

for any distribution μ_0, μ_1 and markov transition kernel P. The bound of term (a) relies on the Girsanov's Theorem (Theorem 7.7 of [19]) which leads to the final bound that

$$(a) \leq C_3 \exp\left[C_3 T\right]\left(\bar{\gamma} + \mathrm{M}^2\right),$$

which is independent of α and the bound of term (b) is controlled through the mixing properties, whose proof is reminiscent of [2]. □

We briefly comment on the intuition of the assumptions in Theorem 6.1. The main assumption (6.1), in addition to the bounded L^{∞} error, can be shown to lead to:

$$\left\|\nabla^{\ell} \log p_t\left(x_t\right)\right\| \leq D_{\ell}\left(1+\left\|x_t\right\|^{\beta_{\ell}}\right)$$

for any t and the certain continuity property of $\nabla^{\ell} \log p_t(\cdot)$ in order to finally satisfy the condition of applying Girsanov's Theorem, and thus bounding the term (a). In addition, since (a) is a composition of the discretization error and the score-matching error, it is possible that we divide the error into two parts which may improve the bound, which not surprisingly is captured by later analysis.

6.2. **TV distance with L^2 error.** The [5] provides a quite similar argument of the DDPM case [13] for the TV distance under a L^2 error. Moreover, [5] indeed divide the error term (a) into two parts corresponding to the discretization error and convergence error, as we commented before. Again the paper assumes that the diffusion term is a constant with $\sigma \equiv I$ (although the paper's author claimed that it could be extended to general case, e.g. VE-SDE or VP-SDE) and $n=d$, the dimension of data is same as the dimension of the brownian motion. The key assumptions of [5] are:

Aspt 1 (Lipschitz score). For all $t \geq 0$, the score $\nabla \ln q_t$ is L-Lipschitz.

Aspt 2 (second moment bound). For some $\eta>0, \mathbb{E}_q\left[\|\cdot\|^{2+\eta}\right]$ is finite.

Aspt 3 (regular starting distribution). For all $k=1, \ldots, N, \mathbb{E}_{q_{k h}}\left[\left\|s_{k h}-\nabla \ln q_{k h}\right\|^2\right] \leq \varepsilon_{\text {score }}^2$

Aspt 4 (score estimation error). The data distribution q has finite KL divergence w.r.t. the standard Gaussian, i.e. $\mathrm{KL}\left(q \| \gamma^d\right)<\infty$.

Again the assumptions 1 and 2 are essentially used to bound the discretization error.

Theorem 6.2. *Let $\mathcal{L}\left(\bar{X}_T\right)$ denotes the true distribution of $\bar{X}_T$ after discretization and score-matching error, and suppose that the step size $h:=T / N$ satisfies $h \lesssim 1 / L$, where $L \geq 1$. Then, it holds that*

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

Notably Corollary 6 of [5] provides a Wasserstein 2 bound given the bounded support assumption.

6.3. **Wasserstein-1 distance with maniford hypothesis.**

6.4. **Ignoring error.** With existing works like [6, 18], assuming an oracle of the score function with no matching error.

7. Min-Max Optimal Rate for the diffusion models

8. Constrained Domains

Discuss other approaches, e.g. [7]; diffusion models in constraint domains [9], on the simplex [8, 24] and reflected diffusion models [20]. These are related to constrained diffusions with the stationary distribution of exponential form.

Appendix A. Lemmas and Proofs

Lemma A.1 (Grönwall's inequality)**.** *Let $\alpha, \beta : [t_0, T] \to \mathbb{R}$ be integrable with*

$$0 \leq \alpha(t) \leq \beta(t) + L \int_{t_0}^{t} \alpha(s) ds$$

for $t \in [t_0, T]$ where $L > 0$. Then

$$\alpha(t) \leq \beta(t) + L \int_{t_0}^{t} e^{L(t-s)} \beta(s) ds$$

for $t \in [t_0, T]$.

Lemma A.2 (Doob's inequality)**.** *A right-continuous martingale $X = \{X_t, t \geq 0\}$ with finite p th-moment satisfies the Doob's inequality:*

$$E\left(\sup_{0 \leq s \leq t} |X_t|^p\right) \leq \left(\frac{p}{p-1}\right)^p E\left(|X_t|^p\right).$$

Lemma A.3 (Theorem 4.5.4 and 4.5.5 of [16])**.** *Suppose that for all $t \in [0, T]$ and $x, y \in \mathbb{R}$, there exists a constant $K > 0$ such that*

$$|b(t,x) - b(t,y)| \leq K|x-y|, |\sigma(t,x) - \sigma(t,y)| \leq K|x-y|$$

and

$$|b(t,x)|^2 \leq K^2\left(1 + |x|^2\right), |\sigma(t,x)|^2 \leq K^2\left(1 + |x|^2\right).$$

Further, assume that

$$E\left|X_0\right|^2 < \infty,$$

then we have the following second moment estimate and uniform second moment estimate of the solution X_t of (1.1) *satisfies*

$$E\left|X_t\right|^2 \leq \left(1 + E\left|X_0\right|^2\right) e^{(2K^2+1)\cdot t} \tag{A.1}$$

for $t \in [0, T]$ and

$$E\left(\sup_{0 \leq s \leq T} |X_s|^2\right) \leq (4T + 16)\left\{E\left|X_0\right|^2 + (1 + E\left|X_0\right|^2) T e^{(2K^2+1)T}\right\}. \tag{A.2}$$

Proof. We first prove (A.1). By Ito's formula:

$$\mathrm{d}|X_t|^2 = 2X_t \cdot (b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t) + |\sigma(t,x)|^2 \mathrm{d}t$$

thus integrating from 0 to t and taking expectation, we have

$$\begin{aligned}
\mathbb{E}|X_t|^2 &= \mathbb{E}|X_0|^2 + \int_0^t \left(2X_s \cdot b(s, X_s) + |\sigma(t,x)|^2\right) \mathrm{d}t \\
&\leq \mathbb{E}|X_0|^2 + \int_0^t \left((K^2+1)\left(1 + |X_s|^2\right) + K^2\left(1 + |X_s|^2\right)\right) \mathrm{d}t \\
&\leq \mathbb{E}|X_0|^2 + (2K^2+1)t + (2K^2+1)\int_0^t \mathbb{E}|X_s|^2 \mathrm{d}s.
\end{aligned}$$

By applying the Grönwall's inequality, we have:

$$\begin{aligned}\mathbb{E}|X_t|^2 &\leq \mathbb{E}|X_0|^2 + (2K^2+1)t + (2K^2+1)\int_0^t e^{(2K^2+1)(t-s)}\left(\mathbb{E}|X_0|^2 + (2K^2+1)s\right)ds \\ &= \mathbb{E}|X_0|^2 + (2K^2+1)t + \mathbb{E}|X_0|^2(e^{(2K^2+1)t} - 1) - (2K^2+1)t - 1 + e^{(2K^2+1)t} \\ &\leq (1+\mathbb{E}|X_0|^2)e^{(2K^2+1)t},\end{aligned}$$

which yields (A.1). To prove (A.2), notice that

$$Y_t := X_t - \int_0^t b(s, X_s)\mathrm{d}s - X_0$$

is a martingale, thus by Doob's inequality:

$$E\left(\sup_{0\leq t\leq T}|Y_t|^2\right) \leq 4E\left(|Y_T|^2\right).$$

By Itô's isometry, we have that the right hand side:

$$\mathbb{E}\left(|Y_T|^2\right) = \mathbb{E}(\int_0^T |\sigma(s, X_s)|^2\mathrm{d}s) \leq \mathbb{E}(\int_0^T K^2(1+|X_s|^2)\mathrm{d}s).$$

By using the growth condition and (A.1):

$$\begin{aligned}\mathbb{E}(\int_0^T K^2(1+|X_s|^2)\mathrm{d}s) &\leq K^2T + K^2\int_0^T (1+\mathbb{E}|X_0|^2)e^{(2K^2+1)t}\mathrm{d}s \\ &\leq K^2T + \frac{1}{2}(1+\mathbb{E}|X_0|^2)\left(e^{(2K^2+1)T} - 1\right) \\ &\leq (1+\mathbb{E}|X_0|^2)\left(e^{(2K^2+1)T} - 1\right)\end{aligned}$$

Also notice that:

$$\begin{aligned}\sup_{0\leq t\leq T}|X_t|^2 &= \sup_{0\leq t\leq T}\left|Y_t + \int_0^t b(s, X_s)\mathrm{d}s + X_0\right|^2 \\ &\leq 4\left(\sup_{0\leq t\leq T}|Y_t|^2 + \sup_{0\leq t\leq T}\left|\int_0^t b(s, X_s)\mathrm{d}s\right|^2\right) + 2|X_0|^2 \\ &\leq 4\left(\sup_{0\leq t\leq T}|Y_t|^2 + \left(\int_0^T |b(s, X_s)|\mathrm{d}s\right)^2\right) + 2|X_0|^2 \\ &\leq 4\left(\sup_{0\leq t\leq T}|Y_t|^2 + T\int_0^T |b(s, X_s)|^2\mathrm{d}s\right) + 2|X_0|^2\end{aligned}$$

Thus taking the expectation on both sides and using the growth condition:

$$\begin{aligned}\mathbb{E}\left(\sup_{0\leq t\leq T}|X_t|^2\right) &\leq 16\mathbb{E}\,|Y_t|^2 + 4T\int_0^T K^2(1+\mathbb{E}|X_s|^2)\mathrm{d}s + 2\mathbb{E}\left(|X_0|^2\right) \\ &\leq (4T+16)\left\{E\,|X_0|^2 + (1+E\,|X_0|^2)Te^{(2K^2+1)T}\right\}\end{aligned}$$

which concludes our proof. □

Lemma A.4. *Let Y_t defined as in* (3.5) *and further denote*

$$R_s = \tag{A.3}$$

Proof. The result is a direct corollary pf [16, Lemma 10.8.1] when $\ell(\alpha) = n(\alpha) = 1$ and $\ell(\alpha) = 1, n(\alpha) = 0$. □

Lemma A.5. *Suppose the same conditions as in Lemma A.3, then we have that Y_t defined in* (3.5) *satisfies*

$$E\left|Y_t\right|^2 \leq \left(1 + E\left|X_0\right|^2\right) e^{(2K^2+1)\cdot t} \tag{A.4}$$

for $t \in [0, T]$ and

$$E\left(\sup_{0 \leq s \leq T} \left|Y_s\right|^2\right) \leq (4T + 16)\left\{E\left|X_0\right|^2 + (1 + E\left|X_0\right|^2)Te^{(2K^2+1)T}\right\} \tag{A.5}$$

Proof. □

References

[1] D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, Berlin, 1985.

[2] D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

[3] A. Block, Y. Mroueh, and A. Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

[4] P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *arXiv:2104.07708*, 2021.

[5] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[6] S. Chen, G. Daras, and A. Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023.

[7] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Neurips*, volume 34, pages 17695–17709, 2021.

[8] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, and C. Durkan. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

[9] N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. Hutchinson. Diffusion models for constrained domains. *arXiv:2304.05364*, 2023.

[10] H. Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pages 156–163. Springer, Berlin, 1985.

[11] H. Föllmer. Time reversal on Wiener space. In *Stochastic processes—mathematics and physics (Bielefeld, 1984)*, volume 1158 of *Lecture Notes in Math.*, pages 119–129. Springer, Berlin, 1986.

[12] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *Ann. Probab.*, 14(4):1188–1205, 1986.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Neurips*, volume 33, pages 6840–6851, 2020.

[14] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

[15] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

[16] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[17] P. E. Kloeden, E. Platen, P. E. Kloeden, and E. Platen. *Stochastic differential equations*. Springer, 1992.

[18] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

[19] R. S. Liptser and A. N. Shiraev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

[20] A. Lou and S. Ermon. Reflected diffusion models. *arXiv:2304.04740*, 2023.

[21] G. N. Milstein and M. V. Tretyakov. *Stochastic numerics for mathematical physics*. Scientific Computation. Springer-Verlag, Berlin, 2004.

[22] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

[23] J. Quastel. Time reversal of degenerate diffusions. In *In and out of equilibrium (Mambucaba, 2000)*, volume 51 of *Progr. Probab.*, pages 249–257. Birkhäuser Boston, Boston, MA, 2002.

[24] P. H. Richemond, S. Dieleman, and A. Doucet. Categorical SDEs with simplex diffusion. *arXiv:2210.14784*, 2022.

[25] A. Schlichting. *The Eyring-Kramers formula for Poincaré and logarithmic Sobolev inequalities*. PhD thesis, Universität Leipzig, 2012. Available at `http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-97965`.

[26] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Neurips*, volume 32, page 1191811930, 2019.

[27] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[28] W. Tang and X. Y. Zhou. Tail probability estimates of continuous-time simulated annealing processes. *Numer. Algebra Control Optim.*, 13(3-4):473–485, 2023.

[29] H. Zhao, W. Tang, and D. D. Yao. Policy optimization for continuous reinforcement learning. *arXiv:2305.18901*, 2023.

Department of Industrial Engineering and Operations Research, Columbia University.
E-mail address: wt2319@columbia.edu